Package 'gausscov'

September 14, 2020

Version 0.0.7

Date 2020-09-12

Title The Gaussian Covariate Method for Variable Selection

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, ``Covariate Selection Based on a Model-free Approach to Linear Regression with Exact Probabilities", 2020, <arXiv:1906.01990>. L. Davies, ``Lasso, Knockoff and Gaussian covariates: A comparison", 2018, <arXiv:1807.09633>.

LazyData true

License GPL-3

Depends R (>= 3.5.0), stats

Encoding UTF-8

RoxygenNote 6.1.1

NeedsCompilation yes

Repository CRAN

Date/Publication 2020-09-14 12:00:02 UTC

R topics documented:

`abcql` *American Business Cycle*

Description

The 22 variables are quarterly data from 1919-1941 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. Each of these is given with 16 lags.

Usage

```
abcql
```

Format

A matrix of size 224 x 353

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

`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

```
a<-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=FALSE in fgeninter, this number plus 1 if intercept=TRUE
`ord`	The order of the interactions

Value

decom The component parts of the interaction.

Examples

```
a<-decomp(7783,14,8)
```

`dentx` *Dental data, the 20 covariates*

Description

These are the covariates in a 8x3x5 ANOVA table together with 7 interaction terms. There were eight different gold alloys, three different methods of preparation and five different dentists who prepared the filling .The covariates 1:7 are the gold used, the covariate 8:9 are the method of condensation, the covariates 10:13 are the dentists. The covariates 14:20 are interaction terms for observations 14, 90, 93,96, 103 119 and 120. have interaction terms. with.

Usage

```
dentx
```

Format

A matrix of dimension 120x20

Source

Towards robust analysis of variance, Seheult, A.H. and Tukey, J.W. (2001)

References

Towards robust analysis of variance. Seheult, A.H. and Tukey, J.W. In Saleh, A.K.M.E., editor, Data Analysis from Statistical Foundations: A Festschrift in Honor of the 75th Birthday of D.A.S. Fraser, (2001), 217–244.

`denty` *Dental data, hardness*

Description

This is the dependent variable, the hardness of a dental gold filling in a 8x3x5 ANOVA table with. There were eight different gold alloys, three different methods of preparation and five different dentists who prepared the filling .

Usage

```
denty
```

Format

A vector of length 120

Source

Towards robust analysis of variance, Seheult, A.H. and Tukey, J.W. (2001)

References

Towards robust analysis of variance. Seheult, A.H. and Tukey, J.W. In Saleh, A.K.M.E., editor, Data Analysis from Statistical Foundations: A Festschrift in Honor of the 75th Birthday of D.A.S. Fraser, (2001), 217–244.

`f1st` *Stepwise selection of covariates*

Description

Stepwise selection of covariates

Usage

```
f1st(y,x,p0=0.01,nu=1,km=0,mx=20,kx=0,sub=F,inr=T,xinr=F)
```

Arguments

`y`	Dependent variable
`x`	Covariates
`p0`	The P-value cut-off
`nu`	The order statistic of Gaussian covariates used for comparison
`km`	The maximum number of included covariates
`mx`	The maximum number covariates for an all subset search
`kx`	The excluded covariates
`sub`	Logical if TRUE best subset selected.to include intercept
`inr`	Logical if TRUE include intercept if not present
`xinr`	Logical if TRUE intercept already presentt

Value

pv The in order selected covariates, the regression coefficients, the P-values, the standard P-values.

res The residuals

stpv The in order stepwise P-values, sum of squared residuals and the percentage sum of squared residuals explained

Examples

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

`f2st` *Repeated stepwise selection of covariates*

Description

Repeated stepwise selection of covariates

Usage

```
f2st(y,x,p0=0.01,nu=1,km=0,mx=20,kx=0,lm=9^9,sub=F,inr=T,xinr=F)
```

Arguments

`y`	Dependent variable
`x`	Covariates
`p0`	The P-value cut-off
`nu`	The order statistic of Gaussian covariates used for comparison
`km`	The maximum number of included covariates
`lm`	The maximum number of linear approximations
`mx`	The maximum number of covariates for an all subset search
`kx`	The excluded covariates
`sub`	Logical if TRUE choose the best subset
`inr`	Logical if TRUE include intercept
`xinr`	Logical if TRUE intercept already included

Value

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

Examples

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

fgeninter *generation of interactions*

Description

Generates all interactions of degree at most ord

Usage

```
fgeninter(x, ord, inr = T)
```

Arguments

x	Covariates
ord	Order of interactions
inr	Logical to include intercept

Value

xx All interactions of order at most ord.

Examples

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

fgentrig *generation of sine and cosine functions*

Description

Generates the sine and cosine functions of order j, j=1,...,m.

Usage

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

Arguments

n	Sample size
m	m sine and m cosine functions

Value

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

Examples

```
trig100<-fgentrig(100,50)[[1]]
```

`fgr1st` *Calculates an independence graph using stepwise selection*

Description

Calculates an independence graph using stepwise selection

Usage

```
fgr1st(x,p0=0.01,km=0,nu=1,nedge=10^6,inr=T,dr=F)
```

Arguments

`x`	The variables
`p0`	Cut-off p-value
`km`	Maximum number selected variables for each node
`nu`	Order statistic
`nedge`	Maximum number of edges
`inr`	Logical, if TRUE include an intercept
`dr`	Logical, if TRUE (a,b) and (b,a), a not equal b, are different edges

Value

ned Number of edges

edg The edges for each node in the graph

Examples

```
data(boston)
a<-fgr1st(boston[,1:13])
```

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,km=0,nu=1,nedge=10^6,inr=T,dr=F)
```

Arguments

x	The variables
p0	Cut-off P-value
km	Maximum number selected variables for each node
nu	The order statistic of Gaussian covariates used for comparison.
nedge	Maximum number of edges
inr	Logical, if TRUE include an intercept
dr	Logical, if TRUE (a,b) and (b,a), a not equal b, are different edges

Value

ned Number of edges

edg List of edges

Examples

```
data(redwine)
a<-fgr2st(redwine[,1:11])
```

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

Description

It sel =TRUE it calls fselect 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

Usage

```
fmch(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 covarites from which to choose
ind	Indices of subset for which all subsets are to be considered
sel	If TRUE calls fselect to removes all subsets of chosen sets
inr	If TRUE to include intercept
xinr	If TRUE intercept already included

Value

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

Examples

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

---

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,inr=T,xinr=F)
```

Arguments

y	The dependent variable
x	The covariates
ind	The subset of the covariates x for which the P-values are required
q	The total number of covariates from which ind was chosen
inr	Logical If TRUE intercept to be included
xinr	If TRUE intercept already included, overrides inr

Value

apv In order the subset ind, the regression coefficients, the P-values, the standard P-values.

res The residuals

Examples

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

`frmch` *Robust selection of covariates using Huber's psi-funtion or Hampel's redescending psi-function based on all subsets*

Description

Calculates all possible subsets and selects those where each included covariate is significant using a robustified version of flmmdch.R

Usage

```
frmch(y,x,cn=1,cnr=c(2,4,8),p0=0.01,q=-1,sg=0,ind=0,sel=T,inr=T,xinr=F,red=F)
```

Arguments

`y`	Dependent variable
`x`	Covariates
`cn`	Constant for Huber's psi-function
`cnr`	Constants for for Hampel's three part redescending psi-function
`p0`	The P-value cut-off
`q`	The numer of covariates available
`sg`	The scale parameter
`ind`	The subset for which the results are required
`sel`	Logical, if TRUE remove all subsets of chosen sets
`inr`	Logical if TRUE include intercept
`xinr`	Logical If TRUE intercept already included
`red`	Logical If true Hampel's three part redescending psi function

Value

nv List of subsets with number of covariates and scale.

Examples

```
data(boston)
a<-frmch(boston[,14],boston[,1:6])
ind<-decode(57,6)
```

`frrg` *Robust regression using Huber's psi-function or Hampel's redescending psi-function without P-values*

Description

Robust regression using Huber's psi-function or Hampel's redescending psi-function without P-values

Usage

```
frrg(y,x,cn=1,cnr=c(2,4,8),sg=0,scale=T,inr=T,xinr=F,red=F)
```

Arguments

`y`	Dependent variable
`x`	Covariates
`cn`	Tuning parameter for Huber's psi-function
`cnr`	Tuning constants for Hampel's three part redescending psi function
`sg`	Scale
`scale`	Logical, if TRUE calculates sg simultaneously, otherwise keeps initial sg
`inr`	Logical if TRUE to include intercept
`xinr`	Logical if TRUE intercept already included
`red`	Logical If TRUE Hampel's three part redescending psi function

Value

beta Regression coefficients

res Residuals

sg Scale

rho Sums of rho, psi and psi1 functions.

Examples

```
data(boston)
a<-frrg(boston[,14],boston[,1:13])
```

`frrgp` *Robust regression using Huber's psi-function or Hampel's three part redescending psi-function providing P-values*

Description

Robust regression using Huber's psi-function or Hampel's three part redescending psi-function providing P-values

Usage

```
frrgp(y,x,cn=1,cnr=c(2,4,8),sg=0,q=-1,ind=0,scale=T,inr=T,xinr=F,red=F)
```

Arguments

`y`	Dependent variable
`x`	Covariates
`cn`	Tuning constant for Huber's psi-function
`cnr`	Thuning constants for Hampel's three part redescending psi function
`sg`	Scale
`q`	The number of covariates available
`ind`	The subset of covariates for which the results are required
`scale`	Logical. If TRUE sclae sg recaculted
`inr`	Logical, TRUE to include intercept
`xinr`	Logical TRUE if x already includes intercept
`red`	Logical It true Hampel's three part redescending psi function

Value

ppi In order the subset ind, the regression coefficients, the P-values, the standard P-values

res Residuals

sig Scale

rho Sums of rho, psi and psi1 functions.

Examples

```
data(boston)
a<-frrgp(boston[,14],boston[,1:13])
```

frst *Robust stepwise selection of covariates*

Description

Robust stepwise selection of covariates

Usage

```
frst(y,x,cn=1,cnr=c(2,4,8),p0=0.01,sg=0,nu=1,km=0,mx=20,kx=0,sub=F,inr=T,xinr=F,red=F)
```

Arguments

y	Dependent variable
x	Covariates
cn	The constnat for Huber's psi-function
cnr	The constants for Hampel's three part redescending psi function
p0	The P-value cut-off
sg	Scale value of residuals
nu	The order for calculating the P-value
km	The maximum number of included covariates
mx	The maximum number of included covariates if the option subset =TRUE is used
kx	The excluded covariates
sub	Logical, if TRUE best subset selected
inr	Logical TRUE to include intercept
xinr	Logical TRUE if intercept already included
red	Logical If true Hampel's three part redescending psi function

Value

pv In order the subset ind, the regression coefficients, the P-values, the standard P-values.

res The residuals

stpv The stepwise regression results: covariate, P-value and scale

Examples

```
data(boston)
a<-frst(boston[,14],boston[,1:13])
```

fselect — *Selects the subsets specified by fmch.*

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 fmch
k	The variables

Value

ind The selected subsets.

Examples

```
nv<-c(650,1962,160,1033,394,1730,577,1839,334)
nv<-matrix(nv,ncol=3)
a<-fselect(nv,11)
```

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

Description

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

Usage

```
fsimords(n, k, p0 , nu, km, nsim = 500)
```

Arguments

n	The dimension of dependent variable
k	The number of covariates
p0	Cut-off p-value
nu	The order statistics
km	Maximum number of selected covariates
nsim	Number of simulations

Value

p Histogram of number of false positives.

mn Mean number of false positives.

ss Standard deviation of number of false positives

Examples

```
a<-fsimords(100,100,0.01,c(5,10),15,nsim=100)
```

`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 Buehlmann, 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 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.

`snspt` *Sunspot data*

Description

The average number of sunspots each month from January 1749 to January 2020: variable 1 year; variable 2, month; variable 3 number of sunspots.

Usage

```
snspt
```

Format

A matrix of size 3253 x 7

Source

WDC-SILSO, Royal Observatory of Belgium, Brussels

Index