Package: SIMEXBoost (via r-universe)

September 12, 2024

Type Package

Title Boosting Method for High-Dimensional Error-Prone Data

Version 0.2.0

Description Implementation of the boosting procedure with the simulation and extrapolation approach to address variable selection and estimation for high-dimensional data subject to measurement error in predictors. It can be used to address generalized linear models (GLM) in Chen (2023) <doi:10.1007/s11222-023-10209-3> and the accelerated failure time (AFT) model in Chen and Qiu (2023) <doi:10.1111/biom.13898>. Some relevant references include Chen and Yi (2021) <doi:10.1111/biom.13331> and Hastie, Tibshirani, and Friedman (2008, ISBN:978-0387848570).

License GPL-2

Encoding UTF-8

Imports MASS

NeedsCompilation yes

LazyData false

Author Bangxu Qiu [aut, cre], Li-Pang Chen [aut]

Maintainer Bangxu Qiu <1135427976@qq.com>

Repository CRAN

Date/Publication 2023-11-16 15:44:02 UTC

Contents

SIMEXBoost-package *Boosting Method for High-Dimensional Error-Prone Data*

Description

Implementation of the boosting procedure with the simulation and extrapolation (SIMEX) approach to address variable selection and estimation for high-dimensional data subject to measurement error in covariates.

Details

This package aims to do variable selection and estimation by using the boosting procedure. An important and ubiquitous feature in the dataset is measurement error in covariates. To handle measurement error effects, we employ the simulation and extrapolation (SIMEX) method. In addition to commonly used generalized linear models (GLM), we also consider the accelerated failure time (AFT) model to fit length-biased and interval-censored survival data.

Author(s)

Bangxu Qiu and Li-Pang Chen

Maintainer: Bangxu Qiu <1135427976@qq.com>

See Also

SIMEXBoost

Boost_VSE *Boosting Method for Variable Selection and Estimation*

Description

The function `Boost_VSE`, named after the Boosting procedure for Variable Selection and Estimation, is used to deal with regression models and data structures that are considered in `ME_Data`.

Usage

```
Boost_VSE(Y,Xstar,type="normal",Iter=200,Lambda=0)
```

Arguments

`Y`	Responses in the dataset. If `type` is specified as "normal", "binary", or "poisson", then `Y` should be a n-dimensional vector; if `type` is given by "AFT-normal" or "AFT-loggamma", then `Y` should be a (n,2) matrix of interval-censored responses, where the first column is the lower bound of an interval-censored response and the second column is the upper bound of an interval-censored response.

`Xstar`	An (n,p) matrix of covariates. They can be error-prone or precisely measured.
`type`	`type` reflects the specification of regression models. "normal" means the linear regression model with the error term generated by the standard normal distribution; "binary" means the logistic regression model; "poisson" means the Poisson regression model. In addition, the accelerated failure time (AFT) model is also considered to fit length-biased and interval-censored survival data. Specifically, "AFT-normal" represents the AFT model with the error term being normal distributions; "AFT-loggamma" represents the AFT model with the error term specified as log-gamma distributions.
`Iter`	The number of iterations for the boosting procedure. The default value is 100.
`Lambda`	A tuning parameter that aims to deal with the collinearity of covariates. "Lambda=0" means that no L2-norm is involved, and it is taken as a default value.

Details

This function aims to address variable selection and estimation for (ultra)high-dimensional data. This function can handle generalized linear models (in particular, linear regression models, logistic regression models, and Poisson regression models) and accelerated failure time models in survival analysis. When the input `Xstar` is precisely measured covariates, the resulting `BetaHat` is the vector of estimators; if the input `Xstar` is error-prone covariates, the resulting `BetaHat` is called "naive" estimator.

Value

`BetaHat`	the estimator obtained by the boosting method.

Author(s)

Bangxu Qiu and Li-Pang Chen

References

Chen, L.-P. (2023). De-noising boosting methods for variable selection and estimation subject to error-prone variables. *Statistics and Computing*, 33:38.

Chen, L.-P. and Qiu, B. (2023). Analysis of length-biased and partly interval-censored survival data with mismeasured covariates. *Biometrics*. To appear. <doi: 10.1111/biom.13898>

Hastie, T., Tibshirani, R. and Friedman, J. (2008). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

See Also

`ME_Data`

Examples

```
##### Example 1: A linear model with precisely measured covariates #####

X1 = matrix(rnorm((20)*400),nrow=400,ncol=20,byrow=TRUE)
```

```
data=ME_Data(X=X1,beta=c(1,1,1,rep(0,dim(X1)[2]-3)),
type="normal",sigmae=diag(0,dim(X1)[2]))

Y<-data$response
Xstar<-data$ME_covariate

Boost_VSE(Y,Xstar,type="normal",Iter=3)


##### Example 2: A linear model with error-prone covariates #####

X1 = matrix(rnorm((20)*400),nrow=400,ncol=20,byrow=TRUE)

data=ME_Data(X=X1,beta=c(1,1,1,rep(0,dim(X1)[2]-3)),
type="normal",sigmae=diag(0.3,dim(X1)[2]))

Y<-data$response
Xstar<-data$ME_covariate

Boost_VSE(Y,Xstar,type="normal",Iter=3)
```

ME_Data *Boosting Method for High-Dimensional Error-Prone Data*

Description

This function aims to generate artificial data with error-prone covariates.

Usage

```
ME_Data(X,beta,type="normal",sigmae,pr0=0.5)
```

Arguments

`X`	An (n,p) matrix of the "unobserved" covariates provided by users.
`beta`	An p-dimensional vector of parameters provided by users.
`type`	A regression model that is specified to generate the response. "normal" means the linear regression model with the error term generated by the standard normal distribution; "binary" means the logistic regression model; "poisson" means the Poisson regression model. In addition, the accelerated failure time (AFT) model is considered to fit length-biased and interval-censored survival data. Specifically, "AFT-normal" generates the length-biased and interval-censored survival data under the AFT model with the error term being normal distributions; "AFT-loggamma" generates the length-biased and interval-censored survival data under the AFT model with the error term being log-gamma distributions.

`sigmae` An (p,p) covariance matrix of the noise term in the classical measurement error model. Given `sigmae` with non-zero entries, one can generate the error-prone covariates. Moreover, if `sigmae` is given by the zero matrix, then the resulting covariate is the original input given by users.

`pr0` A numerical value in an interval (0,1). It is used to determine the censoring rate for the length-biased and interval-censored data. The default value is 0.5.

Details

This function aims to generate artificial data with error-prone covariates. Given generalized linear models (GLM), we generate an n-dimensional vector of responses. Linear regression models, logistic regression models, and Poisson regression models are particularly considered. In survival analysis, accelerated failure time (AFT) models are perhaps commonly used formulations. We use AFT models to generate length-biased and interval-censored responses. In addition to responses generated by specific regression models, we also employ the classical measurement error model to generate the mismeasured covariates.

Value

`response` Responses generated by a specific regression model. `type="normal"` gives a n-dimensional continuous vector; `type="binary"` gives a n-dimensional vector with binary entries; `type="poisson"` gives a n-dimensional vector with entries being counting numbers. In addition, `type="AFT-normal"` and `type="AFT-loggamma"` generates a (n,2) matrix of length-biased and interval-censored responses, where the first column is the lower bound of an interval-censored response and the second column is the upper bound of an interval-censored response.

`ME_covariate` an (n,p) matrix of error-prone covariates.

Author(s)

Bangxu Qiu and Li-Pang Chen

Examples

```
##### Example 1: A linear model with precisely measured covariates ##########
X<-matrix(rnorm((20)*400),nrow=400,ncol=20,byrow=TRUE)
data=ME_Data(X=X,beta=c(1,1,1,rep(0,dim(X)[2]-3)),type="normal",diag(0,dim(X)[2]))
Y<-data$response
Xstar<-data$ME_covariate



##### Example 2: A linear model with error-prone covariates ##########
X<-matrix(rnorm((20)*400),nrow=400,ncol=20,byrow=TRUE)
data=ME_Data(X=X,beta=c(1,1,1,rep(0,dim(X)[2]-3)),type="normal",diag(0.3,dim(X)[2]))
Y<-data$response
Xstar<-data$ME_covariate
```

SIMEXBoost *Boosting Method with SIMEX Correction for High-Dimensional Error-Prone Data*

Description

This function aims to address variable selection and estimation for (ultra)high-dimensional data subject to covariate measurement error, which are particularly considered in ME_Data.

Usage

```
SIMEXBoost(Y,Xstar,zeta=c(0,0.25,0.5,0.75,1),B=500,type="normal",sigmae,Iter=100,
Lambda=0,Extrapolation="linear")
```

Arguments

Y	Responses in the dataset. If type is specified as "normal", "binary", or "poisson", then Y should be a n-dimensional vector; if type is given by "AFT-normal" or "AFT-loggamma", then Y should be a (n,2) matrix of interval-censored responses, where the first column is the lower bound of an interval-censored response and the second column is the upper bound of an interval-censored response.
Xstar	An (n,p) matrix of the error-prone covariates.
zeta	A sequence of values used in the procedure of the SIMEX method. A default sequence is given by c(0,0.25,0.5,0.75,1).
B	The number of repetition in the SIMEX method. The default value is 500.
type	type reflects the specification of regression models. "normal" means the linear regression model with the error term generated by the standard normal distribution; "binary" means the logistic regression model; "poisson" means the Poisson regression model. In addition, the accelerated failure time (AFT) model is also considered to fit length-biased and interval-censored survival data. Specifically, "AFT-normal" represents the AFT model with the error term being normal distributions; "AFT-loggamma" represents the AFT model with the error term specified as log-gamma distributions.
sigmae	An (p,p) covariance matrix of the noise term in the classical measurement error model.
Iter	The number of iterations for the boosting procedure. The default value is 100.
Lambda	A tuning parameter that aims to deal with the collinearity of covariates. Lambda=0 means that no L2-norm is involved, and it is taken as the default value.
Extrapolation	A extrapolation function for the SIMEX method. Two choices are included: "linear" means a linear function; "quadratic" means a quadratic function. The default argument is "linear".

Details

This function aims to address variable selection and estimation for (ultra)high-dimensional data subject to covariate measurement error. In the SIMEX method, inputs of `B`, `zeta`, and `Extrapolation` are user-specific. Normally, larger values of `B` and `zeta` give a more precise estimator, and meanwhile, longer computational times. More detailed descriptions of the SIMEX method can be found in the following references.

Value

`BetaHatCorrect` the estimator obtained by SIMEXBoost.

Author(s)

Bangxu Qiu and Li-Pang Chen

References

Chen, L.-P. (2023). De-noising boosting methods for variable selection and estimation subject to error-prone variables. *Statistics and Computing*, 33:38.

Chen, L.-P. and Qiu, B. (2023). Analysis of length-biased and partly interval-censored survival data with mismeasured covariates. *Biometrics*. To appear. <doi: 10.1111/biom.13898>

Chen, L.-P. and Yi, G. Y. (2021). Analysis of noisy survival data with graphical proportional hazards measurement error models. *Biometrics*, 77, 956–969.

Hastie, T., Tibshirani, R. and Friedman, J. (2008). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

See Also

ME_Data Boost_VSE

Examples

```
##### Example 1: A linear model under default settings #####

X1 = matrix(rnorm((20)*400),nrow=400,ncol=20,byrow=TRUE)

data=ME_Data(X1,beta=c(1,1,1,rep(0,dim(X1)[2]-3)),
type="normal",
sigmae=diag(0.1,dim(X1)[2]))

Y = data$response
Xstar = data$ME_covariate

SIMEXBoost(Y,Xstar,B=2,zeta=c(0,0.5,1),
type="normal",Iter=3,sigmae=diag(0.1,dim(X1)[2]))



##### Example 2: An AFT model #####
```

```
X1 = matrix(rnorm((100)*400),nrow=400,ncol=100,byrow=TRUE)

data=ME_Data(X1,beta=c(1,1,1,rep(0,dim(X1)[2]-3)),pr0=0.3,
type="AFT-loggamma",
sigmae=diag(0.1,dim(X1)[2]))

Y = data$response
Xstar = data$ME_covariate

SIMEXBoost(Y,Xstar,B=2,zeta=c(0,0.5,1),
type="AFT-loggamma",Iter=3,sigmae=diag(0.1,dim(X1)[2]))
```

Index