Package 'StepReg'

October 13, 2024

Version 1.5.4

Title Stepwise Regression Analysis

Date 2024-10-12

Description The stepwise regression analysis is a statistical technique used to identify a subset of predictor variables essential for constructing predictive models. This package performs stepwise regression analysis across various regression models such as linear, logistic, Cox proportional hazards, Poisson, Gamma, and negative binomial regression. It incorporates diverse stepwise regression algorithms like forward selection, backward elimination, and bidirectional elimination alongside the best subset method. Additionally, it offers a wide range of selection criteria, including Akaike Information Criterion (AIC), Sawa Bayesian Information Criterion (BIC), and Significance Levels (SL). We validated the output accuracy of StepReg using public datasets within the SAS software environment. To facilitate efficient model comparison and selection, StepReg allows for multiple strategies and selection metrics to be executed in a single function call. Moreover, StepReg integrates a Shiny application for interactive regression analysis, broadening its accessibility.

License MIT + file LICENSE

BugReports https://github.com/JunhuiLi1017/StepReg/issues

VignetteBuilder knitr

Suggests knitr, testthat, BiocStyle, kableExtra

Imports dplyr, ggplot2, ggrepel, MASS, purrr, stringr, survival, flextable, cowplot, shiny, ggcorrplot, tidyr, summarytools, shinythemes, rmarkdown, DT, shinycssloaders, shinyjs

Encoding UTF-8

RoxygenNote 7.3.1

NeedsCompilation no

Repository CRAN

Author Junhui Li [cre] (<https://orcid.org/0000-0003-3973-1700>),
Junhui Li [aut],
Kai Hu [aut],
Xiaohuan Lu [aut],
Kun Cheng [ctb],
Sushmita N Nayak [ctb],

Cesar Bautista Sotelo [ctb],
Michael A Lodato [ctb],
Robert H Brown [ctb],
Wenxin Liu [aut],
Lihua Julie Zhu [aut]

Maintainer Junhui Li <junhui.li11@umassmed.edu>

Date/Publication 2024-10-12 23:40:12 UTC

Contents

`creditCard` *creditCard*

Description

Cross-section data on the credit history for a sample of applicants for a type of credit card. This dataset is from `CreditCard`

Usage

```
data(creditCard)
```

Format

A data frame containing 1,319 observations on 12 variables.

Details

- card Factor. Was the application for a credit card accepted?
- reports Number of major derogatory reports.
- age Age in years plus twelfths of a year.
- income Yearly income (in USD 10,000).
- share Ratio of monthly credit card expenditure to yearly income.
- expenditure Average monthly credit card expenditure.

- owner Factor. Does the individual own their home?
- selfemp Factor. Is the individual self-employed?
- dependents Number of dependents.
- months Months living at current address.
- majorcards Number of major credit cards held.
- active Number of active credit accounts.

For more information, refer to `CreditCard`

References

Greene, W.H. (2003). Econometric Analysis, 5th edition. Upper Saddle River, NJ: Prentice Hall.

`plot.StepReg` *Plots from a StepReg object*

Description

plot.StepReg visualizes the variable selection procedure using a StepReg object

Usage

```
## S3 method for class 'StepReg'
plot(x, num_digits = 6, ...)
```

Arguments

`x`	StepReg object
`num_digits`	The number of digits to keep when rounding the results. Default is 6.
`...`	Not used

Value

A list of plots comprising the selection detail plot and selection summary plot for each strategy.

Examples

```
## Not run:
data(mtcars)
formula <- mpg ~ .
x <- stepwise(formula = formula,
              data = mtcars,
              type = "linear",
              strategy = c("forward","bidirection","backward"),
              metric = c("AIC","BIC","SL"))
plot(x)

## End(Not run)
```

print.StepReg *Prints from a StepReg object*

Description

print.StepReg prints to console the from an object of class StepReg

Usage

```
## S3 method for class 'StepReg'
print(x, ...)
```

Arguments

x	StepReg object
...	further parameters

Value

formatted dataframe

remission *remission*

Description

A dataset containing the remission and 6 risk factors thought to be related to leukemia remission.

Usage

```
data(remission)
```

Format

A data frame with 27 rows and 7 columns.

Details

- remiss Indicates whether cancer remission occurred. A value of 1 indicates occurrence, while 0 indicates non-occurrence.
- cell Cellularity of the marrow clot section
- smear Smear differential percentage of blasts
- infil Percentage of absolute marrow leukemia cell infiltrate
- li Percentage labeling index of the bone marrow leukemia cells
- blast The absolute number of blasts in the peripheral blood
- temp The highest temperature before the start of treatment

References

Lee, E. T. (1974). "A Computer Program for Linear Logistic Regression Analysis." Computer Programs in Biomedicine 4:80–92.

https://online.stat.psu.edu/stat501/book/export/html/1011

`report` *report from a StepReg object*

Description

report output all tables in StepReg object to a report with format of html, docx, pptx, rtf, and xlsx.

Usage

```
report(x, report_name, format = c("html", "docx", "rtf", "pptx"))
```

Arguments

`x`	StepReg object
`report_name`	report name
`format`	the format of report, choose one or more from 'html', 'docx', 'rtf', 'pptx'. default is 'html'

Examples

```
## Not run:
data(mtcars)
mtcars$yes <- mtcars$wt
formula <- mpg ~ . + 0
x <- stepwise(formula = formula,
              data = mtcars,
              type = "linear",
              strategy = "bidirection",
              metric = c("AIC", "BIC"))
report(x,report_name = "report", format = c("html","docx"))

## End(Not run)
```

StepRegShinyApp *StepReg Shiny App*

Description

StepRegShinyApp is a Shiny application designed for performing stepwise regression analysis. In Step 1, users can upload their dataset, configure settings such as header, separator, and quotes, and select variables for distribution plots. In Step 2, users can choose the regression type (linear, logit, cox, poisson, gamma, or negbin), select dependent and independent variables, specify stepwise strategy (forward, backward, bidirectional, or subset), and set various metrics for model selection. The app dynamically adjusts input options based on the chosen regression type. Additionally, users can specify significant levels for entry and stay in the stepwise process. Finally, they can run the analysis to obtain stepwise regression results and visualize them through summary outputs and plots.

Usage

```
StepRegShinyApp()
```

stepwise *Main wrapper function for stepwise regression*

Description

Select optimal model using various stepwise regression strategies, e.g., Forward Selection, Backward Elimination, Bidirectional Elimination; meanwhile, it also supports Best Subset method. Four types of models are currently implemented: linear regression, logistic regression, Cox regression, Poisson, and Gamma regression. For selection criteria, a.k.a, stop rule, users can choose from AIC, AICc, BIC, HQ, Significant Level, and more.

Usage

```
stepwise(
  formula,
  data,
  type = c("linear", "logit", "cox", "poisson", "gamma", "negbin"),
  include = NULL,
  strategy = c("forward", "backward", "bidirection", "subset"),
 metric = c("AIC", "AICc", "BIC", "CP", "HQ", "adjRsq", "SL", "SBC", "IC(3/2)", "IC(1)"),
  sle = 0.15,
  sls = 0.15,
  test_method_linear = c("Pillai", "Wilks", "Hotelling-Lawley", "Roy"),
  test_method_glm = c("Rao", "LRT"),
  test_method_cox = c("efron", "breslow", "exact"),
  tolerance = 1e-07,
```

```
  weight = NULL,
  best_n = 3,
  num_digits = 6
)
```

Arguments

formula	(formula) The formula used for model fitting by defining the scope of dependent and independent variables. The formula takes the form of a '~' (tilde) symbol, with the response variable(s) on the left-hand side, and the predictor variable(s) on the right-hand side. The 'lm()' function uses this formula to fit a regression model. A formula can be as simple as 'y ~ x'. For multiple predictors, they must be separated by the '+' (plus) symbol, e.g. 'y ~ x1 + x2'. To include an interaction term between variables, use the ':' (colon) symbol: 'y ~ x1 + x1:x2'. Use the '.' (dot) symbol to indicate that all other variables in the dataset should be included as predictors, e.g. 'y ~ .'. In the case of multiple response variables (multivariate), the formula can be specified as 'cbind(y1, y2) ~ x1 + x2'. By default, an intercept term is always included in the models, to exclude it, include '0' or '- 1' in your formula: 'y ~ 0 + x1', 'y ~ x1 + 0', and 'y ~ x1 - 1'.
data	(data.frame) A dataset consisting of predictor variable(s) and response variable(s).
type	(character) The stepwise regression type. Choose from 'linear', 'logit', 'poisson', 'cox', 'gamma' and 'negbin'. Default is 'linear'. More information, see StepReg_vignettes
include	(NULL|character) A character vector specifying predictor variables that will always stay in the model. A subset of the predictors in the dataset.
strategy	(character) The model selection strategy. Choose from 'forward', 'backward', 'bidirectional' and 'subset'. Default is 'forward'. More information, see StepReg_vignettes
metric	(character) The model selection criterion (model fit score). Used for the evaluation of the predictive performance of an intermediate model. Choose from 'AIC', 'AICc', 'BIC', 'CP', 'HQ', 'adjRsq', 'SL', 'SBC', 'IC(3/2)', 'IC(1)'. Default is 'AIC'. More information, see StepReg_vignettes
sle	(numeric) Significance Level to Enter. It is the statistical significance level that a predictor variable must meet to be included in the model. E.g. if 'sle = 0.05', a predictor with a P-value less than 0.05 will 'enter' the model. Default is 0.15.
sls	(numeric) Significance Level to Stay. Similar to 'sle', 'sls' is the statistical significance level that a predictor variable must meet to 'stay' in the model. E.g. if 'sls = 0.1', a predictor that was previously included in the model but whose P-value is now greater than 0.1 will be removed.
test_method_linear	(character) Test method for multivariate linear regression analysis, choose from 'Pillai', 'Wilks', 'Hotelling-Lawley', 'Roy'. Default is 'Pillai'. For univariate regression, 'F-test' will be used.
test_method_glm	(character) Test method for logit, Poisson, Gamma, and negative binomial regression analysis, choose from 'Rao', 'LRT'. Default is 'Rao'. Only "Rao" is

available for strategy = 'subset'.

`test_method_cox`
: (character) Test method for cox regression analysis, choose from 'efron', 'breslow', 'exact'. Default is 'efron'.

`tolerance`
: (numeric) A statistical measure used to assess multicollinearity in a multiple regression model. It is calculated as the proportion of the variance in a predictor variable that is not accounted for by the other predictor variables in the model. Default is 1e-07.

`weight`
: (numeric) A numeric vector specifying the coefficients assigned to the predictor variables. The magnitude of the weight reflects the degree to which each predictor variable contributes to the prediction of the response variable. The range of weight should be from 0 to 1. Values greater than 1 will be coerced to 1, and values less than 0 will be coerced to 0. Default is NULL, which means that all weight are set equal.

`best_n`
: (numeric(integer)) The number of models to be retained in the process output. Default is 3, indicating that only the top 3 best models with the same number of variables are displayed. If all models are displayed, set it to Inf.

`num_digits`
: (numeric(integer)) The number of digits to keep when rounding the results. Default is 6.

Value

A list containing multiple tables will be returned.

- Summary of arguments for model selection: Arguments used in the stepwise function, either default or user-supplied values.
- Summary of variables in dataset: Variable names, types, and classes in dataset.
- Summary of selection process under xxx(strategy) with xxx(metric): Overview of the variable selection process under specified strategy and metric.
- Summary of coefficients for the selected model with xxx(dependent variable) under xxx(strategy) and xxx(metric): Coefficients for the selected models under specified strategy with metric. Please note that this table will not be generated for the strategy 'subset' when using the metric 'SL'.

Author(s)

Junhui Li, Kai Hu, Xiaohuan Lu

References

Alsubaihi, A. A., Leeuw, J. D., and Zeileis, A. (2002). Variable strategy in multivariable regression using sas/iml. , 07(i12).

Darlington, R. B. (1968). Multiple regression in psychological research and practice. Psychological Bulletin, 69(3), 161.

Dharmawansa, P. , Nadler, B. , & Shwartz, O. . (2014). Roy's largest root under rank-one alternatives:the complex valued case and applications. Statistics.

Hannan, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. Journal of the Royal Statistical Society, 41(2), 190-195.

Harold Hotelling. (1992). The Generalization of Student's Ratio. Breakthroughs in Statistics. Springer New York.

Hocking, R. R. (1976). A biometrics invited paper. the analysis and strategy of variables in linear regression. Biometrics, 32(1), 1-49.

Hurvich, C. M., & Tsai, C. (1989). Regression and time series model strategy in small samples. Biometrika, 76(2), 297-307.

Judge, & GeorgeG. (1985). The Theory and practice of econometrics /-2nd ed. The Theory and practice of econometrics /. Wiley.

Mallows, C. L. (1973). Some comments on cp. Technometrics, 15(4), 661-676.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). Multivariate analysis. Mathematical Gazette, 37(1), 123-131.

Mckeon, J. J. (1974). F approximations to the distribution of hotelling's t20. Biometrika, 61(2), 381-383.

Mcquarrie, A. D. R., & Tsai, C. L. (1998). Regression and Time Series Model strategy. Regression and time series model strategy /. World Scientific.

Pillai, K. . (1955). Some new test criteria in multivariate analysis. The Annals of Mathematical Statistics, 26(1), 117-121.

R.S. Sparks, W. Zucchini, & D. Coutsourides. (1985). On variable strategy in multivariate regression. Communication in Statistics- Theory and Methods, 14(7), 1569-1587.

Sawa, T. (1978). Information criteria for discriminating among alternative regression models. Econometrica, 46(6), 1273-1291.

Schwarz, G. (1978). Estimating the dimension of a model. Annals of Statistics, 6(2), pags. 15-18.

Examples

```
## perform multivariate linear stepwise regression with 'bidirection'
## strategy and 'AIC' stop rule, excluding intercept.
data(mtcars)
mtcars$yes <- mtcars$wt
formula <- cbind(mpg,drat) ~ . + 0
stepwise(formula = formula,
         data = mtcars,
         type = "linear",
         strategy = "bidirection",
         metric = "AIC")
## perform linear stepwise regression with 'bidirection' strategy and
## "AIC","SBC","SL","AICc","BIC", and "HQ" stop rule.
formula <- mpg ~ . + 1
stepwise(formula = formula,
         data = mtcars,
         type = "linear",
         strategy = c("forward","bidirection"),
         metric = c("AIC","SBC","SL","AICc","BIC","HQ"))
```

```
## perform logit stepwise regression with 'forward' strategy and significance
## level as stop rule.
data(remission)
formula <- remiss ~ .
stepwise(formula = formula,
         data = remission,
         type = "logit",
         strategy = "forward",
         metric = "SL",
         sle=0.05,
         sls=0.05)
```

---

tobacco *tobacco*

Description

data on chemical components of 25 tobacco leaf

Usage

```
data(tobacco)
```

Format

A data frame containing 25 observations on 9 variables.

Details

- cigarette Rate of cigarette burn in inches per 1000 seconds.
- sugar Percent sugar in the leaf.
- nicotine Percent nicotine.
- nitrogen Percentage of nitrogen.
- chlorine Percentage of chlorine.
- potassium Percentage of potassium.
- phosphorus Percentage of phosphorus
- calcium Factor. Percentage of calcium.
- magnesium Percentage of magnesium.

References

Anderson, R. L. and Bancroft, T. A. (1952), Statistical Theory in Research, McGraw-Hill Book Company, Inc., New York, NY.

vote *Vote for all models*

Description

Votes for all models across all combinations of strategies and metrics

Usage

```
vote(x, ...)
```

Arguments

x	each dataframe from outputlist
...	further parameters

Value

A dataframe with column names "model" and combinations of strategy and metric. The first column represents the model formula, and a checkmark indicates that the corresponding model was supported by the given strategy and metric combination. Please note that for the subset strategy, the "vote" will report the single best model across all numbers of variables under Information Criteria (IC). However, this rule should not be applied to Significance Level (SL) because the F/Rao value is only comparable for models with the same number of variables.

Examples

```
data(mtcars)
formula <- mpg ~ .
x <- stepwise(formula = formula,
              data = mtcars,
              type = "linear",
              strategy = c("forward","backward","subset"),
              metric = c("AIC","BIC"))
vote(x)
```

Index