Package 'glmbb'

August 16, 2020

Version 0.4

Date 2020-08-07

Title All Hierarchical or Graphical Models for Generalized Linear Model

Author Charles J. Geyer <charlie@stat.umn.edu>.

Maintainer Charles J. Geyer <charlie@stat.umn.edu>

Depends R (>= 3.1.1)

Imports digest, stats

ByteCompile TRUE

Description Find all hierarchical models of specified generalized linear model with information criterion (AIC, BIC, or AICc) within specified cutoff of minimum value. Alternatively, find all such graphical models. Use branch and bound algorithm so we do not have to fit all models.

License MIT + file LICENSE

URL https://github.com/cjgeyer/glmbb

NeedsCompilation no

Repository CRAN

Date/Publication 2020-08-16 14:40:08 UTC

R topics documented:

crabs *Horseshoe Crab Mating Data*

Description

Data on horseshoe crabs (*Limulus polyphemus*). Response is number of males surrounding a breeding female, color (factor), condition (factor), weight (quantitative), and width (quantitative) of the female.

Usage

```
data(crabs)
```

Format

A data frame with 173 observations on 6 variables. Individuals (rows of the data frame) are female horseshoe crabs. Variables other than `satell` refer to these females. The variables are

color color. The colors given in Agresti are "light medium", "medium", "dark medium", and "dark". Here they are abbreviated to `light`, `medium`, `dark`, and `darker`, respectively.

spine spine condition. The conditions given in Agresti are "both good", "one worn or broken", and "both worn or broken". Here they are abbreviated to `good`, `middle`, `bad`, respectively.

width carapace width in centimeters

satell number of satellites, which males clustering around the female in addition to the male with which she is breeding.

weight weight in grams.

y shorthand for `as.numeric(satell > 0)`.

Details

Quoting from the abstract of Brockmann (1996). "Horseshoe crabs arrive on the beach in pairs and spawn . . . during . . . high tides. Unattached males also come to the beach, crowd around the nesting couples and compete with attached males for fertilizations. Satellite males form large groups around some couples while ignoring others, resulting in a nonrandom distribution that cannot be explained by local environmental conditions or habitat selection."

Source

Agresti, A. (2013) *Categorical Data Analysis*, Wiley, Hoboken, NJ., Section 4.3.2, http://users.stat.ufl.edu/~aa/cda/data.html

Brockmann, H. J. (1996) Satellite Male Groups in Horseshoe Crabs, *Limulus polyphemus*, *Ethology*, **102**, 1–21.

Examples

```
data(crabs)
gout <- glm(satell ~ color + spine + width + weight, family = poisson,
    data = crabs)
```

glmbb *All Hierarchical or Graphical Models for Generalized Linear Model*

Description

Find all hierarchical submodels of specified GLM with information criterion (AIC, BIC, or AICc) within specified cutoff of minimum value. Alternatively, all such graphical models. Use branch and bound algorithm so we do not have to fit all models.

Usage

```
glmbb(big, little = ~ 1, family = poisson, data,
    criterion = c("AIC", "AICc", "BIC"), cutoff = 10,
    trace = FALSE, graphical = FALSE, BIC.option = c("length", "sum"), ...)
```

Arguments

big	an object of class "formula" specifying the largest model to be considered. Model specified must be hierarchical. (See also glm and formula and 'Details' section below.)
little	a formula specifying the smallest model to be considered. The response may be omitted and if not omitted is ignored (the response is taken from big). Default is ~ 1. Model specified must be nested within the model specified by big.
family	a description of the error distribution and link function to be used in the model. This can be a character string naming a family function, a family function or the result of a call to a family function. (See family for details of family functions.)
data	an optional data frame, list or environment (or object coercible by as.data.frame to a data frame) containing the variables in the models. If not found in data, the variables are taken from environment(big), typically the environment from which glmbb is called.
criterion	a character string specifying the information criterion, must be one of "AIC" (Akaike Information Criterion, the default), "BIC" (Bayes Information Criterion) or "AICc" (AIC corrected for sample size). See section on AICc below.
cutoff	a nonnegative real number. This function finds all hierarchical models that are submodels of big and supermodels of little with information criterion less than or equal to the cutoff plus the minimum information criterion over all these models.
trace	logical. Emit debug info if TRUE.
graphical	logical. If TRUE search only over graphical models rather than hierarchical models.
BIC.option	a character string specifying the sample size n to be used in calculating BIC (ignored if criterion != "BIC"), must be either "length" or "sum" meaning either the length of the response vector (or number of rows if the response "vector" is actually a matrix) or the sum of the response (number of individuals classified if we are doing categorical data analysis). See section about BIC below. May be abbreviated.

...	additional named or unnamed arguments to be passed to glm.

Details

Typical value for big is something like foo ~ bar * baz * qux where foo is the response variable (or matrix when family is binomial or quasibinomial, see glm) and bar, baz, and qux are all the predictors that are considered for inclusion in models.

A model is hierarchical if it includes all lower-order interactions for each term. This is automatically what formulas with all variables connected by stars (*) do, like the example above. But other specifications are possible. For example, foo ~ (bar + baz + qux)^2 specifies the model with all main effects, and all two-way interactions, but no three-way interaction, and this is hierarchical.

A model m_1 is nested within a model m_1 if all terms in m_1 are also terms in m_2. The default little model ~ 1 is nested within every model except those specified to have no intercept by 0 + or some such (see formula).

The interaction graph of a model is the undirected graph whose node set is the predictor variables in the model and whose edge set has one edge for each pair of variables that are in an interaction term. A clique in a graph is a maximal complete subgraph. A model is graphical if it is hierarchical and has an interaction term for the variables in each clique. When graphical = TRUE only graphical models are considered.

Value

An object of class "glmbb" containing at least the following components:

data	the model frame, a data frame containing all the variables.
little	the argument little.
big	the argument big.
criterion	the argument criterion.
cutoff	the argument cutoff.
envir	an R environment object containing all of the fits done.
min.crit	the minimum value of the criterion.
graphical	the argument graphical.

BIC

It is unclear what the sample size, the n in the BIC penalty $n \log(p)$ should be. Before version 0.4 of this package the BIC was taken to be the result of applying R generic function BIC to the fitted object produced by R function glm. This is generally wrong whenever we think we are doing categorical data analysis (Raftery, 1986; Kass and Raftery, 1995). Whether we consider the sampling scheme to be Poisson, multinomial, or product multinomial (and binomial is a special case of product multinomial) the sample size is the total number of individuals classified and is the only thing that is considered as going to infinity in the usual asymptotics for categorical data analysis. This the option BIC.option = "sum" should always be used for categorical data analysis.

AICc

AICc was derived by Hurvich and Tsai only for normal response models. Burnham and Anderson (2002, p. 378) recommend it for other models when no other small sample correction is known, but this is not backed up by any theoretical derivation.

References

Burnham, K. P. and Anderson, D. R. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, second edition. Springer, New York.

Hand, D. J. (1981) Branch and bound in statistical data analysis. *The Statistician*, **30**, 1–13.

Hurvich, C. M. and Tsai, C.-L. (1989) Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.

Kass, R. E. and Raftery, A. E. (1995) Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

Raftery, A. E. (1986) A note on Bayes factors for log-linear contingency table models with vague prior information. *Journal of the Royal Statistical Society, Series B*, **48**, 249–250.

See Also

`family`, `formula`, `glm`, `isGraphical`, `isHierarchical`

Examples

```
data(crabs)
gout <- glmbb(satell ~ (color + spine + width + weight)^3,
    criterion = "BIC", data = crabs)
summary(gout)
```

---

isGraphical *Hierarchical and Graphical Models*

Description

Say whether a formula corresponds to a hierarchical model or a graphical model. Or return a formula for a hierarchical or a graphical model.

Usage

```
asGraphical(formula)
isGraphical(formula)
asHierarchical(formula)
isHierarchical(formula)
```

Arguments

`formula`	an object of class "`formula`".

Details

A model is hierarchical if for every interaction it contains all the main effects and lower-order interactions for variables in that interaction.

The interaction graph of a model is the undirected graph whose node set is the predictor variables in the model and whose edge set has one edge for each pair of variables that are in an interaction term. A clique in a graph is a maximal complete subgraph. A model is graphical if it is hierarchical and has an interaction term for the variables in each clique.

Value

For "is" functions, logical. `TRUE` if and only if the model is hierarchical or graphical, as the case may be.

For "as" functions, a formula for the smallest supermodel of the given model that is hierarchical or graphical, as the case may be.

Examples

```
isHierarchical(~ u * v)
isHierarchical(~ u : v)

isGraphical(~ u * v + u * w)
isGraphical(~ (u + v + w)^2)

asHierarchical(~ u:v + v:w)
asGraphical(~ (u + v + w)^2)
```

`summary.glmbb` *Summarize GLM Model Selection via Branch and Bound*

Description

These functions are all `methods` for class `glmbb` or `summary.glmbb` objects.

Usage

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

## S3 method for class 'summary.glmbb'
print(x, digits = max(3, getOption("digits") - 3),
    ...)
```

Arguments

object	an object of class "glmbb", usually, a result of a call to glmbb.
cutoff	a nonnegative real number. Only report on models having criterion value no larger than the minimum value plus cutoff. This argument may be omitted, in which case object$cutoff is used.
x	an object of class "summary.glmbb", usually, a result of a call to summary.glmbb.
digits	the number of significant digits to use when printing.
...	not used. Required by their generics.

Details

Let criterion denote the vector of criterion (AIC, BIC, or AICc) values for all of the models evaluated in the search. Those with criterion value greater than min(criterion) + cutoff are tossed.

We also define a vector weight by

```
w <- exp(- criterion / 2)
weight <- w / sum(w)
```

except that it is calculated differently to avoid overflow. These are so-called Akaike weights. They may or may not provide some guide as to how to deal with these models. For more see Burnham and Anderson (2002).

Value

summary.glmbb returns an object of class "summary.glmbb", a list with components

results	a data frame having variables
	criterion the vector criterion described in the Details section, in sorted order.
	weight the corresponding Akaike weights.
	formula the corresponding formulas describing the corresponding models.
cutoff.search	the cutoff argument to the call to glmbb that produced object.
cutoff.summary	the cutoff argument to the call to summary.glmbb.
criterion	a character variable giving the name of the criterion (AIC, BIC, or AICc). Not to be confused with results$criterion.

References

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed. Springer-Verlag, New York.

Examples

```
## For examples see those in help(glmbb)
```

`tidy.formula.hierarchical` *Shorten a Hierarchical Formula*

Description

Simplify a formula, assuming it is hierarchical, that is, an interaction implies all lower-order interactions and main effects involving the same variables are in the model.

Usage

```
tidy.formula.hierarchical(formula)
```

Arguments

`formula`	an object of class "`formula`".

Value

A character string coercible to a formula equivalent to the input.

Examples

```
tidy.formula.hierarchical(y ~ u + v + w + u:v + u:w + v:w + u:v:w)
```

Index