Package 'FWDselect'

December 19, 2015

Title Selecting Variables in Regression Models

Version 2.1.0

Date 2015-12-18

Author Marta Sestelo [aut, cre],
Nora M. Villanueva [aut],
Javier Roca-Pardinas [aut]

Maintainer Marta Sestelo <sestelo@uvigo.es>

Description A simple method
to select the best model or best subset of variables using
different types of data (binary, Gaussian or Poisson) and
applying it in different contexts (parametric or non-parametric).

URL http://cran.r-project.org/package=FWDselect

BugReports http://github.com/sestelo/fwdselect/issues

Depends R (>= 3.1.0)

License MIT + file LICENSE

LazyData true

Imports cvTools, mgcv, parallel, graphics, stats

RoxygenNote 5.0.0

NeedsCompilation no

Repository CRAN

Date/Publication 2015-12-19 09:34:15

R topics documented:

`diabetes` *Diabetes data.*

Description

The diabetes data is a data frame with 11 variables and 442 measurements. These are the data used in the Efron et al. (2004) paper. The data has been standardized to have unit L2 norm in each column and zero mean.

Usage

```
diabetes
```

Format

`diabetes` is a data frame with 11 variables (columns). The first column of the data frame contains the response variable (`diabetes$y`) which is a quantitative measure of disease progression one year after baseline. The rest of the columns contain the measurements of the ten explanatory variables (age, sex, body mass index, average blood pressure and six blood serum registers) obtained from each of the 442 diabetes patients.

Source

The orginal data are available in the lars package, see http://cran.r-project.org/web/packages/lars/.

References

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression (with discussion). Annals of Statistics, 32:407–499.

Examples

```
library(FWDselect)
data(diabetes)
head(diabetes)
```

episode *Episode of SO2. Pollution incident data.*

Description

Registered values of SO2 in different temporal instant. Each column of the dataset corresponds with the value obtained by the series of bi-hourly means for SO2 in the instant `t` (5-min temporal instant). The values of this dataset are greater than the maximum value permitted for SO2 atmospheric.

Usage

```
episode
```

Format

`episode` is a data frame with 19 variables (columns).

Y response variable, registered values of SO2 at a specific temporal instant, in microg/m3N. This is the value that we want to predict.

In0 registered values of SO2 at a specific temporal instant, in this case instant zero, in microg/m3N.

In1 registered values of SO2 at a specific temporal instant, in this case 5-min instant temporal before, in microg/m3N.

In2 registered values of SO2 at a specific temporal instant, in this case 10-min instant temporal before, in microg/m3N....

Examples

```
data(episode)
head(episode)
```

FWDselect `FWDselect`*: Selecting Variables in Regression Models.*

Description

This package introduces a simple method to select the best model using different types of data (binary, Gaussian or Poisson) and applying it in different contexts (parametric or non-parametric). The proposed method is a new forward stepwise-based selection procedure that selects a model containing a subset of variables according to an optimal criterion (obtained by cross-validation) and also takes into account the computational cost. Additionally, bootstrap resampling techniques are used to implement tests capable of detecting whether significant effects of the unselected variables are present in the model.

Details

Package:	FWDselect
Type:	Package
Version:	2.1.0
Date:	2015-12-18
License:	MIT + file LICENSE

`FWDselect` is just a shortcut for "Forward selection" and is a very good summary of one of the package's major functionalities, i.e., that of providing a forward stepwise-based selection procedure. This software helps the user select relevant variables and evaluate how many of these need to be included in a regression model. In addition, it enables both numerical and graphical outputs to be displayed. The package includes several functions that enable users to select the variables to be included in linear, generalized linear or generalized additive regression models. Users can obtain the best combinations of q variables by means of the main function `selection`. Additionally, if one wants to obtain the results for more than one size of subset, it is possible to apply the `qselection` function, which returns a summary table showing the different subsets, selected variables and information criterion values. The object obtained when using this last function is the argument required for `plot.qselection`, which provides a graphical output. Finally, to determine the number of variables that should be introduced into the model, only the `test` function needs to be applied.

Author(s)

Marta Sestelo, Nora M. Villanueva and Javier Roca-Pardinas.

References

Burnham, K., Anderson, D. (2002). Model selection and multimodel inference: a practical information-theoretic approach. 2nd Edition Springer.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. Annals of Statistics, 7:1-26.

Efron, B. and Tibshirani, R. J. (1993). An introduction to the Bootstrap. Chapman and Hall, London.

Miller, A. (2002). Subset selection in regression. Champman and Hall.

Sestelo, M., Villanueva, N. M. and Roca-Pardinas, J. (2013). FWDselect: Variable selection algorithm in regression models. Discussion Papers in Statistics and Operation Research, University of Vigo, 13/02.

`plot.qselection` *Visualization of* `qselection` *object*

Description

This function plots the cross-validation information criterion for several subsets of size q chosen by the user.

Usage

```
## S3 method for class 'qselection'
plot(x = object, y = NULL, ylab = NULL, ...)
```

Arguments

x	qselection object.
y	NULL
ylab	NULL
...	Other options.

Value

Simply returns a plot.

Author(s)

Marta Sestelo, Nora M. Villanueva and Javier Roca-Pardinas.

See Also

selection.

Examples

```
library(FWDselect)
data(diabetes)
x = diabetes[ ,2:11]
y = diabetes[ ,1]
obj2 = qselection(x, y, qvector = c(1:9), method = "lm", criterion = "variance", cluster = FALSE)
plot(obj2)
```

---

pollution *Emission of SO2. Pollution incident data.*

Description

Registered values of SO2 in different temporal instant. Each column of the dataset corresponds with the value obtained by the series of bi-hourly means for SO2 in the instant t (5-min temporal instant).

Usage

```
pollution
```

Format

`pollution` is a data frame with 19 variables (columns).

Y response variable, registered values of SO2 at a specific temporal instant, in microg/m3N. This is the value that we want to predict.

In0 registered values of SO2 at a specific temporal instant, in this case instant zero, in microg/m3N.

In1 registered values of SO2 at a specific temporal instant, in this case 5-min instant temporal before, in microg/m3N.

In2 registered values of SO2 at a specific temporal instant, in this case 10-min instant temporal before, in microg/m3N. ...

Examples

```
data(pollution)
head(pollution)
```

---

print.qselection *Short* qselection *summary*

Description

qselection summary

Usage

```
## S3 method for class 'qselection'
print(x = object, ...)
```

Arguments

x	qselection object.
...	Other options.

Value

The function returns a summary table with the subsets of size q, their information criterion values and the chosen variables for each one. Additionally, an asterisk is shown next to the size of subset which minimizes the information criterion.

Author(s)

Marta Sestelo, Nora M. Villanueva and Javier Roca-Pardinas.

See Also

selection.

Examples

```
library(FWDselect)
data(diabetes)
x = diabetes[ ,2:11]
y = diabetes[ ,1]
obj2 = qselection(x, y, qvector = c(1:9), method = "lm", criterion = "variance", cluster = FALSE)
obj2
```

`print.selection` *Short* `selection` *summary*

Description

`selection` summary

Usage

```
## S3 method for class 'selection'
print(x = model, ...)
```

Arguments

`x`	`selection` object.
`...`	Other options.

Value

The function returns the best subset of size q and its information criterion value. In the case of `seconds=TRUE` this information is returned for each alternative model.

Author(s)

Marta Sestelo, Nora M. Villanueva and Javier Roca-Pardinas.

See Also

`selection`.

Examples

```
library(FWDselect)
data(diabetes)
x = diabetes[ ,2:11]
y = diabetes[ ,1]
obj1 = selection(x, y, q = 1, method = "lm", criterion = "variance", cluster = FALSE)
obj1
```

qselection *Selecting variables for several subset sizes*

Description

Function that enables to obtain the best variables for more than one size of subset. Returns a table with the chosen covariates to be introduced into the models and their information criteria. Additionally, an asterisk is shown next to the size of subset which minimizes the information criterion.

Usage

```
qselection(x, y, qvector, criterion = "deviance", method = "lm",
  family = "gaussian", nfolds = 5, cluster = TRUE, ncores = NULL)
```

Arguments

x	A data frame containing all the covariates.
y	A vector with the response values.
qvector	A vector with more than one variable-subset size to be selected.
criterion	The information criterion to be used. Default is the deviance. Other functions provided are the coefficient of determination ("R2"), the residual variance ("variance"), the Akaike information criterion ("aic"), AIC with a correction for finite sample sizes ("aicc") and the Bayesian information criterion ("bic"). The deviance, coefficient of determination and variance are calculated by cross-validation.
method	A character string specifying which regression method is used, i.e., linear models ("lm"), generalized additive models ("glm") or generalized additive models ("gam").
family	A description of the error distribution and link function to be used in the model: ("gaussian"), ("binomial") or ("poisson").
nfolds	Number of folds for the cross-validation procedure, for deviance, R2 or variance criterion.
cluster	A logical value. If TRUE (default), the procedure is parallelized. Note that there are cases without enough repetitions (e.g., a low number of initial variables) that R will gain in performance through serial computation. R takes time to distribute tasks across the processors also it will need time for binding them all together later on. Therefore, if the time for distributing and gathering pieces together is greater than the time need for single-thread computing, it does not worth parallelize.
ncores	An integer value specifying the number of cores to be used in the parallelized procedure. If NULL (default), the number of cores to be used is equal to the number of cores of the machine - 1.

Value

q	A vector of subset sizes.
criterion	A vector of Information criterion values.
selection	Selected variables for each size.

Author(s)

Marta Sestelo, Nora M. Villanueva and Javier Roca-Pardinas.

See Also

selection plot.qselection.

Examples

```
library(FWDselect)
data(diabetes)
x = diabetes[ ,2:11]
y = diabetes[ ,1]
obj2 = qselection(x, y, qvector = c(1:9), method = "lm", criterion = "variance", cluster = FALSE)
obj2
```

---

selection *Selecting a subset of* q *variables*

Description

Main function for selecting the best subset of q variables. Note that the selection procedure can be used with lm, glm or gam functions.

Usage

```
selection(x, y, q, prevar = NULL, criterion = "deviance", method = "lm",
  family = "gaussian", seconds = FALSE, nmodels = 1, nfolds = 5,
  cluster = TRUE, ncores = NULL)
```

Arguments

x	A data frame containing all the covariates.
y	A vector with the response values.
q	An integer specifying the size of the subset of variables to be selected.
prevar	A vector containing the number of the best subset of q-1 variables. NULL, by default.

criterion	The information criterion to be used. Default is the deviance. Other functions provided are the coefficient of determination ("R2"), the residual variance ("variance"), the Akaike information criterion ("aic"), AIC with a correction for finite sample sizes ("aicc") and the Bayesian information criterion ("bic"). The deviance, coefficient of determination and variance are calculated by cross-validation.
method	A character string specifying which regression method is used, i.e., linear models ("lm"), generalized additive models ("glm") or generalized additive models ("gam").
family	A description of the error distribution and link function to be used in the model: ("gaussian"), ("binomial") or ("poisson").
seconds	A logical value. By default, FALSE. If TRUE then, rather than returning the single best model only, the function returns a few of the best models (equivalent).
nmodels	Number of secondary models to be returned.
nfolds	Number of folds for the cross-validation procedure, for deviance, R2 or variance criterion.
cluster	A logical value. If TRUE (default), the procedure is parallelized. Note that there are cases without enough repetitions (e.g., a low number of initial variables) that R will gain in performance through serial computation. R takes time to distribute tasks across the processors also it will need time for binding them all together later on. Therefore, if the time for distributing and gathering pieces together is greater than the time need for single-thread computing, it does not worth parallelize.
ncores	An integer value specifying the number of cores to be used in the parallelized procedure. If NULL (default), the number of cores to be used is equal to the number of cores of the machine - 1.

Value

Best model	The best model. If seconds=TRUE, it returns also the best alternative models.
Variable name	Names of the variable.
Variable number	Number of the variables.
Information criterion	Information criterion used and its value.
Prediction	The prediction of the best model.

Author(s)

Marta Sestelo, Nora M. Villanueva and Javier Roca-Pardinas.

Examples

```
library(FWDselect)
data(diabetes)
x = diabetes[ ,2:11]
y = diabetes[ ,1]
```

```
obj1 = selection(x, y, q = 1, method = "lm", criterion = "variance", cluster = FALSE)
obj1

# second models
obj11 = selection(x, y, q = 1, method = "lm", criterion = "variance",
seconds = TRUE, nmodels = 2, cluster = FALSE)
obj11

# prevar argument
obj2 = selection(x, y, q = 2, method = "lm", criterion = "variance", cluster = FALSE)
obj2
obj3 = selection(x, y, q = 3, prevar = obj2$Variable_numbers,
method = "lm", criterion = "variance", cluster = FALSE)
```

`test` *Bootstrap based test for covariate selection*

Description

Function that applies a bootstrap based test for covariate selection. It helps to determine the number of variables to be included in the model.

Usage

```
test(x, y, method = "lm", family = "gaussian", nboot = 50,
  speedup = TRUE, qmin = NULL, unique = FALSE, q = NULL,
  bootseed = NULL, cluster = TRUE, ncores = NULL)
```

Arguments

`x`	A data frame containing all the covariates.
`y`	A vector with the response values.
`method`	A character string specifying which regression method is used, i.e., linear models (`"lm"`), generalized additive models.
`family`	A description of the error distribution and link function to be used in the model: (`"gaussian"`), (`"binomial"`) or (`"poisson"`).
`nboot`	Number of bootstrap repeats.
`speedup`	A logical value. If `TRUE` (default), the testing procedure is computationally efficient since it considers one more variable to fit the alternative model than the number of variables used to fit the null. If `FALSE`, the fit of the alternative model is based on considering the best subset of variables of size greater than `q`, the one that minimizes an information criterion. The size of this subset must be given by the user filling the argument `qmin`.

qmin	By default NULL. If speedup is FALSE, qmin is an integer number selected by the user. To help you select this argument, it is recommended to visualize the graphical output of the plot function and choose the number q which minimizes the curve.
unique	A logical value. By default FALSE. If TRUE, the test is performed only for one null hypothesis, given by the argument q.
q	By default NULL. If unique is TRUE, q is the size of the subset of variables to be tested.
bootseed	Seed to be used in the bootstrap procedure.
cluster	A logical value. If TRUE (default), the testing procedure is parallelized.
ncores	An integer value specifying the number of cores to be used in the parallelized procedure. If NULL (default), the number of cores to be used is equal to the number of cores of the machine - 1.

Details

In a regression framework, let $X_1, X_2, \ldots, X_p$, a set of p initial variables and Y the response variable, we propose a procedure to test the null hypothesis of q significant variables in the model – q effects not equal to zero– versus the alternative in which the model contains more than q variables. Based on the general model

$$Y = m(\mathbf{X}) + \varepsilon \quad \text{where} \quad m(\mathbf{X}) = m_1(X_1) + m_2(X_2) + \ldots + m_p(X_p)$$

the following strategy is considered: for a subset of size q, considerations will be given to a test for the null hypothesis

$$H_0(q) : \sum_{j=1}^{p} I_{\{m_j \neq 0\}} \leq q$$

vs. the general hypothesis

$$H_1 : \sum_{j=1}^{p} I_{\{m_j \neq 0\}} > q$$

Value

A list with two objects. The first one is a table containing

Hypothesis	Number of the null hypothesis tested
Statistic	Value of the T statistic
pvalue	pvalue obtained in the testing procedure
Decision	Result of the test for a significance level of 0.05

The second argument nvar indicates the number of variables that have to be included in the model.

Note

The detailed expression of the formulas are described in HTML help http://cran.r-project.org/web/packages/FWDselect/FWDselect.pdf

Author(s)

Marta Sestelo, Nora M. Villanueva and Javier Roca-Pardinas.

References

Sestelo, M., Villanueva, N. M. and Roca-Pardinas, J. (2013). FWDselect: an R package for selecting variables in regression models. Discussion Papers in Statistics and Operation Research, University of Vigo, 13/01.

See Also

selection

Examples

```
library(FWDselect)
data(diabetes)
x = diabetes[ ,2:11]
y = diabetes[ ,1]
test(x, y, method = "lm", cluster = FALSE, nboot = 5)

## for speedup = FALSE
# obj2 = qselection(x, y, qvector = c(1:9), method = "lm",
# cluster = FALSE)
# plot(obj2) # we choose q = 7 for the argument qmin
# test(x, y, method = "lm", cluster = FALSE, nboot = 5,
# speedup = FALSE, qmin = 7)
```

Index