Package 'gausscov'

September 11, 2019

Title The Gaussian Covariate Method for Variable Selection

Version 0.0.2

Author Laurie Davies [aut, cre]

Maintainer Laurie Davies <laurie.davies@uni-due.de>

Description

Given the standard linear model the traditional way of deciding whether to include the jth covariate is to apply the F-test to decide whether the corresponding beta coefficient is zero. The Gaussian covariate method is completely different. The question as to whether the beta coefficient is or is not zero is replaced by the question as to whether the covariate is better or worse than i.i.d. Gaussian noise. The P-value for the covariate is the probability that Gaussian noise is better. Surprisingly this can be given exactly and it is the same a the P-value for the classical model based on the F-distribution. The Gaussian covariate P-value is model free, it is the same for any data set. Using the idea it is possible to do covariate selection for a small number of covariates 25 by considering all subsets. Post selection inference causes no problems as the P-values hold whatever the data. The idea extends to stepwise regression again with exact probabilities. In the simplest version the only parameter is a specified cut-off P-value which can be interpreted as the probability of a false positive being included in the final selection. For more information see the website below and the accompanying papers: L. Davies and L. Duembgen, ``A Model-free Approach to Linear Least Squares Regression with Exact Probabilities and Applications to Covariate Selection", 2019, <arXiv:1906.01990>. L. Davies, ``Lasso, Knockoff and Gaussian covariates: A comparison", 2018, <arXiv:1807.09633v4>.

LazyData true

License GPL-3

Depends R (>= 2.10), stats

Encoding UTF-8

RoxygenNote 6.1.1

NeedsCompilation yes

Repository CRAN

Date/Publication 2019-09-11 20:50:03 UTC

R topics documented:

`boston` *Boston data*

Description

This data set is part of the MASS package. The 14 columns are:

crim per capita crime rate by town

zn proportion of residential land zoned for lots over 25.000 sq.ft.

indus proportion of non-residential business acres per town

chas Charles River dummy variable (=1 if tract bounds rive; 0 otherwise)

nox nitrogen oxides concentration (parts per 10 million)

rm average number of rooms per dwelling

age proportion of owner-occupied units built prior to 1940

dis weighted mean of distances to five Boston employment centres

rad index of accessibility to radial highways

tax full-value property-tax rate per $10,000

ptration pupil-teacher ration by town

black 100(Bk-0.63)^2 where Bk is the proportion of blacks by town

lstat lower status of the population (percent)

medv median value of owner occupies homes in $1000s.

Usage

```
boston
```

Format

A 506 x 14 matrix.

Source

R package MASS https://cran.r-project.org/web/packages/available_packages_by_name.html

References

MASS Support Functions and Datasets for Venables and Ripley's MASS

---

colon *Colon data*

Description

The vector "colon.y" consists of 62 persons, 40 with colon cancer (=1) and 22 controls (=0).

Usage

```
colon.y
```

Format

a vector of length 62

Source

http://microarray.princeton. edu/oncology http://stat.ethz.ch/~dettling/bagboost.html

References

Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., and Levine, A. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. PNAS, (1999), 96(12). Booting for tumor classification with gene expression data. Dettling, M. and B\"uhlmann, P., Bioinformatics, (2003),19(9):1061–1069.

colon.x *Colon data*

Description

The measurements of gene expression of 2000 genes.

Usage

```
colon.x
```

Format

A 2000 x 62 matrix

Source

http://microarray.princeton. edu/oncology http://stat.ethz.ch/~dettling/bagboost.html

References

Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., and Levine, A. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. PNAS, (1999), 96(12). Booting for tumor classification with gene expression data. Dettling, M. and B\"uhlmann, P., Bioinformatics, (2003),19(9):1061–1069.

decode *Decodes the number of a subset selected by flmmdch to give the covariates*

Description

Decodes the number of a subset selected by flmmdch to give the covariates

Usage

```
decode(j, k)
```

Arguments

j	The number of the subset
k	The number of covariates

Value

set A binary vector giving the covariates

Examples

```
decode(19,8)
```

`decomp` *decompose a given interaction ic into its component parts*

Description

decompose a given interaction ic into its component parts

Usage

```
decomp(ic, k, ord)
```

Arguments

`ic`	The number of the interaction
`k`	The number of covariates of x if intercept=F in fgeninter, this number plus 1 if intercept=T
`ord`	The order of the interactions

Value

decom The component parts of the interaction.

Examples

```
decomp(7783,14,8)
```

`fgeninter` *generation of interactions*

Description

generation of interactions

Usage

```
fgeninter(x, ord, intercept = TRUE)
```

Arguments

`x`	Covariates
`ord`	Order of interactions
`intercept`	Logical to include intercept

Value

xx All interactions of order at most ord.

Examples

```
data(boston)
bostinter<-fgeninter(boston[,1:13],7)[[1]]
dim(bostinter)
```

`fgraphst` *Calculates an independence graph for a set of variables using stepwise selection*

Description

Calculates an independence graph for a set of variables using stepwise selection

Usage

```
fgraphst(x, alpha, nu = 1, kmax = 0, intercept = TRUE,
  chkintercept = FALSE)
```

Arguments

`x`	The variables
`alpha`	Cut-off p-value
`nu`	Order statistic
`kmax`	Maximum number selected variables for each node
`intercept`	If true intercept included
`chkintercept`	If true intercept included depending on p-value

Value

ned Number of edges

edg The edges for each node in the graph

Examples

```
data(colonx)
colongrph<-fgraphst(colon.x,0.05)
colongrph[[1]]
colongrph[[2]][1:10,]
```

`fgraphstst` *Calculates an independence graph for a set of variables using repeated stepwise selection*

Description

Calculates an independence graph for a set of variables using repeated stepwise selection

Usage

```
fgraphstst(x, alpha, nu = 1, kmax = 0, nedge = 10^6,
  intercept = TRUE, chkintercept = FALSE)
```

Arguments

`x`	The variables
`alpha`	Cut-off p-value
`nu`	Order statistic
`kmax`	Maximum number selected variables for each node
`nedge`	Maximum number of edges
`intercept`	If true intercept included
`chkintercept`	If true intercept included depending on p-value

Value

ned Number of edges

edg The edges for each node in the graph

Examples

```
data(redwine)
redwgrph<-fgraphstst(redwine[,1:11],0.01)
redwgrph[[1]]
redwgrph[[2]]
```

`flmmdch` *Calculates all possible subsets and selects those where each included covariate is significant.*

Description

It select =TRUE it calls fselect which removes all such subsets which are a subset of some other selected subset. The remaining ones are ordered according to the sum of squared residuals

Usage

```
flmmdch(y, x, p0 = 0.01, select = TRUE, intercept = TRUE,
  chkintercept = FALSE)
```

Arguments

`y`	The dependent variable
`x`	The covariates
`p0`	Cut-off p-value for significance
`select`	If true use fselect
`intercept`	If true intercept included
`chkintercept`	Include intercept depending on p-value

Value

nv List of subsets with number of covariates and sum of squared residuals

Examples

```
data(redwine)
flmmdch(redwine[,12],redwine[,1:11])
```

`frobreg` *Robust regression using Huber's psi-function*

Description

Robust regression using Huber's psi-function

Usage

```
frobreg(y, x, cn, cpp = 0, sig = 0, intercept = TRUE)
```

Arguments

y	Dependent variable
x	Covariates
cn	Tuning parameter for Huber's psi-function
cpp	Fisher consistency parameter
sig	Scale
intercept	Logical to include intercept

Value

beta Regression coefficients

res Residuals

sig Scale

Examples

```
data(boston)
bostrob<-frobreg(boston[,14],boston[,1:13],1)
bostrob[[1]]
plot(bostrob[[2]])
```

---

frobstepwise *robust stepwise selection of covariates*

Description

robust stepwise selection of covariates

Usage

```
frobstepwise(y, x, cn, alpha, nu = 1, kmax = 0, intercept = TRUE,
  chkintercept = FALSE, kexk = 0, sig = 0)
```

Arguments

y	Dependent variable
x	Covariates
cn	Parameter fir Huber's psi-function
alpha	The P-value cut-off
nu	The order statistic of Gaussian covariates used for comparison
kmax	The maximum number of included covariates
intercept	Logical to include intercept
chkintercept	Logical to include or exclude intercept dependent on the P-value
kexk	The excluded covariates
sig	Scale value of dependent variable

Value

pv pv[[1]] the included covariates, the P-values; pv[[2]] coefficients of robust linear regression; pv[[3]] residuals; pv[[4]] scale.

Examples

```
data(boston)
bostonrob<-frobstepwise(boston[,14],boston[,1:13],1,0.01,15,intercept=TRUE)
bostonrob[[1]]
bostonrob[[2]]
plot(bostonrob[[3]])
```

---

fselect *Selects the subsets specified by flmmdch. It is called by flmmdch*

Description

All subsets which are a subset of a specified subset are removed. The remaining subsets are ordered by the sum of squares of the residuals

Usage

```
fselect(nv, k)
```

Arguments

nv	The subsets specified by flmmdch
k	The variables

Value

ind The selected subsets.

Examples

```
data(redwine)
nv<-flmmdch(redwine[,12],redwine[,1:11])[[1]]
fselect(nv,13)
```

`fsimords` *Simulates the number of false positives for given dimensions (n,k) and given order statistic nu*

Description

Simulates the number of false positives for given dimensions (n,k) and given order statistic nu

Usage

```
fsimords(n, k, alpha, nu, kmax, nsim = 100)
```

Arguments

`n`	The dimension of dependent variable
`k`	The number of covariates
`alpha`	Cut-off p-value
`nu`	Order statistic
`nsim`	Number of simulations
`kmax`	Maximum number of false positives

Value

res Histogram of number of false positives.

mn Mean number of false positives.

ss Standard deviation of number of false positives

Examples

```
fsimords(100,1000,0.05,3,6)
```

`fstepstepwise` *Repeated stepwise selection of covariates*

Description

Repeated stepwise selection of covariates

Usage

```
fstepstepwise(y, x, alpha, nu = 1, kmax = 0, kexk = 0,
  lmax = 10^10, intercept = TRUE, chkintercept = FALSE,
  misclass = FALSE)
```

Arguments

y	Dependent variable
x	Covariates
alpha	The P-value cut-off
nu	The order statistic of Gaussian covariates used for comparison
kmax	The maximum number of included covariates
kexk	The excluded covariates
lmax	The maximum number of linear approximations
intercept	Logical to include intercept
chkintercept	Logical to include or exclude intercept dependent on the P-value
misclass	Logical giving the number of misclassifications if appropriate, eg for binary y

Value

pv In order, the number of linear approximation, the included covariates, the P-values, sum of squared residuals and if appropriate number of misclassifications.

Examples

```
data(leukemiax)
data(leukemiay)
fstepstepwise(ly.original,lx.original,0.01,lmax=10,misclass=TRUE)
```

---

fstepwise *Stepwise selection of covariates*

Description

Stepwise selection of covariates

Usage

```
fstepwise(y, x, alpha, nu = 1, kmax = 0, kexk = 0,
  intercept = TRUE, chkintercept = FALSE, misclass = FALSE)
```

Arguments

y	Dependent variable
x	Covariates
alpha	The P-value cut-off
nu	The order statistic of Gaussian covariates used for comparison
kmax	The maximum number of included covariates
kexk	The excluded covariates
intercept	Logical to include intercept
chkintercept	Logical to include or exclude intercept dependent on the P-value
misclass	Logical The number of misclassifications if appropriate, eg for binary y

Value

pv The selected covariates in order together with P-values, sum of squared residuals and if appropriate number of misclassifications.

Examples

```
data(leukemiax)
data(leukemiay)
fstepwise(ly.original,lx.original,0.01,misclass=TRUE)
```

`lx.original` *Leukemia data*

Description

The measurements of gene expression of 3571 genes.

Usage

```
lx.original
```

Format

A 72 x 3571 matrix

Source

http://stat.ethz.ch/~dettling/bagboost.html

References

Boosting for tumor classification with gene expression data. Dettling, M. and B\"uhlmann, P. Bioinformatics, 2003,19(9):1061–1069.

`ly.original` *Leukemia data*

Description

The 72 persons, 25 with leukemia (=1) and 47 controls (=0).

Usage

```
ly.original
```

Format

A column vector of length 72

Source

http://stat.ethz.ch/~dettling/bagboost.html

References

Boosting for tumor classification with gene expression data. Dettling, M. and B\"uhlmann, P. Bioinformatics, 2003,19(9):1061–1069.

`prostate.x` *Prostate cancer data*

Description

The measurements of gene expression of 6033 genes of 102 persons

Usage

```
prostate.x
```

Format

A 6033 x 102 matrix

Source

https://stat.ethz.ch/~dettling/bagboost.html

References

Booting for tumor classification with gene expression data. Dettling, M. and B\"uhlmann, P., Bioinformatics, (2003),19(9):1061–1069.

prostate.y *Prostate cancer data*

Description

A vector of length 102 indicating healthy (=0) and cancer patients (=1)

Usage

```
prostate.y
```

Format

A 0-1 vector of length 102

Source

https://stat.ethz.ch/~dettling/bagboost.html

References

Booting for tumor classification with gene expression data. Dettling, M. and B\"uhlmann, P., Bioinformatics, (2003),19(9):1061–1069.

redwine *Redwine data*

Description

The the subjective quality of wine on an integer scale from 1-10 (variable 12) together with 11 physicochemical properties

Usage

```
redwine
```

Format

A matrix of size 1599 x 12

Source

https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/

References

Modeling wine preferences by data mining from physicochemical properties, Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J., Decision Support Systems, Elsevier, 2009,47(4):547–553.

`stackloss` *Stack loss data*

Description

Stack loss (column 4) together with Air.Flow, Water.Temp, Acid.Conc. columns 1-3 respectively.

Usage

```
stackloss
```

Format

A 21 x 4 matrix

Source

R

References

Brownlee, K. A. (1960, 2nd ed. 1965) Statistical Theory and Methodology in Science and Engineering. New York: Wiley. pp. 491–500.

Index