Package 'stima'

June 3, 2019

Title Simultaneous Threshold Interaction Modeling Algorithm

Version 1.2.1

Date 2018-08-29

Maintainer Elise Dusseldorp <elise.dusseldorp@fsw.leidenuniv.nl>

Description Regression trunk model estimation proposed by Dusseldorp and Meulman (2004) <doi:10.1007/bf02295641>
and Dusseldorp, Conversano, Van Os (2010) <doi:10.1198/jcgs.2010.06089>, integrating a regression tree
and a multiple regression model.

Depends R (>= 3.0.2), rpart

Imports graphics, stats

License GPL-2

LazyLoad yes

RoxygenNote 6.1.0

NeedsCompilation yes

Author Elise Dusseldorp [aut, cre, cph],
Claudio Conversano [aut, cph],
Cor Ninaber [ctb],
Kristof Meers [ctb],
Peter Neufeglise [trl],
Juan Claramunt [ctb]

Repository CRAN

Date/Publication 2019-06-03 10:13:19 UTC

R topics documented:

`stima-package` *Simultaneous Threshold Interaction Modeling Algorithm*

Description

This package enables you to estimate a regression trunk model. The core function is `stima`, which is also the name of the algorithm. The default model is a regression trunk model. A regression trunk model is an integration of a regression tree and a multiple regression model. Currently, the classification trunk model is being developed.

Details

Package:	stima
Type:	Package
Version:	1.1
Date:	2013-11-08
License:	GPL-02
LazyLoad:	yes

The most important functions are `stima`, and `stima.control`.

Author(s)

Elise Dusseldorp, Peter Neufeglise, and Claudio Conversano, with contributions of Kristof Meers and Cor Ninaber

Maintainer: elise.dusseldorp@tno.nl

References

Dusseldorp, E. Conversano, C., and Van Os, B.J. (2010). Combining an additive and tree-based regression model simultaneously: STIMA. *Journal of Computational and Graphical Statistics, 19(3)*, 514-530.

`boston` *Boston house-price data*

Description

The response is the median value of owner-occupied homes measured for each of 506 cencus tracts in the Boston area.

Usage

```
data(boston)
```

Format

A data frame with 506 observations on the following 16 variables.

`c.medv` numeric response variable: median value of owner-occupied homes measured in 1000's USD

`chas` a factor with levels `"lontano"` and `"vicino"`, indicating if a suburb tracts the bound of Charles river (= "lontano") or not

`long` a numeric variable: longitude

`latid` a numeric variable: latitude of census tract

`crim` a numeric variable: per capita crime rate per town

`zn` a numeric variable: proportion of residential land zoned for lots over 25,000 sq.ft.

`indus` a numeric variable: proportion of non-retail business acres per town

`nox` a numeric variable: nitric oxides concentration (parts per 10 million)

`rm` a numeric variable: average number of rooms per dwelling

`age` a numeric variable: proportion of owner-occupied units built prior to 1940

`dis` a numeric variable: weighted distances to five Boston employment centers

`rad` a numeric variable: index of accessibility to radial highways

`tax` a numeric variable: full-value property-tax rate per 10,000 USD

`ptratio` a numeric variable: pupil-teacher ratio by town

`b` a numeric variable: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town

`lstat` a numeric variable: percentage lower status of the population

Source

Statlib website: http://lib.stat.cmu.edu/datasets

References

Harrison, D. and Rubinfeld, D.L. (1978). Hedonic prices and the demand for clean air. *J. Environ. Economics & Management, 5,* 81-102.

`employee` *Employee Dataset*

Description

A dataset with information on background characteristics and salary of 473 employees.

Usage

```
data(employee)
```

Format

A data frame with 473 observations on the following 9 variables:

`salary` a numeric variable, used as response variable: current salary in US dollars

`age` a numeric variable: age in years

`edu` a numeric variable: educational level in years

`startsal` a numeric variable: beginning salary in US dollars

`jobtime` a numeric variable: months since hire

`prevexp` a numeric variable: previous work experience in months

`minority` a factor variable: minority classification with levels `min`, indicating minority, and `no_min`, no minority

`gender` a factor variable: gender type with levels `f`, indicating female, and `m`, indicating male

`jobcat` a factor variable: type of job with levels `Clerical`, `Custodial`, and `manager`

Source

This is an example dataset from the statistical software program SPSS, Version 20.0. If you use this dataset, refer to IBM Corp. (2011), see references. The dataset is used as a benchmark dataset in Dusseldorp, Conversano, and Van Os (2010).

References

IBM Corp. (2011). *IBM SPSS Statistics for Windows, Version 20.0*. Armonk, NY: IBM Corp.

Dusseldorp, E. Conversano, C., and Van Os, B.J. (2010). Combining an additive and tree-based regression model simultaneously: STIMA. *Journal of Computational and Graphical Statistics, 19(3)*, 514-530.

plot.rt *Function to plot a regression trunk.*

Description

Results in a plot a regression trunk.

Usage

```
## S3 method for class 'rt'
plot(x,digits=2,...)
```

Arguments

x	an object of class rt, typically the result of stima using the option model="regtrunk".
digits	number of decimal places used in the plot. Default value is 2.
...	additional arguments to be passed.

Details

The output is a plot of a regression trunk. Exception: If the first splitting predictor is categorical with more than 2 categories, the output will be multiple plots: for each category one plot of a regression trunk.

Note

The number of digits of the mean y value displayed in each node can be adjusted using the command options(digits =..) before the plot command.

Known bug: If a splitting variable (not the first one) in the regression trunk is categorical, the values of the categories are not displayed in the plot.

See Also

stima,stima.control,summary.rt

Examples

```
data(employee)
fit1<-stima(employee,2,first=3,vfold=0)

##adjust the number of decimal places used in the plot
plot(fit1,digits=1)

#categorical first split
fit2<-stima(employee,3,first=9,vfold=0)
plot(fit2)
#click on the plot to see the next one
```

```
#for each category of variable "jobcat" the subtree is shown in a separate plot
```

`prune.rt` *Pruning of a regression trunk.*

Description

Determines the optimally pruned size of the regression trunk by applying the *c**standard error rule to the results from the cross-validation procedure.

Usage

```
## S3 method for class 'rt'
prune(tree, data, c.par = NULL,...)
```

Arguments

`tree`	a tree of class `rt`, that is, a regression trunk. This is the result of `stima` using the option `model="regtrunk"`. To be able to prune, it is a prerequisite that the cross-validation procedure was performed with `stima`.
`data`	the dataset that was used to create the regression trunk.
`c.par`	the pruning parameter (*c*) that will be used in the *c*SE* rule. In the default option, the pruning function uses the best value of *c*, as recommended by Dusseldorp, Conversano & Van Os (2010). This best value depends on the sample size of the included dataset.
`...`	additional arguments to be passed.

Value

The function returns the pruned regression trunk, and the corresponding regression trunk model. The output is an object of class `rt`. If the pruning rule resulted in the root node, no object is returned.

References

Dusseldorp, E. Conversano, C., and Van Os, B.J. (2010). Combining an additive and tree-based regression model simultaneously: STIMA. *Journal of Computational and Graphical Statistics, 19(3)*, 514-530.

See Also

`stima`,`summary.rt`

Examples

```
#Example with employee data
data(employee)
#a regression trunk with a maximum of three splits is grown
#variable used for the first split (edu) is third variable in the dataset
#twofold cross-validation is performed to save time in the example,
#tenfold cross-validation is recommended

emprt1<-stima(employee,3,first=3,vfold=2)
summary(emprt1)
#prune the regression trunk
emprt1_pr<-prune(emprt1,data=employee)
```

`stima` *Simultaneous Threshold Interaction Modeling Algorithm*

Description

This function fits a regression trunk model (default option) using the simultaneous threshold interaction modeling algorithm. The algorithm fits a regression tree and a multiple regression model simultaneously.

Usage

```
stima(data, maxsplit, model = "regtrunk", first = NULL, vfold = 10,
CV = 1, Save = FALSE, control = NULL, printoutput = TRUE)
```

Arguments

`data`	a data frame with one continuous response variable and multiple predictors (categorical or continuous). IMPORTANT: The first column is treated as the response variable, the remaining columns as predictors.
`maxsplit`	the maximum number of splits.
`model`	the default model is a regression trunk model. The classification trunk model is under development.
`first`	the column number in the data frame of the predictor that is used for the first split of the regression trunk. The default option automatically selects the predictor for the first split.
`vfold`	the number of sets to be used in the cross-validation. The default value is 10, which means 10-fold cross-validation. If `vfold = 0`, no cross-validation is performed.
`CV`	the number of times the cross-validation procedure is performed. The default is once. If `CV = 5` and `vfold = 10`, five times a tenfold cross-validation is performed.

Save	if Save = TRUE, the new data are saved and added to the output of the rt-object. The data include indicator variables of the terminal nodes (regions) of the regression trunk.
control	options controlling details of the algorithm. For default options see stima.control.
printoutput	if TRUE, output will be printed while running the function.

Value

an object of class rt, which is a list containing at least the following components

call	the matched call.
trunk	the fitted regression trunk. MeanResponse is the mean response value of the observations in that particular node (this is not the predicted response value).
splitsequence	the number of the nodes that are split.
goffull	goodness-of-fit estimates of the full regression trunk model estimated after 1 split through the model estimated after the maximum number of splits.
full	the estimated full regression trunk model after the maximum number of splits. Coefficient = estimated unstandardized regression coefficient; Std. Coef. = standardized regression coefficient.

References

Dusseldorp, E. & Meulman, J. J. (2004). The regression trunk approach to discover treatment covariate interaction. *Psychometrika, 69,* 355-374.

Dusseldorp, E. Conversano, C., and Van Os, B.J. (2010). Combining an additive and tree-based regression model simultaneously: STIMA. *Journal of Computational and Graphical Statistics, 19(3)*, 514-530.

See Also

stima.control, summary.rt, prune.rt, plot.rt and help("stima-package")

Examples

```
#Example with Boston Housing dataset from paper in JCGS
data(boston)
#grow a full regression trunk with automatic first split selection
#and maximum number of splits = 10, with: bostonrt<-stima(boston,10)
#NB. This analysis will take a long time (about one hour)
#inspect the output with: summary(bostonrt)
#prune the tree with: prune(bostonrt,data=boston)
#the pruned regression trunk has 7 splits
#to save time in the example, we select the splitting candidates beforehand,
#and we grow a tree with a maximum of 4 splits:
contr<-stima.control(predtrunk=c(8,9,16))
bostonrt_pr<-stima(boston,4,first=16,vfold=0,Save=TRUE,control = contr)
summary(bostonrt_pr)
```

```
#inspect the coefficients of the final regression trunk model
round(bostonrt_pr$full,digits=2)
#inspect the new data including the indicator variables referring
#to the terminal nodes
bostonrt_pr$newdata
```

`stima.control` *Control options for the stima function*

Description

The output are various parameters that control aspects of the simultaneaous threshold interaction algorithm

Usage

```
stima.control(minbucket = NULL, crit = "f2", mincrit = 0.001,
predtrunk = NULL, ref = 1, sel = "none", ksel = 2, predsel = NULL,
cvvec = NULL, seed = 3)
```

Arguments

`minbucket`	the minimum number of observations in a terminal node. The default is the square root of the total sample size.
`crit`	the type of statistic to be used in the partitioning criterion. The default for the regression trunk model is the effect size `"f2"` which equals the relative increase in variance accounted for. Other options are `"R2change"` which is the absolute increase in variance accounted for, or `"F-value"` which is the *F*-statistic of the anova test.
`mincrit`	the minimum node deviance before growing stops.
`predtrunk`	a row vector that indicates the column numbers in the data frame of the predictors that can be used in the regression trunk. The default action uses all predictors as available splitting candidates; NB. this column number can not be 1, because the first column is the response variable.
`ref`	a number referring to the region of the regression trunk that will be used as reference category in the regression trunk model. The default value is 1, referring to *R1*.
`sel`	if `sel = "backward"`, the full regression trunk model is reduced using a backward selection procedure; if `sel = "manual"`, one needs to give a specification of predsel.
`ksel`	the multiple of the number of degrees of freedom used for the penalty in the backward selection procedure. The default value is 2, which gives the genuine AIC: `ksel = log(n)` is sometimes referred to as BIC or SBC.
`predsel`	row vector that indicates the column numbers in the `newdata` set (obtained by `Save = TRUE` in `stima`) of the predictors to be used in the final regression trunk model.

cvvec	index vector for the rows of the dataframe that will be used in each cross-validation set. The default option is a random division into "vfold" sets.
seed	an integer between 0 and 1023 that will be used in set.seed(). The default value equals 3.

Value

a list containing the parameters.

References

Dusseldorp, E. Conversano, C., and Van Os, B.J. (2010). Combining an additive and tree-based regression model simultaneously: STIMA. *Journal of Computational and Graphical Statistics, 19(3)*, 514-530.

See Also

stima,summary.rt,plot.rt,prune.rt

Examples

```
##Adjust the stopping rule in a minimum of 5 observations in a terminal node
data(employee)
contr1<-stima.control(minbucket=5)


##Adjust the seed used to create an index vector for the 10fold cross-validation
##With seed=3, the result equals the one reported in the online Appendix D of
##the paper in the Journal of Computational and Graphical Statistics
##NB. To save time in the example, the splitting candidates of the regression
##trunk(i.e., edu and jobtime) are selected with predtrunk=c(3,5),
##where 3 and 5 denote the column numbers in the dataset

contr2<-stima.control(sel="backward",seed=3,predtrunk=c(3,5))
emprt2<-stima(employee,2,first=3,control=contr2)
summary(emprt2)


##Apply a manual selection of predictors to be used in the pruned model

contr3<-stima.control(sel="manual",predsel=c(2,3,4,5,6,8))
```

summary.rt *Summarizing Regression Trunk Model Fits from stima*

Description

summary method for class "rt" (i.e. a regression trunk)

Usage

```
## S3 method for class 'rt'
summary(object, digits = 3,...)
```

Arguments

object	an object of class rt, usually a result of a call to stima using the default option:model="regtrunk"
digits	the number of decimals to used in the output.
...	Additional arguments to be passed

Value

The function summary.rt returns the goodness-of-fit summary of the estimated regression trunk model, using the components "goffull" and, if available, "gofsel".

full	goodness-of-fit estimates of the full regression trunk model estimated after 1 split through the model estimated after the maximum number of splits. f2 = the effect size of the indicator variable added to the model after a split. RE = apparent error; SE = standard error of RE; REcv = cross-validated error; SEcv = standard error of REcv. If available: REcvm = Average cross-validated error; SEcvm = standard error of REcvm.
selected	goodness-of-fit estimates of the selected regression trunk model (if applicable).

See Also

stima.control,stima,plot.rt

Index