Package 'PGEE'

August 16, 2016

Type Package

Title Penalized Generalized Estimating Equations in High-Dimension

Version 1.4

Date 2016-08-16

Author Gul Inan (Visiting scholar, University of Minnesota), Jianhui Zhou (Associate Professor, University of Virginia) and Lan Wang (Professor, University of Minnesota)

Maintainer Gul Inan <inanx002@umn.edu>

Description Fits penalized generalized estimating equations to longitudinal data with high-dimensional covariates.

License GPL (>= 2)

Depends MASS, mvtnorm

NeedsCompilation no

Repository CRAN

Date/Publication 2016-08-16 09:59:11

R topics documented:

PGEE-package *Penalized Generalized Estimating Equations*

Description

This package fits penalized generalized estimating equations to longitudinal data with high-dimensional covariates through accommodating SCAD-penalty function into generalized estimating equations.

Details

This package consists of three functions. The function PGEE fits penalized generalized estimating equations to the data. But, before that, the tuning parameter should be estimated through the function CVfit. On the other hand, the function MGEE fits unpenalized generalized estimating equations to the data.

Author(s)

Gul Inan, Jianhui Zhou and Lan Wang

Maintainer: Gul Inan

References

Wang, L., Zhou, J., and Qu, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics*, **68**, 353–360.

CVfit *Function to compute cross-validated tuning parameter value*

Description

This function computes cross-validated tuning parameter value for longitudinal data with working independence structure.

Usage

```
CVfit(formula, id, data, family, scale.fix, scale.value, fold, lambda.vec, pindex,
eps, maxiter, tol)
```

Arguments

formula	A formula expression in the form of response ~ predictors.
id	A vector for identifying subjects/clusters.
data	A data frame which stores the variables in formula with id variable.
scale.fix	A logical variable; if true, the scale parameter is fixed at the value of scale.value. The default value is FALSE.
scale.value	If scale.fix = TRUE, this assignes a numeric value to which the scale parameter should be fixed.
family	A family object in PGEE.
fold	The number of folds used in cross-validation.
lambda.vec	A vector of tuning parameters that will be used in the cross-validation.
pindex	An index vector showing the parameters which are not subject to penalization. The default value is NULL. However, in case of a model with intercept, the intercept parameter should be never penalized.
eps	A numerical value for the epsilon used in minorization-maximization algorithm. The default value is 10^-6.
maxiter	The number of iterations that is used in the estimation algorithm. The default value is 25.
tol	The tolerance level that is used in the estimation algorithm. The default value is 10^-3.

Value

An object class of CVfit.

References

Wang, L., Zhou, J., and Qu, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics*, **68**, 353–360.

See Also

PGEE

MGEE *Function to fit generalized estimating equations*

Description

This function fits a generalized estimating equation model to longitudinal data.

Usage

```
MGEE(formula, id, data, na.action = NULL, family = gaussian(link = "identity"),
corstr = "independence", Mv = NULL, beta_int = NULL, R = NULL, scale.fix = FALSE,
scale.value = 1, maxiter = 25, tol = 10^-3, silent = FALSE)
```

Arguments

formula	A formula expression in the form of `response ~ predictors`.
id	A vector for identifying subjects/clusters.
data	A data frame which stores the variables in `formula` with `id` variable.
na.action	A function to remove missing values from the data. Only `na.omit` is allowed here.
family	A `family` object: a list of functions and expressions for defining `link` and `variance` functions. Families supported in `MGEE` are `binomial`, `gaussian`, `gamma` and `poisson`. The `links`, which are not available in gee, is not available here. The default family is `gaussian`.
corstr	A character string, which specifies the type of correlation structure. Structures supported in `MGEE` are `"AR-1"`,`"exchangeable"`, `"fixed"`, `"independence"`, `"stat_M_dep"`,`"non_stat_M_dep"`, and `"unstructured"`. The default `corstr` type is `"independence"`.
Mv	If either `"stat_M_dep"`, or `"non_stat_M_dep"` is specified in `corstr`, then this assigns a numeric value for `Mv`. Otherwise, the default value is `NULL`.
beta_int	User specified initial values for regression parameters. The default value is `NULL`.
R	If `corstr = "fixed"` is specified, then `R` is a square matrix of dimension maximum cluster size containing the user specified correlation. Otherwise, the default value is `NULL`.
scale.fix	A logical variable; if true, the scale parameter is fixed at the value of `scale.value`. The default value is `FALSE`.
scale.value	If `scale.fix = TRUE`, this assignes a numeric value to which the scale parameter should be fixed.
maxiter	The number of iterations that is used in the estimation algorithm. The default value is `25`.
tol	The tolerance level that is used in the estimation algorithm. The default value is `10^-3`.
silent	A logical variable; if true, the regression parameter estimates at each iteration are printed. The default value is `FALSE`.

Value

An object class of `MGEE` representing the fit.

Note

The structures `"non_stat_M_dep"` and `"unstructured"` are valid only when the data is balanced.

References

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

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

See Also

PGEE

PGEE *Function to fit penalized generalized estimating equations*

Description

This function fits a penalized generalized estimating equation model to longitudinal data.

Usage

```
PGEE(formula, id, data, na.action = NULL, family = gaussian(link = "identity"),
corstr = "independence", Mv = NULL, beta_int = NULL, R = NULL, scale.fix = FALSE,
scale.value = 1, lambda, pindex = NULL, eps = 10^-6, maxiter = 30, tol = 10^-3,
silent = FALSE)
```

Arguments

formula	A formula expression in the form of `response ~ predictors`.
id	A vector for identifying subjects/clusters.
data	A data frame which stores the variables in `formula` with `id` variable.
na.action	A function to remove missing values from the data. Only `na.omit` is allowed here.
family	A `family` object: a list of functions and expressions for defining `link` and `variance` functions. Families supported in `PGEE` are `binomial`, `gaussian`, `gamma` and `poisson`. The `links`, which are not available in gee, is not available here. The default family is `gaussian`.
corstr	A character string, which specifies the type of correlation structure. Structures supported in `PGEE` are `"AR-1"`,`"exchangeable"`, `"fixed"`, `"independence"`, `"stat_M_dep"`,`"non_stat_M_dep"`, and `"unstructured"`. The default `corstr` type is `"independence"`.
Mv	If either `"stat_M_dep"`, or `"non_stat_M_dep"` is specified in `corstr`, then this assigns a numeric value for `Mv`. Otherwise, the default value is `NULL`.
beta_int	User specified initial values for regression parameters. The default value is `NULL`.
R	If `corstr = "fixed"` is specified, then `R` is a square matrix of dimension maximum cluster size containing the user specified correlation. Otherwise, the default value is `NULL`.

scale.fix	A logical variable; if true, the scale parameter is fixed at the value of scale.value. The default value is FALSE.
lambda	A numerical value for the penalization parameter of the scad function, which is estimated via cross-validation.
pindex	An index vector showing the parameters which are not subject to penalization. The default value is NULL. However, in case of a model with intercept, the intercept parameter should be never penalized.
eps	A numerical value for the epsilon used in minorization-maximization algorithm. The default value is 10^-6.
scale.value	If scale.fix = TRUE, this assignes a numeric value to which the scale parameter should be fixed.
maxiter	The number of iterations that is used in the estimation algorithm. The default value is 25.
tol	The tolerance level that is used in the estimation algorithm. The default value is 10^-3.
silent	A logical variable; if true, the regression parameter estimates at each iteration are printed. The default value is FALSE.

Value

An object class of PGEE representing the fit.

References

Wang, L., Zhou, J., and Qu, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics*, **68**, 353–360.

See Also

CVfit, MGEE

Examples

```
# Consider an example similar to example 1
# in Wang et al. (2012).

# required R package
library(mvtnorm)
# number of subjects
n <- 200
# number of covariates
pn <- 10
# number of time points
m <- 4

# vector if subject ids
id.vect <- rep(1:n, each = m)

# covariance matrix of (pn-1) number of continuous covariates
```

```
X.sigma <- matrix(0,(pn-1),(pn-1))
{
for (i in 1:(pn-1))
X.sigma[i,] <- 0.5^(abs((1:(pn-1))-i))
}

# generate matrix of covariates
x.mat <- as.matrix(rmvnorm(n*m, mean = rep(0,(pn-1)), X.sigma))
x.mat <- cbind(rbinom(n*m,1, 0.5), x.mat)

# true values
beta.true <- c(2,3,1.5,2,rep(0,6))
sigma2 <- 1
rho <- 0.5
R <- matrix(rho,m,m)+diag(rep(1-rho,m))

# covariance matrix of error
SIGMA <- sigma2*R
error <- rmvnorm(n, mean = rep(0,m),SIGMA)

# generate longitudinal data with continuous outcomes
y.temp <- x.mat%*%beta.true
y.vect <- y.temp+as.vector(t(error))

mydata <- data.frame(id.vect,y.vect,x.mat)
colnames(mydata) <- c("id","y",paste("x",1:length(beta.true),sep = ""))

###Input Arguments for CVfit fitting###
library(PGEE)
formula <- "y ~.-id-1"
data <- mydata
family <- gaussian(link = "identity")
lambda.vec <- seq(0.1,1,0.1)

## Not run:
cv <- CVfit(formula = formula, id = data[,1], data = data, family = family,
fold = 4, lambda.vec = lambda.vec, pindex = NULL, eps = 10^-6, maxiter = 30,
tol = 10^-3)

names(cv)
cv$lam.opt

## End(Not run)

lambda <- 0.1 #this value obtained through CVfit

# analyze the data through penalized generalized estimating equations

myfit1 <- PGEE(formula = formula, id = data[,1], data = data, na.action = NULL,
family = family, corstr = "exchangeable", Mv = NULL,
beta_int = c(rep(0,length(beta.true))), R = NULL, scale.fix = FALSE,
scale.value = 1, lambda = lambda, pindex = NULL, eps = 10^-6, maxiter = 30,
tol = 10^-3, silent = FALSE)
```

```
summary(myfit1)

# analyze the data through unpenalized generalized estimating equations

myfit2 <- MGEE(formula = formula, id = data[,1], data = data, na.action = NULL,
family = family, corstr = "exchangeable", Mv = NULL,
beta_int = c(rep(0,length(beta.true))), R = NULL, scale.fix = FALSE,
scale.value = 1, maxiter = 30, tol = 10^-3, silent = FALSE)

summary(myfit2)
```

---

yeastG1 *Yeast cell-cycle gene expression data*

Description

A yeast cell-cycle gene expression data set collected in the CDC15 experiment of Spellman et al. (1998) where genome-wide mRNA levels of 6178 yeast open reading frames (ORFs) in a two cell-cycle period were measured at M/G1-G1-S-G2-M stages. However, to better understand the phenomenon underlying cell-cycle process, it is important to identify transcription factors (TFs) that regulate the gene expression levels of cell cycle-regulated genes. In this study, we presented a subset of 283 cell-cycled-regularized genes observed over 4 time points at G1 stage and the standardized binding probabilities of a total of 96 TFs obtained from a mixture model approach of Wang et al. (2007) based on the ChIP data of Lee et al. (2002).

Usage

```
data("yeastG1")
```

Details

A data frame with 1132 observations (283 cell-cycled-regularized genes observed over 4 time points) with 99 variables (e.g., id, y, time, and 96 TFs).

References

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., et al. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, **298**, 799–804.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of Cell*, **9**, 3273–3297.

Wang, L., Chen, G., and Li, H. (2007). Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics*, **23**, 1486–1494.

Wang, L., Zhou, J., and Qu, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data anaysis. *Biometrics*, **68**, 353–360.

Examples

```
## Not run:
library(PGEE)
# load data
data(yeastG1)
data <- yeastG1
# get the column names
colnames(data)[1:9]
# see some portion of yeast G1 data
head(data,5)[1:9]

# define the input arguments
formula <- "y ~.-id"
family <- gaussian(link = "identity")
lambda.vec <- seq(0.01,0.2,0.01)
# find the optimum lambda
cv <- CVfit(formula = formula, id = data[,1], data = data, family = family, scale.fix = TRUE,
scale.value = 1, fold = 4, lambda.vec = lambda.vec, pindex = c(1,2), eps = 10^-6,
maxiter = 30, tol = 10^-6)
# print the results
print(cv)

# see the returned values by CVfit
names(cv)
# get the optimum lambda
cv$lam.opt

#fit the PGEE model
myfit1 <- PGEE(formula = formula, id = data[,1], data = data, na.action = NULL,
family = family, corstr = "independence", Mv = NULL,
beta_int = c(rep(0,dim(data)[2]-1)), R = NULL, scale.fix = TRUE,
scale.value = 1, lambda = cv$lam.opt, pindex = c(1,2), eps = 10^-6,
maxiter = 30, tol = 10^-6, silent = FALSE)

# get the values returned by myfit object
names(myfit1)
# get the values returned by summary(myfit) object
names(summary(myfit1))
# see a portion of the results returned by summary(myfit1)
# $coefficients
head(summary(myfit1)$coefficients,7)

# see the variables which have non-zero coefficients
index1 <- which(abs(summary(myfit1)$coef[,"Estimate"]) > 10^-3)
index1

# see the PGEE summary statistics of these non-zero variables
summary(myfit1)$coef[index1,]

# fit the GEE model
myfit2 <- MGEE(formula = formula, id = data[,1], data = data, na.action = NULL,
family = family, corstr = "independence", Mv = NULL,
```

```
beta_int = c(rep(0,dim(data)[2]-1)), R = NULL, scale.fix = TRUE,
scale.value = 1, maxiter = 30, tol = 10^-6, silent = FALSE)

# get the GEE summary statistics of the variables that turned out to be
#non-zero in PGEE analysis
summary(myfit2)$coef[index1,]

# see the significantly associated TFs in PGEE analysis
which(abs(summary(myfit1)$coef[index1,][,"Robust z"]) > 1.96)

# see the significantly associated TFs in both PGEE and GEE analyses
which(abs(summary(myfit2)$coef[index1,][,"Robust z"]) > 1.96)


## End(Not run)
```

Index