Package ‘StepReg’

January 14, 2022

Title Stepwise Regression Analysis

Version 1.4.3

Date 2022-01-10

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

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

Description Three most common types of stepwise regression including linear regression, logistic regression and cox proportional hazard regression can be performed to select best model with methods of forward selection, backward elimination, bidirectional selection and best subset selection. A widely used selection criteria are available for variable selection.

License GPL (>= 2)

Imports survival

Encoding UTF-8

RoxygenNote 7.1.2

NeedsCompilation no

Repository CRAN

Date/Publication 2022-01-14 09:22:49 UTC

R topics documented:

modelFitStat 2
stepwise 3
stepwiseCox 6
stepwiseLogit 8

Index **11**

`modelFitStat` *Fit Model Statistics*

Description

Fit Model Statistics with least square or likelihood method to return an information criteria value

Usage

```
modelFitStat(ic, fit, method = c("LeastSquare", "Likelihood"), cox = FALSE)
```

Arguments

`ic`	Information criteria, including AIC, AICc, BIC, CP, HQ, HQc, Rsq, adjRsq and SBC
`fit`	Object of linear model or general linear model
`method`	Method to calculate information criteria value, including 'LeastSquare' and 'Likelihood'
`cox`	Compute model fit statistics for cox regression or not, where partial likelihood value will be used instead of the ordinary.

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

```
data(mtcars)
fit <- lm(mpg~wt+qsec+vs+am+gear+carb,data=mtcars)
modelFitStat("AIC",fit,"LeastSquare")
```

`stepwise` *Stepwise Linear Model Regression*

Description

Stepwise linear regression analysis selects model based on information criteria and F or approximate F test with 'forward', 'backward', 'bidirection' and 'score' model selection method.

Usage

```
stepwise(
  formula,
  data,
  include = NULL,
  selection = c("forward", "backward", "bidirection", "score"),
 select = c("AIC", "AICc", "BIC", "CP", "HQ", "HQc", "Rsq", "adjRsq", "SL", "SBC"),
  sle = 0.15,
  sls = 0.15,
  multivarStat = c("Pillai", "Wilks", "Hotelling-Lawley", "Roy"),
  weights = NULL,
  best = NULL
)
```

Arguments

`formula` Model formulae. The models fitted by the lm functions are specified in a compact symbolic form. The basic structure of a formula is the tilde symbol (~) and at least one independent (righthand) variable. In most (but not all) situations, a single dependent (lefthand) variable is also needed. Thus we can construct a formula quite simple formula (y ~ x). Multiple independent variables by simply separating them with the plus (+) symbol (y ~ x1 + x2). Variables in the formula are removed with a minus(-) symbol (y ~ x1 - x2). One particularly useful feature is the . operator when modelling with lots of variables (y ~ .). The %in% operator indicates that the terms on its left are nested within those on the right. For example y ~ x1 + x2 %in% x1 expands to the formula y ~ x1 + x1:x2. A model with no intercept can be specified as y ~ x - 1 or y ~ x + 0 or y ~ 0 + x. Multivariate multiple regression can be specified as cbind(y1,y2) ~ x1 + x2.

`data` Data set including dependent and independent variables to be analyzed

`include` Force vector of effects name to be included in all models.

`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 requests specifies the best-subset selection method, which uses the branch-and-bound technique to efficiently search for subsets of model effects that best predict the response variable.

`select` Specify the criterion that uses to determine the order in which effects enter and leave at each step of the specified selection method including "AIC","AICc","BIC","CP","HQ","HQc","Rs and "SL".

`sle` Specify the significance level for entry, default is 0.15

`sls` Specify the significance level for staying in the model, default is 0.15

`multivarStat` Statistic for multivariate regression analysis, including Wilks' lamda ("Wilks"), Pillai Trace ("Pillai"), Hotelling-Lawley's Trace ("Hotelling"), Roy's Largest Root ("Roy")

`weights` Numeric vector to provide a weight for each observation in the input data set. 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.

`best` Control the number of models displayed in the output, default is NULL, which means all possible model will be displayed.

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.

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 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. . (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 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

```
data(mtcars)
mtcars$yes <- mtcars$wt
formula <- cbind(mpg,drat) ~ . + 0
stepwise(formula=formula,
         data=mtcars,
         include=NULL,
         selection="bidirection",
         select="AIC",
         sle=0.15,
         sls=0.15,
         multivarStat="Pillai",
         weights=NULL,
         best=NULL)
```

stepwiseCox *Stepwise Cox Proportional Hazards Regression*

Description

Stepwise Cox regression analysis selects model based on information criteria and significant test with 'forward', 'backward', 'bidirection' and 'score' variable selection method.

Usage

```
stepwiseCox(
  formula,
  data,
  include = NULL,
  selection = c("forward", "backward", "bidirection", "score"),
  select = c("SL", "AIC", "AICc", "SBC", "HQ", "HQc", "IC(3/2)", "IC(1)"),
  sle = 0.15,
  sls = 0.15,
  method = c("efron", "breslow", "exact"),
  weights = NULL,
  best = NULL
)
```

Arguments

formula	Model formulae. The models fitted by the coxph functions are specified in a compact symbolic form. The basic structure of a formula is the tilde symbol (~) and at least one independent (righthand) variable. In most (but not all) situations, a single dependent (lefthand) variable is also needed. Thus we can construct a formula quite simple formula (y ~ x). Multiple independent variables by simply separating them with the plus (+) symbol (y ~ x1 + x2). Variables in the formula are removed with a minus(-) symbol (y ~ x1 - x2). One particularly useful feature is the . operator when modelling with lots of variables (y ~ .). The %in% operator indicates that the terms on its left are nested within those on the right. For example y ~ x1 + x2 %in% x1 expands to the formula y ~ x1 + x1:x2.
data	Data set including dependent and independent variables to be analyzed
include	Force 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 requests best subset selection.

`select`	Specify the criterion that uses to determine the order in which effects enter and leave at each step of the specified selection method including AIC, AICc, SBC, IC(1), IC(3/2), HQ, HQc and Significant Levels(SL)
`sle`	Specify the significance level for entry, default is 0.15
`sls`	Specify the significance level for staying in the model, default is 0.15
`method`	Specify the method for tie handling. If there are no tied death times all the methods are equivalent. Nearly all Cox regression programs use the Breslow method by default, but not this one. The Efron approximation is used as the default here, it is more accurate when dealing with tied death times, and is as efficient computationally. The “exact partial likelihood is equivalent to a conditional logistic model, and is appropriate when the times are a small set of discrete values.
`weights`	Numeric vector to provide a weight for each observation in the input data set. 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.
`best`	Control the number of models displayed in the output, default is NULL which means all possible model will be displayed

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

```
lung <- survival::lung
my.data <- na.omit(lung)
my.data$status1 <- ifelse(my.data$status==2,1,0)
data <- my.data
formula = Surv(time, status1) ~ . - status

stepwiseCox(formula,
data,
include=NULL,
selection=c("bidirection"),
select="HQ",
method=c("efron"),
sle=0.15,
sls=0.15,
weights=NULL,
best=NULL)
```

`stepwiseLogit` *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(
  formula,
  data,
  include = NULL,
  selection = c("forward", "backward", "bidirection", "score"),
  select = c("SL", "AIC", "AICc", "SBC", "HQ", "HQc", "IC(3/2)", "IC(1)"),
  sle = 0.15,
```

```
    sls = 0.15,
    sigMethod = c("Rao", "LRT"),
    weights = NULL,
    best = NULL
  )
```

Arguments

`formula`	Model formulae. The models fitted by the glm functions are specified in a compact symbolic form. The basic structure of a formula is the tilde symbol (~) and at least one independent (righthand) variable. In most (but not all) situations, a single dependent (lefthand) variable is also needed. Thus we can construct a formula quite simple formula (y ~ x). Multiple independent variables by simply separating them with the plus (+) symbol (y ~ x1 + x2). Variables in the formula are removed with a minus(-) symbol (y ~ x1 - x2). One particularly useful feature is the . operator when modelling with lots of variables (y ~ .). The %in% operator indicates that the terms on its left are nested within those on the right. For example y ~ x1 + x2 %in% x1 expands to the formula y ~ x1 + x1:x2. A model with no intercept can be specified as y ~ x - 1 or y ~ x + 0 or y ~ 0 + x.
`data`	Data set including dependent and independent variables to be analyzed
`include`	Force 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 requests best subset selection.
`select`	Specify the criterion that uses to determine the order in which effects enter and leave at each step of the specified selection method including AIC, AICc, SBC, IC(1), IC(3/2), HQ, HQc and Significant Levels(SL)
`sle`	Specify the significance level for entry, default is 0.15
`sls`	Specify the significance level for staying in the model, default is 0.15
`sigMethod`	Specify the method of significant test for variable to be entered in the model. "Rao" and "LRT" cab be chosen for Rao's efficient score test and likelihood ratio test.
`weights`	Numeric vector to provide a weights for each observation in the input data set. 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 weights vector, each observation has a default weights of 1.
`best`	Control the number of models displayed in the output, default is NULL which means all possible model will be displayed

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

```
formula=vs ~ .
stepwiseLogit(formula,
              data=mtcars,
              include=NULL,
              selection="bidirection",
              select="SL",
              sle=0.15,
              sls=0.15,
              sigMethod="Rao",
              weights=NULL,
              best=NULL)
```

Index

∗ **logistic**
stepwiseLogit, 8
∗ **regression**
stepwise, 3
stepwiseLogit, 8
∗ **stepwise**
stepwise, 3
stepwiseLogit, 8

modelFitStat, 2

stepwise, 3
stepwiseCox, 6
stepwiseLogit, 8