Package 'rpms'

June 2, 2017

Type Package

Title Recursive Partitioning for Modeling Survey Data

Version 0.2.1

Date 2017-06-02

Maintainer daniell toth <danielltoth@yahoo.com>

Description Fits a linear model to survey data in each node obtained by recursively partitioning the data. The splitting variables and splits selected are obtained using a procedure which adjusts for complex sample design features used to obtain the data. Likewise the model fitting algorithm produces design-consistent coefficients to the least squares linear model between the dependent and independent variables. The first stage of the design is accounted for in the provided variance estimates. The main function returns the resulting binary tree with the linear model fit at every end-node. The package provides a number of functions and methods for these trees.

License CC0

Depends R (>= 2.10)

Imports Rcpp (>= 0.12.3)

LinkingTo Rcpp, RcppArmadillo

RoxygenNote 6.0.1

NeedsCompilation yes

LazyData true

Author daniell toth [aut, cre]

Repository CRAN

Date/Publication 2017-06-02 15:34:34 UTC

R topics documented:

`rpms-package` *Recursive Partitioning for Modeling Survey Data (rpms)*

Description

This package provides a function `rpms` to produce an `rpms` object and method functions that operate on them. The `rpms` object is a representation of a regression tree achieved by recursively partioning the dataset, fitting the specified linear model on each node seperately. The recursive partitioning algorithm has an unbiased variable selection and accounts for the sample design. The algorithm accounts for one-stage of stratification and clustering as well as unequal probability of selection. This version does not handle missing values, so only complete cases of a dataset are used.

`CE` *CE Consumer expenditure data (first quarter of 2014)*

Description

A dataset containing consumer unit characteristics, assets and expenditure data from the Bureau of Labor Statistics' Consumer Expenditure Survey public use interview data file.

Usage

```
CE
```

Format

A data frame with 6483 rows and 61 variables:

Location and sample-design variables

NEWID Consumer unit identifying variable

PSU Primary Sampling Unit code

CID Cluster Identifier for all clusters, (defined using PSU, REGION, STATE, and POPSIZE)

FINLWT21 Final sample weight to make inference to total population

POPSIZE Population size of PSU 1-biggest 5-smallest

REGION Region code: 1 Northeast; 2 Midwest; 3 South; 4 West

STATE State FIPS code

BLS_URBN Urban = 1, Rural = 2

SMSASTAT Is CU in a MSA: 1 Yes; 2 No

Household variables

HH_CU_Q Number of CU in household

FAM_TYPE CU code based on relationship of members to reference person (children incldue blood-related, step and adopted): 1 Married Couple only; 2 Married Couple, children (oldest < 6 years old); 3 Married Couple, children (oldest 6 to 17 years old); 4 Married Couple, children (oldest > 17 years old); 5 All other Married Couple CUs 6 One parent (male), children (at least one child < 18 years old); 7 One parent (female), children (at least one child < 18 years old); 8 Single consumers; 9 Other CUs

FAM_SIZE Number of members in CU

AS_COMP1 Number of males >16 yrs old

AS_COMP2 Number of females >16 yrs old

PERSLT18 Number of people <18 yrs old

PERSOT64 Number of people >64 yrs old

NO_EARNR Number of earners

Housing and transportation

CUTENURE Housing tenure: 1 Owned with mortgage; 2 Owned without mortgage 3 Owned mortgage not reported; 4 Rented; 5 Occupied without payment of cash rent; 6 Student housing

ROOMSQ Number of rooms, including finished living areas and excluding all baths

BATHRMQ Number of complete bathrooms in the unit

BEDROOMQ Tota number of bedrooms in the unit

VEHQ Number of owned vehicles

VEHQL Number of leased vehicles

Reference person

AGE_REF Age of reference person

EDUC_REF Education level of reference person coded: 00 None; 10 1st-8th Grade; 11 some HS; 12 HS; 13 Some college; 14 AA degree; 15 Bachelors degree; 16 Advanced degree

REF_RACE Race code of reference person: 1 White; 2 Black; 3 Native American; 4 Asian; 5 Pacific Islander; 6 Multi-race

SEX_REF Male = 1; Female = 2

HISP_REF Hispanic = 1; Not Hispanic = 2

HIGH_EDU Highest level education level attained by anyone within CU: 00 None; 10 1st-8th Grade; 11 some HS; 12 HS; 13 Some college; 14 AA degree; 15 Bachelors degree; 16 Advanced degree

Labor status variables

INC_HRS1 Number of hours usually worked per week by reference person

INCNONW1 Reason reference person did not work during the past 12 months: 1 Retired; 2 Home maker; 3 School; 4 health; 5 Unable to find work; 6 Doing something else

Income variables

FINCBTAX Amount of CU income before taxes in past 12 months

FINCBTAX_I Imputation indicator for FINCBTAX: 0-reported; 1-imputed

FINCBT_X Flag for FINCBTAX: D-reported value; T-top coded

FINCATAX Amount of CU income after taxes in past 12 months

FINCAT_X Flag for FINCATAX: D-reported value; T-top coded

FSALARYX Amount of wage and salary income, before deductions, received by all CU members in past 12 months

FSALARY_I Imputation indicator for FSALARYX: 0-reported; 1-imputed

FSAL_RYX Flag for FSAL_RYX: D-reported; T-top coded

FRRETIRX Amount of Social Security and Railroad Retirement income

FRRET_I Imputation indicator for FRRETIRX: 0-reported; 1-imputed

WELFAREM Total amount of income from public assistance or welfare, including money from job training grants, received by ALL CU members during the past 12 months

WELFARE_I Imputation indicator for WELFAREM: 0-reported; 1-imputed

NETRENTX Amount of net rental income

NETRENT_I Imputation indicator for NETRENTX: 0-reported; 1-imputed

NETR_NTX Flag for NETRENTX: A-valid blank; C-refusal; D-reported value; T-value is top coded

ROYESTX Amount income received in royalties, estates and trusts

ROYESTX_I Imputation indicator for ROYESTX: 0-reported; 1-imputed

ROYESTX_ Flag for ROYESTX: A-valid blank; C-refusal; D-reported value; T-value is top coded

Assets

IRAYRX value of all retirement accounts one year ago today

IRAYRX_ Flag for IRAYRX: A-valid blank; C-refusal; D-reported value; T-value is top coded

LIQUDYRX Value of all checking, savings, money market accounts, and certificates of deposit one year ago today

LIQU_YRX Flag for LIQUDYRX: A-valid blank; C-refusal; D-reported value; T-value is top coded

STOCKB Range of total value of all directly-held stocks, bonds, and mutual funds: 1 $0 - $1999; 2 $2,000 - $9,999; 3 $10,000 - $49,999; 4 $50,000 - $199,999; 5 $200,000 - $449,999; 6 $450,000 and over

Expenditures

FINDRETX Amount of money put in an individual retirement plan, such as an IRA or Keogh, by all CU members in past 12 months

FIND_ETX Flag for FINDRETX: D-reported; T-top coded

TOTXEST Estimated total taxes paid

FDHOMECQ Expenditure on food at home this quarter

FDAWAYCQ Expenditure on food away from home this quarter

ALCBEVCQ Expenditure on alcoholic beverages this quarter

GASMOCQ Expenditure on gasoline and motor oil this quarter

#end describe

Source

http://www.bls.gov/cex/pumd_data.htm

See Also

For more information see http://www.bls.gov/cex/2015/csxintvw.pdf

---

`end_nodes` *end_nodes*

Description

Get vector end-node labels

Usage

```
end_nodes(t1)
```

Arguments

`t1`	`rpms` object

Value

vector of lables of end-nodes.

in_node *in_node*

Description

Get index of elements in dataframe that are in the specified end-node of an rpms object

Usage

```
in_node(node, t1, data)
```

Arguments

node	integer label of the desired end-node.
t1	rpms object
data	dataframe containing the variables used for the recursive partitioning.

Value

vector of indexes for observations in the end-node.

Examples

```
# model linear fit between retirement contributions and amount of income
r1 <-rpms(FINDRETX~EDUC_REF+AGE_REF+BLS_URBN+REGION, data=CE,
     e_equ=FINDRETX~FINCBTAX, clusters=~CID)

if(2 %in% end_nodes(r1))
  summary(CE$FSALARYX[in_node(node=2, r1, data=CE)])

if(6 %in% end_nodes(r1))
  summary(CE$FSALARYX[in_node(node=6, r1, data=CE)])
```

node_plot *node_plot*

Description

plots end-node of object of class rpms

Usage

```
node_plot(t1, node, data, variable = NA, ...)
```

Arguments

t1	rpms object
node	integer label of the desired end-node.
data	data.frame that includes variables used in rp_equ, e_equ, and design information
variable	string name of variable in data to use as x-axis in plot
...	further arguments passed to plot function.

Examples

```
{

# model linear fit between retirement contributions and amount of income
r1 <-rpms(FINDRETX~EDUC_REF+AGE_REF+BLS_URBN+REGION, data=CE,
     e_equ=FINDRETX~FINCBTAX, clusters=~CID)

# plot node 2 if it is in tree
if(2 %in% end_nodes(r1))
  node_plot(r1, node=2, data=CE)

#' # plot last end-node

if(6 %in% end_nodes(r1))
  node_plot(r1, node=6, data=CE)


}
```

---

predict.rpms *predict.rpms*

Description

Predicted values based on rpms object

Usage

```
## S3 method for class 'rpms'
predict(object, newdata, ...)
```

Arguments

object	Object inheriting from rpms
newdata	data frame with variables to use for predicting new values.
...	further arguments passed to or from other methods.

Value

vector of predicticed values for each row of newdata

Examples

```
{
# get rpms model of mean retirement contribution by several factors
r1 <- rpms(FINDRETX~EDUC_REF+AGE_REF+BLS_URBN+REGION, data = CE)

# first 10 predicted means
predict(r1, CE[1:10, ])

}
```

`print.rpms` *print.rpms*

Description

print method for class `rpms`

Usage

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

Arguments

`x`	`rpms` object
`...`	further arguments passed to or from other methods.

`qtree` *qtree*

Description

Code to write a latex qtree plot takes a rpm frame and returns latex code to produce qtree uses linearize as a guide Produces text code to produce tree structure in tex document Requires using LaTex packages and the following commands in preamble of LaTex doc: usepackage{lscape} usepackage{tikz-qtree}

Usage

```
qtree(t1, title = "rpms", label = NA, caption = "", digits = 2,
  scale = 1)
```

Arguments

t1	rpms object created by rpms function
title	string for the top node of the tree
label	string used for labeling the tree figure
caption	string used for caption
digits	integer number of displayed digits
scale	factor for scaling size of tree

Examples

```
{
# get rpms model of mean retirement contribution by several factors
r1 <-rpms(FINDRETX~EDUC_REF+AGE_REF+BLS_URBN+REGION, data=CE,
     e_equ=FINDRETX~FINCBTAX, clusters=~CID)

# get Latex code
qtree(r1)

}
```

---

rpms *rpms*

Description

main function producing a regression tree using variables from rp_equ to partition the data and fit the model e_equ on each node. Currently only uses data with complete cases.

Usage

```
rpms(rp_equ, data, weights = ~1, strata = ~1, clusters = ~1, e_equ = ~1,
  e_fn = "survLm", l_fn = NULL, bin_size = NULL, perm_reps = 500L,
  pval = 0.05)
```

Arguments

rp_equ	formula containing all variables for partitioning
data	data.frame that includes variables used in rp_equ, e_equ, and design information
weights	formula or vector of sample weights for each observation
strata	formula or vector of strata labels
clusters	formula or vector of cluster labels
e_equ	formula for modeling data in each node
e_fn	string name of function to use for modeling (only "survLm" is operational)
l_fn	loss function (does nothing yet)

bin_size	numeric minimum number of observations in each node
perm_reps	integer specifying the number of permuations
pval	numeric p-value used to reject null hypothesis in permutation test

Value

object of class "rpms"

Examples

```
{
# model mean of retirement contributions with a binary tree while accounting
# for clusterd data

rpms(FINDRETX~EDUC_REF+AGE_REF+BLS_URBN+REGION, data = CE,  clusters=~CID)


# model linear fit between retirement contributions and amount of income
# with a regression tree while accounting for clusterd data

rpms(FINDRETX~EDUC_REF+AGE_REF+BLS_URBN+REGION, data=CE,
     e_equ=FINDRETX~FINCBTAX, clusters=~CID)

}
```

---

survLm_model *Fit a linear model using data collected from a complex sample*

Description

Fit a linear model using data collected from a complex sample

Usage

```
survLm_model(y, X, weights, strata, clusters)
```

Arguments

y	A vector of values
X	The design matrix of the linear model
weights	A vector of sample weights for each observation
strata	A vector of strata labels
clusters	A vector of cluster labels

Value

list containing coefficients, covariance matrix and the residuals

Index