Package ‘StepReg’

February 24, 2020

Type Package

Title Stepwise Regression Analysis

Version 1.4.0

Date 2020-02-22

Author Junhui Li,Xiaohuan Lu,Kun Cheng,Wenxin Liu

Maintainer Junhui Li <jjunhui11@cau.edu.cn>

StepReg-package

License GPL (>= 2)
Imports Rcpp (>= 0.12.13), statmod (>= 1.4.33), ResourceSelection (>= 0.3-5)
LinkingTo Rcpp, RcppEigen
Depends R (>= 2.10)
NeedsCompilation yes
Repository CRAN
Date/Publication 2020-02-24 12:50:09 UTC

R topics documented:

StepReg-package ................................................................. 2
bestsubset ............................................................... 4
ModelFitStat ................................................................. 6
stepOne ................................................................. 7
stepwise ................................................................. 9
stepwiselogit ............................................................ 11

Index 14

StepReg-package       Stepwise Regression Analysis

Description


Details

<table>
<thead>
<tr>
<th>Package:</th>
<th>StepReg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type:</td>
<td>Package</td>
</tr>
<tr>
<td>Version:</td>
<td>1.4.0</td>
</tr>
<tr>
<td>Date:</td>
<td>2020-2-22</td>
</tr>
<tr>
<td>License:</td>
<td>GPL (&gt;= 2)</td>
</tr>
</tbody>
</table>

Author(s)

Junhui Li, Xiaohuan Lu, Kun Cheng, Wenxin Liu
Maintainer: Junhui Li <junhuili@cau.edu.cn>

References

sas/iml., 07(i12).
of the Royal Statistical Society, 41(2), 190-195.
Springer New York.
John Wiley & Sons.
Judge, & George G. (1985). The Theory and practice of econometrics /-2nd ed. The Theory and 
practice of econometrics /. Wiley.
37(1), 123-131.


---

**bestsubset**

**Best subset selection**

**Description**

This function uses information criteria to find a specified number of best models containing one, two, or three variables, and so on, up to the single model containing all of the explanatory variables.

**Usage**

```r
bestsubset(data, y, exclude = NULL, include = NULL, Class = NULL, weights = c(rep(1, nrow(data))), select = "SBC", tolerance = 1e-07, best = 5)
```

**Arguments**

- **data**
  - Data set including dependent and independent variables to be analyzed
- **y**
  - Numeric or character vector for dependent variables
- **exclude**
  - Numeric or character vector for independent variables removed from stepwise regression analysis
- **include**
  - Forces the effects vector listed in the data to be included in all models. The selection methods are performed on the other effects in the data set
- **Class**
  - Class effect variable
- **weights**
  - The weights names numeric vector to provide a weight for each observation in the input data set. And note that weights should be ranged from 0 to 1, while negative numbers are forcibly converted to 0, and numbers greater than 1 are forcibly converted to 1. If you do not specify a weight vector, each observation has a default weight of 1.
- **select**
  - Specifies the criterion that uses to calculate all models including Akaike Information Criterion(AIC), the Corrected form of Akaike Information Criterion(AICc), Bayesian Information Criterion(BIC), Schwarz criterion(SBC), Hannan and Quinn Information Criterion(HQ), R-square statistic(Rsq), adjusted R-square statistic(adjRsq) and Mallows Cp statistic(CP)
- **tolerance**
  - Tolerance value for multicollinearity, default is 1e-7
- **best**
  - Controls the number of models displayed in the output, default is 5
Author(s)
Junhui Li

References

Examples
set.seed(4)
dfY <- data.frame(matrix(c(rnorm(20,0,2),c(rep(1,10),rep(2,10)),rnorm(20,2,3)),20,3))
colnames(dfY) <- paste("Y",1:3,sep="")
dfX <- data.frame(matrix(c(rnorm(100,0,2),rnorm(100,2,1))),20,10)
colnames(dfX) <- paste("X",1:10,sep="")
yx <- cbind(dfY,dfX)
bestsubset(yx,y="Y1",exclude="Y3",include="Y2",Class="Y2",
weights=c(rep(0.5,2),rep(1,18)),select="SBC",tolerance=1e-7,best=5)
ModelFitStat

*calculate model fit statistics*

Description

calculate model fit statistics adjRsq, AIC, AICc, BIC, CP, HQ, HQc, Rsq and SBC

Usage

ModelFitStat(Stattype,SSE,SST,n,nY,p,sigmaVal)

Arguments

| Stattype  | Model fit statistics adjRsq, AIC, AICc, BIC, CP, HQ, HQc, Rsq and SBC |
| SSE      | Sum of squares of error                                           |
| SST      | Total sum of squares corrected for the mean for the dependent variable |
| n        | Number of observation                                             |
| nY       | Number of dependent variable                                      |
| p        | Number of independent variable in the model including the intercept |
| sigmaVal | Estimate of pure error variance from fitting the full model       |

Author(s)

Junhui Li Xiaohuan Lu

References


---

**stepOne**

*Optimized residual models*

**Description**

Get the optimized residual models statistics with forward or backward direction in only one step

**Usage**

```r
stepOne(findIn, p, n, sigma, tolerance, Ftrace, criteria, Y, X1, X0, k, SST)
```

**Arguments**

- `findIn`: Logical value, if FALSE then add independent variable to regression model, otherwise remove independent variable from regression model.
- `p`: The number of independent variable entered in regression.
- `n`: The sample size.
- `sigma`: Pure error variance from full regression model for Bayesian information criterion (BIC).
- `tolerance`: Tolerance value for multicollinearity.
- `Ftrace`: Statistic of multivariate regression including Wilks’ lambda, Pillai trace and Hotelling-lawley trace.
- `criteria`: Information criterion including AIC, AICc, BIC, SBC, HQ, HQc and SL.
- `Y`: Data set for dependent variable.
- `X1`: Data set for independent variables not in regression model.
- `X0`: Data set for independent variables entered in regression model.
- `k`: Forces the first k effects entered in regression model, and the selection methods are performed on the other effects in the data set.
- `SST`: Total sum of squares corrected for the mean for the dependent variable.
Details

This function can compute probability value or information criteria statistics with multivariate and
univariate regression using least square method

Value

<table>
<thead>
<tr>
<th>PIC</th>
<th>P value or Information Criteria statistic value</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEQ</td>
<td>Pointer for independent variable enter or eliminate</td>
</tr>
<tr>
<td>SSE</td>
<td>Maximum or minimum of SSE</td>
</tr>
<tr>
<td>RkCh</td>
<td>Rank changed or not</td>
</tr>
</tbody>
</table>

Author(s)

Junhui Li

References

using sas/iml. , 07(i12).
of the Royal Statistical Society, 41(2), 190-195.
Springer New York.
Judge, & George G. (1985). The Theory and practice of econometrics /-2nd ed. The Theory and
practice of econometrics /. Wiley.
37(1), 123-131.
381-383.
and time series model selection /. World Scientific.
R.S. Sparks, W. Zucchini, & D. Coutsourides. (1985). On variable selection in multivariate regres-
Econometrica, 46(6), 1273-1291.
Stepwise Regression

Description

Stepwise regression analysis can be performed with univariate and multivariate based on information criteria specified, which includes 'forward', 'backward' and 'bidirection' direction model selection method. Also continuous variables nested within class effect and weighted stepwise are considered.

Usage

stepwise(data, y, exclude, include, Class, weights, selection, select, sle, sls, tolerance, Trace, Choose)

Arguments

data  Data set including dependent and independent variables to be analyzed
y     Numeric or character vector for dependent variables
exclude Numeric or character vector for independent variables removed from stepwise regression analysis
include Forces the effects vector listed in the data to be included in all models. The selection methods are performed on the other effects in the data set
Class  Class effect variable
weights The weights names numeric vector to provide a weight for each observation in the input data set. And note that weights should be ranged from 0 to 1, while negative numbers are forcibly converted to 0, and numbers greater than 1 are forcibly converted to 1. If you do not specify a weight vector, each observation has a default weight of 1.
selection Model selection method including "forward", "backward" and "bidirection", forward selection starts with no effects in the model and adds effects, backward selection starts with all effects in the model and removes effects, while bidirection regression is similar to the forward method except that effects already in the model do not necessarily stay there
select Specifies the criterion that uses to determine the order in which effects enter and/or leave at each step of the specified selection method including Akaike Information Criterion(AIC), the Corrected form of Akaike Information Criterion(AICc), Bayesian Information Criterion(BIC), Schwarz criterion(SBC), Hannan and Quinn Information Criterion(HQ), R-square statistic(Rsq), adjusted R-square statistic(adjRsq), Mallows Cp statistic(CP) and Significant Levels(SL)
sle    Specifies the significance level for entry
sls    Specifies the significance level for staying in the model
tolerance Tolerance value for multicollinearity, default is 1e-7
Trace Statistic for multivariate regression analysis, including Wilks' lambda ("Wilks"), Pillai Trace ("Pillai") and Hotelling-Lawley’s Trace ("Hotelling")

Choose Chooses from the list of models at the steps of the selection process the model that yields the best value of the specified criterion. If the optimal value of the specified criterion occurs for models at more than one step, then the model with the smallest number of parameters is chosen. Choose method includes AIC, AICc, BIC, HQ, HQc, SBC, Rsq, adjRsq, CP and NULL, if you do not specify the Choose option, then the model selected is the model at the final step in the selection process

Details

Multivariate regression and univariate regression can be detected by parameter 'y', where numbers of elements in 'y' is more than 1, then multivariate regression is carried out otherwise univariate regression

Author(s)

Junhui Li

References


**Examples**

```r
set.seed(4)
dfY <- data.frame(matrix(c(rnorm(20,0,2),c(rep(1,10),rep(2,10)),rnorm(20,2,3)),20,3))
colnames(dfY) <- paste("Y",1:3,sep="")
dfX <- data.frame(matrix(c(rnorm(100,0,2),rnorm(100,2,1)),20,10))
colnames(dfX) <- paste("X",1:10,sep="")
yx <- cbind(dfY,dfX)
tol <- 1e-7
Trace <- "Pillai"
sle <- 0.15
sls <- 0.15
# weights vector
w0 <- c(rep(0.5,2),rep(1,18))
w2 <- c(rep(0.5,3),rep(1,14),0.5,1,0.5)

# univariate regression with select = 'SBC' & choose = 'AIC' and select = 'CP' & choose = NULL
# without forced effect and continuous variable nested in class effect
stepwise(yx, y="Y1", exclude="Y3", include=NULL, Class=NULL, w0, selection="backward", select="SBC", sle, sls, tol, Trace, Choose='AIC')
stepwise(yx, y="Y1", exclude="Y3", include=NULL, Class=NULL, w0, selection="bidirection", select="CP", sle, sls, tol, Trace, NULL)

# univariate regression with select='AICc' & choose='HQc' and select='BIC' & choose = NULL
# with forced effect and continuous variable nested in class effect
stepwise(yx, y="Y1", exclude="Y3", include="Y2", Class="Y2", w2, selection="forward", select='AICc', sle, sls, tol, Trace, 'HQc')
stepwise(yx, y="Y1", exclude="Y3", include="Y2", Class="Y2", w2, selection="bidirection", 'BIC', sle, sls, tol, Trace, NULL)

# multivariate regression with select='HQ' & choose='BIC'
# with forced effect and continuous variable nested in class effect
stepwise(yx, y=c("Y1","Y3"), exclude=NULL, include="Y2", Class="Y2", w2, selection="bidirection", select='HQ', sle, sls, tol, Trace, 'BIC')
```

**Stepwise Logistic Regression**

**Description**

Stepwise logistic regression analysis selects model based on information criteria and Wald or Score test with 'forward', 'backward', 'bidirection' and 'score' model selection method.
Usage

    stepwiselogit(data, y, exclude = NULL, include = NULL, selection = "bidirection", select = "SL", sle = 0.15, sls = 0.15, goft = TRUE)

Arguments

data: Data set including dependent and independent variables to be analyzed
y: Numeric or character vector for dependent variables
exclude: Numeric or character vector for independent variables removed from stepwise regression analysis
include: Forces the effects vector listed in the data to be included in all models. The selection methods are performed on the other effects in the data set
selection: Model selection method including "forward", "backward", "bidirection" and 'score'. Forward selection starts with no effects in the model and adds effects, backward selection starts with all effects in the model and removes effects, while bidirection regression is similar to the forward method except that effects already in the model do not necessarily stay there, and score method uses information criteria to find a specified number of best models containing one, two, or three variables, and so on, up to the single model containing all of the explanatory variables.
select: Specifies the criterion that uses to determine the order in which effects enter and leave at each step of the specified selection method including Akaike Information Criterion(AIC), the Corrected form of Akaike Information Criterion(AICc), Schwarz criterion(SBC), and Significant Levels(SL)
sle: Specifies the significance level for entry, default is 0.15
sls: Specifies the significance level for staying in the model, default is 0.15
goft: performs the Hosmer and Lemeshow goodness-of-fit test for the case of a binary response model

Value

- RegressionModelsSelectedbyInformationCriterion: summary of regression models selected by information criterion
- SummaryOfSelection: summary of selection process
- AnalysisOfMaximumLikelihoodEstimate: analysis of maximum likelihood estimate for the selected model
- GoodnessOfTeste: Hosmer and Lemeshow goodness of fit (GOF) test

Author(s)

Junhui Li
References


Examples

```r
set.seed(1)
yd <- data.frame(sample(c(0,1),30,replace=TRUE))
colnames(yd) <- "remiss"
set.seed(4)
xd <- data.frame(matrix(c(round(rnorm(100,0,2),2),round(rnorm(140,2,4),2),
sample(c(1,0),30,replace=TRUE),sample(1:80,30,replace=TRUE)),30,10))
colnames(xd) <- c(paste("X",1:8,sep=""),"gender","age")
yx <- cbind(yd,xd)
y <- "remiss"
stepwiselogit(yx,y,selection="bidirection",select="IC(3/2)")
```
Index

*Topic **best subset selection**
  bestsubset, 4
*Topic **model fit statistics**
  ModelFitStat, 6
*Topic **package**
  StepReg-package, 2
*Topic **stepwise logistic regression**
  stepwiselogit, 11
*Topic **stepwise regression**
  stepOne, 7
  stepwise, 9

bestsubset, 4

ModelFitStat, 6

stepOne, 7
StepReg (StepReg-package), 2
StepReg-package, 2
stepwise, 9
stepwiselogit, 11