Package 'fsdac'

June 13, 2021

Version 0.5-14

Date 2021-06-10

Title Demo Calling MATLAB Coder Translated to C Functions from R: addt, FSRbsb, FSRfan, LXS, FSR, FSRmdr, ...

Author Valentin Todorov <valentin@todorov.at>

Maintainer Valentin Todorov <valentin@todorov.at>

Depends R (>= 3.0.2)

Suggests robustbase

Description This is a demo package to show how FSDA functions in C code obtained from the MATLAB coder can be called from R (using the R functions .C and/or .Call).

License GPL (>= 2)

RoxygenNote 7.1.1

Archs i386, x64

R topics documented:

addt *Produces the t-test for an additional explanatory variable*

Description

Produces the t-test for an additional explanatory variable.

Usage

```
addt(y, x, w, intercept = TRUE, la, nocheck = FALSE, trace)
```

Arguments

y	A vector with n elements that contains the response variable.
x	An n x p data matrix (n observations and p variables) of explanatory variables (also called 'regressors') of dimension n x (p-1) where p denotes the number of explanatory variables including the intercept. Rows of X represent observations, and columns represent variables. By default, there is a constant term in the model, unless you explicitly remove it using input option intercept, so do not include a column of 1s in X. Missing values (NAs) and infinite values (Inf's) are allowed, since observations (rows) with missing or infinite values will automatically be excluded from the computations.
w	the added variable, a vector containing the additional explanatory variable whose t-test must be computed.
intercept	Indicator for the constant term (intercept) in the fit, defaults to intercept=TRUE.
la	transformation parameter. It specifies for which Box Cox transformation parameter is necessary to compute the t statistic for the additional variable. If la is missing (default) no transformation is used. For example la=0.5 tests square root transformation
nocheck	Wheather to check the input arguments. If nocheck=TRUE no check is performed on matrix y and matrix X. Notice that y and X are left unchanged. In other words the additional column of ones for the intercept is not added. By default nocheck=FALSE.
trace	Whether to print intermediate results. Default is trace=FALSE.

Value

An object of class addt, will be returned which is basically a list containing two matrices, Un and BB:

1. b estimate of the slope for additional explanatory variable
2. S2add estimate of s^2 of the model which contains the additional explanatory variable
3. Tadd t-statistic for additional explanatory variable
4. pval p-value of the t-statistic

Author(s)

FSDA team, <valentin.todorov@chello.at>

References

Atkinson A.C. and Riani M. (200), *Robust Diagnostic Regression Analysis* Springer Verlag, New York.

Examples

```
data(wool)
XX <- wool
y <- log(XX[, ncol(XX)])
w <- XX[, ncol(XX)-1]
X <- XX[, 1:(ncol(XX)-2), drop=FALSE]
## Intercept=TRUE (default)
##  - has effect only when nocheck=FALSE (default)
(out <- addt(y, X, w, intercept=TRUE, nocheck=FALSE))
```

`fishery` *Fishery data.*

Description

The fishery data consist of 677 transactions of a fishery product in Europe. For each transaction the Value in 1000 euro and the quantity in Tons are reported. Data extracted from monthly aggregates (flows) of trade declarations (Riani et al. 2008). The dataset is formed by 677 flows of a fishery product imported in the European Union from a third country in a period of one year. Among the many variables available we provide:

1. `x`: the quantity of the trade flow;
2. `y`: the value of the trade flow;

By regressing the variable `value` against the `quantity` one can see that the dataset is characterized by the presence of a mixture of linear groups, which roughly correspond to the clusters indicated by the subject matter expert. Riani et al. (2008) have shown how the FS can estimate such a mixture, allocate the units to the components of the mixture and identify in the dataset possible outliers, i.e. units that do not belong to any component. The three identified components are consistent with the clusters identified by the subject matter experts. The dataset is one among thousands of similar datasets that have to be analyzed automatically, for which there is no subject matter classification available.

Usage

```
data(fishery)
```

Format

A data frame with 677 rows and 2 variables

FSR *Computes forward search estimator in linear regression*

Description

FSR computes forward search estimator in linear regression.

Usage

```
FSR(
  y,
  x,
  intercept = TRUE,
  lms = 1,
  bsbmfullrank = TRUE,
  bonflev,
  alpha = 0.5,
  h,
  nsamp = 1000,
  threshoutX,
  weak = FALSE,
  init,
  msg = TRUE,
  nocheck = FALSE,
  trace = FALSE
)
```

Arguments

y	A vector with n elements that contains the response variable.
x	An n x p data matrix (n observations and p variables) of explanatory variables (also called 'regressors') of dimension n x (p-1) where p denotes the number of explanatory variables including the intercept. Rows of X represent observations, and columns represent variables. By default, there is a constant term in the model, unless you explicitly remove it using input option intercept, so do not include a column of 1s in X. Missing values (NaN's) and infinite values (Inf's) are allowed, since observations (rows) with missing or infinite values will automatically be excluded from the computations.
intercept	Indicator for the constant term (intercept) in the fit, defaults to intercept=TRUE.
lms	estimation method, If lms=1 (default) Least Median of Squares is computed, else if lms=2 or lms='lts' fast LTS with all default options is used else if lms is a scalar different from 1 and 2 standard LTS is used (without concentration steps).
bsbmfullrank	how to deal with singular x matrix. This option tells what to do in case when a subset at step m (say bsbm) produces a singular x. In other words, this options controls what to do when rank(X[bsbm,] is smaller then number of explanatory variables. If bsbmfullrank=TRUE (default) these units (whose number is say mnofullrank) are constrained to enter the search in the final n-mnofullrank steps else the search continues using as estimate of beta at step m the estimate of beta found in the previous step.

bonflev — signal to use to identify outliers. This option is used if the distribution of the data is strongly non-normal and, thus, the general signal detection rule based on consecutive exceedances cannot be used. In this case bonflev can be:

1. a value smaller than 1 which specifies the confidence level for a signal and a stopping rule based on the comparison of the minimum MD with a Bonferroni bound. For example if bonflev=0.99 the procedure stops when the trajectory exceeds for the first time the 99% bonferroni bound.
2. a value greater than 1. In this case the procedure stops when the residual trajectory exceeds for the first time this value.

alpha — the percentage (roughly) of squared residuals whose sum will be minimized, by default alpha=0.5. In general, alpha must be between 0.5 and 1.

h — The number of observations that have determined the least trimmed squares estimator, scalar. h is an integer greater or equal than p but smaller then n. Generally h=[0.5*(n+p+1)] (default value).

nsamp — number of subsamples which will be extracted to find the robust estimator. If nsamp=0 all subsets will be extracted. They will be n choose p.

Remark: if the number of all possible subset is <1000 the default is to extract all subsets otherwise just 1000.

threshoutX — threshold to bound the effect of high leverage units. If the design matrix X contains several high leverage units (that is units which are very far from the bulk of the data), it may happen that the best subset of LXS may include some of these units, or it may happen that these units have a deletion residual which is very small due to their extremely high value of hi. threshoutX=1 imposes the constraints that:

1. the extracted subsets which contain at least one unit declared as outlier in the X space by FSM using a Bonferronized confidence level of 0.99 are removed from the list of candidate subsets to find the LXS solution.
2. imposes the constraint that $h_i(m^*)$ cannot exceed $10 \times p/m$.

weak — indicator to use a different decision rule to detect the signal and flag outliers. If weak=FALSE (default) FSRcore values are used, if weak=TRUE 'stronger' quantiles are used as a decision rule to trim outliers and VIOM outliers are the ones entering the search after the first signal.

init — Search initialization. It specifies the initial subset size to start monitoring units forming subset. By default, init=p+1, if the sample size is smaller than 40 or init=min(3*p+1,floor(0.5*(n+p+1))), otherwise.

msg — Level of output to display. It controls whether to display or not messages on the screen. If msg=TRUE (default) messages are displayed on the screen about step of the fwd search else no message is displayed on the screen.

nocheck — Wheather to check the input arguments. If nocheck=TRUE no check is performed on matrix y and matrix X. Notice that y and X are left unchanged. In other words the additional column of ones for the intercept is not added. By default nocheck=FALSE.

trace — Whether to print intermediate results. Default is trace=FALSE.

Value

An object of class FSR will be returned which is basically a list containing the following elements:

1. coefficients: FSR coefficient estimates, including the intercept when intercept=TRUE.

2. residuals: a vector containing the standardized residuals from the regression.
3. fitted.values: a vector containing the fitted values.
4. scale: scale estimate of the residuals.
5. outliers: a vector containing the list of the units declared as outliers using confidence level specified in conflev.
6. mdr a (n-init) x 2 matrix in which
   (a) 1st col = fwd search index
   (b) 2nd col = value of minimum deletion residual in each step of the fwd search
7. Un a (n-init) x 11 matrix which contains the unit(s) included in the subset at each step of the fwd search.

   REMARK: in every step the new subset is compared with the old subset. Un contains the unit(s) present in the new subset but not in the old one. Un[1,2], for example, contains the unit included in step init+1. Un[nrow(Un),2] contains the units included in the final step of the search.
8. nout a 2 x 5 matrix containing the number of times mdr went out of particular quantiles. The first row contains the quantiles c(1,99,99.9,99.99,99.999). The second row contains the frequency distribution.

Author(s)

FSDA team, <valentin@todorov.at>

References

Riani, M., Atkinson, A.C. and Cerioli, A. (2009), Finding an unknown number of multivariate outliers, *Journal of the Royal Statistical Society Series B*, **71**, pp. 201–221.

Examples

```
data(hbk, package="robustbase")
XX <- hbk
y <- XX[, ncol(XX)]
X <- XX[, 1:(ncol(XX)-1), drop=FALSE]

(out <- FSR(y, X))
```

FSRbsb *Returns the units belonging to the subset in each step of the forward search*

Description

Returns the units belonging to the subset in each step of the forward search.

Usage

```
FSRbsb(
  y,
  x,
  bsb,
  init,
  intercept = TRUE,
  nocheck = FALSE,
  bsbsteps,
  msg = TRUE,
  trace
)
```

Arguments

y	A vector with n elements that contains the response variable.
x	A data matrix (n observations and p-1 variables) of explanatory variables (also called 'regressors') of dimension n x (p-1) where p denotes the number of explanatory variables including the intercept. Rows of x represent observations, and columns represent variables. By default, there is a constant term in the model, unless you explicitly remove it using the input option intercept. In such case (intercept=FALSE) a column of 1s will not be added to x. Missing values (NA's) and infinite values (Inf's) are allowed, since observations (rows) with missing or infinite values will automatically be excluded from the computations.
bsb	List of the units forming the initial subset. If bsb = 0 then the procedure starts with p units randomly chosen else if bsb is not 0 the search will start with m0=length(bsb).
init	Search initialization. It specifies the point where to initialize the search and start monitoring required diagnostics. If it is not specified by default it will be equal to: • p+1, if the sample size is smaller than 40; • min(3*p+1,floor(0.5*(n+p+1))), otherwise.
intercept	Indicator for the constant term (intercept) in the fit, defaults to intercept=TRUE.
nocheck	Wheather to check the input arguments. If nocheck=TRUE no check is performed on matrix y and matrix X. Notice that y and X are left unchanged. In other words the additional column of ones for the intercept is not added. By default nocheck=FALSE.
bsbsteps	It specifies for which steps of the fwd search it is necessary to save the units forming subset. If bsbsteps=0 we store the units forming the subset in all steps. If bsbsteps=c() or omitted, the default is to store the units forming the subset in all steps if n <= 5000, else to store the units forming the subset at steps init and steps which are multiple of 100. For example, as default, if n = 753 and init = 6, the units forming the subset are stored for m=init,100,200,300,400,500 and 600.
msg	Level of output to display. It controls whether to display or not messages on the screen. If msg=TRUE (default) messages are displayed on the screen about step of the fwd search else no message is displayed on the screen.
trace	Whether to print intermediate results. Default is trace=FALSE.

Value

An object of class FSRbsb, will be returned which is basically a list containing two matrices, Un and BB:

1. Un Units included in each step;
2. BB Units belonging to search in each step or selected steps.

Un is an (n-init) x 11 matrix which contains the unit(s) included in the subset at each step of the search. REMARK: in every step the new subset is compared with the old subset. Un contains the unit(s) present in the new subset but not in the old one. Un[1,2] for example contains the unit included in step init+1. Un[nrow(Un),2] contains the units included in the final step of the search.

BB is an n-by-(n-init+1) or n-by-length(bsbsteps) matrix which contains the units belonging to the subset at each step (or in selected steps as specified by optional vector bsbsteps) of the forward search. More precisely:

1. BB[,1] contains the units forming the subset in step bsbsteps[1];
2. ....;
3. BB[,ncol(BB)] contains the units forming the subset in step bsbsteps[length(bsbsteps)];

Row 1 of matrix BB is referred to unit 1; .....; Row n of matrix BB is referred to unit n;

Units not belonging to subset are denoted with NaN.

Author(s)

FSDA team, <valentin.todorov@chello.at>

References

Atkinson A.C., Riani M. and Cerioli A. (2004), *Exploring multivariate data with the forward search* Springer Verlag, New York.

Examples

```
data(fishery)

y <- fishery[,2, drop=FALSE]
X <- fishery[,1, drop=FALSE]
bsb <- c(7, 431)                    # found by LTS

out <- FSRbsb(y, X, bsb)            # call 'FSRbsb' with all default parameters
```

FSRfan *Monitors the values of the score test statistic for each lambda*

Description

The transformations for negative and positive responses were determined by Yeo and Johnson (2000) by imposing the smoothness condition that the second derivative of zYJ (λ) with respect to y be smooth at y = 0. However some authors, for example Weisberg (2005), query the physical interpretability of this constraint which is oftern violated in data analysis. Accordingly, Atkinson et al. (2019) and (2020) extend the Yeo-Johnson transformation to allow two values of the transformations parameter: λ_N for negative observations and λ_P for non-negative ones.

FSRfan monitors:

1. the t test associated with the constructed variable computed assuming the same transformation parameter for positive and negative observations fixed. In short we call this test, "global score test for positive observations".
2. the t test associated with the constructed variable computed assuming a different transformation for positive observations keeping the value of the transformation parameter for negative observations fixed. In short we call this test, "test for positive observations".
3. the t test associated with the constructed variable computed assuming a different transformation for negative observations keeping the value of the transformation parameter for positive observations fixed. In short we call this test, "test for negative observations".
4. the F test for the joint presence of the two constructed variables described in points 2) and 3).
5. the F likelihood ratio test based on the MLE of λ_P and λ_N. In this case the residual sum of squares of the null model bsaed on a single trasnformation parameter λ is compared with the residual sum of squares of the model based on data transformed data using MLE of λ_P and λ_N.

Usage

```
FSRfan(
  y,
  x,
  intercept = TRUE,
  family = c("BoxCox", "YJ", "YJpn", "YJall"),
  la,
  lms,
  alpha = 0.75,
  h,
  nsamp = 1000,
  init,
  msg = TRUE,
  nocheck = FALSE,
  trace
)
```

Arguments

y	A vector with n elements that contains the response variable.

x — An n x p data matrix (n observations and p variables) of explanatory variables (also called 'regressors') of dimension n x (p-1) where p denotes the number of explanatory variables including the intercept.

Rows of X represent observations, and columns represent variables. By default, there is a constant term in the model, unless you explicitly remove it using input option intercept, so do not include a column of 1s in X. Missing values (NaN's) and infinite values (Inf's) are allowed, since observations (rows) with missing or infinite values will automatically be excluded from the computations.

intercept — Indicator for the constant term (intercept) in the fit, defaults to intercept=TRUE.

family — string which identifies the family of transformations which must be used. Possible values are c('BoxCox','YJ','YJpn','YJall'). Default is 'BoxCox'. The Box-Cox family of power transformations equals $(y^\lambda - 1)/\lambda$ for λ not equal to zero, and $\log(y)$ if $\lambda = 0$. The Yeo-Johnson (YJ) transformation is the Box-Cox transformation of $y + 1$ for nonnegative values, and of $|y| + 1$ with parameter $2 - \lambda$ for y negative. Remember that BoxCox can be used only if input y is positive. Yeo-Johnson family of transformations does not have this limitation. If family='YJpn' Yeo-Johnson family is applied but in this case it is also possible to monitor (in the output arguments Scorep and Scoren) the score test for positive and negative observations respectively. If family='YJall', it is also possible to monitor the joint F test for the presence of the two constructed variables for positive and negative observations.

la — a vector with the values of the transformation parameter for which the score test is to be computed. By default la=c(-1,-0.5,0,0.5,1), i.e. the five most common values of lambda.

lms — how to find the initlal subset to initialize the search. If lms=1 (default) Least Median of Squares (LMS) is computed, else Least trimmed of Squares (LTS)is computed. If, lms is matrix of size p -1 + intercept X length(la) it contains in column j=1,...,lenght(la) the list of units forming the initial subset for the search associated with la(j). In this case the input option nsamp is ignored.

alpha — the percentage (roughly) of squared residuals whose sum will be minimized, by default alpha=0.5. In general, alpha must between 0.5 and 1.

h — The number of observations that have determined the least trimmed squares estimator, scalar. h is an integer greater or equal than p but smaller then n. Generally h=[0.5*(n+p+1)] (default value).

nsamp — number of subsamples which will be extracted to find the robust estimator. If nsamp=0 all subsets will be extracted. They will be n choose p.

Remark: if the number of all possible subset is <1000 the default is to extract all subsets otherwise just 1000. If nsamp is a matrix of size r-by-p, it contains in the rows the subsets which sill have to be extracted. For example, if p=3 and nsamp=c(2,4,9; 23,45,49; 90,34,1) the first subset is made up of units c(2,4,9), the second subset of units c(23,45,49) and the third subset of units c(90 34 1).

init — Search initialization. It specifies the initial subset size to start monitoring units forming subset. By default, init=p+1, if the sample size is smaller than 40 or init=min(3*p+1,floor(0.5*(n+p+1))), otherwise.

msg — Level of output to display. It controls whether to display or not messages on the screen. If msg=TRUE (default) messages are displayed on the screen about step of the fwd search else no message is displayed on the screen.

nocheck — Wheather to check the input arguments. If nocheck=TRUE no check is performed on matrix y and matrix X. Notice that y and X are left unchanged. In other

words the additional column of ones for the intercept is not added. By default nocheck=FALSE.

trace — Whether to print intermediate results. Default is trace=FALSE.

Value

An object of class FSRfan will be returned which is basically a list containing the following elements:

1. la vector containing the values of lambda for which fan plot is constructed
2. bs matrix of size p X length(la) containing the units forming the initial subset for each value of lambda
3. Score a matrix containing the values of the score test for each value of the transformation parameter:
   - 1st col = fwd search index;
   - 2nd col = value of the score test in each step of the fwd search for la[1]
   - ...
4. Scorep matrix containing the values of the score test for positive observations for each value of the transformation parameter.
   Note: this output is present only if input option family='YJpn' or family='YJall'.
5. Scoren matrix containing the values of the score test for negative observations for each value of the transformation parameter.
   Note: this output is present only if input option 'family' is 'YJpn' or 'YJall'.
6. Scoreb matrix containing the values of the score test for the joint presence of both constructed variables (associated with positive and negative observations) for each value of the transformation parameter. In this case the reference distribution is the F with 2 and subset_size -p degrees of freedom.
   Note: this output is present only if input option family='YJall'.
7. Un a three-dimensional array containing length(la) matrices of size retnUn=(n-init) X retpUn=11. Each matrix contains the unit(s) included in the subset at each step in the search associated with the corresponding element of la.
   REMARK: at each step the new subset is compared with the old subset. Un contains the unit(s) present in the new subset but not in the old one.

Author(s)

FSDA team, <valentin.todorov@chello.at>

References

Atkinson, A.C. and Riani, M. (2000), *Robust Diagnostic Regression Analysis* Springer Verlag, New York.

Atkinson, A.C. and Riani, M. (2002), Tests in the fan plot for robust, diagnostic transformations in regression, *Chemometrics and Intelligent Laboratory Systems*, **60**, pp. 87–100.

Atkinson, A.C. Riani, M. and Corbellini A. (2019), The analysis of transformations for profit-and-loss data, *Journal of the Royal Statistical Society, Series C, "Applied Statistics"*, **69**, pp. 251–275. doi: 10.1111/rssc.12389

Atkinson, A.C. Riani, M. and Corbellini A. (2021), The Box-Cox Transformation: Review and Extensions, *Statistical Science*, **36**(2), pp. 239–255. doi: 10.1214/20STS778.

Examples

```
data(wool)
XX <- wool
y <- XX[, ncol(XX)]
X <- XX[, 1:(ncol(XX)-1), drop=FALSE]

##out <- FSRfan(y, X)                    # call 'FSRfan' with all default parameters
```

FSRmdr *Computes the minimum deletion residual and other basic linear regression quantities in each step of the search*

Description

Computes the minimum deletion residual and other basic linear regression quantities in each step of the search

Usage

```
FSRmdr(
  y,
  x,
  bsb,
  intercept = TRUE,
  init,
  nocheck = FALSE,
  bsbsteps,
  bsbmfullrank = TRUE,
  constr,
  threshlevoutX,
  internationaltrade = FALSE,
  msg = TRUE,
  trace = FALSE
)
```

Arguments

y	A vector with n elements that contains the response variable.
x	A data matrix (n observations and p-1 variables) of explanatory variables (also called 'regressors') of dimension n x (p-1) where p denotes the number of explanatory variables including the intercept. Rows of x represent observations, and columns represent variables. By default, there is a constant term in the model, unless you explicitly remove it using the input option intercept. In such case (intercept=FALSE) a column of 1s will not be added to x. Missing values (NA's) and infinite values (Inf's) are allowed, since observations (rows) with missing or infinite values will automatically be excluded from the computations.
bsb	List of the units forming the initial subset. If bsb = 0 then the procedure starts with p units randomly chosen else if bsb is not 0 the search will start with m0=length(bsb).

intercept — Indicator for the constant term (intercept) in the fit, defaults to intercept=TRUE.

init — Search initialization. It specifies the point where to initialize the search and start monitoring required diagnostics. If it is not specified by default it will be equal to:

- p+1, if the sample size is smaller than 40;
- min(3*p+1,floor(0.5*(n+p+1))), otherwise.

nocheck — Wheather to check the input arguments. If nocheck=TRUE no check is performed on matrix y and matrix X. Notice that y and X are left unchanged. In other words the additional column of ones for the intercept is not added. By default nocheck=FALSE.

bsbsteps — It specifies for which steps of the forward search it is necessary to save the units forming subset. If bsbsteps=0 we store the units forming the subset in all steps. If bsbsteps=c() or omitted, the default is to store the units forming the subset in all steps if n <= 5000, else to store the units forming the subset at steps init and steps which are multiple of 100. For example, as default, if n = 753 and init = 6, the units forming the subset are stored for m=init,100,200,300,400,500 and 600.

bsbmfullrank — how to deal with singular x matrix. This option tells what to do in case when a subset at step m (say bsbm) produces a singular x. In other words, this options controls what to do when rank(X[bsbm,] is smaller then number of explanatory variables. If bsbmfullrank=TRUE (default) these units (whose number is say mnofullrank) are constrained to enter the search in the final n-mnofullrank steps else the search continues using as estimate of beta at step m the estimate of beta found in the previous step.

constr — controls the constrained search. A vector of length r which contains a list of units which are forced to join the search in the last r steps. The default is an empty vector which means that no constraint is imposed. For example constr=1:10 forces the first 10 units to join the subset in the last 10 steps

threshlevoutX — threshold to bound the effect of high leverage units in the computation of deletion residuals. In the computation of the quantity $h_i(m^*) = x_i^T\{X(m^*)^T X(m^*)\}^{-1}x_i$, $i \notin S_*^{(m)}$, units which are very far from the bulk of the data (represented by $X(m^*)$) will have a huge value of $h_i(m^*)$ and consequently of the deletion residuals. In order to tackle this problem it is possible to put a bound to the value of $h_i(m^*)$. For example threshlevoutX=r imposes the contrainst that $h_i(m^*)$ cannot exceed $r \times p/m$. The default value is to leave threshlevoutX empty, which means that no threshold is imposed.

internationaltrade — criterion for updating the subset. If internationaltrade=TRUE (default is internationaltrade=FALSE) the residuals which have large of the final column of X (generally quantity) are reduced. Note that this guarantees that leverge units which have a large value of X will tend to stay in the subset. This option is particularly useful in the context of international trade data where we regress the value value=price*Q on quantity Q. In other words, we use the residuals as if we were regressing y/X (i.e., the price) on the vector of ones.

msg — Level of output to display. It controls whether to display or not messages on the screen. If msg=TRUE (default) messages are displayed on the screen about step of the fwd search else no message is displayed on the screen.

trace — Whether to print intermediate results. Default is trace=FALSE.

Details

Let $S_*^{(m)} \in \mathcal{M}$ be the optimum subset of size m, for which the matrix of regressors is $X(m^*)$. Least squares applied to this subset yields parameter estimates $\hat{\beta}(m^*)$ and $s^2(m^*)$, the mean square estimate of σ^2 on $m - p$ degrees of freedom. The residuals can be calculated for all observations including those not in $S_*^{(m)}$. The n resulting least squares residuals are

$$e_i(m^*) = y_i - x_i^T \hat{\beta}(m^*).$$

The search moves forward with the subset $S_*^{(m+1)}$ consisting of the observations with the $m + 1$ smallest absolute values of $e_i(m^*)$. When $m < n$ the estimates of the parameters are based on only those observations giving the central m residuals.

To test for outliers the deletion residual is calculated for the $n - m$ observations not in $S_*^{(m)}$. These residuals are

$$r_i^*(m^*) = \frac{y_i - x_i^T \hat{\beta}(m^*)}{\sqrt{s^2(m^*)\{1 + h_i(m^*)\}}} = \frac{e_i(m^*)}{\sqrt{s^2(m^*)\{1 + h_i(m^*)\}}},$$

where $h_i(m^*) = x_i^T \{X(m^*)^T X(m^*)\}^{-1} x_i$; the leverage of each observation depends on $S_*^{(m)}$. Let the observation nearest to those constituting $S_*^{(m)}$ be i_{min} where

$$i_{\text{min}} = \arg\min |r_i^*(m^*)| \text{ for } i \notin S_*^{(m)},$$

the observation with the minimum absolute deletion residual among those not in $S_*^{(m)}$. This function computes $r_i^*(m^*)$ for $m^* = init, init + 1, \ldots, n - 1$.

Value

An object of class FSRmdr, will be returned which is basically a list containing five matrices:

1. mdr Monitoring of the minimum deletion residual at each step;
2. Un Units included in each step;
3. BB Units belonging to search in each step or selected steps.
4. Bols OLS coefficients at each step;
5. S2 S2 and R2 at each step;

mdr is a n-init x 2 matrix which contains the monitoring of the minimum deletion residual at each step of the forward search. The first column is the forward search index (from init to n-1) and the second column contains the minimum deletion residual. REMARK: if in a certain step of the search the matrix is singular, this procedure checks how many observations produce a singular matrix. In this case mdr is a column vector which contains the list of units for which the matrix X is non singular.

Un is an (n-init) x 11 matrix which contains the unit(s) included in the subset at each step of the search. REMARK: in every step the new subset is compared with the old subset. Un contains the unit(s) present in the new subset but not in the old one. Un[1,2] for example contains the unit included in step init+1. Un[nrow(Un),2] contains the units included in the final step of the search.

BB is an n-by-(n-init+1) or n-by-length(bsbsteps) matrix which contains the units belonging to the subset at each step (or in selected steps as specified by optional vector bsbsteps) of the forward search. More precisely:

1. BB[,1] contains the units forming the subset in step bsbsteps[1];
2. ....;

3. `BB[,ncol(BB)]` contains the units forming the subset in step `bsbsteps[length(bsbsteps)]`;

Row 1 of matrix `BB` is referred to unit 1; .....; Row `n` of matrix `BB` is referred to unit `n`;

Units not belonging to subset are denoted with NaN.

`Bols` is a `n-init+1` x `p+1` matrix containing the monitoring of the estimated beta coefficients in each step of the forward search (`p` includes the intercept, if selected).

`S2` is a `n-init+1` x `3` matrix containing the monitoring of S2 (2nd column) and R2 (third column) in each step of the forward search.

Author(s)

FSDA team, <valentin.todorov@chello.at>

References

Atkinson, A.C. and Riani, M. (2000), *Robust Diagnostic Regression Analysis* Springer Verlag, New York.

Atkinson, A.C. and Riani, M. (2006), Distribution theory and simulations for tests of outliers in regression, *Journal of Computational and Graphical Statistics*, **15**, pp. 460–476.

Riani, M. and Atkinson, A.C. (2007), Fast calibrations of the forward search for testing multiple outliers in regression, *Advances in Data Analysis and Classification*, **1**, pp. 123–141.

Examples

```
data(fishery)

y <- fishery[,2, drop=FALSE]
X <- fishery[,1, drop=FALSE]
bsb <- c(7, 431)                    # found by LTS
out <- FSRmdr(y, X, bsb)            # call 'FSRmdr' with all default parameters
plot(out)                           # a very simple plot
```

LTSts *Extention of the LTS estimator to time series*

Description

The function LTSts extends the LTS estimator to time series. It is possible to set a model with a trend (up to third order), a seasonality (constant or of varying amplitude and with a different number of harmonics) and a level shift (in this last case it is possible to specify the window in which level shift has to be searched for).

Usage

```
LTSts(
  y,
  intercept = TRUE,
  model,
  alpha = 0.5,
```

```
  h,
  lts,
  conflev = 0.975,
  msg = TRUE,
  nbestindexes = 3,
  nocheck = FALSE,
  nsamp,
  lshiftlocref,
  reftolALS = 0.001,
  refstepsALS = 50,
  SmallSampleCor = 2,
  trace = FALSE
)
```

Arguments

y	A vector with T elements that contains the time series to analize.
intercept	Indicator for the constant term (intercept) in the fit, defaults to intercept=TRUE.
model	model type. A list which specifies the model which will be used. The list contains the following control elements:

Value	Description
s	length of seasonal period. For monthly data s=12 (default), for quartely data s=4, etc.
trend	order of the trend component
seasonal	integer specifying number of frequencies, i.e. harmonics, in the seasonal component. Possible values for seas
X	matrix of size T-by-nexpl containing the values of nexpl extra covariates which are likely to affect y
lshift	numeric, greater or equal to 0 or equal to -1, or a vector of positive integer values which specifies whether it is
ARp	a number greater or equal to 0 which specifies the length of the autoregressive component. The default value is

Remark: the default model is for monthly data with a linear trend (2 parameters) + seasonal component with just one harmonic (2 parameters), no additional explanatory variables and no level shift, i.e. model=list(s=12,trend=1,seasonal=1,X=NULL,lshift

alpha	the percentage (roughly) of squared residuals whose sum will be minimized, by default alpha=0.5. In general, alpha must be between 0.5 and 1.
h	the number of observations that determined the least trimmed squares estimator, an integer greater than p (number of columns of matrix X including the intercept) but smaller then n. If the purpose is outlier detection than h does not have to be smaller than 0.5*(T+p+1). The default is h=0.75*T. Note that if h is supplied the input argument alpha will be ignored.
lts	a list which controls a number of options of the estimation procedure. The list contains the following control elements:

Value	Description
refsteps	defines the number of concentration steps (default is refsteps=2). If refsteps=0, this means "raw-su
reftol	default value of tolerance for the refining steps. The default is reftol=1e-6
bestr	defines the number of 'best betas' to remember from the subsamples. These will be later iterated until c
refstepsbestr	defines the maximum number of refining steps for each best subset (the default is refstepsbestr=50).
reftolbestr	the default value of tolerance for the refining steps for each of the best subsets. The default is reftolbe

Remark: if lts is missing, all default values of the object lts will be used.

conflev	confidence level which is used to declare outliers. It could be 95,0.975,0.99 (individual alpha) or 1-0.05/n,1-0.025/n,1-0.01/n (simultaneous alpha). The default is conflev=0.975.
msg	Level of output to display. It controls whether to display or not messages on the screen. If msg=TRUE (default) messages are displayed on the screen about step of the fwd search else no message is displayed on the screen.
nbestindexes	For each tentative level shift solution, it is interesenting to understand whether the best solutions of the btarget function come from subsets associated with the current level shift solution or from the best solutions from previous tentative level shift position. The indexes from 1 to lts$bestr/2 are associated with subsets just extracted. The indexes from lts$bestr/2+1 to lts$bestr are associated with best solutions from previous tentative level shift.The default is nbestindexes=3.
nocheck	Wheather to check the input arguments. If nocheck=TRUE no check is performed on matrix y and matrix X. Notice that y and X are left unchanged. In other words the additional column of ones for the intercept is not added. By default nocheck=FALSE.
nsamp	A vector of length 1 or 2 which controls the number of subsamples which will be extracted to find the robust estimator. If lshift > 0 then nsamp[1] controls the number of subsets which have to be extracted to find the solution for t=lshift and nsamp[2] controls the number of subsets which have to be extracted to find the solution for t=lshift+1,lshift+2,...,T-lshift. Note that nsamp[2] is generally smaller than nsamp[1] because in order to compute the best solution for t=lshift+1,lshift+2,...,T-lshift we use the lts$bestr/2 best solutions from previous t (after shifting by one the position of the level shift in the estimator of beta). If lshift > 0 the default value is nsamp=c(500 250). If lshift > 0 and nsamp is supplied as a scalar the default is to extract [nsamp/2] subsamples for t=lshift+1,lshift+2,.... Therefore, for example, in order to extract 600 subsamples for t=lshift and 300 subsamples for t= lshift+1 ... you can use nsamp=600 or nsamp=c(600,300). If nsamp=0 all subsets will be extracted.
lshiftlocref	a list with parameters for local shift refinement. The list contains the following control elements:

Value	Description
wlength	a number greater than 0 which identifies the length of the window. By default wlength=15, i.e., the tentative le
typeres	a number which identifies the type of residuals to consider. If typerres=1, the local residuals sum of squares is
huberc	tuning constant for the Huber estimator just in case that lshiftlocref$typeres=1. The default is huberc=2.

reftolALS	value of tolerance for the refining steps inside the ALS routine. The default is reftolALS=1e-03.
refstepsALS	maximum number of iterations inside the ALS routine (default is refstepsALS=50).
SmallSampleCor	small sample correction factor to control the empirical size of the test. Can be 1, 2, 3 or 4 and the default is SmallSampleCor=2.
trace	Whether to print intermediate results. Default is trace=FALSE.

Value

An object of class LTSts will be returned which is basically a list containing the following elements:

1. h the number of observations that have determined the initial LTS estimator, i.e. the value of h.
2. coefficients: LTS (LMS) coefficient estimates, including the intercept when intercept=TRUE.
3. bs: a vector containing the units with the smallest p+k squared residuals before the reweighting step, where p is the total number of parameters in the model and p+k is the smallest number of units such that the design matrix is full rank. bs can be used to initialize the forward search.
4. residuals: a vector containing the standardized residuals from the regression.
5. scale: scale estimate of the residuals.
6. weights: a vector containing weights. The elements of this vector are 0 or 1. These weights identify the h observations which are used to compute the final LTS (LMS) estimate. sum(weights)=h if there is not a perfect fit, otherwise sum(weights) can be greater than h.
7. outliers: a vector containing the list of the units declared as outliers using confidence level specified in conflev.
8. conflev: confidence level which is used to declare outliers.
9. singsub: number of subsets wihtout full rank. Notice that if this number is greater than 0.1*(number of subsamples) a warning is produced on the screen.
10. y: the responce variable.
11. X: the predictor matrix.

Author(s)

FSDA team, <valentin.todorov@chello.at>

References

Rousseeuw, P.J., Perrotta, D., Riani, M. and Hubert, M. (2019), Robust Monitoring of Many Time Series with Application to Fraud Detection, "Econometrics and Statistics", **9**, pp. 108–121. doi: 10.1016/j.ecosta.2018.05.001.

Examples

```
data(heart, package="robustbase")

## Default method works with 'x'-matrix and y-var and all default optional arguments
heart.x <- data.matrix(heart[, 1:2]) # the X-variables
heart.y <- heart[,"clength"]
(out <- LXS(heart.y, heart.x))

data(stackloss)
LXS(stackloss$stack.loss, stackloss[, -4])
```

LXS *Computes the Least Median of Squares (LMS) or Least Trimmed Squares (LTS) estimators*

Description

LXS computes the Least Median of Squares (LMS) or Least Trimmed Squares (LTS) estimators.

Usage

```
LXS(
  y,
  x,
  intercept = TRUE,
  lms = 1,
  rew = FALSE,
  bonflevoutX,
  alpha = 0.5,
  h,
  conflev = 0.975,
  nsamp,
  nomes = FALSE,
  msg = TRUE,
  nocheck = FALSE,
  csave = FALSE,
  trace = FALSE
)
```

Arguments

y	A vector with n elements that contains the response variable.
x	An n x p data matrix (n observations and p variables) of explanatory variables (also called 'regressors') of dimension n x (p-1) where p denotes the number of explanatory variables including the intercept. Rows of X represent observations, and columns represent variables. By default, there is a constant term in the model, unless you explicitly remove it using input option intercept, so do not include a column of 1s in X. Missing values (NaN's) and infinite values (Inf's) are allowed, since observations (rows) with missing or infinite values will automatically be excluded from the computations.
intercept	Indicator for the constant term (intercept) in the fit, defaults to intercept=TRUE.
lms	estimation method, If lms=1 (default) Least Median of Squares is computed, else if lms=2 or lms='lts' fast LTS with all default options is used else if lms is a scalar different from 1 and 2 standard lts is used (without concentration steps) else if lms is a list or a named vector fast lts (with concentration steps) is used. In this case the user can control the following options: 1. refsteps - number of refining iterations on each subsample (default is refsteps=3). If refsteps=0, this means 'raw-subsampling' without iterations. 2. reftol - default value for the tolerance used for the refining steps (default is reftol=1e-6).

3. bestr - number of 'best betas' to remember from the subsamples. These will be later iterated until convergence (default is bestr=5).
4. refstepsbestr - number of refining iterations for each best subset (default is refstepsbestr=50).
5. reftolbestr - default value for the tolerance for the refining steps for each of the best subsets (default is reftolbestr=1e-8).

rew — wheather to reweight the LTS/LMS estomates. If rew=TRUE the reweighted version of LTS/LMS is used and the output quantities refer to the reweighted version, else no reweighting is performed (default).

bonflevoutX — control outlier detection in the designmatrix. If the design matrix X contains several high leverage units (that is units which are very far from the bulk of the data), it may happen that the best subset may include some of these units. If boflevoutX is not missing, outlier detection procedure FSM() is applied to the design matrix X, using name/pair option bonflev=bonflevoutX. The extracted subsets which contain at least one unit declared as outlier in the X space by FSM() are removed (more precisely they are treated as singular subsets) from the list of candidate subsets to find the LXS() solution. By default (bonflevoutX is missing) FSM() is not invoked.

alpha — the percentage (roughly) of squared residuals whose sum will be minimized, by default alpha=0.5. In general, alpha must be between 0.5 and 1.

h — The number of observations that have determined the least trimmed squares estimator, scalar. h is an integer greater or equal than p but smaller then n. Generally h=[0.5*(n+p+1)] (default value).

conflev — confidence level which is used to declare outliers. It could be 95,0.975,0.99 (individual alpha) or 1-0.05/n,1-0.025/n,1-0.01/n (simultaneous alpha). The default is conflev=0.975.

nsamp — number of subsamples which will be extracted to find the robust estimator. If nsamp=0 all subsets will be extracted. They will be n choose p.

Remark: if the number of all possible subset is <1000 the default is to extract all subsets otherwise just 1000.

nomes — controls whether to display or not on the screen messages about estimated time to compute LMS (LTS). If nomes=TRUE no message about estimated time to compute LMS (LTS) is displayed, else if nomes=FALSE (default), a message about estimated time is displayed.

msg — Level of output to display. It controls whether to display or not messages on the screen. If msg=TRUE (default) messages are displayed on the screen about step of the fwd search else no message is displayed on the screen.

nocheck — Wheather to check the input arguments. If nocheck=TRUE no check is performed on matrix y and matrix X. Notice that y and X are left unchanged. In other words the additional column of ones for the intercept is not added. By default nocheck=FALSE.

csave — wheather to return the optional matrix C containing the indexes of the subsamples extracted for computing the estimate (the so called elemental sets).

trace — Whether to print intermediate results. Default is trace=FALSE.

Value

An object of class lts or lms will be returned which is basically a list containing the following elements:

1. `rew`: wheather reweighting was applied. If `rew=TRUE`, all subsequent output refers to reweighted estimates.
2. `coefficients`: LTS (LMS) coefficient estimates, including the intercept when `intercept=TRUE`.
3. `bs`: a vector containing the units forming the subset associated with bLMS (bLTS).
4. `residuals`: a vector containing the standardized residuals from the regression.
5. `scale`: scale estimate of the residuals.
6. `weights`: a vector containing weights. The elements of this vector are 0 or 1. These weights identify the `h` observations which are used to compute the final LTS (LMS) estimate. `sum(weights)=h` if there is not a perfect fit, otherwise `sum(weights)` can be greater than `h`.
7. `h`: the number of observations that have determined the LTS (LMS) estimator, i.e. the value of h.
8. `outliers`: a vector containing the list of the units declared as outliers using confidence level specified in `conflev`.
9. `conflev`: confidence level which is used to declare outliers.
10. `singsub`: number of subsets wihtout full rank. Notice that if this number is greater than `0.1*(number of subsamples)` a warning is produced on the screen.
11. `y`: the responce variable.
12. `X`: the predictor matrix.
13. `C`: the matrix containing the indexes of the subsamples extracted for computing the estimate (the so called elemental sets) (only of `csave=TRUE`).

Author(s)

FSDA team, <valentin.todorov@chello.at>

References

Rousseeuw P.J. and Leroy A.M. (1987), *Robust regression and outlier detection*, Wiley.

Examples

```
data(heart, package="robustbase")

## Default method works with 'x'-matrix and y-var and all default optional arguments
heart.x <- data.matrix(heart[, 1:2]) # the X-variables
heart.y <- heart[,"clength"]
(out <- LXS(heart.y, heart.x))

data(stackloss)
LXS(stackloss$stack.loss, stackloss[, -4])
```

multiple_regression *Multiple regression data showing the effect of masking (Atkinson and Riani, 2000).*

Description

There are 60 observations on a response y with the values of three explanatory variables. The scatter plot matrix of the data shows y increasing with each of x1, x2 and x3. The plot of residuals against fitted values shows no obvious pattern. However the FS finds that there are 6 masked outliers.

Usage

```
data(multiple_regression)
```

Format

A data frame with 60 rows and 4 variables The variables are as follows:

- X1
- X2
- X3
- y

@references Atkinson, A. C., and Riani, M. (2000). *Robust Diagnostic Regression Analysis*. Springer-Verlag, New York.

myc *Square Numbers*

Description

Squares a vector of numbers.

Usage

```
myc(x)
```

Arguments

x	a vector of type "numeric".

Value

the vector of squares of elements of x.

Examples

```
myc(1:5)
```

`wool` *Wool data (Box and Cox, 1964)*

Description

Number of cycles to failure of samples of worsted yarn in a 33 experiment (Box and Cox, 1964). The wool data give the number of cycles to failure of a worsted yarn under cycles of repeated loading. The results are from a single 33 factorial experiment. The three factors and their levels are:

- x1: length of test specimen (25, 30, 35 cm) - x2: amplitude of loading cycle (8, 9, 10 mm) - x3: load (40, 45, 50 g).

The number of cycles to failure ranges from 90, for the shortest specimen subject to the most severe conditions, to 3,636 for observation 19 which comes from the longest specimen subjected to the mildest conditions. In their analysis Box and Cox(1964) recommend that the data be fitted after the log transformation of y. The FS plots explain the effect of the ordering of the data during the FS on the estimates of regression coefficients and the error variance and on a score statistic for transformation of the response.

Usage

```
data(wool)
```

Format

A data frame with 27 rows and 4 variables The variables are as follows:

- length
- amplitude
- load
- cycles

@references Atkinson, A. C., and Riani, M. (2000). *Robust Diagnostic Regression Analysis*. Springer-Verlag, New York.

Index