Package 'gausscov'

February 2, 2023

Version 1.0.2

Date 2023-2-1

Title The Gaussian Covariate Method for Variable Selection

Author Laurie Davies [aut, cre]

Maintainer Laurie Davies <pldavies44@cantab.net>

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 web site below and the accompanying papers: L. Davies and L. Duembgen, ``Covariate Selection Based on a Model-free Approach to Linear Regression with Exact Probabilities", 2022, <arxiv:2202.01553>. L. Davies, ``Linear Regression, Covariate Selection and the Failure of Modelling", 2022, <arXiv:2112.08738>.

LazyData true

License GPL-3

Depends R (>= 3.5.0), stats

Encoding UTF-8

RoxygenNote 6.1.1

NeedsCompilation yes

Repository CRAN

Date/Publication 2023-02-02 18:10:02 UTC

R topics documented:

abcq . . . 2
boston . . . 3
f1st . . . 4
f2st . . . 5
f3st . . . 6
f3sti . . . 7
fasb . . . 8
fdecode . . . 9
fgeninter . . . 9
fgentrig . . . 10
fgr1st . . . 10
fgr2st . . . 11
flag . . . 12
fpval . . . 13
fselect . . . 13
fundr . . . 14
fvauto . . . 15
leukemia . . . 16
mel-temp . . . 16
redwine . . . 17
simgpval . . . 17
snspt . . . 18

Index **19**

---

abcq *American Business Cycle*

Description

The 22 variables are quarterly data from 1919-1941 and 1947-1983 of the variables GNP72, CPRATE, CORPYIELD, M1, M2, BASE, CSTOCK, WRICE67, PRODUR72, NONRES72, IRES72, DBUSI72, CDUR72, CNDUR72, XPT72, MPT72, GOVPUR72,NCSPDE72, NCSBS72, NCSCON72, CCSPDE72 and CCSBS72.

Usage

```
abcq
```

Format

A matrix of size 240 x 22

Source

http://data.nber.org/data/abc/

`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

f1st *Stepwise selection of covariates*

Description

Stepwise selection of covariates

Usage

```
f1st(y,x,p0=0.01,kmn=0,kmx=0,mx=21,kex=0,sub=T,inr=T,xinr=F,qq=0)
```

Arguments

y	Dependent variable
x	Covariates
p0	The P-value cut-off
kmn	The minimum number of included covariates irrespective of cut-off P-value
kmx	The maximum number of included covariates irrespective of cut-off P-value.
mx	The maximum number covariates for an all subset search
kex	The excluded covariates
sub	Logical if TRUE best subset selected
inr	Logical if TRUE include intercept if not present
xinr	Logical if TRUE intercept already present
qq	The number of covariates to choose from. If qq=0 the number of covariates of x is used.

Value

pv In order the included covariates, the regression coefficient values, the Gaussian P-values, the standard P-values and the proportional reduction in the sum of squared residuals due to this covariate

res The residuals

stpv The in order stepwise P-values, sum of squared residuals and the proportional reduction in the sum of squared residuals due to this covariate.

Examples

```
data(boston)
bostint<-fgeninter(boston[,1:13],2)[[1]]
a<-f1st(boston[,14],bostint,kmn=10,sub=TRUE)
```

f2st *Repeated stepwise selection of covariates*

Description

Repeated stepwise selection of covariates

Usage

```
f2st(y,x,p0=0.01,kmn=0,kmx=0,kex=0,mx=21,lm=9^9,

sub=T,inr=T,xinr=F,qq=0)
```

Arguments

y	Dependent variable
x	Covariates
p0	The P-value cut-off
kmn	The minimum number of included covariates irrespective of cut-off P-value
kmx	The maximum number of included covariates irrespective of cut-off P-value.
kex	The excluded covariates
mx	The maximum number of covariates for an all subset search
lm	The maximum number of linear approximations
sub	Logical if TRUE select the best subset
inr	Logical if TRUE include an intercept
xinr	Logical if TRUE intercept already included
qq	The number of covariates to choose from. If qq=0 the number of covariates of x is used.

Value

pv In order the linear approximation, the included covariates, the regression coefficient values, the Gaussian P-values, the standard P-values and the proportional reduction in the sum of squared residuals due to this covariate.

Examples

```
data(boston)
bostint<-fgeninter(boston[,1:13],2)[[1]]
a<-f2st(boston[,14],bostint,lm=3)
```

f3st *Stepwise selection of covariates*

Description

Stepwise selection of covariates

Usage

```
f3st(y,x,m,kmn=10,p0=0.01,kmx=0,mx=21,lm=100,kex=0,sub=T,inr=T,xinr=F,qq=0,kexmx=100)
```

Arguments

y	Dependent variable
x	Covariates
m	The number of iterations
kexmx	The maximum number of covariates in an approximation
p0	The P-value cut-off
kmn	The minimum number of included covariates irrespective of cut-off P-value
kmx	The maximum number of included covariates irrespective of cut-off P-value.
mx	The maximum number covariates for an all subset search
lm	The maximum number of approximations.
kex	The excluded covariates
sub	Logical if TRUE best subset selected
inr	Logical if TRUE include intercept if not present
xinr	Logical if TRUE intercept already present
qq	The number of covariates to choose from. If qq=0 the number of covariates of x is used.

Value

covch The sum of squared residuals and the selected covariates ordered in increasing size of sum of squared residuals.

lai The number of rows of covch

Examples

```
data(leukemia)
a<-f3st(leukemia[[1]],leukemia[[2]],m=2,kmn=5,sub=TRUE,kexmx=5)
```

f3sti *Selection of covariates with given excluded covariates*

Description

Selection of covariates with given excluded covariates

Usage

```
f3sti(y,x,covch,ind,m,kexmx=100,p0=0.01,kmn=0,kmx=0,

 mx=21,lm=1000,kex=0,sub=T,inr=T,xinr=F,qq=0,lm0=0)
```

Arguments

y	Dependent variable
x	Covariates
covch	Sum of squared residuals and selected covariates
ind	The excluded covariates
m	Number of iterations
kexmx	The maximum number of covariates in an approximation.
p0	The P-value cut-off
kmn	The minimum number of included covariates irrespective of cut-off P-value
kmx	The maximum number of included covariates irrespective of cut-off P-value.
mx	The maximum number covariates for an all subset search
lm	The maximum number of approximations.
kex	The excluded covariates
sub	Logical if TRUE best subset selected
inr	Logical if TRUE include intercept if not present
xinr	Logical if TRUE intercept already present
qq	The number of covariates to choose from. If qq=0 the number of covariates of x is used.
lm0	The current number of approximations

Value

ind1 The excluded covariates

covch The sum of squared residuals and the selected covariates ordered in increasing size of sum of squared residuals

lm0 The current number of approximations.

Examples

```
data(leukemia)
covch=c(2.023725,1182,1219,2888,0)
covch<-matrix(covch,nrow=1,ncol=5)
ind<-c(1182,1219,2888)
ind<-matrix(ind,nrow=3,ncol=1)
m<-1
a<-f3sti(leukemia[[1]],leukemia[[2]],covch,ind,m,kexmx=5)
```

---

fasb *Calculates all subsets where each included covariate is significant.*

Description

It sel =TRUE it calls fselect.R and 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. Subsets can be decoded with decode.R.

Usage

```
fasb(y,x,p0=0.01,q=-1,ind=0,sel=T,inr=T,xinr=F)
```

Arguments

y	The dependent variable
x	The covariates
p0	Cut-off p-value for significance
q	The number of covariates from which to choose. Equals number of covariates minus length of ind if q=-1.
ind	The indices of a subset of covariates for which all subsets are to be considered
sel	If TRUE calls fselect.R to removes all subsets of chosen sets
inr	If TRUE to include intercept
xinr	If TRUE intercept already included

Value

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

Examples

```
data(redwine)
nvv<-fasb(redwine[,12],redwine[,1:11])
```

`fdecode` *Decodes the number of a subset selected by fasb.R to give the covariates*

Description

Decodes the number of a subset selected by fasb.R to give the covariates

Usage

```
fdecode(ns, k)
```

Arguments

`ns`	The number of the subset
`k`	The number of covariates

Value

ind The list of covariates

set A binary vector giving the covariates

Examples

```
a<- fdecode(19,8)
```

`fgeninter` *Generation of interactions*

Description

Generates all interactions of degree at most ord

Usage

```
fgeninter(x,ord)
```

Arguments

`x`	Covariates
`ord`	Order of interactions

Value

xx All interactions of order at most ord.

intx Decomposes a given interaction covariate of xx

Examples

```
data(boston)
bostint<-fgeninter(boston[,1:13],2)
```

`fgentrig` *Generation of sine and cosine functions*

Description

Generates sin(pi*j*(1:n)/n) (odd) and cos(pi*j*(1:n)/n) (even) for j=1,...,m for a given sample size n.

Usage

```
fgentrig(n,m)
```

Arguments

n	Sample size
m	Maximum order of sine and cosine functions

Value

x The functions sin(pi*j*(1:n)/n) (odd) and cos(pi*j*(1:n)/n) (even) for j=1,...,m.

Examples

```
trig<-fgentrig(36,36)
```

`fgr1st` *Calculates a dependence graph using Gaussian stepwise selection*

Description

Calculates an independence graph using Gaussian stepwise selection

Usage

```
fgr1st(x,p0=0.01,ind=0,kmn=0,kmx=0,nedge=10^5,inr=T,xinr=F)
```

Arguments

x	The matrix of covariates
p0	Cut-off P-value
ind	Restricts the dependent nodes to this subset
kmn	The minimum number selected variables for each node irrespective of cut-off P-value
kmx	The maximum number selected variables for each node irrespective of cut-off P-value
nedge	Maximum number of edges
inr	Logical, if TRUE include an intercept
xinr	Logical, if TRUE intercept already included

Value

ned Number of edges

edg List of edges together with P-values for each edge and proportional reduction of sum of squared residuals.

Examples

```
data(boston)
a<-fgr1st(boston[,1:13],ind=3:6)
```

---

fgr2st *Calculates an independence graph using repeated stepwise selection*

Description

Calculates a dependency graph using repeated Gaussian stepwise selection

Usage

```
fgr2st(x,p0=0.01,ind=0,kmn=0,kmx=0,nedge=10^5,inr=T,xinr=F)
```

Arguments

x	Matrix of covariates
p0	Cut-off P-value
ind	Restricts the dependent nodes to this subset
kmn	The minimum number of selected variables for each node irrespective of cut-off P-value
kmx	The maximum number of selected variables for each node irrespective of cut-off P-value
nedge	Maximum number of edges
inr	Logical, if TRUE include an intercept
xinr	Logical, if TRUE intercept already included

Value

ned Number of edges

edg List of edges giving nodes (covariates), the approximations for each node, the covariates in the approximation and the corresponding P-values.

Examples

```
data(redwine)
a<-fgr2st(redwine[,1:11],ind=4:8)
```

`flag` *Calculation of lagged covariates*

Description

Calculation of lagged covariates

Usage

```
flag(x,n,i,lag)
```

Arguments

`x`	The covariates
`n`	The sample size
`i`	The dependent variable
`lag`	The maximum lag

Value

y The ith covariate of x without a lag, the dependent varaible.

xl The covariates with lags from 1 :lag starting with the first covariate.

Examples

```
data(abcq)
abcql<-flag(abcq,240,1,16)
a<-f1st(abcql[[1]],abcql[[2]])
```

fpval — *Calculates the regression coefficients, the P-values and the standard P-values for the chosen subset ind*

Description

Calculates the regression coefficients, the P-values and the standard P-values for the chosen subset ind.

Usage

```
fpval(y,x,ind,q=-1,inr=T,xinr=F)
```

Arguments

y	The dependent variable
x	The covariates
ind	The indices of the subset of the covariates whose P-values are required
q	The total number of covariates from which ind was chosen. If q=-1 the number of covariates of x minus length ind plus 1 is taken.
inr	Logical If TRUE intercept to be included
xinr	If TRUE intercept already included

Value

apv In order the subset ind, the regression coefficients, the Gaussian P-values, the standard P-values and the proportion of sum of squares explained.

res The residuals

Examples

```
data(boston)
a<-fpval(boston[,14],boston[,1:13],c(1,2,4:6,8:13))
```

fselect — *Selects the subsets specified by fasb.R and frasb.R.*

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 (fasb.R) or the scale (frasb.R)

Usage

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

Arguments

nv	The subsets specified by fasb.R or frasb.R
k	The variables

Value

ind The selected subsets.

Examples

```
b<-fasb(redwine[,12],redwine[,1:5 ],sel=FALSE)[[1]]
a<-fselect(b,11)[[1]]
b[a,]
```

---

fundr *Converts directed into an undirected graph*

Description

Conversion of a directed graph into an undirected graph

Usage

```
fundr(gr)
```

Arguments

gr	A directed graph

Value

gr The undirected graph

Examples

```
data(boston)
grb<-fgr1st(boston[,1:13])
grbu<-fundr(grb[[2]][,1:2])
```

`fvauto` *Vector autoregressive approximation*

Description

Vector autoregressive approximation

Usage

```
fvauto(x,n,omx,p0=0.01,kmn=0,kmx=0,mx=21,kex=0,sub=T,inr=TRUE)
```

Arguments

`x`	Variable
`n`	Sample size
`omx`	Maximum lag
`p0`	The P-value cut-off
`kmn`	Minimum number of included covariates irrespective of cut-off P-value
`kmx`	Maxmum number of included covariates irrespective of cut-off P-value
`mx`	The maximum number covariates for an all subset search
`kex`	The excluded covariates
`sub`	Logical, if TRUE best subset selected
`inr`	Logical, if TRUE include intercept if not present

Value

res The selected lagged variables for each variable res2 The regression coefficients and P-values res4 The residuals

Examples

```
data(abcq)
 a<-fvauto(abcq,240,16)
```

leukemia *Leukemia data*

Description

Dataset of $n = 72$ persons indicating presence or absence of leukemia and $q = 3571$ gene expressions of the 72 persons

Usage

```
data(leukemia)
```

Format

itemlekemia[[1]]0-1 vector of length giving presence or absence of leukemia itemleukemia[[2]]72x3571 matrix giving the gene exprassions of the 72 persons

Source

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

References

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

mel-temp *Melbourne minimum temperature*

Description

The daily minimum temperature in Melbourne for the years 1981-1990.

Usage

```
mel_temp
```

Format

A vector of length 3650

Source

https://www.kaggle.com/paulbrabban/daily-minimum-temperatures-in-melbourne

redwine *Redwine data*

Description

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.

simgpval *Simulates Gaussian P-values*

Description

Simulates Gaussian P-values

Usage

```
simgpval(y,x,i,nsim, plt=TRUE)
```

Arguments

y	Dependent variable
x	Covariates
i	The chosen covariate
nsim	Number of simulations
plt	Logical, if TRUE the F P-values of the Gaussian covariates are ordered and plotted

Value

pvg P-value of x_i and relative frequency

Examples

```
data(snspt)
a<-flag(snspt,3253,1,12)
simgpval(a[[1]],a[[2]],7,10,plt=FALSE)
```

`snspt` *Sunspot data*

Description

The average number of sunspots each month from January 1749 to January 2020

Usage

```
snspt
```

Format

A vector of size 3253

Source

WDC-SILSO, Royal Observatory of Belgium, Brussels

Index

∗ **datasets**
abcq, 2
boston, 3
leukemia, 16
mel-temp, 16
redwine, 17
snspt, 18

abcq, 2

boston, 3

f1st, 4
f2st, 5
f3st, 6
f3sti, 7
fasb, 8
fdecode, 9
fgeninter, 9
fgentrig, 10
fgr1st, 10
fgr2st, 11
flag, 12
fpval, 13
fselect, 13
fundr, 14
fvauto, 15

leukemia, 16

mel-temp, 16
mel_temp *(*mel-temp*)*, 16

redwine, 17

simgpval, 17
snspt, 18