Package 'FastRCS'

October 18, 2015

Type Package

Title Fits the FastRCS Robust Multivariable Linear Regression Model

Version 0.0.7

Date 2014-01-13

Depends R (>= 3.1.1), matrixStats

Suggests mvtnorm

LinkingTo Rcpp, RcppEigen

SystemRequirements C++11

Description The FastRCS algorithm of Vakili and Schmitt (2014) for robust fit of the multivariable linear regression model and outliers detection.

License GPL (>= 2)

LazyLoad yes

Author Kaveh Vakili [aut, cre]

Maintainer Kaveh Vakili <vakili.kaveh.email@gmail.com>

NeedsCompilation yes

Repository CRAN

Date/Publication 2015-10-18 17:19:56

R topics documented:

FastRCS-package *Code to compute the FastRCS regression outlyingness index.*

Description

Uses the FastRCS algorithm to compute the RCS outlyingness index of regression.

Details

Package:	FastRCS
Type:	Package
Version:	0.1.1
Date:	2013-01-13
Suggests:	mvtnorm
License:	GPL (>= 2)
LazyLoad:	yes

Index:

Author(s)

Kaveh Vakili [aut, cre], Maintainer: Kaveh Vakili <vakili.kaveh.email@gmail.com>

References

Vakili, K. and Schmitt, E. (2014). Finding Regression Outliers With FastRCS. (http://arxiv.org/abs/1307.4834)

FastRCS *Computes the FastRCS outlyingness index for regression.*

Description

Computes a fast and robust regression model for a n by p matrix of multivariate continuous regressors and a single dependent variable.

Usage

```
FastRCS(x,y,nSamp,alpha=0.5,seed=1,intercept=1)
```

Arguments

x	A numeric n (n>5*p) by p (p>1) matrix or data frame. Should not contain an intercept.
y	A numeric nvector.
nSamp	a positive integer giving the number of resamples required; "nSamp" may not be reached if too many of the p-subsamples, chosen out of the observed vectors, are in a hyperplane. If "nSamp" is omitted, it is calculated so that the probability of getting at least one uncontaminated starting point is always at least 99 percent when there are n/2 outliers.
alpha	numeric parameter controlling the size of the active subsets, i.e., "h=quanf(alpha,n,p)". Allowed values are between 0.5 and 1 and the default is 0.5.
seed	starting value for random generator. A positive integer. Default is seed = 1
intercept	If true, a model with constant term will be estimated; otherwise no constant term will be included. Default is intercept=TRUE.

Details

The current version of FastRCS includes the use of a C-step procedure to improve efficiency (Rousseeuw and van Driessen (1999)). C-steps are taken after the raw subset is found and before reweighting. In experiments, we found that carrying C-Steps starting from the members of $rawBest improves the speed of convergence without increasing the bias of the final estimates. FastRCS is regression and affine equivariant and thus consistent at the elliptical model (Grubel and Rock (1990)).

Value

nSamp	The value of nSamp used.
alpha	The value of alpha used.
obj	The value of the FastRCS objective function (the I-index) obtained for H*.
rawBest	The index of the h observation with smallest outlyingness indexes.
rawDist	The distances of the observations to the model defined by rawBest.
best	The index of the J observation with outlyingness smaller than the rejection threshold.
coefficients	The vector of coefficients of the hyperplane fitted to the members of rewbest.
fitted.values	the fitted mean values: cbind(1,x)%*%rew$coefficients.
residuals	the residuals, that is response minus fitted values.
rank	the numeric rank of the fitted linear model.
weights	(only for weighted fits) the specified weights.
df.residual	the residual degrees of freedom.
scale	(robust) scale estimate of the reweighted residuals.

Author(s)

Kaveh Vakili

References

Grubel, R. and Rocke, D. M. (1990). On the cumulants of affine equivariant estimators in elliptical families. Journal of Multivariate Analysis, Vol. 35, p. 203–222. Journal of Multivariate Analysis

Rousseeuw, P. J., and van Driessen, K. (2006). Computing lts regression for large data sets. Data mining and Knowledge Discovery, 12, 29–45

Vakili, K. and Schmitt, E. (2014). Finding Regression Outliers With FastRCS. (http://arxiv.org/abs/1307.4834)

Examples

```
## testing outlier detection
set.seed(123)
n<-100
p<-3
x0<-matrix(rnorm(n*p),nc=p)
y0<-rnorm(n)
z<-c(rep(0,30),rep(1,70))
x0[1:30,]<-matrix(rnorm(30*p,5,1/100),nc=p)
y0[1:30]<-rnorm(30,10,1/100)
ns<-FRCSnumStarts(p=p,eps=0.4);
results<-FastRCS(x=x0,y=y0,alpha=0.5,nSamp=ns)
z[results$best]

## testing outlier detection, different value of alpha
set.seed(123)
n<-100
p<-3
x0<-matrix(rnorm(n*p),nc=p)
y0<-rnorm(n)
z<-c(rep(0,20),rep(1,80))
x0[1:20,]<-matrix(rnorm(20*p,5,1/100),nc=p)
y0[1:20]<-rnorm(20,10,1/100)
ns<-FRCSnumStarts(p=p,eps=0.25);
results<-FastRCS(x=x0,y=y0,alpha=0.75,nSamp=ns)
z[results$best]



#testing exact fit
set.seed(123)
n<-100
p<-3
x0<-matrix(rnorm(n*p),nc=p)
y0<-rep(1,n)
z<-c(rep(0,30),rep(1,70))
x0[1:30,]<-matrix(rnorm(30*p,5,1/100),nc=p)
y0[1:30]<-rnorm(30,10,1/100)
ns<-FRCSnumStarts(p=p,eps=0.4);
```

```
results<-FastRCS(x=x0,y=y0,alpha=0.5,nSamp=ns,seed=1)
z[results$rawBest]
results$obj

#testing regression equivariance
n<-100
p<-3
x0<-matrix(rnorm(n*(p-1)),nc=p-1)
y0<-rnorm(n)
ns<-FRCSnumStarts(p=p,eps=0.4);
y1<-y0+cbind(1,x0)%*%rep(-1,p)
results1<-FastRCS(y=y0,x=x0,nSamp=ns,seed=1)$coefficients
results2<-FastRCS(y=y1,x=x0,nSamp=ns,seed=1)$coefficients
results1+rep(-1,p)
#should be the same:
results2
```

FRCSnumStarts *Computes the number of starting p-subsets*

Description

Computes the number of starting p-subsets so that the desired probability of selecting at least one clean one is achieved. This is an internal function not intended to be called by the user.

Usage

```
FRCSnumStarts(p,gamma=0.99,eps=0.5)
```

Arguments

p	number of dimensions of the data matrix X.
gamma	desired probability of having at least one clean starting p-subset.
eps	suspected contamination rate of the sample.

Value

An integer number of starting p-subsets.

Author(s)

Kaveh Vakili

Examples

```
FRCSnumStarts(p=3,gamma=0.99,eps=0.4)
```

Lemons *Sales Data for the Chrysler Town \& Country*

Description

Sales data for the Chrysler Town \& Country.

Usage

```
Lemons
```

Format

`VehBCost` Acquisition cost paid for the vehicle at time of purchase.

`MMRAcquisitionAuctionAveragePrice` Acquisition price for this vehicle in average condition at time of purchase.

`MMRAcquisitonRetailCleanPrice` Acquisition price for this vehicle in the above Average condition at time of purchase.

`MMRAcquisitionRetailAveragePrice` Acquisition price for this vehicle in the retail market in average condition at time of purchase.

`MMRAcquisitonRetailCleanPrice` Acquisition price for this vehicle in the retail market in above average condition at time of purchase.

`MMRCurrentAuctionAveragePrice` Acquisition price for this vehicle in average condition as of current day.

`MMRCurrentAuctionCleanPrice` Acquisition price for this vehicle in above condition as of current day.

`MMRCurrentRetailAveragePrice` Acquisition price for this vehicle on the retail market in average condition as of current day.

`MMRCurrentRetailCleanPrice` Acquisition price for this vehicle on the retail market in above average condition as of current day.

`WarrantyCost` Warranty price (term=36month and millage=36K).

`VehOdo` The vehicle's odometer reading.

Examples

```
data(Lemons)
alpha<-0.5
p<-ncol(Lemons)
ns<-FRCSnumStarts(p=p,eps=(1-alpha)*4/5)
Fit<-FastRCS(x=Lemons[,-1],y=Lemons[,1],nSamp=ns,seed=1)
plot(Fit)
```

plot.FastRCS *Robust Diagnostic Plots For FastRCS*

Description

Shows the robust Score distances versus robust Orthogonal distances and their respective cutoffs, for the an object of class FastRCS.

Usage

```
## S3 method for class 'FastRCS'
plot(x,col="black",pch=16,...)
```

Arguments

x	For the plot() method, a FastRCS object, typically result of FastRCS.
col	A specification for the default plotting color. Vector of values are recycled.
pch	Either an integer specifying a symbol or a single character to be used as the default in plotting points. Note that only integers and single-character strings can be set as a graphics parameter. Vector of values are recycled.
...	Further arguments passed to the plot function.

Details

This function produces the robust standardized, residuals as well as an indicative cut-off (under normal model). This tool is a diagnostic plot for robust regression and can be used used to reveal the outliers.

See Also

FastRCS

Examples

```
set.seed(123)
n<-100
p<-3
x0<-matrix(rnorm(n*p),nc=p)
y0<-rnorm(n)
z<-c(rep(0,30),rep(1,70))
x0[1:30,]<-matrix(rnorm(30*p,5,1/100),nc=p)
y0[1:30]<-rnorm(30,10,1/100)
ns<-FRCSnumStarts(p=p,eps=0.4);
results<-FastRCS(x=x0,y=y0,alpha=0.5,nSamp=ns)
plot(results)
```

quanf *Converts alpha values to h-values*

Description

FastRCS selects the subset of size h that minimizes the I-index criterion. The function quanf determines the size of h based on the rate of contamination the user expects is present in the data. This is an internal function not intended to be called by the user.

Usage

```
quanf(n,p,alpha)
```

Arguments

n	Number of rows of the data matrix.
p	Number of columns of the data matrix.
alpha	Numeric parameter controlling the size of the active subsets, i.e., "h=quanf(alpha,n,p)". Allowed values are between 0.5 and 1 and the default is 0.5.

Value

An integer number of the size of the starting p-subsets.

Author(s)

Kaveh Vakili

Examples

```
quanf(p=3,n=500,alpha=0.5)
```

Index