Package ‘smoothic’

December 15, 2022

Type Package

Title Variable Selection Using a Smooth Information Criterion

Version 1.0.0

Depends MASS, numDeriv, R (>= 3.5.0)

Maintainer Meadhbh O'Neill <meadhbh.oneill@ul.ie>

Description Implementation of the SIC epsilon-telescope method, either
using single or multi-parameter regression. Includes classical regression
with normally distributed errors and robust regression, where the errors are from
the Laplace distribution. The ``smooth generalized normal distribution'' is used,
where the estimation of an additional shape parameter allows the user to move
smoothly between both types of regression. See O'Neill and Burke (2022)
``Robust Distributional Regression with Automatic Variable Selection'' for more details.
<arXiv:2212.07317>. This package also contains the data analyses from O'Neill and
Burke (2021). ``Variable Selection Using a Smooth Information Criterion for
Multi-Parameter Regression Models''. <arXiv:2110.02643>.

License GPL-3

URL https://github.com/meadhbh-oneill/smoothic,
https://meadhbh-oneill.github.io/smoothic/

Encoding UTF-8

LazyData true

RoxygenNote 7.2.2

Suggests dplyr, ggplot2, knitr, rmarkdown, tidyr

VignetteBuilder knitr

NeedsCompilation no

Author Meadhbh O'Neill [aut, cre],
Kevin Burke [aut]

Repository CRAN

Date/Publication 2022-12-15 17:30:02 UTC

R topics documented:

`bostonhouseprice` *Boston House Price Data (Original)*

Description

Original data, which come from a study by Harrison Jr and Rubinfeld (1978), examining the association between median house prices in a particular community with various community characteristics. See `bostonhouseprice2` for the corrected version, with additional variables.

Usage

```
bostonhouseprice
```

Format

A data frame with 506 rows and 9 variables:

crime crimes committed per capita

rooms average number of rooms per house

radial index of accessibility to radial highways

stratio average student-teacher ratio of schools in the community

lowstat percentage of the population that are "lower status"

lnox log(annual average nitrogen oxide concentration (pphm))

lproptax log(property tax per $1000)

ldist log(weighted distances to five employment centres in the Boston region)

lprice log(median house price ($))

Source

`https://CRAN.R-project.org/package=wooldridge`

References

Harrison Jr, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. Journal of environmental economics and management, 5(1):81-102.

Wooldridge, J. M. (2015). Introductory econometrics: A modern approach. Cengage learning.

`bostonhouseprice2` *Boston House Price Data (Corrected Version)*

Description

Corrected data, which come from a study by Harrison Jr and Rubinfeld (1978), examining the association between median house prices in a particular community with various community characteristics. See `bostonhouseprice` for the original version.

Usage

```
bostonhouseprice2
```

Format

A data frame with 506 rows and 13 variables:

crim per capita crime rate by town

zn proportion of residential land zoned for lots over 25,000 sq.ft

indus proportion of non-retail business acres per town

rm average number of rooms per dwelling

age proportion of owner-occupied units built prior to 1940

rad index of accessibility to radial highways

ptratio pupil-teacher ratio by town

lnox log(nitric oxides concentration (parts per 10 million))

ldis log(weighted distances to five Boston employment centres)

ltax log(full-value property-tax rate per USD 10,000)

llstat log(percentage of lower status of the population)

chast Charles River dummy variable (=1 if tract bounds river; 0 otherwise)

lcmedv log(corrected median value of owner-occupied homes in USD 1000's)

Source

`https://CRAN.R-project.org/package=mlbench`

References

Harrison Jr, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. Journal of environmental economics and management, 5(1):81-102.

Leisch F, Dimitriadou E (2021). mlbench: Machine Learning Benchmark Problems. R package version 2.1-3.

`diabetes` *Diabetes Data*

Description

Data relating to a study of disease progression one year after baseline.

Usage

```
diabetes
```

Format

A data frame with 442 rows and 11 variables:

AGE age of the patient

SEX sex of the patient

BMI body mass index of the patient

BP blood pressure of the patient

S1 blood serum measurement 1

S2 blood serum measurement 2

S3 blood serum measurement 3

S4 blood serum measurement 4

S5 blood serum measurement 5

S6 blood serum measurement 6

Y quantitative measure of disease progression one year after baseline

Source

https://CRAN.R-project.org/package=lars

References

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. The Annals of Statistics.

`pcancer` *Prostate Cancer Data*

Description

Data, which come from a study by Stamey et al. (1989), examining the correlation between the level of prostate-specific antigen (PSA) and various clinical measures in men who were about the receive a radical prostatectomy.

Usage

```
pcancer
```

Format

A data frame with 97 rows and 9 variables:

lcavol log(cancer volume (cm^3))

lweight log(prostate weight (g))

age age of the patient

lbph log(amount of benign prostatic hyperplasia (cm^2))

svi presence of seminal vesicle invasion (1=yes, 0=no)

lcp log(capsular penetration (cm))

gleason Gleason score

pgg45 percentage of Gleason scores four of five

lpsa log(PSA (ng/mL))

Source

https://web.stanford.edu/~hastie/ElemStatLearn/datasets/prostate.data

References

Stamey, T. A., Kabalin, J. N., McNeal, J. E., Johnstone, I. M., Freiha, F., Redwine, E. A., and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. ii. radical prostatectomy treated patients. The Journal of urology, 141(5):1076-1083.

smoothic *Variable Selection Using a Smooth Information Criterion (SIC)*

Description

Implements the SIC ϵ-telescope method, either using single or multi-parameter regression. Returns estimated coefficients, estimated standard errors (SEE) and the value of the penalized likelihood function. Note that the function will scale the predictors to have unit variance, however, the final estimates are converted back to their original scale.

Usage

```
smoothic(
  formula,
  data,
  family = "sgnd",
  model = "mpr",
  lambda = "log(n)",
  epsilon_1 = 10,
  epsilon_T = 1e-04,
  steps_T = 100,
  zero_tol = 1e-05,
  max_it = 10000,
  optimizer = "nlm",
  kappa,
  tau,
  stepmax_nlm
)
```

Arguments

`formula` An object of class "`formula`": a two-sided object with response on the left hand side and the model variables on the right hand side.

`data` A data frame containing the variables in the model; the data frame should be unstandardized.

`family` The family of the model, default is `family = "sgnd"` for the "Smooth Generalized Distribution" where the shape parameter kappa is also estimated. Classical regression with normally distributed errors is performed when `family = "normal"`. If `family = "laplace"`, this corresponds to a robust regression with errors from the Laplace distribution.

`model` The type of regression to be implemented, either `model = "mpr"` for multi-parameter regression, or `model = "spr"` for single parameter regression (i.e., classical normal linear regression). Defaults to `model="mpr"`.

`lambda` Value of penalty tuning parameter. Suggested values are `"log(n)"` and `"2"` for the BIC and AIC respectively. Defaults to `lambda ="log(n)"` for the BIC case.

`epsilon_1` Starting value for ϵ-telescope. Defaults to 10.

`epsilon_T` Final value for ϵ-telescope. Defaults to `1e-04`.

`steps_T` Number of steps in ϵ-telescope. Defaults to 100.

`zero_tol` Coefficients below this value are treated as being zero. Defaults to `1e-05`.

`max_it` Maximum number of iterations to be performed before the optimization is terminated. Defaults to `1e+04`.

`optimizer` The optimization procedure to be used. Defaults to `optimizer = "nlm"`, where the `nlm` function from the **stats** package is used. This tends to be more stable than the manually coded Newton-Raphson procedure that is used when `optimizer = "manual"`.

`kappa` Optional user-supplied positive kappa value (> 0.2 to avoid computational issues) if `family = "sgnd"`. If supplied, the shape parameter kappa will be fixed to this value in the optimization. If not supplied, kappa is estimated from the data.

`tau` Optional user-supplied positive smoothing parameter value in the "Smooth Generalized Normal Distribution" if `family = "sgnd"` or `family = "laplace"`. If not supplied, then `tau = "0.15"`. Smaller values of `tau` bring the approximation closer to the absolute value function, but this can cause the optimization to become unstable. Some issues with standard error calculation with smaller values of `tau` when using the Laplace distribution in the robust regression setting.

`stepmax_nlm` Optional maximum allowable scaled step length (positive scalar) to be passed to `nlm` if `optimizer = "nlm"`. If not supplied, default values in `nlm` are used.

Value

A list with estimates and estimated standard errors.

- `coefficients` - vector of coefficients.
- `see` - vector of estimated standard errors.
- `model` - the matched type of model which is called.
- `plike` - value of the penalized likelihood function.
- `kappa` - value of the estimated/fixed shape parameter kappa if `family = "sgnd"`.

Author(s)

Meadhbh O'Neill

References

O'Neill, M. and Burke, K. (2021) Variable Selection Using a Smooth Information Criterion for Multi-Parameter Regression Models. <arXiv:2110.02643>

O'Neill, M. and Burke, K. (2022) Robust Distributional Regression with Automatic Variable Selection. <arXiv:2212.07317>

Examples

```
# Sniffer Data --------------------
# MPR Model ----
results <- smoothic(
  formula = y ~ .,
  data = sniffer,
  family = "normal",
  model = "mpr"
)
summary(results)
```

`sniffer` *Sniffer Data*

Description

Data examining the factors that impact the amount of hydrocarbon vapour released when gasoline is pumped into a tank.

Usage

```
sniffer
```

Format

A data frame with 125 rows and 5 variables:

tanktemp initial tank temperature (degrees F)

gastemp temperature of the dispensed gasoline (degrees F)

tankpres initial vapour pressure in the tank (psi)

gaspres vapour pressure of the dispensed gasoline (psi)

y hydrocarbons emitted (g)

Source

https://CRAN.R-project.org/package=alr4

References

Weisberg, S. (2014). Applied Linear Regression, 4th edition. Hoboken NJ: Wiley.

summary.smoothic *Summarising Smooth Information Criterion (SIC) Fits*

Description

summary method class “smoothic”

Usage

```
## S3 method for class 'smoothic'
summary(object, ...)
```

Arguments

object	an object of class “smoothic” which is the result of a call to smoothic.
...	further arguments passed to or from other methods.

Value

A list containing the following components:

- model - the matched model from the smoothic object.
- coefmat - a typical coefficient matrix whose columns are the estimated regression coefficients, estimated standard errors (SEE) and p-values.
- plike - value of the penalized likelihood function.

Author(s)

Meadhbh O’Neill

Examples

```
# Sniffer Data --------------------
# MPR Model ----
results <- smoothic(
  formula = y ~ .,
  data = sniffer,
  family = "normal",
  model = "mpr"
)
summary(results)
```

Index