

# Package ‘gausscov’

May 21, 2024

**Version** 1.1.3

**Date** 2024-5-21

**Title** The Gaussian Covariate Method for Variable Selection

**Author** Laurie Davies [aut, cre]

**Maintainer** Laurie Davies <pldavies44@cantab.net>

**Description** The standard linear regression theory whether frequentist or Bayesian is based on an 'assumed (revealed?) truth' (John Tukey) attitude to models. This is reflected in the language of statistical inference which involves a concept of truth, for example confidence intervals, hypothesis testing and consistency. The motivation behind this package was to remove the word true from the theory and practice of linear regression and to replace it by approximation. The approximations considered are the least squares approximations. An approximation is called valid if it contains no irrelevant covariates. This is operationalized using the concept of a Gaussian P-value which is the probability that pure Gaussian noise is better in term of least squares than the covariate. The precise definition given in the paper, it is intuitive and requires only four simple equations. Its overwhelming advantage compared with a standard F P-value is that it is exact and valid whatever the data. In contrast F P-values are only valid for specially designed simulations. Given this a valid approximation is one where all the Gaussian P-values are less than a threshold  $p_0$  specified by the statistician, in this package with the default value 0.01. This approximations approach is not only much simpler it is overwhelmingly better than the standard model based approach. The will be demonstrated using six real data sets, four from high dimensional regression and two from vector autoregression. The simplicity and superiority of Gaussian P-values derive from their universal exactness and validity. This is in complete contrast to standard F P-values which are valid only for carefully designed simulations. The function `f1st` is the most important function. It is a greedy forward selection procedure which results in either just one or no approximations which may however not be valid. If the size is less than a threshold with default value 21 then an all subset procedure is called which returns the best valid subset. A good default start is `f1st(y,x,kmn=15)` The best function for returning multiple approximations is `f3st` which repeatedly calls `f1st`. 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", 2202, <doi:10.48550/arXiv.2202.01553>. L. Davies, "An Approximation Based Theory of Linear Regression", 2402, <doi:10.48550/arXiv.2402.09858>.

**LazyData** true

**License** GPL-3

**Depends** R (>= 3.5.0), stats

**Encoding** UTF-8

**RoxygenNote** 6.1.1

**NeedsCompilation** yes

**Repository** CRAN

**Date/Publication** 2024-05-21 20:20:12 UTC

## R topics documented:

abcq . . . . .	2
boston . . . . .	3
decode . . . . .	4
f1st . . . . .	4
f2st . . . . .	5
f3st . . . . .	6
f3sti . . . . .	7
fasb . . . . .	8
fgeninter . . . . .	9
fgentrig . . . . .	9
fgr1st . . . . .	10
flag . . . . .	11
fpval . . . . .	11
fundr . . . . .	12
leukeia . . . . .	13
mel-temp . . . . .	13
redwine . . . . .	14
simgpval . . . . .	14
snspt . . . . .	15
vardata . . . . .	15
<b>Index</b>	<b>16</b>

---

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, C STOCK, WRICE67, PRODUR72, NONRES72, IRES72, DBUSI72, CDUR72, CNDUR72, XPT72, MPT72, GOVPUR72, NCS PDE72, NCSBS72, NCSCON72, CC-SPDE72 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 river; 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 occupied 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](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	<i>Decodes the number of a subset selected by fasb.R to give the covariates</i>
--------	---

---

**Description**

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

**Usage**

```
decode(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<- decode(19,8)
```

---

f1st	<i>Stepwise selection of covariates</i>
------	---

---

**Description**

Stepwise selection of covariates

**Usage**

```
f1st(y, x, p0=0.01, kmn=0, kmx=0, kex=0, mx=21, sub=T, inr=T, xinr=F, qq=-1)
```

**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 covariates for an all subset search
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=-1 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.

res The residuals

stp The covariates in order of selection, Gaussian P-values and sum of squared residuals.

### 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=-1)
```

### 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=-1 the number of covariates of x is used.

**Value**

pv In order the linear approximation, the included covariates, the Gaussian P-values.

**Examples**

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

f3st

*Stepwise selection of covariates***Description**

Stepwise selection of covariates

**Usage**

```
f3st(y,x,m,p0=0.01,kmn=0,kmx=0,kex=0,mx=21,sub=T,inr=T,xinr=F,qq=-1,kexmx=100)
```

**Arguments**

y	Dependent variable
x	Covariates
m	The number of iterations
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 covariates for an all subset search
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=-1 the number of covariates of x is used.
kexmx	The maximum number of covariates in an approximation.

**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=10)
```

---

 f3sti

*Selection of covariates with given excluded covariates*


---

**Description**

Selection of covariates with given excluded covariates

**Usage**

```
f3sti(y,x,covch,ind,m,p0=0.01,kmn=0,kmx=0,kex=0,mx=21,sub=T,inr=F,xinr=F,qq=-1,kexmx=100)
```

**Arguments**

y	Dependent variable
x	Covariates
covch	Sum of squared residuals and selected covariates
ind	The excluded covariates
m	Number of iterations
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 covariates for an all subset search
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=-1 the number of covariates of x is used.
kexmx	The maximum number of covariates in an approximation.

**Value**

ind1 The excluded covariates

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

**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**

The subset are ordered according to the sum of squared residuals. Subsets can be decoded with decode.R.

**Usage**

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

**Arguments**

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

**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])
```

---

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)[[1]]
```

---

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  $j=1,\dots,m$  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  $j=1,\dots,m$ .

**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,mx=21,nedge=10^5,inr=T,xinr=F,qq=-1)
```

**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
mx	Maximum number of selected covariates for each node for all subset search
nedge	Maximum number of edges
inr	Logical, if TRUE include an intercept
xinr	Logical, if TRUE intercept already included
qq	The number of covariates to choose from. If qq=-1 the number of covariates of x is used

**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)
```

---

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

**Examples**

```
data(abcq)
abcql<-flag(abcq,240,1,16)
a<-fst(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,inr=T,xinr=F,qq=-1)
```

**Arguments**

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

**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))
```

---

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])
```

---

leukeia

*Leukemia data set*

---

**Description**

Dataset of  $n = 72$  persons indicating presence or absence of leukemia (variable 3572) and  $q = 3571$  gene expressions of the 72 persons (variables 1 to 3571)

**Usage**

```
data(leukemia)
```

**Format**

**y** 0-1 data of individuals with and without leukemia.  
**x** covariates of the level of 3571 genes.

**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	<i>Redwine data</i>
---------	---------------------

---

**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	<i>Simulates Gaussian P-values</i>
----------	------------------------------------

---

**Description**

Simulates Gaussian P-values

**Usage**

simgpval(y, x, i, nsim, qq=-1, plt=TRUE)

**Arguments**

y	Dependent variable
x	Covariates
i	The chosen covariate
nsim	The number of simulations
qq	The total number of covariates available
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)
snspt<-matrix(snspt,nrow=3253,ncol=1)
a<-flag(snspt,3253,1,12)
simgpval(a[[1]],a[[2]],7,10,plt=FALSE)
```

---

snspt	<i>Sunspot data</i>
-------	---------------------

---

**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

---

vardata	<i>USA economics data</i>
---------	---------------------------

---

**Description**

United States economic data taken from the FRED-MD macroeconomic database with the NAs removed. 182 indices each of length 256

**Usage**

```
vardata
```

**Format**

A matrix of size 256 X 182

**Source**

<https://research.stlouisfed.org/econ/mccracken/fred-databases>

# Index

## \* datasets

- abcq, 2
- boston, 3
- leukeia, 13
- mel-temp, 13
- redwine, 14
- snspt, 15
- vardata, 15

abcq, 2

boston, 3

decode, 4

f1st, 4

f2st, 5

f3st, 6

f3sti, 7

fasb, 8

fgeninter, 9

fgentrig, 9

fgr1st, 10

flag, 11

fpval, 11

fundr, 12

leukeia, 13

leukemia (leukeia), 13

mel-temp, 13

mel\_temp (mel-temp), 13

redwine, 14

simgpval, 14

snspt, 15

vardata, 15