

# Package 'bestglm'

January 2, 2012

**Type** Package

**Title** Best Subset GLM

**Version** 0.33

**Date** 2011-11-03

**Author** A.I. McLeod and Changjiang Xu

**Maintainer** A.I. McLeod <aimcleod@uwo.ca>

**Depends** R (>= 2.0.0), leaps, lars, ElemStatLearn

**Suggests** xtable

**Description** Best subset glm using AIC, BIC, EBIC, BICq or Cross-Validation. For the normal case, the 'leaps' is used. Otherwise, a slower exhaustive search. The 'xtable' package is needed for vignette 'SimExperimentBICq.Rnw' accompanying this package.

**LazyLoad** yes

**LazyData** yes

**Classification/ACM** G.3, G.4, I.5.1

**Classification/MSC** 62M10, 91B84

**License** GPL (>= 2)

**URL** <http://www.stats.uwo.ca/faculty/aim>

**Repository** CRAN

**Date/Publication** 2011-11-03 21:30:43

**R topics documented:**

bestglm-package	2
AirQuality	3
asbinary	4
bestglm	5
CVd	9
CVDH	11
CVHTF	12
Detroit	13
Fires	15
LOOCV	17
manpower	18
MontesinhoFires	19
oneSdRule	20
print.bestglm	21
SAheart	22
Shao	23
summary.bestglm	24
znuclear	25
zprostate	26
<b>Index</b>	<b>28</b>

---

bestglm-package	<i>bestglm: Best Subset GLM</i>
-----------------	---------------------------------

---

**Description**

Provides new information criterion BICq as well as AIC, BIC and EBIC for selecting the best model. Additionally, various CV algorithms are also provided.

**Details**

Package:	bestglm
Type:	Package
Version:	0.33
Date:	2011-11-03
License:	GLP 2.0 or greater
LazyData:	yes
LazyLoad:	yes

bestglm is the main function. All other functions are utility functions and are not normally invoked. Many examples are provided in the vignettes accompanying this package. The vignettes are produced using the R package Sweave and so R scripts can easily be extracted.

The R package xtable is needed for the vignette in SimExperimentBICq.Rnw.

**Author(s)**

A.I. McLeod and Changjiang Xu

**References**

Xu, C. and McLeod, A.I. (2009). Bayesian Information Criterion with Bernoulli Prior.

**See Also**

[leaps](#)

**Examples**

```
data(zprostate)
train<-(zprostate[zprostate[,10],,])[,-10]
#Best subset using AIC
bestglm(train, IC="AIC")
#Best subset using BIC
bestglm(train, IC="BIC")
#Best subset using EBIC
bestglm(train, IC="BICg")
#Best subset using BICg with g=0.5 (tuning parameter)
bestglm(train, IC="BICg", t=0.5)
#Best subset using BICq. Note BICq with q=0.25 is default.
bestglm(train, IC="BICq")
#Best subset using BICq with q=0.5 (equivalent to BIC)
bestglm(train, IC="BICq", t=0.5)
#Remark: set seed since CV depends on it
set.seed(123321123)
bestglm(train, IC="CV", t=10)
#using HTF method
bestglm(train, IC="CV", CVArgs=list(Method="HTF", K=10, REP=1))
#Best subset, logistic regression
data(SAheart)
bestglm(SAheart, IC="BIC", family=binomial)
#Best subset, factor variables with more than 2 levels
data(AirQuality)
#subset
bestglm(AirQuality, IC="BICq")
```

---

AirQuality

*Daily ozone pollution with meteorological and date inputs*

---

**Description**

This dataset was derived from the R built-in dataset ‘airquality’ by adding date information and deleting all missing values. This dataset is referred to as ‘environmental’ in Cleveland (1993).

**Usage**

```
data(AirQuality)
```

**Format**

A data frame with 111 observations on the following 6 variables.

Solar.R input, a numeric vector

Wind input, a numeric vector

Temp input, a numeric vector

month input, a factor with levels May Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar Apr

weekday input, a factor with levels Sunday Monday Tuesday Wednesday Thursday Friday Saturday

Ozone output, a numeric vector

**Details**

Cleveland (1993, Chapter 5) presents an insightful analysis using co-plots and the scatterplot matrix. Several interesting interactions are noted. For a fixed 'Wind', the effect of 'Solar.R' changes as 'Temp' increases. And for a fixed 'Temp', as 'Wind' decreases, the effect of 'Solar.R' is less.

**Source**

[airquality](#)

**References**

Cleveland, W.S. (1993). Visualizing Data.

**Examples**

```
data(AirQuality)
#Example 1. Find best model
bestglm(AirQuality, IC="BIC")
```

---

asbinary

*Binary representation of non-negative integer*

---

**Description**

A non-negative integer is represented as a binary number. The digits, 0 or 1, of this number are returned in a vector.

**Usage**

```
to.binary(n, k = ceiling(logb(n+1,base=2)))
```

**Arguments**

n	a non-negative integers
k	number of digits to be returned.

**Value**

A vector of length k. The first element is the least significant digit.

**Author(s)**

A.I. McLeod

**Examples**

```
to.binary(63)
to.binary(64)
#sometimes we want to pad result with 'leading' 0's
to.binary(63, k=20)
to.binary(64, k=20)
```

---

 bestglm

---

*Best Subset GLM using Information Criterion or Cross-Validation*


---

**Description**

Best subset selection using 'leaps' algorithm (Furnival and Wilson, 1974) or complete enumeration (Morgan and Tatar, 1972). Complete enumeration is used for the non-Gaussian and for the case where the input matrix contains factor variables with more than 2 levels. The best fit may be found using the information criterion IC: AIC, BIC, EBIC, or BICq. Alternatively, with IC='CV' various types of cross-validation may be used.

**Usage**

```
bestglm(Xy, family = gaussian, IC = "BIC", t = "default", CVArgs = "default", qLevel = 0.99, TopModels =
```

**Arguments**

Xy	Dataframe containing the design matrix X and the output variable y. All columns must be named.
family	One of the glm distribution functions. The glm function is not used in the Gaussian case. Instead for efficiency either 'leaps' is used or when factor variables are present with more than 2 levels, 'lm' may be used.
IC	Information criteria to use: "AIC", "BIC", "BICg", "BICq", "LOOCV", "CV".
t	adjustable parameter for BICg, BICq or CV. For BICg, default is g=t=1. For BICq, default is q=t=0.25. For CV, default the delete-d method with d=ceil(n(1-1/(log n - 1))) and REP=t=1000. The default value of the parameter may be changed by changing t.

CVArgs	Used when IC is set to 'CV'. The default is use the delete-d algorithm with $d = \text{ceil}(n(1-1/(\log n - 1)))$ and $t=100$ repetitions. Note that the number of repetitions can be changed using $t$ . More generally, CVArgs is a list with 3 named components: Method, K, REP, where Method is one of \"HTF\", \"DH\", \"d\" corresponding to using the functions CVHTM (Hastie et al., 2009, K-fold CV), CVDH (adjusted K-fold CV, Davison and Hartigan, 1997) and CVd (delete-d CV with random subsamples, Shao, 1997).
qLevel	the alpha level for determining interval for best q. Larger alpha's result in larger intervals.
TopModels	Finds the best TopModels models.
method	Method used in leaps algorithm for searching for the best subset.
intercept	Default TRUE means the intercept term is always included. If set to FALSE, no intercept term is included. If you want only include the intercept term when it is significant then set IncludeInterceptQ=FALSE and include a column of 1's in the design matrix.
weights	weights
nvmax	maximum number of independent variables allowed. By default, all variables
RequireFullEnumerationQ	Use exhaustive search algorithm instead of 'leaps'
...	Optional arguments which are passed to lm or glm

## Details

In the Gaussian case, the loglikelihood may be written  $\log L = -(n/2)\log(RSS/n)$ , where RSS is the residual sum-of-squares and  $n$  is the number of observations. When the function 'glm' is used, the log-likelihood,  $\log L$ , is obtained using 'logLik'. The penalty for EBIC and BICq depends on the tuning parameter argument,  $t$ . The argument  $t$  also controls the number of replications used when the delete-d CV is used as default. In this case, the parameter  $d$  is chosen using the formula recommended by Shao (1997). See [CVd](#) for more details.

In the binomial GLM, nonlogistic, case the last two columns of  $Xy$  are the counts of 'success' and 'failures'.

Cross-validation may also be used to select the best subset. When cross-validation is used, the best models of size  $k$  according to the log-likelihood are compared for  $k=0,1,\dots,p$ , where  $p$  is the number of inputs. Cross-validation is not available when there are categorical variables since in this case it is likely that the training sample may not contain all levels and in this case we can't predict the response in the validation sample. In the case of GLM, the \"DH\" method for CV is not available.

Usually it is a good idea to keep the intercept term even if it is not significant. See discussion in vignette.

Cross-validation is not available for models with no intercept term or when `force.in` is non-null or when `nvmax` is set to less than the full number of independent variables.

Please see the package vignette for more details and examples.

**Value**

A list with class attribute 'bestglm' and named components:

BestModel	An lm-object representing the best fitted regression.
Title	A brief title describing the algorithm used: CV(K=K), CVadj(K=K), CVd(d=K). The range of q for an equivalent BICq model is given.
Subsets	The best subsets of size, $k=0,1,\dots,p$ are indicated as well the value of the log-likelihood and information criterion for each best subset. In the case of categorical variables with more than 2 levels, the degrees of freedom are also shown.
qTable	Table showing range of q for choosing each possible subset size. Assuming intercept=TRUE, $k=1$ corresponds to model with only an intercept term and $k=p+1$ , where p is the number of input variables, corresponds to including all variables.
Bestq	Optimal q
ModelReport	A list with components: NullModel, LEAPSQ, glmQ, gaussianQ, NumDF, CategoricalQ, Bestk.
BestModels	Variables in the TopModels best list

Methods function 'print.bestglm' and 'summary.bestglm' are provided.

**Author(s)**

C. Xu and A.I. McLeod

**References**

- Xu, C. and McLeod, A.I. (2009). Bayesian Information Criterion with Bernoulli Prior.
- Chen, J. and Chen, Z. (2008). Extended Bayesian Information Criteria for Model Selection with Large Model Space. *Biometrika* 2008 95: 759-771.
- Furnival, G.M. and Wilson, R. W. (1974). Regressions by Leaps and Bounds. *Technometrics*, 16, 499-511.
- Morgan, J. A. and Tatar, J. F. (1972). Calculation of the Residual Sum of Squares for All Possible Regressions. *Technometrics* 14, 317-325.
- Miller, A. J. (2002), *Subset Selection in Regression*, 2nd Ed. London, Chapman and Hall.
- Shao, Jun (1997). An Asymptotic Theory for Linear Model Selection. *Statistica Sinica* 7, 221-264.

**See Also**

[glm](#), [lm](#), [leaps](#) CVHTF, CVDH, CVd

**Examples**

```
#Example 1.
#White noise test.
set.seed(123321123)
p<-25 #number of inputs
```

```

n<-100 #number of observations
X<-matrix(rnorm(n*p), ncol=p)
y<-rnorm(n)
Xy<-as.data.frame(cbind(X,y))
names(Xy)<-c(paste("X",1:p,sep=""),"y")
bestAIC <- bestglm(Xy, IC="AIC")
bestBIC <- bestglm(Xy, IC="BIC")
bestEBIC <- bestglm(Xy, IC="BICg")
bestBICq <- bestglm(Xy, IC="BICq")
NAIC <- length(coef(bestAIC$BestModel))-1
NBIC <- length(coef(bestBIC$BestModel))-1
NEBIC <- length(coef(bestEBIC$BestModel))-1
NBICq <- length(coef(bestBICq$BestModel))-1
ans<-c(NAIC, NBIC, NEBIC, NBICq)
names(ans)<-c("AIC", "BIC", "BICg", "BICq")
ans
# AIC  BIC EBIC BICq
#  3    1   0   0

#Example 2. bestglm with BICq
#Find best model. Default is BICq with q=0.25
data(znuclear) #standardized data.
#Rest of examples assume this dataset is loaded.
out<-bestglm(znuclear, IC="BICq")
out
#The optimal range for q
out$Bestq
#The possible models that can be chosen
out$qTable
#The best models for each subset size
out$Subsets
#The overall best models
out$BestModels
#
#Example 3. Normal probability plot, residuals, best model
ans<-bestglm(znuclear, IC="BICq")
e<-resid(ans$BestModel)
qqnorm(e, ylab="residuals, best model")
#
#To save time, none of the remaining examples are run
## Not run:
#Example 4. bestglm, using EBIC, g=1
bestglm(znuclear, IC="BICg")
#EBIC with g=0.5
bestglm(znuclear, IC="BICg", t=0.5)
#
#Example 5. bestglm, CV
data(zprostate)
train<-(zprostate[zprostate[,10],,])[,-10]
#the default CV method takes too long, set t=10 to do only
# 10 replications instead of the recommended 1000
bestglm(train, IC="CV", t=10)
bestglm(train, IC="CV", CVArgs=list(Method="HTF", K=10, REP=1))

```

```

#Compare with DH Algorithm. Normally set REP=100 is recommended.
bestglm(train, IC="CV", CVArgs=list(Method="DH", K=10, REP=1))
#Compare LOOCV
bestglm(train, IC="LOOCV")
#
#Example 6. Optimal q for manpower dataset
data(manpower)
out<-bestglm(manpower)
out$Bestq
#
#Example 7. Factors with more than 2 levels
data(AirQuality)
bestglm(AirQuality)
#
#Example 8. Logistic regression
data(SAheart)
bestglm(SAheart, IC="BIC", family=binomial)
#BIC agrees with backward stepwise approach
out<-glm(chd~., data=SAheart, family=binomial)
step(out, k=log(nrow(SAheart)))
#but BICq with q=0.25
bestglm(SAheart, IC="BICq", t=0.25, family=binomial)
#
#Cross-validation with glm
#make reproducible results
set.seed(33997711)
#takes about 15 seconds and selects 5 variables
bestglm(SAheart, IC="CV", family=binomial)
#about 6 seconds and selects 2 variables
bestglm(SAheart, IC="CV", CVArgs=list(Method="HTF", K=10, REP=1), family=binomial)
#Will produce an error -- NA
\dontrun{bestglm(SAheart, IC="CV", CVArgs=list(Method="DH", K=10, REP=1), family=binomial)}
\dontrun{bestglm(SAheart, IC="LOOCV", family=binomial)}
#
#Example 9. Model with no intercept term
X<-matrix(rnorm(200*3), ncol=3)
b<-c(0, 1.5, 0)
y<-X%*%b + rnorm(40)
Xy<-data.frame(as.matrix.data.frame(X), y=y)
bestglm(Xy, intercept=FALSE)

## End(Not run)

```

**Description**

The delete-d method for cross-validation uses a random sample of d observations as the validation sample. This is repeated many times.

**Usage**

```
CVd(X, y, d = ceiling(n * (1 - 1/(log(n) - 1))), REP = 100, family = gaussian, ...)
```

**Arguments**

X	training inputs
y	training output
d	size of validation sample
REP	number of replications
family	glm family
...	optional arguments passed to glm or lm

**Details**

Shao (1993, 1997) suggested the delete-d algorithm implemented in this function. In this algorithm, a random sample of  $d$  observations are taken as the validation sample. This random sampling is repeated REP times. Shao (1997, p.234, eqn. 4.5 and p.236) suggests  $d = n(1 - 1/(\log n - 1))$ . This is obtained by taking  $\lambda_n = \log n$  on page 236 (Shao, 1997). As shown in the table Shao's recommended choice of the  $d$  parameter corresponds to validation samples that are typically much larger than used in 10-fold or 5-fold cross-validation. LOOCV corresponds to  $d=1$  only!

n	d	K=10	K=5
50	33	5	10
100	73	10	20
200	154	20	40
500	405	50	100
1000	831	100	200

**Value**

Vector of two components comprising the cross-validation MSE and its sd based on the MSE in each validation sample.

**Author(s)**

A.I. McLeod and C. Xu

**References**

Shao, Jun (1993). Linear Model Selection by Cross-Validation. Journal of the American Statistical Association 88, 486-494.

Shao, Jun (1997). An Asymptotic Theory for Linear Model Selection. Statistica Sinica 7, 221-264.

**See Also**

[bestglm](#), [CVHTF](#), [CVDH](#), [LOOCV](#)

**Examples**

```
#Example 1. delete-d method
#For the training set, n=67. So 10-fold CV is like using delete-d
#with d=7, approximately.
data(zprostate)
train<-(zprostate[zprostate[,10],,])[-10]
X<-train[,1:2]
y<-train[,9]
set.seed(123321123)
CvD(X, y, d=7, REP=10)
#should set to 1000. Used 10 to save time in example.
```

CVDH

*Adjusted K-fold Cross-Validation***Description**

An adjustment to K-fold cross-validation is made to reduce bias.

**Usage**

```
CVDH(X, y, K = 10, REP = 1)
```

**Arguments**

X	training inputs
y	training output
K	size of validation sample
REP	number of replications

**Details**

Algorithm 6.5 (Davison and Hinkley, p.295) is implemented.

**Value**

Vector of two components comprising the cross-validation MSE and its sd based on the MSE in each validation sample.

**Author(s)**

A.I. McLeod and C. Xu

**References**

Davison, A.C. and Hinkley, D.V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press.

**See Also**

[bestglm](#), [CVHTF](#), [CVd](#), [LOOCV](#)

**Examples**

```
#Example 1. Variability in 10-fold CV with Davison-Hartigan Algorithm.
#Plot the CVs obtained by using 10-fold CV on the best subset
#model of size 2 for the prostate data. We assume the best model is
#the model with the first two inputs and then we compute the CV's
#using 10-fold CV, 100 times. The result is summarized by a boxplot as well
#as the sd.
NUMSIM<-10
data(zprostate)
train<-(zprostate[zprostate[,10],,])[-10]
X<-train[,1:2]
y<-train[,9]
cvs<-numeric(NUMSIM)
set.seed(123321123)
for (isim in 1:NUMSIM)
  cvs[isim]<-CVDH(X,y,K=10,REP=1)[1]
summary(cvs)
```

---

 CVHTF

---

*K-fold Cross-Validation*


---

**Description**

K-fold cross-validation.

**Usage**

```
CVHTF(X, y, K = 10, REP = 1, family = gaussian, ...)
```

**Arguments**

X	training inputs
y	training output
K	size of validation sample
REP	number of replications
family	glm family
...	optional arguments passed to glm or lm

**Details**

HTF (2009) describe K-fold cross-validation. The observations are partitioned into K non-overlapping subsets of approximately equal size. Each subset is used as the validation sample while the remaining K-1 subsets are used as training data. When  $K = n$ , where n is the number of observations the algorithm is equivalent to leave-one-out CV. Normally  $K = 10$  or  $K = 5$  are used. When  $K < n - 1$ , there may be many possible partitions and so the results of K-fold CV may vary somewhat depending on the partitions used. In our implementation, random partitions are used and we allow for many replications. Note that in the Shao's delete-d method, random samples are used to select the validation data whereas in this method the whole partition is selected as random. This is accomplished using, `fold <- sample(rep(1:K,length=n))`. Then `fold` indicates each validation sample in the partition.

**Value**

Vector of two components comprising the cross-validation MSE and its sd based on the MSE in each validation sample.

**Author(s)**

A.I. McLeod and C. Xu

**References**

Hastie, T., Tibshirani, R. and Friedman, J. (2009). The Elements of Statistical Learning. 2nd Ed. Springer-Verlag.

**See Also**

[bestglm](#), [CvD](#), [CVDH](#), [LOOCV](#)

**Examples**

```
#Example 1. 10-fold CV
data(zprostate)
train<-(zprostate[zprostate[,10],,],-10]
X<-train[,1:2]
y<-train[,9]
CVHTF(X,y,K=10,REP=1)[1]
```

**Description**

For convenience we have labelled the input variables 1 through 11 to be consistent with the notation used in Miller (2002). Only the first 11 variables were used in Miller's analyses. The best fitting subset regression with these 11 variables, uses only 3 inputs and has a residual sum of squares of 6.77 while using forward selection produces a best fit with 3 inputs with residual sum of squares 21.19. Backward selection and stagewise methods produce similar results. It is remarkable that there is such a big difference. Note that the usual forward and backward selection algorithms may fail since the linear regression using 11 variables gives essentially a perfect fit.

**Usage**

```
data(Detroit)
```

**Format**

A data frame with 13 observations on the following 14 variables.

FTP. 1 Full-time police per 100,000 population  
UEMP. 2 Percent unemployed in the population  
MAN. 3 Number of manufacturing workers in thousands  
LIC. 4 Number of handgun licences per 100,000 population  
GR. 5 Number of handgun registrations per 100,000 population  
CLEAR. 6 Percent homicides cleared by arrests  
WM. 7 Number of white males in the population  
NMAN. 8 Number of non-manufacturing workers in thousands  
GOV. 9 Number of government workers in thousands  
HE. 10 Average hourly earnings  
WE. 11 Average weekly earnings  
ACC Death rate in accidents per 100,000 population  
ASR Number of assaults per 100,000 population  
HOM Number of homicides per 100,000 of population

**Details**

The data were originally collected and discussed by Fisher (1976) but the complete dataset first appeared in Gunst and Mason (1980, Appendix A). Miller (2002) discusses this dataset throughout his book. The data were obtained from StatLib.

**Source**

<http://lib.stat.cmu.edu/datasets/detroit>

## References

- Fisher, J.C. (1976). Homicide in Detroit: The Role of Firearms. *Criminology*, vol.14, 387-400.
- Gunst, R.F. and Mason, R.L. (1980). *Regression analysis and its application: A data-oriented approach*. Marcel Dekker.
- Miller, A. J. (2002). *Subset Selection in Regression*. 2nd Ed. Chapman & Hall/CRC. Boca Raton.

## Examples

```
#Detroit data example
data(Detroit)
#As in Miller (2002) columns 1-11 are used as inputs
p<-11
#For possible comparison with other algorithms such as LARS
# it is preferable to work with the scaled inputs.
#From Miller (2002, Table 3.14), we see that the
#best six inputs are: 1, 2, 4, 6, 7, 11
X<-as.data.frame(scale(Detroit[,c(1,2,4,6,7,11)]))
y<-Detroit[,ncol(Detroit)]
Xy<-cbind(X,HOM=y)
#Use backward stepwise regression with BIC selects full model
out <- lm(HOM~., data=Xy)
step(out, k=log(nrow(Xy)))
#
#Same story with exhaustive search algorithm
out<-bestglm(Xy, IC="BIC")
out
#But many coefficients have p-values that are quite large considering
# the selection bias. Note: 1, 6 and 7 are all about 5% only.
#We can use BICq to reduce the number of variables.
#The qTable let's choose q for other possible models,
out$qTable
#This suggest we try q=0.05 or q=0.0005
bestglm(Xy,IC="BICq", t=0.05)
bestglm(Xy,IC="BICq", t=0.00005)
#It is interesting that the subset model of size 2 is not a subset
# itself of the size 3 model. These results agree with Miller (2002, Table 3.14).
#
#Using delete-d CV with d=4 suggests variables 2,4,6,11
set.seed(1233211)
bestglm(Xy, IC="CV", CVArgs=list(Method="d", K=4, REP=50))
```

---

Fires

*Forest fires in Montesinho natural park. Standardized inputs.*

---

## Description

The forest fire data were collected during January 2000 to December 2003 for fires in the Montesinho natural park located in the northeast region of Portugal. The response variable of interest

was area burned in ha. When the area burned as less than one-tenth of a hectare, the response variable as set to zero. In all there were 517 fires and 247 of them recorded as zero. The region was divided into a 10-by-10 grid with coordinates X and Y running from 1 to 9. The categorical variable xyarea indicates the region in this grid for the fire.

### Usage

```
data(Fires)
```

### Format

A data frame with 517 observations on the following 12 variables. All quantitative variables have been standardized.

xyarea a factor with 36 levels  
month an ordered factor with 12 levels  
day an ordered factor with 7 levels  
FFMC fine fuel moisture code  
DMC Duff moisture code  
DC drought code  
ISI initial spread index  
temp average ambient temperature  
RH a numeric vector  
wind wind speed  
rain rainfall  
lburned  $\log(x+1)$ , x is burned area with x=0 for small fires

### Details

The original data may be found at the website below as well as an analysis. The quantitative variables in this dataset have been standardized. For convenience, the original data is provided in [MontesinhoFires](#).

### Source

<http://archive.ics.uci.edu/ml/datasets/Forest+Fires>

### References

P. Cortez and A. Morais, 2007. A Data Mining Approach to Predict Forest Fires using Meteorological Data. In J. Neves, M. F. Santos and J. Machado Eds., *New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence*, December, Guimaraes, Portugal, pp. 512-523, 2007.

### See Also

[MontesinhoFires](#)

**Examples**

```
data(Fires)
names(Fires)
#ANOVA for xyarea is significant at 1.1%.
summary(aov(lburned~xyarea, data=Fires))
```

---

LOOCV

*Leave-one-out cross-validation*

---

**Description**

An observation is removed and the model is fit to the remaining data and this fit used to predict the value of the deleted observation. This is repeated,  $n$  times, for each of the  $n$  observations and the mean square error is computed.

**Usage**

```
LOOCV(X, y)
```

**Arguments**

X	training inputs
y	training output

**Details**

LOOCV for linear regression is exactly equivalent to the PRESS method suggested by Allen (1971) who also provided an efficient algorithm.

**Value**

Vector of two components comprising the cross-validation MSE and its sd based on the MSE in each validation sample.

**Author(s)**

A.I. McLeod and C. Xu

**References**

Hastie, T., Tibshirani, R. and Friedman, J. (2009). The Elements of Statistical Learning. 2nd Ed.  
Allen, D.M. (1971). Mean Square Error of Prediction as a Criterion for Selecting Variables. Technometrics, 13, 469 -475.

**See Also**

[bestglm](#), [CVD](#), [CVDH](#), [CVHTF](#)

**Examples**

```

#Example. Compare LOO CV with K-fold CV.
#Find CV MSE's for LOOCV and compare with K=5, 10, 20, 40, 50, 60
#Takes about 30 sec
data(zprostate)
train<-(zprostate[zprostate[,10],,])[-10]
X<-train[,1:2]
y<-train[,9]
CVL00<-LOOCV(X,y)
KS<-c(5,10,20,40,50,60)
nKS<-length(KS)
cvs<-numeric(nKS)
set.seed(1233211231)
for (iK in 1:nKS)
  cvs[iK]<-CVDH(X,y,K=KS[iK],REP=10)[1]
boxplot(cvs)
abline(h=CVL00, lwd=3, col="red")
title(sub="Boxplot of CV's with K=5,10,20,40,50,60 and LOO CV in red")

```

---

manpower

*Hospital manpower data*


---

**Description**

The goal of this study is to predict the manpower requirement as given in the output variable Hours given the five other input variables. Data is from Table 3.8 of Myers (1990). See also Examples 3.8, 4.5, 8.8.

**Usage**

```
data(manpower)
```

**Format**

A data frame with 17 observations. The output variable is Hours and the inputs are Load, Xray, BedDays, AreaPop and Stay. The site 1 through 17 is indicated by the row name.

Load a numeric vector

Xray a numeric vector

BedDays a numeric vector

AreaPop a numeric vector

Stay a numeric vector

Hours a numeric vector

**Details**

This data illustrates the multicollinearity problem and the use of VIF to identify it. It provides an illustrative example for ridge regression and more modern methods such as lasso and lars.

**Source**

Myers (1990) indicates the source was "Procedures and Analysis for Staffing Standards Development: Data/Regression Analysis Handbook", Navy Manpower and Material Analysis Center, San Diego, 1979.

**References**

Myers, R. (1990). Classical and Modern Regression with Applications. The Duxbury Advanced Series in Statistics and Decision Sciences. Boston: PWS-KENT Publishing Company.

**Examples**

```
data(manpower)
```

---

MontesinhoFires	<i>Forest fires in Montesinho natural park</i>
-----------------	--

---

**Description**

The forest fire data were collected during January 2000 to December 2003 for fires in the Montesinho natural park located in the northeast region of Portugal. The response variable of interest was area burned in ha. When the area burned as less than one-tenth of a hectare, the response variable as set to zero. In all there were 517 fires and 247 of them recorded as zero. The region was divided into a 10-by-10 grid with coordinates X and Y running from 1 to 9.

**Usage**

```
data(MontesinhoFires)
```

**Format**

A data frame with 517 observations on the following 13 variables.

X X coordinate for region, 0-10

Y X coordinate for region, 0-10

month an ordered factor with 12 levels

day an ordered factor with 7 levels

FFMC fine fuel moisture code

DMC Duff moisture code

DC drought code

ISI initial spread index

temp average ambient temperature

RH a numeric vector

wind wind speed

rain rainfall

burned area burned in hectares

## Details

This is the original data taken from the website below.

## Source

<http://archive.ics.uci.edu/ml/datasets/Forest+Fires>

## References

P. Cortez and A. Morais, 2007. A Data Mining Approach to Predict Forest Fires using Meteorological Data. In J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, Guimaraes, Portugal, pp. 512-523, 2007.

## See Also

[Fires](#)

## Examples

```
data(MontesinhoFires)
names(MontesinhoFires)
data(Fires)
names(Fires)
#Anova for month
summary(aov(burned~month, data=MontesinhoFires))
```

---

oneSdRule

*Utility function. Implements the 1-sd rule.*

---

## Description

The CV and its standard deviation are provided for a range of models ordered by the number of parameters estimated.

## Usage

```
oneSdRule(CVout)
```

## Arguments

CVout            A matrix with two columns. First column is the CV and second, its sd. Row ordering is from fewest parameter to most.

## Value

The row corresponding to the best model.

**Author(s)**

A.I. McLeod and C. Xu

**References**

Hastie, T., Tibshirani, R. and Friedman, J. (2009). The Elements of Statistical Learning. 2nd Ed. Springer-Verlag.

**Examples**

```
CV<-c(1.4637799,0.7036285,0.6242480,0.6069406,0.6006877,0.6005472,0.5707958,
      0.5907897,0.5895489)
CVsd<-c(0.24878992,0.14160499,0.08714908,0.11376041,0.08522291,0.11897327,0.07960879,0.09235052,0.12860983)
CVout <- matrix(c(CV,CVsd), ncol=2)
oneSdRule(CVout)
```

---

print.bestglm                    *Print method for 'bestglm' object*

---

**Description**

A brief description of the best fit is given.

**Usage**

```
## S3 method for class 'bestglm'
print(x, ...)
```

**Arguments**

x	Output from the bestglm function
...	optional arguments

**Value**

No value. Output to terminal only.

**Author(s)**

A.I. McLeod and C. Xu

**See Also**

[bestglm](#), [summary.bestglm](#)

**Examples**

```
data(znuclear)
bestglm(znuclear)
```

---

SAheart

*South African Heart Disease Data*

---

### Description

A retrospective sample of males in a heart-disease high-risk region of the Western Cape, South Africa.

### Usage

```
data(SAheart)
```

### Format

A data frame with 462 observations on the following 10 variables.

**sbp** systolic blood pressure

**tobacco** cumulative tobacco (kg)

**ldl** low density lipoprotein cholesterol

**adiposity** a numeric vector

**famhist** family history of heart disease, a factor with levels Absent Present

**typea** type-A behavior

**obesity** a numeric vector

**alcohol** current alcohol consumption

**age** age at onset

**chd** response, coronary heart disease

### Details

A retrospective sample of males in a heart-disease high-risk region of the Western Cape, South Africa. There are roughly two controls per case of CHD. Many of the CHD positive men have undergone blood pressure reduction treatment and other programs to reduce their risk factors after their CHD event. In some cases the measurements were made after these treatments. These data are taken from a larger dataset, described in Rousseauw et al, 1983, South African Medical Journal.

### Source

Rousseauw, J., du Plessis, J., Benade, A., Jordaan, P., Kotze, J. and Ferreira, J. (1983). Coronary risk factor screening in three rural communities, South African Medical Journal 64: 430–436.

### Examples

```
data(SAheart)
str(SAheart)
summary(SAheart)
```

Shao

*Simulated Regression Data***Description**

Data a simulation study reported by Shao (1993, Table 1). The linear regression model Shao (1993, Table 2) reported 4 simulation experiments using 4 different values for the regression coefficients:

$$y = 2 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + e,$$

where  $e$  is an independent normal error with unit variance.

The four regression coefficients for the four experiments are shown in the table below,

Experiment	$\beta_2$	$\beta_3$	$\beta_4$	$\beta_5$
1	0	0	4	0
2	0	0	4	8
3	9	0	4	8
4	9	6	4	8

The table below summarizes the probability of correct model selection in the experiment reported by Shao (1993, Table 2). Three model selection methods are compared: LOOCV (leave-one-out CV), CV(d=25) or the delete-d method with d=25 and APCV which is a very efficient computation CV method but specialized to the case of linear regression.

Experiment	LOOCV	CV(d=25)	APCV
1	0.484	0.934	0.501
2	0.641	0.947	0.651
3	0.801	0.965	0.818
4	0.985	0.948	0.999

The CV(d=25) outperforms LOOCV in all cases and it also outforms APCV by a large margin in Experiments 1, 2 and 3 but in case 4 APCV is slightly better.

**Usage**

```
data(Shao)
```

**Format**

A data frame with 40 observations on the following 4 inputs.

x2 a numeric vector

x3 a numeric vector

x4 a numeric vector

x5 a numeric vector

**Source**

Shao, Jun (1993). Linear Model Selection by Cross-Validation. Journal of the American Statistical Association 88, 486-494.

**Examples**

```
#In this example BICq(q=0.25) selects the correct model but BIC does not
data(Shao)
X<-as.matrix.data.frame(Shao)
b<-c(0,0,4,0)
set.seed(123321123)
#Note: matrix multiplication must be escaped in Rd file
y<-X%*%b+rnorm(40)
Xy<-data.frame(Shao, y=y)
bestglm(Xy)
bestglm(Xy, IC="BICq")
```

---

summary.bestglm      *summary of 'bestglm' object*

---

**Description**

An analysis of deviance and a likelihood-ratio test with p-value. The p-value is greatly exaggerated due to selection.

**Usage**

```
## S3 method for class 'bestglm'
summary(object, SubsetsQ=FALSE, ...)
```

**Arguments**

object	Output from the bestglm function
SubsetsQ	List best subsets of each size
...	optional arguments

**Value**

No value. Output to terminal only.

**Author(s)**

A.I. McLeod and C. Xu

**See Also**

[bestglm](#), [print.bestglm](#)

**Examples**

```

data(znuclear)
summary(bestglm(znuclear))
#
#find statistical significance of overall regression
data(Fires)
summary(bestglm(Fires, IC="BICq", t=1))

```

znuclear

*Nuclear plant data. Quantitative inputs logged and standardized.***Description**

Data on 32 nuclear power plants. The response variable is cost and there are ten covariates.

**Usage**

```
data(znuclear)
```

**Format**

A data frame with 32 observations on the following 12 variables. All quantitative variables, except date, have been logged and standardized to have mean 0 and variance 1.

date Quantitative covariate. The date on which the construction permit was issued. The data are measured in years since January 1 1990 to the nearest month.

T1 Quantitative covariate. The time between application for and issue of the construction permit.

T2 Quantitative covariate. The time between issue of operating license and construction permit.

capacity Quantitative covariate. The net capacity of the power plant (MWe).

PR Binary covariate. Value 1, indicates the prior existence of a LWR plant at the same site.

NE Binary covariate, located in North-East USA

CT Binary covariate, presence of cooling tower

BW Binary covariate, where 1 indicates that the nuclear steam supply system was manufactured by Babcock-Wilcox.

N Quantitative covariate. The cumulative number of power plants constructed by each architect-engineer.

PT Binary covariate, partial turnkey guarantee.

cost Outcome. The capital cost of construction in millions of dollars adjusted to 1976 base.

**Details**

Davison (2003) explores fitting models to this data using forward and backward stepwise regression. In this modelling logs of quantitative variables are used. We have also standardized this data to facilitate comparison with other techniques such as LARS and principal component regression.

Davison and Hinkley (1997, Example 6.8, 6.10, 6.12) use this data in a series of examples. Example 6.8: estimation of prediction error. Example 6.10: prediction error using cross-validation and bootstrapping. Example 6.12: subset model selection using cross-validation.

**Source**

Obtained from the CRAN package boot.

**References**

Davison, A. C. (2003). *Statistical Models*. Cambridge: Cambridge University Press.

Davison, A.C. and Hinkley, D.V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press.

**Examples**

```
data(znuclear)
bestglm(znuclear, IC="BICq")
```

---

zprostate

*Prostate cancer data. Standardized.*

---

**Description**

Data with 8 inputs and one output used to illustrate the prediction problem and regression in the textbook of Hastie, Tibshirani and Freedman (2009).

**Usage**

```
data(zprostate)
```

**Format**

A data frame with 97 observations, 9 inputs and 1 output. All input variables have been standardized.

lcavol log-cancer volume

lweight log prostate weight

age age in years

lbph log benign prostatic hyperplasia

svi seminal vesicle invasion

lcp log of capsular penetration

gleason Gleason score

pgg45 percent of Gleascores 4/5

lpsa Outcome. Log of PSA

train TRUE or FALSE

**Details**

A study of 97 men with prostate cancer examined the correlation between PSA (prostate specific antigen) and a number of clinical measurements: lcavol, lweight, lbph, svi, lcp, gleason, pgg45

**Source**

<http://www-stat-class.stanford.edu/~tibs/ElemStatLearn/>

**References**

Hastie, Tibshirani & Friedman. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Ed. Springer.

**Examples**

```
#Example 1. Prostate data. Table 3.3 HTF.
data(zprostate)
#full dataset
trainQ<-zprostate[,10]
train <-zprostate[trainQ,-10]
test <-zprostate[!trainQ,-10]
ans<-lm(lpsa~., data=train)
sig<-summary(ans)$sigma
yHat<-predict(ans, newdata=test)
yTest<-zprostate$lpsa[!trainQ]
TE<-mean((yTest-yHat)^2)
#subset
ansSub<-bestglm(train, IC="BICq")$BestModel
sigSub<-summary(ansSub)$sigma
yHatSub<-predict(ansSub, newdata=test)
TESub<-mean((yTest-yHatSub)^2)
m<-matrix(c(TE,sig,TESub,sigSub), ncol=2)
dimnames(m)<-list(c("TestErr","Sd"),c("LS","Best"))
m
```

# Index

## \*Topic **arith**

asbinary, 4

## \*Topic **datasets**

AirQuality, 3

Detroit, 13

Fires, 15

manpower, 18

MontesinhoFires, 19

SAheart, 22

Shao, 23

znuclear, 25

zprostate, 26

## \*Topic **models**

bestglm, 5

CVd, 9

CVDH, 11

CVHTF, 12

LOOCV, 17

print.bestglm, 21

summary.bestglm, 24

## \*Topic **package**

bestglm-package, 2

## \*Topic **regression**

bestglm, 5

CVd, 9

CVDH, 11

CVHTF, 12

LOOCV, 17

print.bestglm, 21

summary.bestglm, 24

AirQuality, 3

airquality, 4

asbinary, 4

bestglm, 5, 10, 12, 13, 17, 21, 24

bestglm-package, 2

CVd, 6, 7, 9, 12, 13, 17

CVDH, 7, 10, 11, 13, 17

CVHTF, 7, 10, 12, 12, 17

Detroit, 13

Fires, 15, 20

glm, 7

leaps, 3, 7

lm, 7

LOOCV, 10, 12, 13, 17

manpower, 18

MontesinhoFires, 16, 19

oneSdRule, 20

print.bestglm, 21, 24

SAheart, 22

Shao, 23

summary.bestglm, 21, 24

to.binary(asbinary), 4

znuclear, 25

zprostate, 26