Package 'cpmBigData'

February 13, 2023

Title Fitting Semiparametric Cumulative Probability Models for Big Data

Version 0.0.1

Description A big data version for fitting cumulative probability models using the orm() function from the 'rms' package. See Liu et al. (2017) <DOI:10.1002/sim.7433> for details.

Depends R (>= 4.0.0)

License GPL (>= 2)

Encoding UTF-8

RoxygenNote 7.2.3

Imports rms (>= 6.2-0),Hmisc (>= 4.3-0),doParallel (>= 1.0.11),parallel (>= 3.5.2),foreach (>= 1.2.0),iterators (>= 1.0.0),SparseM (>= 1.77),benchmarkme (>= 1.0.4)

NeedsCompilation no

Author Chun Li [cre, aut], Guo Chen [aut]

Maintainer Chun Li <cli77199@usc.edu>

Repository CRAN

Date/Publication 2023-02-13 09:00:03 UTC

R topics documented:

ormBD *Cumulative Probability Model for Big Data*

Description

Fits cumulative probability models (CPMs) for big data. CPMs can be fit with the orm() function in the 'rms' package. When the sample size or the number of distinct values is very large, fitting a CPM may be very slow or infeasible due to demand on CPU time or storage. This function provides three alternative approaches. In the divide-and-combine approach, the data are evenly divided into subsets, a CPM is fit to each subset, followed by a final step to aggregate all the information. In the binning and rounding approaches, a new outcome variable is defined and a CPM is fit to the new outcome variable. In the binning approach, the outcomes are ordered and then grouped into equal-quantile bins, and the median of each bin is assigned as the new outcome for the observations in the bin. In the rounding approach, the outcome variable is either rounded to a decimal place or a power of ten, or rounded to significant digits.

Usage

```
ormBD(
  formula,
  data,
  subset = NULL,
  na.action = na.delete,
  target_num = 10000,
  approach = c("binning", "rounding", "divide-combine"),
  rd_type = c("skewness", "signif", "decplace"),
  mem_limit = 0.75,
  log = NULL,
  model = FALSE,
  x = FALSE,
  y = FALSE,
  method = c("orm.fit", "model.frame", "model.matrix"),
  ...
)
```

Arguments

formula	a formula object
data	data frame to use. Default is the current frame.
subset	logical expression or vector of subscripts defining a subset of observations to analyze
na.action	function to handle NAs in the data. Default is 'na.delete', which deletes any observation having response or predictor missing, while preserving the attributes of the predictors and maintaining frequencies of deletions due to each variable in the model. This is usually specified using options(na.action="na.delete").

target_num	the desired number of observations in a subset for the 'divide-and-combine' method; the target number of bins for the 'binning' method; the desired number of distinct outcome values after rounding for the 'rounding' method. Default to 10,000. Please see Details.
approach	the type of method to analyze the data. Can take value 'binning', 'rounding', and 'divide-combine'. Default is 'binning'.
rd_type	the type of round, either rounding to a decimal place or a power of ten (rd_type = 'decplace') or to significant digits (rd_type = 'signif'). Default is 'skewness', which is to determine the rounding type according to the skewness of the outcome: 'decplace' if skewness < 2 and 'signif' otherwise.
mem_limit	the fraction of system memory to be used in the 'divide-and-combine' method. Default is 0.75, which is 75 percent of system memory. Range from 0 to 1.
log	a parameter for parallel::makeCluster() when the 'divide-and-combine' method is used. See the help page for makeCluster for more detail.
model	a parameter for orm(). Explicitly included here so that the 'divide-and-combine' method gives the correct output. See the help page for orm for more detail.
x	a parameter for orm(). Explicitly included here so that the 'divide-and-combine' method gives the correct output. See the help page for orm for more detail.
y	a parameter for orm(). Explicitly included here so that the 'divide-and-combine' method gives the correct output. See the help page for orm for more detail.
method	a parameter for orm(). Explicitly included here so that the 'divide-and-combine' method gives the correct output. See the help page for orm for more detail.
...	other arguments that will be passed to orm

Details

In the divide-and-combine approach, the data are evenly divided into subsets. The desired number of observations in each subset is specified by 'target_num'. As this number may not evenly divide the whole dataset, a number closest to it will be determined and used instead. A CPM is fit for each subset with the orm() function. The results from all subsets are then aggregated to compute the final estimates of the intercept function alpha and the beta coefficients, their standard errors, and the variance-covariance matrix for the beta coefficients.

In the binning approach, observations are grouped into equal-quantile bins according to their outcome. The number of bins are specified by 'target_num'. A new outcome variable is defined to takes value median[y, y in B] for observations in bin B. A CPM is fit with the orm() function for the new outcome variable.

In the rounding approach, by default the outcome is rounded to a decimal place or a power of ten unless the skewness of the outcome is greater than 2, in which case the outcome is rounded to significant digits. The desired number of distinct outcomes after rounding is specified by 'target_num'. Because rounding can yield too few or too many distinct values compared to the target number specified by 'target_num', a refinement step is implemented so that the final number of distinct rounded values is close to 'target_num'. Details are in Li et al. (2021). A CPM is fit with the orm() function for the new rounded outcome.

Value

The returned object has class 'ormBD'. It contains the following components in addition to those mentioned under the optional arguments and those generated by orm().

`call`	calling expression
`approach`	the type of method used to analyze the data
`target_num`	the 'target_num' argument in the function call
`...`	others, same as for `orm`

Author(s)

Guo Chen
Department of Computer and Data Sciences
Case Western Reserve University

Chun Li
Department of Population and Public Health Sciences
University of Southern California

References

Liu et al. "Modeling continuous response variables using ordinal regression." Statistics in Medicine, (2017) 36:4316-4335.

Li et al. "Fitting semiparametric cumulative probability models for big data." (2023) (submitted)

See Also

`orm` `na.delete` `get_ram` `registerDoParallel` `SparseM.solve`

Examples

```
## generate a small example data and run one of the three methods
set.seed(1)
n <- 200
x1 = rnorm(n); x2 = rnorm(n)
tmpdata = data.frame(x1 = x1, x2 = x2, y = rnorm(n) + x1 + 2*x2)
modbinning <- ormBD(y ~ x1 + x2, data = tmpdata, family = loglog,
                    approach = "binning", target_num = 100)
## modrounding <- ormBD(y ~ x1 + x2, data = tmpdata, family = loglog,
##                      approach = "rounding", target_num = 100)
## moddivcomb <- ormBD(y ~ x1 + x2, data = tmpdata, family = loglog,
##                      approach = "divide-combine", target_num = 100)
```

Index