Package ‘BSR’

June 21, 2013

Title BSR

Version 0.0.1

Date 2013-06-18

Author Faming Liang, William Wheeler, Kai Yu

Description Bayesian Subset Regression

Maintainer Kai Yu `<yuka@mail.nih.gov>`

License GPL-2

Archs i386, x64

R topics documented:

`BSR` *BSR*

Description

BSR is a wrapper function to combine the functions BSR.samc, BSR.sample.model and BSR.prediction

Usage

```
BSR(train.data, test.data, outdir, response.var, id.var, vars0, groups=NULL,
     interactions=NULL, op=NULL)
```

Arguments

`train.data`	Data frame containing only the variables to be used in the analysis. See details.
`test.data`	Data frame for testing that contains the same variables in the same order as `train.data`.
`outdir`	Directory where the output files will be written. If NULL, then the working directory will be used.
`response.var`	Variable name of the case-control response coded as 0=control, 1=case.
`id.var`	Variable name of the id variable or NULL if `train.data` and `test.data` do not contain an id variable.
`vars0`	Character vector of variables to always include in the models. Must be of length >= 1.
`groups`	NULL or a NAMED list of character vectors defining the groups. Each group of variables will always appear together in a model. For example, if "var1" and "var2" are dummy variables for a 3 level categorical variable, then `groups`=list(group1=c("var1", "var2")) will force variables var1 and var2 to always appear together in a model. If NULL, then each variable is its own group. Any variable in the data frames other than `response.var` and `id.var` which is not part of a group definition is treated as a group by itself. Example: `groups`=list(grp1=c("var1", "var2"), grp2=c("x1", "x5", "x34")). The default is NULL.
`interactions`	NULL or a list of character vectors of GROUP names that must be in the model whenever the first GROUP name in the vector is in the model. The GROUP names must be a subset of the names in the `groups` list above, so that the `groups` argument must be specified to use this argument. This argument is used primarily for forcing the main effect variables of an interaction to be in the model whenever the interaction variable(s) are in the model. For instance, suppose that grp1=c("x1", "x2") is a group of dummy variables for a 3-level categorical variable, and that grp2=c("z1", "z2", "z3") is a group of dummy variables for a 4-level categorical variable. The interaction of the two categorical variables would be as group of 6 dummy variables grp3=c("x1z1", "x1z2", "x1z3", "x2z1", "x2z2", "x2z3"). If the `groups` argument is defined as `groups`=list(grp1=c("x1", "x2"), grp2=c("z1", "z2", "z3"), grp3=c("x1z1", "x1z2", "x1z3", "x2z1", "x2z2", "x2z3")), then the `interactions` argument can be defined as `interactions`=list(c("grp3", "grp1", "grp2")). Note that the first group in the vector is the group for the interaction variables. The default is NULL.
`op`	List of options. See `details` for all possible options.

Details

The input data frames `train.data` and `test.data` will use ALL variables in the data frame as covariates except `response.var` and `id.var`. All variables except for `id.var` must be numeric and continuous. Any interactions, dummy variables for categorical variables, and a column of ones for the intercept must already exist in the data frames before calling this function.

Options list:

Below are the names for the options list `op`. All names have default values if they are not specified.

- `out.string` String to be appended to output file names. The default is "".

- `train.op` Options list for training. See `BSR.samc`. Note that the options `lowE` and `maxE` must be specified in this list. The default is NULL.
- `resample.op` Options list for resampling. See `BSR.sample.model`. Note that the following options have no effect when calling the `BSR` function: `stepscale`, `lowE`, `maxE`, `scale`, `rho`, `maxvar`, `maxbeta`, `gamma`, `in_hist` and `in_models`. The default is NULL.
- `prediction.op` Options list for prediction. See `BSR.prediction`. The default is NULL.

Value

A list containing the prediction results (see `BSR.prediction`).

Author(s)

Faming Liang, William Wheeler, Kai Yu

References

Liang, F., Song, Q., and Yu, K. (2013). Bayesian Subset Modeling for High Dimensional Generalized Linear Models. J. Amer. Statist. Assoc., in press

See Also

`BSR.sample.model`, `BSR.prediction`, `BSR.samc`

Examples

```
# Create data
set.seed(123)
n    <- 100
m    <- 50
y    <- rbinom(n, 1, 0.5)
id   <- paste("sub_", 1:n, sep="")
int  <- rep(1, n)
z1   <- rbinom(n, 1, 0.5)
z2   <- rbinom(n, 1, 0.5)
z3   <- runif(n)
z4   <- runif(n)
data <- data.frame(y, id, int, z1, z2, z3, z4)

train.data <- data[1:m, ]
test.data  <- data[(m+1):n, ]
outdir     <- "c:/temp/"
vars0      <- "int"  # The intercept variable

# Options for a small test run
train.op      <- list(lowE=0, maxE=1000, total_iteration=100, warm=10)
resample.op   <- list(total_iteration=100, warm=10)
prediction.op <- list(n.burn=2)
op            <- list(train.op=train.op, resample.op=resample.op,
                      prediction.op=prediction.op, out.string="test")

# Not run
```

```
#ret <- BSR(train.data, test.data, outdir, "y", "id", vars0, groups=NULL,
#         interactions=NULL,  op=op)
```

`BSR.prediction` *BSR.prediction*

Description

Uses models generated from BSR.sample.model to predict output for the training and testing datasets, and to estimate the prediction error based on the training (biased) and testing datasets.

Usage

```
BSR.prediction(train.data, test.data, outdir, response.var, id.var,
                model.file, gof.file, best.file, op=NULL)
```

Arguments

`train.data`	See `BSR`.
`test.data`	See `BSR`.
`outdir`	See `BSR`.
`response.var`	See `BSR`.
`id.var`	See `BSR`.
`model.file`	Output file "out2" from `BSR.sample.model`.
`gof.file`	Output file "out1" from `BSR.sample.model`.
`best.file`	Output file "out_best" from `BSR.sample.model`.
`op`	List of options. See `details` for all possible options.

Details

Options list:

Below are the names for the options list `op`. All names have default values if they are not specified.

- `n.burn` Specifies the first `n.burn` models from the output to remove. The default is 100.
- `out.string` Character string to include in the output files. The default is "".

Value

A list of the output file names for the predicted training and testing data. Each file contains the colums "PRED.AVG", "PRED.BEST", `response.var` and `id.var`. The columns "PRED.AVG" and "PRED.BEST" are the average and best predicted values for each subject.

Author(s)

Faming Liang, William Wheeler, Kai Yu

See Also

`BSR.sample.model`, `BSR.samc`, `BSR`

Examples

```
# Create data
set.seed(123)
n    <- 100
m    <- 50
y    <- rbinom(n, 1, 0.5)
id   <- paste("sub_", 1:n, sep="")
int  <- rep(1, n)
z1   <- rbinom(n, 1, 0.5)
z2   <- rbinom(n, 1, 0.5)
z3   <- runif(n)
z4   <- runif(n)
data <- data.frame(y, id, int, z1, z2, z3, z4)

train.data <- data[1:m, ]
test.data  <- data[(m+1):n, ]
outdir     <- "c:/temp/"
vars0      <- "int"  # The intercept variable

# Options for a small test run
train.op      <- list(total_iteration=100, warm=10)
resample.op   <- list(total_iteration=100, warm=10)
prediction.op <- list(n.burn=2)
op            <- list(train.op=train.op, resample.op=resample.op,
                      prediction.op=prediction.op, out.string="test")

#ret1 <- BSR.samc(train.data, outdir, "y", "id", vars0, groups=NULL,
#       interactions=NULL,  op=train.op)

#resample.op$in_models <- ret1$out_models
#resample.op$in_hist   <- ret1$out_hist
#ret2 <- BSR.sample.model(train.data, outdir, "y", "id", vars0, groups=NULL,
#       interactions=NULL,  op=resample.op)

#model.file <- ret2$out2
#gof.file   <- ret2$out1
#best.file  <- ret2$out_best
#ret <- BSR.prediction(train.data, test.data, outdir, "y", "id",
#                 model.file, gof.file, best.file, op=prediction.op)
```

`BSR.samc` *BSR.samc*

Description

Apply SAMC to estimate the posterior distribution of the model. Four output files are generated. The .log file contains the summary of the model posterior distribution, such as the size, the best model for a given size, and the probability of a variable included in a model. The .hist file contains the probability for each partition region estimated from the SAMC algorithm. The .out1 file contains the summary for each (biased) sampled model. The .out2 file contains the estimated coefficients for each model listed in the .out1 file. The .model file contains results for all the models that have been considered.

Usage

```
BSR.samc(data, outdir, response.var, id.var, vars0, groups=NULL,
           interactions=NULL, op=NULL)
```

Arguments

`data`	See BSR.
`outdir`	See BSR.
`response.var`	See BSR.
`id.var`	See BSR.
`vars0`	See BSR.
`groups`	See BSR.
`interactions`	See BSR.
`op`	List of options. See `details` for all possible options.

Details

The input `data` frame will use all variables in the data frame as covariates except `response.var` and `id.var`. All variables except for `id.var` must be numeric and continuous. Rows with missing values will be removed from the analysis.

Options list:

Below are the names for the options list `op`. All names have default values if they are not specified.

- `stepscale` This is the value called t_0 in equation (26) from the paper by Liang (2013). The default is 1000.
- `total_iteration` Number of iterations. The default is 500000.
- `warm` The number of warm-up iterations. The default is 10000.
- `savestep` Number to output every `savestep` iterations. The default is 1.
- `lowE, maxE, scale` These options are used for defining the sample space partition according to the energy function. The lower and upper boundaries are defined by `lowE` and `maxE`. The option `scale` is used to control the number of subregions defined as `(maxE-lowE)*scale`. The default value of `scale` is 1, but `lowE` and `maxE` must be provided.
- `rho` The parameter for the SAMC algorithm. The default is 1.
- `temperature` The parameter for the MCMC move. The default is 1.
- `maxvar` Maximum number of variables in the regression model. The default is 20.
- `maxbeta` Maximum allowed value for abs(beta) in the regression model. If the model has any abs(beta) > `maxbeta`, then the model is rejected. The default is 10.
- `seed` The default seed is computed from the time.
- `gamma` The penalty parameter. The default is 0.5.
- `model_maxn` Maximum number of models to save. The default is 100000.
- `out.string` Character string to include in the output files. The default is "".
- `delete` 0 or 1 to delete the temporary files written to `outdir`. The default is 0.

Value

A list of the output files created in `outdir` and if the input data frame was modified, then the modified data frame is also returned.

Author(s)

Faming Liang, William Wheeler, Kai Yu

See Also

`BSR.sample.model`, `BSR.prediction`, `BSR`

Examples

```
# Create data
set.seed(123)
n    <- 100
m    <- 50
y    <- rbinom(n, 1, 0.5)
id   <- paste("sub_", 1:n, sep="")
int  <- rep(1, n)
z1   <- rbinom(n, 1, 0.5)
z2   <- rbinom(n, 1, 0.5)
z3   <- runif(n)
z4   <- runif(n)
data <- data.frame(y, id, int, z1, z2, z3, z4)

outdir     <- "c:/temp/"
vars0      <- "int"  # The intercept variable

# Options for a small test run
op <- list(lowE=0, maxE=1000, total_iteration=100, warm=10)

# Not run
#ret <- BSR.samc(data, outdir, "y", "id", vars0, groups=NULL,
#        interactions=NULL,  op=op)
```

`BSR.sample.model` *BSR.sample.model*

Description

Generates models from the posterior distribution estimated from BSR.samc. The .out1 file contains the summary for each sampled model. The .out2 file contains the estimated coefficients for each model listed in the .out1 file. The .best file contains the best model among the sampled models.

Usage

```
BSR.sample.model(data, outdir, response.var, id.var, vars0, groups=NULL,
                 interactions=NULL, op=NULL)
```

Arguments

data	See BSR.
outdir	See BSR.
response.var	See BSR.
id.var	See BSR.
vars0	See BSR.
groups	See BSR.
interactions	See BSR.
op	List of options. See details for all possible options.

Details

Options list:

The same options as in BSR.samc can be specified, except that the following options have no effect: stepscale, lowE, maxE, scale and rho. The options maxvar, maxbeta and gamma should to be set to the same values for when the BSR.samc function was called. There are also two additional options below for op.

- in_models NULL or the "out_models" file from BSR.samc. The default is NULL.
- in_hist The "out_hist" file from BSR.samc. This must be specified.

Value

A list of the output files created in outdir and if the input data frame was modified, then the modified data frame is also returned.

Author(s)

Faming Liang, William Wheeler, Kai Yu

See Also

BSR.samc, BSR.prediction, BSR

Examples

```
# Create data
set.seed(456)
n    <- 100
m    <- 50
y    <- rbinom(n, 1, 0.5)
id   <- paste("sub_", 1:n, sep="")
int  <- rep(1, n)
z1   <- rbinom(n, 1, 0.5)
z2   <- rbinom(n, 1, 0.5)
z3   <- runif(n)
z4   <- runif(n)
data <- data.frame(y, id, int, z1, z2, z3, z4)

outdir      <- "c:/temp/"
```

```
vars0      <- "int"  # The intercept variable

# Options for a small test run
op <- list(total_iteration=100, warm=10)

# Not run
#ret <- BSR.sample.model(data, outdir, "y", "id", vars0, groups=NULL,
#        interactions=NULL,  op=op)
```

Index