Package ‘DetR’

October 12, 2022

Type Package

Title Suite of Deterministic and Robust Algorithms for Linear Regression

Version 0.0.5

Date 2018-05-13

Suggests mvtnorm

Imports robustbase, MASS, pcaPP

Depends R (>= 3.1.1),

LinkingTo Rcpp (>= 0.10.5), RcppEigen (>= 0.3.2.2)

SystemRequirements C++11

Description DetLTS, DetMM (and DetS) Algorithms for Deterministic, Robust Linear Regression.

License GPL (>= 2)

LazyLoad yes

NeedsCompilation yes

Author Kaveh Vakili [aut, cre],
Valentin Todorov [ctb] (modified code originally from the R package robustbase: function ltscheckout, LTScnp2 and LTScnp2.rew and from robustbase:::.detmcd()),
Peter Filzmoser [ctb] (translations of the code for computing the Qn found in package pcaPP),
Heinrich Fritz [ctb] (translations of the code for computing the Qn found in package pcaPP),
Klaudius Kalcher [ctb] (translations of the code for computing the Qn found in package pcaPP),
Kjell Konis [ctb] (translations of the code scaleTau2 found in package robustbase),
Martin Maechler [ctb] (translations of the code scaleTau2 found in package robustbase),
Matias Salibian-Barrera [ctb] (modified code for the FastS from the authors's website),
Peter Rousseeuw [ctb] (modified code originally from the R package

robustbase: function ltscheckout, LTScnp2 and LTScnp2.rew and from
robustbase:::.detmcd()),
Katrien van Driessen [ctb] (modified code originally from the R package
robustbase: function ltscheckout, LTScnp2 and LTScnp2.rew and from
robustbase:::.detmcd())

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

Repository CRAN

Date/Publication 2018-05-19 11:35:38 UTC

R topics documented:

DetR-package 2
chis2009 3
DetLTS 4
DetMM 6
inQn 8
inUMCD 9
OGKCStep 10
quanf 11
test_function 12

Index **14**

`DetR-package` *Deterministic and Robust Algorithms for Regression.*

Description

This packages contains various robust and deterministic algorithms for linear regression.

Details

Package:	DetR
Type:	Package
Version:	0.0.1
Date:	2012-09-19
Depends:	matrixStats, robustbase, MASS
License:	GPL (>= 2)
LazyLoad:	yes

Index:

```
DetR-package            Robust and Deterministic Algothms for Linear
                         Regression
```

```
DetLTS              DetLTS algorithm (deterministic counterpart of FastLTS).
OGKCStep            Tests of OGK+Csteps.
DetMM               DetMM algorithm (deterministic counterpart of FastMM).
test_function unit test functions.
```

Author(s)

Kaveh Vakili [aut, cre], using translation and modifications of codes from other packages (see Desrciption and the individual fuctions' helpfiles)

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

`chis2009` *CHIS 2009 Adult Health Survey Data*

Description

The `chis2009` data frame has 17179 rows and 26 columns.

Usage

```
chis2009
```

Format

This data frame contains the following columns:

`ab1` GENERAL HEALTH CONDITION

`ac13` NUMBER OF TIMES DRANK FRUIT-FLAV LAST MONTH - UNIT

`ac14` NUMBER OF TIMES ATE ICE CREAM/FROZEN DESSERTS LAST MONTH

`ad41w` NUMBER OF TIMES WALKED AT LEAST 10 MIN FOR LEISURE PAST 7 DAYS

`ad42w` AVERAGE LENGTH OF TIME WALKED FOR LEISURE

`ae2` NUMBER OF TIMES ATE FRUIT IN PAST MO

`ae27` NUMBER OF DAYS MODERATE PHYSICAL ACTIVITY IN PAST WEEK

`ae27a` TIME PER DAY OF MODERATE PHYSICAL ACTIVITY

`ae3` NUMBER OF TIMES ATE FRNCH FRIES, HME FRIES, HSH BRWNS IN PAST MO

`ae7` NUMBER OF TIMES ATE VEGETABLES IN PAST MO

`ah5` NUMBER OF TIMES SAW MD IN PAST 12 MOS

`ak3` NUMBER OF USUAL HRS WORKED PER WEEK

`ak7` LENGTH OF TIME WORKING AT MAIN JOB

`distress` SERIOUS PSYCHOLOGICAL DISTRESS

`aheduc` EDUCATIONAL ATTAINMENT

`timead` LENGTH OF TIME LIVED AT CURRENT ADDRESS (IN MONTHS) `ak10_p` RESPONDENT'S EARNINGS LAST MONTH

`ak22_p` HOUSEHOLD'S TOTAL ANNUAL INC

`heighm_p` HEIGHT: METERS

`srage_p` AGE

`wt18k_p` WEIGHT AT 18: KILOS

`sug_past` UNADJUSTED DAILY TEASPOONS OF ADDED SUGAR IN PASTRIES

`sug_bev` UNADJUSTED DAILY TEASPOONS OF ADDED SUGAR IN ALL BEVERAGES

`fv_nobns` DAILY CUP EQUIVALENTS OF FRUITS AND VEGETABLES EXCLUDING BEANS

`sugar2` DAILY TEASPOONS OF ADDED SUGAR

`Weight` WEIGHT: KG

Details

The 2009 California Health Interview Survey (CHIS 2009). The CHIS is a population based telephone survey of California's population. The survey aims to collect extensive information on health status, health conditions, health related behaviors, health insurance coverage as well as access to health care services. Within each household, separate interviews are conducted with a randomly selected adult (age 18 and over). The dataset consists of 536 features measured for 47614 respondents.

Source

CHIS California Health Interview Survey. Los Angeles (CA). UCLA Center for Health Policy Research. `http://www.chis.ucla.edu/`.

`DetLTS` *Robust and Deterministic Linear Regression via DetLTS*

Description

Function to compute the DetLTS estimates of regression.

Usage

```
DetLTS(x, y, intercept = 1, alpha = 0.75, h = NULL, scale_est = "scaleTau2")
```

Arguments

`x` Matrix of design variables. Never contains an intercept.

`y` Vector of responses.

`intercept` A boolean indicating whether the regression contains an intercept.

`alpha` numeric parameter controlling the size of the subsets over which the determinant is minimized, i.e., alpha*n observations are used for computing the determinant. Allowed values are between 0.5 and 1 and the default is 0.75. Can be a vector.

`h` Integer in `[ceiling((n+p+1)/2)`,n) which determines the number of observations which are awarded weight in the fitting process. Can be a vector. If both `h` and `alpha` are set to non default values, `alpha` will be ignored.

`scale_est` A character string specifying the variance functional. Possible values are "Qn" or "scaleTau2".

Value

The function `DetLTS` returns a list with as many components as there are elements in the `h`. Each of the entries is a list containing the following components:

`crit` the value of the objective function of the LTS regression method, i.e., the sum of the h smallest squared raw residuals.

`coefficients` vector of coefficient estimates (including the intercept by default when `intercept=TRUE`), obtained after reweighting.

`best` the best subset found and used for computing the raw estimates, with `length(best)` `== quan = h.alpha.n(alpha,n,p)`.

`fitted.values` vector like `y` containing the fitted values of the response after reweighting.

`residuals` vector like `y` containing the residuals from the weighted least squares regression.

`scale` scale estimate of the reweighted residuals.

`alpha` same as the input parameter `alpha`.

`quan` the number h of observations which have determined the least trimmed squares estimator.

`intercept` same as the input parameter `intercept`.

`cnp2` a vector of length two containing the consistency correction factor and the finite sample correction factor of the final estimate of the error scale.

`raw.coefficients`
vector of raw coefficient estimates (including the intercept, when `intercept=TRUE`).

`raw.scale` scale estimate of the raw residuals.

`raw.resid` vector like `y` containing the raw residuals from the regression.

`raw.cnp2` a vector of length two containing the consistency correction factor and the finite sample correction factor of the raw estimate of the error scale.

`lts.wt` vector like `y` containing weights that can be used in a weighted least squares. These weights are 1 for points with reasonably small residuals, and 0 for points with large residuals.

`raw.weights` vector containing the raw weights based on the raw residuals and raw scale.

`method` character string naming the method (Least Trimmed Squares).

Author(s)

Vakili Kaveh using translation of the C code from pcaPP (by Peter Filzmoser, Heinrich Fritz, Klaudius Kalcher, see citation("pcaPP")) for the Qn and scaleTau2 (Original by Kjell Konis with substantial modifications by Martin Maechler) from robustbase (see citation("scaleTau2")) as well as R code from function ltsReg in package robustbase (originally written by Valentin Todorov valentin.todorov@chello.at, based on work written for S-plus by Peter Rousseeuw and Katrien van Driessen from University of Antwerp, see citation("ltsReg")).

References

Vakili K. (2016). A study and implementation of robust estimators for multivariate and functional data (Doctoral dissertation).

Maronna, R.A. and Zamar, R.H. (2002) Robust estimates of location and dispersion of high-dimensional datasets; *Technometrics* **44**(4), 307–317.

Rousseeuw, P.J. and Croux, C. (1993) Alternatives to the Median Absolute Deviation; *Journal of the American Statistical Association* , **88**(424), 1273–1283.

Peter J. Rousseeuw (1984), Least Median of Squares Regression. *Journal of the American Statistical Association* **79**, 871–881.

P. J. Rousseeuw and A. M. Leroy (1987) *Robust Regression and Outlier Detection.* Wiley.

P. J. Rousseeuw and K. van Driessen (1999) A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **41**, 212–223.

Pison, G., Van Aelst, S., and Willems, G. (2002) Small Sample Corrections for LTS and MCD. *Metrika* **55**, 111-123.

Examples

```
n<-100
h<-c(55,76,89)
set.seed(123)# for reproducibility
x0<-matrix(rnorm(n*2),nc=2)
y0<-rnorm(n)
out1<-DetLTS(x0,y0,h=h)
```

DetMM *Robust and Deterministic Linear Regression via DetMM*

Description

Function to compute the DetMM estimates of regression.

Usage

```
 DetMM(x,y,intercept=1,alpha=0.75,h=NULL,scale_est="scaleTau2",tuning.chi=1.54764,
tuning.psi=4.685061)
```

Arguments

x	Matrix of design variables. Never contains an intercept.
y	Vector of responses.
intercept	A boolean indicating whether the regression contains an intercept.
alpha	numeric parameter controlling the size of the subsets over which the determinant is minimized, i.e., alpha*n observations are used for computing the determinant. Allowed values are between 0.5 and 1 and the default is 0.75. Can be a vector.

`h` Integer in `[ceiling((n+p+1)/2)`,n) which determines the number of observations which are awarded weight in the fitting process. Can be a vector. If both `h` and `alpha` are set to non default values, `alpha` will be ignored.

`scale_est` A character string specifying the variance functional. Possible values are "Qn" or "scaleTau2".

`tuning.chi` tuning constant vector for the bi-weight chi used for the ISteps.

`tuning.psi` tuning constant vector for the bi-weight psi used for the MSteps.

Value

The function `DetLTS` returns a list with as many components as there are elements in the `h`. Each of the entries is a list containing the following components:

`coefficients` The estimate of the coefficient vector

`scale` The scale as used in the M steps.

`residuals` Residuals associated with the estimator.

`converged` `TRUE` if the IRWLS iterations have converged.

`iter` number of IRWLS iterations

`rweights` the “robustness weights” $\psi(r_i/S)/(r_i/S)$.

`fitted.values` Fitted values associated with the estimator.

`DetS` A similar list that contains the results of (initial) returned by DetS

Author(s)

Vakili Kaveh using translation of the C code from pcaPP (by Peter Filzmoser, Heinrich Fritz, Klaudius Kalcher, see citation("pcaPP")) for the Qn and scaleTau2 (Original by Kjell Konis with substantial modifications by Martin Maechler) from robustbase (see citation("scaleTau2")). This function calls lmrob in package robustbase.

References

Maronna, R.A. and Zamar, R.H. (2002) Robust estimates of location and dispersion of high-dimensional datasets; *Technometrics* **44**(4), 307–317.

Rousseeuw, P.J. and Croux, C. (1993) Alternatives to the Median Absolute Deviation; *Journal of the American Statistical Association* , **88**(424), 1273–1283.

Croux, C., Dhaene, G. and Hoorelbeke, D. (2003) *Robust standard errors for robust estimators*, Discussion Papers Series 03.16, K.U. Leuven, CES.

Koller, M. (2012), Nonsingular subsampling for S-estimators with categorical predictors, *ArXiv e-prints*, arXiv:1208.5595v1.

Koller, M. and Stahel, W.A. (2011), Sharpening Wald-type inference in robust regression for small samples, *Computational Statistics & Data Analysis* **55**(8), 2504–2515.

Maronna, R. A., and Yohai, V. J. (2000). Robust regression with both continuous and categorical predictors. *Journal of Statistical Planning and Inference* **89**, 197–214.

Rousseeuw, P.J. and Yohai, V.J. (1984) Robust regression by means of S-estimators, In *Robust and Nonlinear Time Series*, J. Franke, W. Hardle and R. D. Martin (eds.). Lectures Notes in Statistics 26, 256–272, Springer Verlag, New York.

Salibian-Barrera, M. and Yohai, V.J. (2006) A fast algorithm for S-regression estimates, *Journal of Computational and Graphical Statistics*, **15**(2), 414–427.

Yohai, V.J. (1987) High breakdown-point and high efficiency estimates for regression. *The Annals of Statistics* **15**, 642–65.

Examples

```
## generate data
set.seed(1234)  # for reproducibility
n<-100
h<-c(55,76,89)
set.seed(123)
x0<-matrix(rnorm(n*2),nc=2)
y0<-rnorm(n)
out1<-DetMM(x0,y0,h=h)
```

inQn *Test function for the qn*

Description

Test function for the qn used in DetR.

Usage

```
inQn(x)
```

Arguments

x	Vector of 2 or more numbers. Should contain no ties.

Value

the value of the qn estimator of scale.

Author(s)

Kaveh Vakili. Calls code translated from the cde for computing the Qn found in package pcaPP (by Peter Filzmoser, Heinrich Fritz, Klaudius Kalcher , see citation("pcaPP")).

References

see pcaPP::qn and citation("pcaPP").

Examples

```
set.seed(123) #for reproductibility
x<-rnorm(101)
inQn(x)
#should be the same:
pcaPP::qn(x)
```

---

inUMCD *Test function for unimcd*

Description

Test function for the unimcd used in DetR.

Usage

```
inUMCD(x)
```

Arguments

x	Vector of 2 or more numbers. Should contain no ties.

Value

the value of the unimcd estimator of scale.

Author(s)

Kaveh Vakili

References

Rousseeuw, P. J. (1984), Least Median of Squares Regression, Journal of the American Statistical Association,79, 871–880.

Examples

```
set.seed(123) #for reproductibility
x<-rnorm(101)
inUMCD(x)
```

OGKCStep *Robust and Deterministic Linear Regression via OGKCStep*

Description

Function to find the OGKCStep ('best') H-subset.

Usage

```
OGKCStep(x0, scale_est, alpha=0.5)
```

Arguments

x0	Matrix of continuous variables.
alpha	numeric parameter controlling the size of the subsets over which the determinant is minimized, i.e., alpha*n observations are used for computing the determinant. Allowed values are between 0.5 and 1 and the default is 0.5.
scale_est	A character string specifying the variance functional. Possible values are Qn or scaleTau2.

Value

best	the best subset found and used for computing the raw estimates, with length(best) == quan = h.alpha.n(alpha,n,p).

Author(s)

Large part of the the code are from function .detmcd in package robustbase , , see citation("robustbase")

References

Maronna, R.A. and Zamar, R.H. (2002) Robust estimates of location and dispersion of high-dimensional datasets; *Technometrics* **44**(4), 307–317.

Rousseeuw, P.J. and Croux, C. (1993) Alternatives to the Median Absolute Deviation; *Journal of the American Statistical Association* , **88**(424), 1273–1283.

Peter J. Rousseeuw (1984), Least Median of Squares Regression. *Journal of the American Statistical Association* **79**, 871–881.

P. J. Rousseeuw and A. M. Leroy (1987) *Robust Regression and Outlier Detection.* Wiley.

P. J. Rousseeuw and K. van Driessen (1999) A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **41**, 212–223.

Pison, G., Van Aelst, S., and Willems, G. (2002) Small Sample Corrections for LTS and MCD. *Metrika* **55**, 111–123.

Hubert, M., Rousseeuw, P. J. and Verdonck, T. (2012) A deterministic algorithm for robust location and scatter. *Journal of Computational and Graphical Statistics* **21**, 618–637.

Examples

```
n<-100
set.seed(123)# for reproducibility
x0<-matrix(rnorm(n*2),nc=2)
out1<-OGKCStep(x0,alpha=0.5,scale_est=pcaPP::qn)

#comparaison with DetMCD:

#a) create data

set.seed(123456)
Simulation<-DetR:::fx01()
#should be \approx 10
sqrt(min(mahalanobis(Simulation$Data[Simulation$label==0,],rep(0,ncol(Simulation$Data)),
Simulation$Sigma_u))/qchisq(0.975,df=ncol(Simulation$Data)))
a0<-eigen(Simulation$Sigma_u)
Su_ih<-(a0$vector)%*%diag(1/sqrt(a0$values))%*%t(a0$vector)
#run algorithms
A0<-robustbase::covMcd(Simulation$Data,nsamp='deterministic',scalefn=pcaPP::qn,alpha=0.5)
A1<-OGKCStep(Simulation$Data,alpha=0.5,scale_est=pcaPP::qn)
#getbiases algorithms
SB<-eigen(Su_ih%*%var(Simulation$Data[A1,])%*%Su_ih)$values
log10(SB[1]/SB[ncol(Simulation$Data)-1])
SB<-eigen(Su_ih%*%var(Simulation$Data[A0$best,])%*%Su_ih)$values
log10(SB[1]/SB[ncol(Simulation$Data)-1])
```

quanf *Converts alpha values to h-values*

Description

DetLTS selects the subset of size h that minimizes the log-determinant 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)
```

test_function *Test functions for DetR*

Description

Functions to test the cpp codes in the package.

Usage

```
test_function()
```

Details

This is a series of R functions that, together, implement the c++ codes used in the package and which can be used to test those.

Author(s)

Vakili Kaveh.

Examples

```
n<-100
p<-5
#set.seed(123) #for repoducibility.
Z<-matrix(rnorm(n*(p+1)),nc=p+1)
x<-Z[,1:p]
y<-Z[,p+1]
datao<-cbind(x,y)
alpha<-0.6;
test_R_0<-DetR:::test_fxOGK(x0=x,y0=y,cent_est='scaleTau2_test',scal_est='scaleTau2_test',
alpha=alpha)
h<-DetR:::quanf(alpha,n=n,p=p+1) #intercept=1
test_cpp<-DetR:::fxOGK(Data=datao,scale_est=”scaleTau2”,intercept=1,h=h,doCsteps=1)
####should be the same
sort(test_cpp$bestRaw)
sort(as.numeric(test_R_0$bestRaw))
#############
test_R_1<-DetR:::test_Cstep(x=x,y=y,h=h,z0=test_R_0$bestRaw)
####should be the same
sort(test_R_1$bestCStep)
sort(test_cpp$bestCStep[1:h])
```

```
####################################
n<-100
p<-5
set.seed(123) #for repoducibility.
Z<-matrix(rnorm(n*(p+1)),nc=p+1)
x<-Z[,1:p]
y<-Z[,p+1]
datao<-cbind(x,y)
alpha<-0.6;
test_R_0<-DetR:::test_fxOGK(x0=x,y0=y,cent_est='median',scal_est='qn',
alpha=alpha)
h<-DetR:::quanf(alpha,n=n,p=p+1) #intercept=1
test_cpp<-DetR:::fxOGK(Data=datao,scale_est="qn",intercept=1,h=h,doCsteps=1)
####should be the same
sort(test_cpp$bestRaw)
sort(as.numeric(test_R_0$bestRaw))
#############
test_R_1<-DetR:::test_Cstep(x=x,y=y,h=h,z0=test_R_0$bestRaw)
####should be the same
sort(test_R_1$bestCStep)
sort(test_cpp$bestCStep[1:h])
```

Index

∗ **datasets**
- chis2009, 3

∗ **deterministic**
- DetLTS, 4
- DetMM, 6
- OGKCStep, 10
- test_function, 12

∗ **multivariate**
- DetLTS, 4
- DetMM, 6
- inQn, 8
- inUMCD, 9
- OGKCStep, 10
- quanf, 11
- test_function, 12

∗ **package**
- DetR-package, 2

∗ **robust**
- DetLTS, 4
- DetMM, 6
- inQn, 8
- inUMCD, 9
- OGKCStep, 10
- quanf, 11
- test_function, 12

chis2009, 3

DetLTS, 4
DetMM, 6
DetR-package, 2

h.alpha.n, *5*, *10*

inQn, 8
inUMCD, 9

length, *5*, *10*

OGKCStep, 10

quanf, 11

test_function, 12