Package ‘HCmodelSets’

July 21, 2025

Type Package

Title Regression with a Large Number of Potential Explanatory Variables

Version 1.1.3

Author H. H. Hoeltgebaum

Maintainer H. Battey <`h.battey@imperial.ac.uk`>

BugReports `https://github.com/hhhelfer/HCmodelSets/issues`

Description Software for performing the reduction, exploratory and model selection phases of the procedure proposed by Cox, D.R. and Battey, H.S. (2017) <doi:10.1073/pnas.1703764114> for sparse regression when the number of potential explanatory variables far exceeds the sample size. The software supports linear regression, likelihood-based fitting of generalized linear regression models and the proportional hazards model fitted by partial likelihood.

License GPL-2 | GPL-3

Encoding UTF-8

LazyData true

LazyDataCompression xz

Depends R (>= 3.5.0), mvtnorm, ggplot2, survival,

RoxygenNote 7.1.0

NeedsCompilation no

Suggests R.rsp

VignetteBuilder R.rsp

Repository CRAN

Date/Publication 2023-03-15 18:00:09 UTC

Contents

`DGP` *Data generating process used by Battey, H. S. & Cox, D. R. (2018).*

Description

This function generates realizations of random variables as described in the simple example of Battey, H. S. & Cox, D. R. (2018).

Usage

```
DGP(s,a,sigStrength,rho,n,noise=NULL,var,d,intercept,type.response="N",DGP.seed=NULL,
    scale=NULL,shape=NULL,rate=NULL)
```

Arguments

`s`	Number of signal variables.
`a`	Number of noise variables correlated with signal variables.
`sigStrength`	Signal strength.
`rho`	Correlation among signal variables and noise variables correlated with signal variables.
`n`	Sample size.
`noise`	Variance of the observations around the true regression line.
`var`	Variance of the potential explanatory variables.
`d`	Number of potential explanatory variables.
`intercept`	Expected value of the response variable when all potential explanatory variables are at zero. It is only considered when type.response="N".
`type.response`	Generates gaussian ("N") or survival ("S") data from a proportional hazards model with Weibull baseline hazard.
`DGP.seed`	Seed for the random number generator.
`scale`	scale parameter of the proportional hazards model with Weibull baseline hazard.
`shape`	shape parameter of the proportional hazards model with Weibull baseline hazard.
`rate`	rate parameter of the exponential distribution of censoring times. If not provided, uncensored data are generated.

Value

`X`	The simulated design matrix.
`Y`	The simulated response variable.
`TRUE.idx`	Indices of the variables in the true model.
`status`	If type.response="S", provides the status from survival data.

Acknowledgement

The work was supported by the UK Engineering and Physical Sciences Research Council under grant number EP/P002757/1.

Author(s)

Hoeltgebaum, H. H.

References

Cox, D. R. and Battey, H. S. (2017). Large numbers of explanatory variables, a semi-descriptive analysis. *Proceedings of the National Academy of Sciences*, 114(32), 8592-8595.

Battey, H. S. and Cox, D. R. (2018). Large numbers of explanatory variables: a probabilistic assessment. *Proceedings of the Royal Society of London, A.*, 474(2215), 20170631.

Hoeltgebaum, H., & Battey, H. S. (2019). HCmodelSets: An R Package for Specifying Sets of Well-fitting Models in High Dimensions. *The R Journal*, 11(2), 370-379.

Examples

```
## Generates DGP

## Generates a random DGP
dgp = DGP(s=5, a=3, sigStrength=1, rho=0.9, n=100, intercept=5, noise=1,
          var=1, d=1000, DGP.seed = 2018)
```

`Exploratory.Phase` *Perform the Exploratory phase on the hypercube dimension reduction proposed by Cox, D. R. & Battey, H. S. (2017)*

Description

This function performs the exploratory phase on the variables retained through the reduction phase, returning any significant squared and interaction terms.

Usage

```
Exploratory.Phase(X, Y, list.reduction, family=gaussian,
                    signif=0.01, silent=TRUE, Cox.Hazard = FALSE)
```

Arguments

`X`	Design matrix.
`Y`	Response vector.
`list.reduction`	Indices of retained variables from the reduction phase.
`family`	A description of the error distribution and link function to be used in the model. For glm this can be a character string naming a family function, a family function or the result of a call to a family function. See `family` for more details.
`signif`	Significance level for the assessment of squared and interaction terms. The default is 0.01.
`silent`	By default, silent=TRUE. If silent=FALSE the user can decide upon the exclusion of individual interaction terms.
`Cox.Hazard`	If TRUE fits proportional hazards regression model. The family argument will be ignored if Cox.Hazard=TRUE.

Value

`mat.select.SQ`	Indices of variables with significant squared terms.
`mat.select.INTER`	Indices of the pairs of variables with significant interaction terms.

Acknowledgement

The work was supported by the UK Engineering and Physical Sciences Research Council under grant number EP/P002757/1.

Author(s)

Hoeltgebaum, H. H.

References

Cox, D. R., and Battey, H. S. (2017). Large numbers of explanatory variables, a semi-descriptive analysis. *Proceedings of the National Academy of Sciences*, 114(32), 8592-8595.

Battey, H. S. and Cox, D. R. (2018). Large numbers of explanatory variables: a probabilistic assessment. *Proceedings of the Royal Society of London, A.*, 474(2215), 20170631.

Hoeltgebaum, H., & Battey, H. S. (2019). HCmodelSets: An R Package for Specifying Sets of Well-fitting Models in High Dimensions. *The R Journal*, 11(2), 370-379.

See Also

`Reduction.Phase`

Examples

```
## Generates a random DGP
dgp = DGP(s=5, a=3, sigStrength=1, rho=0.9, n=100, intercept=5, noise=1,
          var=1, d=1000, DGP.seed = 2018)

#Reduction Phase using only the first 70 observations
outcome.Reduction.Phase =  Reduction.Phase(X=dgp$X[1:70,],Y=dgp$Y[1:70],
                                          family=gaussian, seed.HC = 1012)

# Exploratory Phase using only the first 70 observations, choosing the variables which
# were selected at least two times in the third dimension reduction

idxs = outcome.Reduction.Phase$List.Selection$`Hypercube with dim 2`$numSelected1
outcome.Exploratory.Phase =  Exploratory.Phase(X=dgp$X[1:70,],Y=dgp$Y[1:70],
                                              list.reduction = idxs,
                                              family=gaussian, signif=0.01)
```

`LymphomaData` *Lymphoma patients data set.*

Description

Data set of lymphoma patients used in the study of Alizadeh et al. (2000) and also Simon et al. (2011).

Usage

```
data(LymphomaData)
```

Format

`patient.data` A list with survival times, staus and covariates from patients.

Value

`x`	Covariates from patients.
`time`	Survival times.
`status`	Patient status.

References

Alizadeh, A. A., et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature, 403(6769), p.503.

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. Journal of statistical software, 39(5), 1.

Examples

```
data(LymphomaData)
x <- t(patient.data$x)
y <- patient.data$time
```

`ModelSelection.Phase` *Construct sets of well-fitting models as proposed by Cox, D. R. & Battey, H. S. (2017)*

Description

This function tests low dimensional subsests of the set of retained variables from the reduction phase and any squared or interaction terms suggested at the exploratory phase. Lists of well-fitting models of each dimension are returned.

Usage

```
ModelSelection.Phase(X,Y, list.reduction, family=gaussian,
                     signif=0.01, sq.terms=NULL, in.terms=NULL,
                     modelSize=NULL, Cox.Hazard = FALSE)
```

Arguments

`X`	Design matrix.
`Y`	Response vector.
`list.reduction`	Indices of variables that where chosen at the reduction phase.
`family`	A description of the error distribution and link function to be used in the model. For glm this can be a character string naming a family function, a family function or the result of a call to a family function. See `family` for more details.
`signif`	Significance level of the likelihood ratio test against the comprehensive model. The default is 0.01.
`sq.terms`	Indices of squared terms suggested at the exploratory phase (See `Exploratory.Phase`).
`in.terms`	Indices of pairs of variables suggested at the exploratory phase (See `Exploratory.Phase`).
`modelSize`	Maximum size of the models to be tested. Curently the maximum is 7. If not provided a default is used.
`Cox.Hazard`	If TRUE fits proportional hazards regression model. The family argument will be ignored if Cox.Hazard=TRUE.

Value

`goodModels`	List of models that are in the confidence set of size 1 to modelSize. An interaction term between, say, variables x_1 and x_2 is displayed as “x_1 * x_2”; a squared term in, say, variable x_1 is displayed as “x_1 ^2”. If an interaction term is present without the corresponding main effects, the main effects should be added.

Acknowledgement

The work was supported by the UK Engineering and Physical Sciences Research Council under grant number EP/P002757/1.

Author(s)

Hoeltgebaum, H. H.

References

Cox, D. R. and Battey, H. S. (2017). Large numbers of explanatory variables, a semi-descriptive analysis. *Proceedings of the National Academy of Sciences*, 114(32), 8592-8595.

Battey, H. S. and Cox, D. R. (2018). Large numbers of explanatory variables: a probabilistic assessment. *Proceedings of the Royal Society of London, A.*, 474(2215), 20170631.

Hoeltgebaum, H., & Battey, H. S. (2019). HCmodelSets: An R Package for Specifying Sets of Well-fitting Models in High Dimensions. *The R Journal*, 11(2), 370-379.

See Also

`Reduction.Phase`, `Exploratory.Phase`

Examples

```
## Generates a random DGP
dgp = DGP(s=5, a=3, sigStrength=1, rho=0.9, n=100, intercept=5, noise=1,
          var=1, d=1000, DGP.seed = 2018)

#Reduction Phase using only the first 70 observations
outcome.Reduction.Phase =  Reduction.Phase(X=dgp$X[1:70,],Y=dgp$Y[1:70],
                                           family=gaussian, seed.HC = 1012)

# Exploratory Phase using only the first 70 observations, choosing the variables which
# were selected at least two times in the third dimension reduction

idxs = outcome.Reduction.Phase$List.Selection$`Hypercube with dim 2`$numSelected1
outcome.Exploratory.Phase =  Exploratory.Phase(X=dgp$X[1:70,],Y=dgp$Y[1:70],
                                               list.reduction = idxs,
                                               family=gaussian, signif=0.01)

# Model Selection Phase using only the remainer observations
sq.terms = outcome.Exploratory.Phase$mat.select.SQ
in.terms = outcome.Exploratory.Phase$mat.select.INTER

MS = ModelSelection.Phase(X=dgp$X[71:100,],Y=dgp$Y[71:100], list.reduction = idxs,
                          sq.terms = sq.terms,in.terms = in.terms, signif=0.01)
```

`Reduction.Phase` *Reduction by successive traversal of hypercubes proposed by Cox, D. R. & Battey, H. S. (2017)*

Description

This function traverses successively lower dimensional hypercubes, discarding variables according to the appropriate decision rules. It provides the number and indices of variables selected at each stage.

Usage

```
Reduction.Phase(X,Y,family=gaussian,
                dmHC=NULL,vector.signif=NULL,seed.HC = NULL, Cox.Hazard = FALSE)
```

Arguments

`X`	Design matrix.
`Y`	Response vector.
`family`	A description of the error distribution and link function to be used in the model. For glm this can be a character string naming a family function, a family function or the result of a call to a family function. See `family` for more details.
`dmHC`	Dimension of the hypercube to be used in the first-stage reduction. This version supports dimensions 2,3,4 and 5. If not specified a sensible value is calculated and used.
`vector.signif`	Vector of decision rules to be used at each stage of the reduction. The first value makes reference to the decision rule for the highest dimensional hypercube and so on. If values are less than 1, this specifies a significance level of a test. All variables significant at this level in at least half the analyses in which they appear will be retained. If the value is 1 or 2, variables are retained if they are among the 1 or 2 most significant in at least half the analyses in which they appear. If unspecified a default rule is used.
`seed.HC`	Seed for randomization of the variable indices in the hypercube. If not provided, the variables are arranged according to their original order.
`Cox.Hazard`	If TRUE fits proportional hazards regression model. The family argument will be ignored if Cox.Hazard=TRUE.

Value

`Matrix.Selection`	The number of variables selected at each reduction of the hypercube.
`List.Selection`	The indices of the variables retained through each stage of the reduction phase.

Acknowledgement

The work was supported by the UK Engineering and Physical Sciences Research Council under grant number EP/P002757/1.

Author(s)

Hoeltgebaum, H. H.

References

Cox, D. R. and Battey, H. S. (2017). Large numbers of explanatory variables, a semi-descriptive analysis. *Proceedings of the National Academy of Sciences*, 114(32), 8592-8595.

Battey, H. S. and Cox, D. R. (2018). Large numbers of explanatory variables: a probabilistic assessment. *Proceedings of the Royal Society of London, A.*, 474(2215), 20170631.

Hoeltgebaum, H., & Battey, H. S. (2019). HCmodelSets: An R Package for Specifying Sets of Well-fitting Models in High Dimensions. *The R Journal*, 11(2), 370-379.

Examples

```
## Generates a random DGP
dgp = DGP(s=5, a=3, sigStrength=1, rho=0.9, n=100, intercept=5, noise=1,
          var=1, d=1000, DGP.seed = 2018)

#Reduction Phase using only the first 70 observations
outcome.Reduction.Phase =  Reduction.Phase(X=dgp$X[1:70,],Y=dgp$Y[1:70],
                                           family=gaussian, seed.HC = 1012)

# Not run, using vector.signif argument
# Fixing a decision rule of getting the 2 most significant in the first reduction
# and in the subsequent reduction, only those variables significant at 0.001 level
# outcome.Reduction.Phase =  Reduction.Phase(X=dgp$X[1:70,],Y=dgp$Y[1:70],
#                            vector.signif = c(2,0.001), family=gaussian, dmHC = 3)
```

Index