Package: LIC (via r-universe)

November 4, 2024

Title The LIC Criterion for Optimal Subset Selection

Version 0.0.2

Description The LIC criterion is to determine the most informative subsets so that the subset can retain most of the information contained in the complete data. The philosophy of the package is described in Guo G. (2022) <doi:10.1080/02664763.2022.2053949>.

License MIT + file LICENSE

Encoding UTF-8

LazyData true

RoxygenNote 7.1.2

Imports stats

Suggests testthat (>= 3.0.0)

Config/testthat/edition 3

NeedsCompilation no

Author Guangbao Guo [aut, cre], Yue Sun [aut], Guoqi Qian [aut], Qian Wang [aut]

Maintainer Guangbao Guo <ggb11111111@163.com>

Depends R (>= 3.5.0)

Date/Publication 2022-03-28 06:00:02 UTC

Repository https://guangbaog.r-universe.dev

RemoteUrl https://github.com/cran/LIC

RemoteRef HEAD

RemoteSha 44972f592c29fb52373ae7e5faacfe3d3d1d8baf

Contents

airfoil 2
estate 3
gt2015 4

LIC 5
Opt1 6
Opt2 6
OSA 7
OSM 8

Index **9**

`airfoil` *Airfoil self-noise*

Description

The Airfoil self-noise data set

Usage

```
data("airfoil")
```

Format

A data frame with 1503 observations on the following 6 variables.

`V1` a numeric vector

`V2` a numeric vector

`V3` a numeric vector

`V4` a numeric vector

`V5` a numeric vector

`V6` a numeric vector

Details

The data set contains 1503 data points, including the 6 variables. Among them, the scaled sound pressure level is the dependent variable and the other five are independent variables.

Source

The Airfoil Self-Noise data set is from the NASA data set in UCI database.

References

T.F. Brooks, D.S. Pope, and A.M. Marcolini. Airfoil self-noise and prediction. Technical report, NASA RP-1218, July 1989.

Examples

```
data(airfoil)
## maybe str(airfoil) ; plot(airfoil) ...
```

estate — *Real estate valuation*

Description

The real estate valuation data set.

Usage

```
data("estate")
```

Format

A data frame with 414 observations on the following 8 variables.

- `No` a numeric vector
- `X1.transaction.date` a numeric vector
- `X2.house.age` a numeric vector
- `X3.distance.to.the.nearest.MRT.station` a numeric vector
- `X4.number.of.convenience.stores` a numeric vector
- `X5.latitude` a numeric vector
- `X6.longitude` a numeric vector
- `Y.house.price.of.unit.area` a numeric vector

Details

Real estate valuation data set contains information about 414 real estate prices of 5 independent variables. The dependent variable is the price per unit area.

Source

The data set is from Xindian District, New Taipei City, Taiwan.

References

Yeh, I. C., & Hsu, T. K. (2018). Building real estate valuation models with comparative approach through case-based reasoning. Applied Soft Computing, 65, 260-271.

Examples

```
data(estate)
## maybe str(estate) ; plot(estate) ...
```

gt2015 *Gas turbine NOx emission*

Description

The gas turbine NOx emission data set.

Usage

```
data("gt2015")
```

Format

A data frame with 7384 observations on the following 11 variables.

`AT` a numeric vector

`AP` a numeric vector

`AH` a numeric vector

`AFDP` a numeric vector

`GTEP` a numeric vector

`TIT` a numeric vector

`TAT` a numeric vector

`TEY` a numeric vector

`CDP` a numeric vector

`CO` a numeric vector

`NOX` a numeric vector

Details

To predict nitrogen oxide emissions, we use the gas turbine NOx emission data set in UCI database, which contains 36,733 instances of 11,733 sensor measurements. The pollutant emission factors of gas turbines include 9 variables. We select 7,200 data points in 2015.

Source

The gas turbine NOx emission data set is from UCI database.

References

NA

Examples

```
data(gt2015)
## maybe str(gt2015) ; plot(gt2015) ...
```

LIC *The LIC criterion is to determine the most informative subsets so that the subset can retain most of the information contained in the complete data.*

Description

The LIC criterion is to determine the most informative subsets so that the subset can retain most of the information contained in the complete data.

Usage

```
LIC(X, Y, alpha, K, nk)
```

Arguments

X	is a design matrix
Y	is a random response vector of observed values
alpha	is the significance level
K	is the number of subsets
nk	is the sample size of subsets

Value

MUopt,Bopt,MAEMUopt,MSEMUopt,opt,Yopt

Examples

```
set.seed(12)
X=matrix(data=sample(1:3,1200*5, replace = TRUE) ,nrow=1200,ncol=5)
b=sample(1:3,5, replace = TRUE)
e= rnorm(1200, 0, 1)
Y=X%*%b+e
alpha=0.05
K=10
nk=1200/K
LIC(X,Y,alpha,K,nk)
```

`Opt1` *The Opt1 chooses the optimal index subset based on minimized interval length.*

Description

The Opt1 chooses the optimal index subset based on minimized interval length.

Usage

```
Opt1(X, Y, alpha, K, nk)
```

Arguments

`X`	is a design matrix
`Y`	is a random response vector of observed values
`alpha`	is the significance level
`K`	is the number of subsets
`nk`	is the sample size of subsets

Value

MUopt1,Bopt1,MAEMUopt1,MSEMUopt1,opt1,Yopt1

Examples

```
set.seed(12)
X=matrix(data=sample(1:3,1200*5, replace = TRUE) ,nrow=1200,ncol=5)
b=sample(1:3,5, replace = TRUE)
e= rnorm(1200, 0, 1)
Y=X%*%b+e
alpha=0.05
K=10
nk=1200/K
Opt1(X,Y,alpha,K,nk)
```

`Opt2` *The Opt2 chooses the optimal index subset based on maximized information sub-matrix.*

Description

The Opt2 chooses the optimal index subset based on maximized information sub-matrix.

Usage

```
Opt2(X, Y, alpha, K, nk)
```

Arguments

X	is a design matrix
Y	is a random response vector of observed values
alpha	is the significance level
K	is the number of subsets
nk	is the sample size of subsets

Value

MUopt2,Bopt2,MAEMUopt2,MSEMUopt2,opt2,Yopt2

Examples

```
set.seed(12)
X=matrix(data=sample(1:3,1200*5, replace = TRUE) ,nrow=1200,ncol=5)
b=sample(1:3,5, replace = TRUE)
e= rnorm(1200, 0, 1)
Y=X%*%b+e
alpha=0.05
K=10
nk=1200/K
Opt2(X,Y,alpha,K,nk)
```

OSA *The OSA gives a simple average estimatoris by averaging all these least squares estimators.*

Description

The OSA gives a simple average estimatoris by averaging all these least squares estimators.

Usage

```
OSA(X, Y, alpha, K, nk)
```

Arguments

X	is a design matrix
Y	is a random response vector of observed values
alpha	is the significance level
K	is the number of subsets
nk	is the sample size of subsets

Value

MUA,BetaA,MAEMUA,MSEMUA

Examples

```
set.seed(12)
X=matrix(data=sample(1:3,1200*5, replace = TRUE) ,nrow=1200,ncol=5)
b=sample(1:3,5, replace = TRUE)
e= rnorm(1200, 0, 1)
Y=X%*%b+e
alpha=0.05
K=10
nk=1200/K
OSA(X,Y,alpha,K,nk)
```

OSM *The OSM is a median processing method for the central processor.*

Description

The OSM is a median processing method for the central processor.

Usage

```
OSM(X, Y, alpha, K, nk)
```

Arguments

X	is a design matrix
Y	is a random response vector of observed values
alpha	is the significance level
K	is the number of subsets
nk	is the sample size of subsets

Value

MUM,BetaM,MAEMUM,MSEMUM

Examples

```
set.seed(12)
X=matrix(data=sample(1:3,1200*5, replace = TRUE) ,nrow=1200,ncol=5)
b=sample(1:3,5, replace = TRUE)
e= rnorm(1200, 0, 1)
Y=X%*%b+e
alpha=0.05
K=10
nk=1200/K
OSM(X,Y,alpha,K,nk)
```

Index

* **datasets**
    airfoil, 2
    estate, 3
    gt2015, 4

airfoil, 2

estate, 3

gt2015, 4

LIC, 5

Opt1, 6
Opt2, 6
OSA, 7
OSM, 8