Package: epilogi (via r-universe)

October 11, 2024

Type Package

Title The 'epilogi' Variable Selection Algorithm for Continuous Data

Version 1.1

Date 2024-09-10

Author Michail Tsagris [aut, cre]

Maintainer Michail Tsagris <mtsagris@uoc.gr>

Depends R (>= 4.0)

Imports Rfast, stats

Description The 'epilogi' variable selection algorithm is implemented for the case of continuous response and predictor variables. The relevant paper is: Lakiotaki K., Papadovasilakis Z., Lagani V., Fafalios S., Charonyktakis P., Tsagris M. and Tsamardinos I. (2023). ``Automated machine learning for Genome Wide Association Studies". Bioinformatics. <doi:10.1093/bioinformatics/btad545>.

License GPL (>= 2)

NeedsCompilation no

Repository CRAN

Date/Publication 2024-09-10 12:20:02 UTC

Contents

epilogi-package *The 'epilogi' Variable Selection Algorithm for Continuous Data.*

Description

The 'epilogi' Variable Selection Algorithm for Continuous Data.

Details

Package:	epilogi
Type:	Package
Version:	1.1
Date:	2024-09-10
License:	GPL-2

Maintainers

Michail Tsagris <mtsagris@uoc.gr>.

Author(s)

Michail Tsagris <mtsagris@uoc.gr>.

References

Lakiotaki K., Papadovasilakis Z., Lagani V., Fafalios S., Charonyktakis P., Tsagris M. and Tsamardinos I. (2023). Automated machine learning for Genome Wide Association Studies. Bioinformatics.

Tsagris M., Papadovasilakis Z., Lakiotaki K. and Tsamardinos I. (2022). The γ-OMP algorithm for feature selection with application to gene expression data. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 19(2): 1214–1224.

epilogi *The epilogi Variable Selection Algorithm for Continuous Data.*

Description

The epilogi Variable Selection Algorithm for Continuous Data.

Usage

```
epilogi(y, x, tol = 0.01, alpha = 0.05)
```

Arguments

y	A vector with the continuous response variable.
x	A matrix with the continuous predictor variables.
tol	The tolerance value for the algortihm to terminate. This takes values greater than 0 and it refers to the change between two successive R^2-adjusted values.
alpha	The significance level to deem a predictor variable is statistically equivalent to a selected variable.

Details

The ϵpilogi variable selection algorithm (Lakiotaki et al., 2023) is a generalisation of the γ-OMP algorithm (Tsagris et al. 2022). It applies the aforementioned algorithm with the addition that it returns possible statistically equivalent predictor(s) for each selected predictor. Once a variable is selected the algorithm searches for possible equivalent predictors using the partial correlation between the residuals.

The heuristic method to consider two predictors R and C informationally equivalent given the current selected predictor S is determined as follows: first, the residuals r of the model using S are computed. Then, if the following two conditions hold R and C are considered equivalent: Ind(R; r | C) and Ind(r ; C | R), where Ind(R; r | C) denotes the conditional independence of R with r given C. When linearity is assumed, the test can be implemented by testing for significance the corresponding partial correlation. The tests Ind return a p-value and independence is accepted when it is larger than a threshold (significance value, argument alpha). Intuitively, R and C are heuristically considered equivalent, if C is known, then R provides no additional information for the residuals r, and if R is known, then C provides no additional information for r.

Value

A list including:

runtime	The runtime of the algorithm.
result	A matrix with two columns. The selected predictor(s) and the adjusted R^2-values.
equiv	A list with the equivalent predictors (if any) corresponding to each selected predictor.

Author(s)

Michail Tsagris.

R implementation and documentation: Michail Tsagris <mtsagris@uoc.gr>.

References

Lakiotaki K., Papadovasilakis Z., Lagani V., Fafalios S., Charonyktakis P., Tsagris M. and Tsamardinos I. (2023). Automated machine learning for Genome Wide Association Studies. Bioinformatics.

Tsagris M., Papadovasilakis Z., Lakiotaki K. and Tsamardinos I. (2022). The γ-OMP algorithm for feature selection with application to gene expression data. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 19(2): 1214–1224.

See Also

pcor.equiv

Examples

```
#simulate a dataset with continuous data
set.seed(1234)
n <- 500
x <- matrix( rnorm(n * 50, 0, 30), ncol = 50 )

#define a simulated class variable
y <- 2 * x[, 1] - 1.5 * x[, 2] + x[, 3] + rnorm(n, 0, 15)

# define some simulated equivalences
x[, 4] <- x[, 1] + rnorm(n, 0, 1)
x[, 5] <- x[, 2] + rnorm(n, 0, 1)

epilogi(y, x, tol = 0.05)
```

---

pcor.equiv *Equivalence test using partial correlation*

Description

Equivalence test using partial correlation.

Usage

```
pcor.equiv(res, y, x, alpha = 0.05)
```

Arguments

res	A vector with the residuals of the linear model.
y	A vector with a selected predictor.
x	A matrix with other predictors.
alpha	The significance level to check for predictors from x that are equivalent to y.

Value

A vector with 0s and 1s. 0s indicate that the predictors are not equivalent, while 1s indicate the equivalent predictors.

Author(s)

Michail Tsagris.

R implementation and documentation: Michail Tsagris <mtsagris@uoc.gr>.

See Also

epilogi

Examples

```
#simulate a dataset with continuous data
set.seed(1234)
n <- 500
x <- matrix( rnorm(n * 50, 0, 30), ncol = 50 )

#define a simulated class variable
y <- 2 * x[, 1] - 1.5 * x[, 2] + x[, 3] + rnorm(n, 0, 15)

# define some simulated equivalences
x[, 4] <- x[, 1] + rnorm(n, 0, 1)
x[, 5] <- x[, 2] + rnorm(n, 0, 1)


b <- epilogi(y, x, tol = 0.05)
sel <- b$result[2, 1]
## standardise the y and x first
y <- (y - mean(y)) / Rfast::Var(y, std = TRUE)
x <- Rfast::standardise(x)

res <- resid( lm(y ~ x[, sel] ) )
sela <- b$result[2:3, 1]
pcor.equiv(res, x[, sela[2]], x[, -sela] )
## bear in mind that this gives the third variable after removing the first two,
## so this is essentially the 5th variable in the "x" matrix.
```

Index