Package 'StepReg'

May 31, 2019

Type Package

Title Stepwise Regression Analysis

Version 1.2.0

Date 2019-05-30

Author Junhui Li,Kun Cheng,Wenxin Liu

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

Description Stepwise regression analysis for variable selection can be used to get the best candidate final regression model in univariate or multivariate regression analysis with the 'forward' and 'stepwise' steps. Procedure can use Akaike information criterion, corrected Akaike information criterion, Bayesian information criterion, Hannan and Quinn information criterion, corrected Hannan and Quinn information criterion, Schwarz criterion and significance levels as selection criteria. Multicollinearity detection in regression model are performed by checking tolerance value. Continuous variables nested within class effect and weighted stepwise regression are also considered in this package.

License GPL (>= 2)

Imports Rcpp (>= 0.12.13)

LinkingTo Rcpp,RcppEigen

Depends R (>= 2.10)

NeedsCompilation yes

Repository CRAN

Date/Publication 2019-05-31 08:40:03 UTC

R topics documented:

StepReg-package *Stepwise Regression Analysis*

Description

Stepwise regression analysis for variable selection can be used to get the best candidate final regression model in univariate or multivariate regression analysis with the 'forward' and 'stepwise' steps. Procedure uses Akaike information criterion, corrected Akaike information criterion, Bayesian information criterion, Hannan and Quinn information criterion, corrected Hannan and Quinn information criterion, Schwarz criterion and significance levels as selection criteria. Multicollinearity detection in regression model are performed by checking tolerance value. Continuous variables nested within class effect and weighted stepwise regression are also considered in this package.

Details

Package:	StepReg
Type:	Package
Version:	1.2.0
Date:	2019-05-30
License:	GPL (>= 2)

Author(s)

Junhui Li,Kun Cheng,Wenxin Liu

Maintainer: Junhui Li <junhuili@cau.edu.cn>

References

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

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

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 selection of variables in linear regression. Biometrics, 32(1), 1-49.

Hurvich, C. M., & Tsai, C. (1989). Regression and time series model selection 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 Selection. Regression and time series model selection /. World Scientific.

Pillai, K. C. S. (2006). Pillai's Trace. Encyclopedia of Statistical Sciences. John Wiley & Sons, Inc.

R.S. Sparks, W. Zucchini, & D. Coutsourides. (1985). On variable selection 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.

`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

References

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

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

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 selection of variables in linear regression. Biometrics, 32(1), 1-49.

Hurvich, C. M., & Tsai, C. (1989). Regression and time series model selection 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 Selection. Regression and time series model selection /. World Scientific.

Pillai, K. C. S. (2006). Pillai's Trace. Encyclopedia of Statistical Sciences. John Wiley & Sons, Inc.

R.S. Sparks, W. Zucchini, & D. Coutsourides. (1985). On variable selection 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.

`optimization` *Optimized residual models*

Description

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

Usage

```
optimization(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 regressoin 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

PIC	P value or Information Criteria statistic value
SEQ	Pointer for independent variable enter or eliminate
SSE	Maximum or minimum of SSE
RkCh	Rank changed or not

Author(s)

Junhui Li

References

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

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

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.

Hurvich, C. M., & Tsai, C. (1989). Regression and time series model selection 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.

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 Selection. Regression and time series model selection /. World Scientific.

Pillai, K. C. S. (2006). Pillai's Trace. Encyclopedia of Statistical Sciences. John Wiley & Sons, Inc.

R.S. Sparks, W. Zucchini, & D. Coutsourides. (1985). On variable selection 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.

`stepwise` *Stepwise Regression*

Description

'stepwise' function is used to do univariate and multivariate stepwise regression analysis and it includes 'forward' and 'stepwise' direction model selection method. Continuous variables nested within class effect and weighted stepwise are also considered. Besides, some common information criteria can be specified.

Usage

```
stepwise(data, y, notX, 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
`notX`	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" and "stepwise",forward selection starts with no effects in the model and adds effects, while stepwise 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' lamda ("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 regreesion

Author(s)

Junhui Li

References

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

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

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 selection of variables in linear regression. Biometrics, 32(1), 1-49.

Hurvich, C. M., & Tsai, C. (1989). Regression and time series model selection 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 Selection. Regression and time series model selection /. World Scientific.

Pillai, K. C. S. (2006). Pillai's Trace. Encyclopedia of Statistical Sciences. John Wiley & Sons, Inc.

R.S. Sparks, W. Zucchini, & D. Coutsourides. (1985). On variable selection 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

```
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)

#for univariate regression
y <- c("Y1")
notX <- c("Y3")
#for multivariate regression you can use this
ym <- c("Y1","Y3")
notXm <- NULL
#*  with continuous variable nested in class effect
ClassY2 <- c("Y2")
#*  without continuous variable nested in class effect
Class0 <- NULL
# without forced effect in regression model
inc0 <- NULL
# force the 'Y2' into the regression model
incY2 <- c("Y2")
sele <- 'stepwise'
```

```
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, notX, inc0, Class0, w0, sele, "SBC", sle, sls, tol, Trace, 'AIC')
stepwise(yx, y, notX, inc0, Class0, w0, sele, "CP", sle, sls, tol, Trace, NULL)

#univariate regression with select='AICc'  & choose='HQc' and select='Rsq' & choose = NULL
#with forced effect and continuous variable nested in class effect
stepwise(yx, y, notX, incY2, ClassY2, w2, sele, 'AICc', sle, sls, tol, Trace, 'HQc')
stepwise(yx, y, notX, incY2, ClassY2, w2, sele, 'Rsq', sle, sls, tol, Trace, NULL)

#multivariate regression with select='HQ'  & choose='BIC'
#with forced effect and continuous variable nested in class effect
stepwise(yx, ym, notXm, incY2, ClassY2, w2, sele, 'HQ', sle, sls, tol, Trace, 'BIC')
```

Index