Package ‘reams’

February 20, 2015

Type Package
Title Resampling-Based Adaptive Model Selection
Version 0.1
Date 2011-12-06
Author Philip Reiss <phil.reiss@nyumc.org> and Lei Huang
<huangracer@gmail.com>
Maintainer Tao Zhang <tao-zhang-1@uiowa.edu>
Depends R (>= 2.9.0), leaps, mgcv
Description Resampling methods for adaptive linear model selection.
These can be thought of as extensions of the Akaike information
criterion that account for searching among candidate models.
License GPL (>= 2)
LazyLoad yes
Repository CRAN
Date/Publication 2012-10-29 08:59:35
NeedsCompilation no

R topics documented:

 reams-package .................................................. 2
 bestmods .......................................................... 2
 cic ................................................................. 3
 cvic ............................................................... 5
 eic ................................................................. 6
 ic.min ............................................................. 7
 scoremods ......................................................... 9
 xy ................................................................. 10

Index 12
Resampling-based adaptive model selection

Description
Resampling methods for adaptive linear model selection. These can be thought of as extensions of the Akaike information criterion that account for searching among candidate models. A number of functions in the package depend crucially on the `leaps` package, whose authors are gratefully acknowledged.

Details
For a complete list of functions type `library(help=reams)`.

Author(s)
Philip Reiss <phil.reiss@nyumc.org> and Lei Huang <huangracer@gmail.com>
Maintainer: Tao Zhang <tao-zhang-1@uiowa.edu>

Find best submodels of a full linear model

Description
This function inputs a table of models produced by `scoremods`, picks out the best models according to a specified information criterion, and (optionally) generates a graphical representation of these models.

Usage
```
bestmods(tbl, ic = "aic", nmods = 10, plot = TRUE,
        labels = dimnames(tbl)[[2]][1:attr(tbl,"npred")],
        cex.axis = 1, las = 2 - all(labels==1:attr(tbl,"npred")),
        xlab = if (las==1) "Predictors" else "",
        ylab = "Criterion value", main = ic, ...
)
```

Arguments
- `tbl` a table of the kind outputted by `scoremods`.
- `ic` the information criterion used to score the models. By default, this is "eic". "AIC", "AICc" and "CVIC" are also available.
- `nmods` maximum number of lowest-scoring models to retain.
- `plot` logical value indicating whether to plot the criterion values for the best models.
- `labels` labels for the predictors, used along the horizontal axis of the plot.
cic

cex.axis, las, xlab, ylab, main
  graphical parameters for the plot; see par.
  ...
  additional graphical parameters passed to plot.

Details

Only models with criterion value equal to or less than that of the null (intercept-only) model are retained, even if there are fewer than nmods such models. If the null model is among the best nmods models and plot = TRUE, the plot includes a dotted line representing the null model.

The defaults for las and labels are intended to make the horizontal axis look sensible, whether or not names for the predictors are provided in tbl. See the example below.

Value

A table consisting of the rows of tbl referring to the models with lowest value of criterion ic.

Author(s)

Philip Reiss <phil.reiss@nyumc.org> and Lei Huang <huangracer@gmail.com>

See Also

scoremods

Examples

data(swiss)
modtable = scoremods(swiss$Fertility, swiss[, -1], nboot=100)

par(mfrow = 1:2)
bestmods(modtable)

# The predictor names may render the above table too wide to be read easily. To remove them, set names = 1:5 in the above call
# to scoremods. Alternatively, modify modtable as follows:
modtable.nonames = modtable
dimnames(modtable.nonames)[[2]][1:5] = 1:5
bestmods(modtable.nonames, main="Same, minus names")

Description

Computes the covariance inflation criterion (CIC) of Tibshirani and Knight (1999) for submodels of a full linear model.
Usage

cic(y, X, nperms = 499, covests = NULL, nullcic = NULL)

Arguments

y
  outcome vector

X
  model matrix. This should not include an intercept column; such a column is added by the function.

nperms
  number of permuted data sets to generate.

covests
  sum of the null-hypothesis covariances between the outcomes and the fitted values for the best linear model of each size. If NULL, covariance is estimated from permuted data.

nullcic
  CIC for the intercept-only model.

Value

A list with components

leaps
  all-subsets regression object (for the unpermuted data) returned by function leaps in package leaps.

covests
  sum of the (estimated) null-hypothesis covariances between the outcomes and the fitted values for the best linear model of each size.

enp
  effective number of parameters for models of each size, as defined by Tibshirani and Knight (1999).

cic
  CIC for each of the models given in the leaps component.

nullcic
  CIC for the intercept-only model.

best
  vector of logicals indicating which predictors are included in the minimum-CIC model.

Author(s)

Philip Reiss <phil.reiss@nyumc.org> and Lei Huang <huangracer@gmail.com>

References


See Also

leaps (in the package of the same name)

Examples

data(swiss)
cicobj = cic(swiss$Fertility, swiss[, -1])
cicobj$best
Description
A model selection criterion proposed by Reiss et al. (2012), which employs cross-validation to estimate the overoptimism associated with the best candidate model of each size.

Usage
\[
\text{cvic}(y, X, \text{nfold} = \text{length}(y), \text{pvec} = 1: (\text{ncol}(X) + 1))
\]

Arguments
- \text{y} \quad \text{outcome vector}
- \text{X} \quad \text{model matrix. This should not include an intercept column; such a column is added by the function.}
- \text{nfold} \quad \text{number of "folds" (validation sets). The sample size must be divisible by this number.}
- \text{pvec} \quad \text{vector of possible dimensions of the model to consider: by default, ranges from 1 (intercept only) to ncol(}X\text{) + 1 (full model).}

Details
CVIC is similar to corrected AIC (Sugiura, 1978; Hurvich and Tsai, 1989), but instead of the nominal model dimension, it substitutes a measure of effective degrees of freedom (edf) that takes best-subset selection into account. The "raw" edf is obtained by cross-validation. Alternatively, one can refine the edf via constrained monotone smoothing, as described by Reiss et al. (2011).

Value
A list with components
- \text{nlogsig2hat} \quad \text{value of the first (non-penalty) term of the criterion, i.e., sample size times log of MLE of the variance, for best model of each dimension in pvec.}
- \text{cv.pen} \quad \text{cross-validation penalty, as described by Reiss et al. (2011).}
- \text{edf, edf.mon} \quad \text{effective degrees of freedom, before and after constrained monotone smoothing.}
- \text{cvic} \quad \text{CVIC based on the raw edf.}
- \text{cvic.mon} \quad \text{CVIC based on edf to which constrained monotone smoothing has been applied.}
- \text{best, best.mon} \quad \text{vectors of logicals indicating which columns of the model matrix are included in the CVIC-minimizing model, without and with constrained monotone smoothing.}

Author(s)
Lei Huang <huangracer@gmail.com> and Philip Reiss <phil.reiss@nyumc.org>
References


See Also

`leaps` in package `leaps` for best-subset selection; `pcls` in package `mgcv` for the constrained monotone smoothing.

Examples

```r
# Predicting fertility from provincial socioeconomic indicators
data(swiss)
cvicobj <- cvic(swiss$fertility, swiss[, -1])
cvicobj$best
cvicobj$bestNmon
```

Extended (bootstrap) information criterion

Description

Model selection by an extended information criterion (EIC), based on nonparametric bootstrapping, was introduced by Ishiguro et al. (1997). This function implements the extension by Reiss et al. (2012) to adaptive linear model selection.

Usage

```r
eic(y, X, nboot, pvec = 1:(ncol(X) + 1), say.which = FALSE, reuse = FALSE)
```

Arguments

- `y` outcome vector
- `X` model matrix. This should not include an intercept column; such a column is added by the function.
- `nboot` number of bootstrap samples.
- `pvec` vector of possible dimensions of the model to consider: by default, ranges from 1 (intercept only) to ncol(X) + 1 (full model).
- `say.which` logical: should the predictors selected for each bootstrap sample be reported?
- `reuse` logical: should the best full-data model of each size be reused in calculating the overoptimism estimate, as opposed to reselecting the best model of each size for each training set?
ic.min

Value

A list with components

- `nlogsig2hat`: value of the first (non-penalty) term of the criterion, i.e., sample size times log MLE of the variance, for best model of each dimension in `pvec`.
- `penalty`: the second (penalty) term of the criterion.
- `eic`: the EIC, i.e., the sum of the previous two components.
- `best`: a vector of logicals indicating which columns of the model matrix are included in the EIC-minimizing model.

Author(s)

Philip Reiss <phil.reiss@nyumc.org> and Lei Huang <huangracer@gmail.com>

References


Examples

```r
# Predicting fertility from provincial socioeconomic indicators
data(swiss)
eicobj <- eic(swiss$fertility, swiss[, -1], nboot=100)
eicobj$best
```

describe ic.min

AIC, corrected AIC and BIC for all-subsets linear regression

Description

Given an outcome vector and model matrix, this function finds the submodel(s) minimizing the Akaike (1973, 1974) information criterion (AIC), a corrected version thereof (Sugiura, 1978; Hurvich and Tsai, 1989), and the Bayesian information criterion (BIC; Schwarz, 1978).

Usage

```r
ic.min(y, X, pvec = 1:(nrow(X) + 1))
```
Arguments

\( y \)  
outcome vector

\( X \)  
model matrix. This should not include an intercept column; such a column is added by the function.

\( \text{pvec} \)  
vector of possible dimensions of the model to consider: by default, ranges from 1 (intercept only) to ncol(\( X \)) + 1 (full model).

Value

A list with components

- nlogsig2hat: value of the first (non-penalty) term of the criterion, i.e., sample size times log of MLE of the variance, for best model of each dimension in pvec.
- aic: lowest AIC for models of each dimension.
- aicc: lowest corrected AIC for models of each dimension.
- bic: lowest BIC for models of each dimension.
- best.aic, best.aicc, best.bic: vectors of logicals indicating which columns of the model matrix are included in the model minimizing AIC, corrected AIC, or BIC.

Author(s)

Philip Reiss <phil.reiss@nyumc.org> and Lei Huang <huangracer@gmail.com>

References


Examples

# Predicting fertility from provincial socioeconomic indicators
data(swiss)
ic.min(swiss$fertility, swiss[, -1])
scoremods

Score best subsets by information criteria

Description

This function uses the `leaps` package to find the best models of each size, and scores each according to AIC, corrected AIC, BIC, EIC and CVIC.

Usage

```r
scoremods(y, X, nboot, nfold=length(y), names=NULL)
```

Arguments

- `y`: outcome vector
- `X`: model matrix. This should not include an intercept column; such a column is added by the function.
- `nboot`: number of bootstrap samples or subsamples.
- `nfold`: number of folds cross validation conduct.
- `names`: vector of names for the columns of `X`. If `NULL`, `names(X)` is used.

Value

A matrix. The first `ncol(X)` columns, essentially the which component of an object outputted by `leaps`, identify which predictors are in each of the best models. The remaining columns provide the AIC, corrected AIC, BIC, EIC, and CVIC for each model. The matrix has an attribute "npred" giving the number of candidate predictors, i.e., `ncol(X)`.

Author(s)

Philip Reiss <phil.reiss@nyumc.org> and Lei Huang <huangracer@gmail.com>

References


See Also

- `bestmods`
- `leaps` (in the package of the same name)

Examples

```r
## see example for bestmods
```
xy

Random generation of linear model matrix and outcomes

Description

This function can be used for simulations to evaluate the performance of linear model selection with independent predictors.

Usage

xy(n, p.all, p.true, R2, beta0 = 5,
    yname = paste("y", p.true, sep = ""),
    xname = paste("x", p.true, p.all, sep = ")

Arguments

n          sample size.

p.all      maximum model dimension, i.e., number of candidate predictors plus 1.

p.true     true model dimension, i.e., number of predictors with nonzero coefficients plus 1.

R2         coefficient of determination for the true model.

beta0      true model intercept; in some contexts this value may be arbitrary.

yname      name for the generated outcome vector.

xname      name for the generated model matrix.

Details

xy simulates entries of a model matrix independently from the standard normal distribution, then simulates outcomes whose mean is simply beta0 plus the sum of the first p.true - 1 predictors. The errors are normal with mean 0 and standard deviation chosen so as to attain the given R2; see Tibshirani & Knight (1999), p. 538.

Value

A list with components X (model matrix, without intercept column) and y (outcome vector).

Author(s)

Philip Reiss <phil.reiss@nyumc.org> and Lei Huang <huangracer@gmail.com>

References

Examples

# Generate 40 vectors of 8 candidate predictors, of which
# (the first) 2 have nonzero coefficients, along with 40 outcomes,
# with $R^2=.8$
# tmp = xy(40, 9, 3, .8)

# As a side effect, the above created objects y5 and X59,
# equal to tmp$y and tmp$X respectively.
# The following lines can then be used to examine how different
# information criteria fare at identifying the true model as "best".
# ic.min(y3, x39)
eic(y3, x39, nboot=100)
cvic(y3, x39)
Index

*Topic package
  reams-package, 2

bestmods, 2, 9

cic, 3
cvic, 5

eic, 6

ic.min, 7

leaps, 6

par, 3
pcls, 6
plot, 3

reams (reams-package), 2
reams-package, 2

scoremods, 2, 3, 9

xy, 10