The hier.part Package

November 20, 2005

Version 1.0-1
Date 2005/11/20
Title Hierarchical Partitioning
Author Chris Walsh <Chris.Walsh@sci.monash.edu.au> and Ralph Mac Nally <Ralph.MacNally@sci.monash.edu.au>
Depends R(>= 1.9.0), gtools
Description Variance partition of a multivariate data set
Maintainer Chris Walsh <Chris.Walsh@sci.monash.edu.au>
License GPL

R topics documented:

all.regs ................................................................. 1
combos .................................................................... 3
hier.part ................................................................. 4
hier.part.data ........................................................... 6
hier.part.internal ....................................................... 7
partition .................................................................... 7
rand.hp .................................................................... 9

Index 11

all.regs Goodness of fit measures for a regression hierarchy

Description

Calculates goodness of fit measures for regressions of a single dependent variable to all combinations of N independent variables

Usage

all.regs(y, xcan, family = "gaussian", gof = "RMSPE", print.vars = FALSE)
Arguments

y         a vector containing the dependent variables
xcan      a dataframe containing the n independent variables
family    family argument of glm
gof       Goodness-of-fit measure. Currently "RMSPE", Root-mean-square 'prediction' error, "logLik", Log-Likelihood or "Rsqu", R-squared
print.vars if FALSE, the function returns a vector of goodness-of-fit measures. If TRUE, a data frame is returned with first column listing variable combinations and the second column listing goodness-of-fit measures.

Details

This function calculates goodness of fit measures for the entire hierarchy of models using all combinations of N dependent variables, and returns them as an ordered list ready for input into the function partition. This function requires the gtools package in the gregmisc bundle.

Value

gfs       If print.vars is FALSE, a vector of goodness of fit measures for all combinations of independent variables in the hierarchy or, if print.vars is TRUE, a data frame listing all combinations of independent variables in the first column in ascending order, and the corresponding goodness of fit measure for the model using those variables

Author(s)

Chris Walsh (Chris.Walsh@sci.monash.edu.au).

References


See Also

`hier.part`, `partition`, `rand.hp`

Examples

```r
#linear regression of log(electrical conductivity) in streams
#against seven independent variables describing catchment
#characteristics (from Hatt et al. 2004)
data(urbanwq)
env <- urbanwq[,2:8]
all.regs(urbanwq$lec, env, fam = "gaussian", gof = "Rsqu", print.vars = TRUE)

#logistic regression of an amphipod species occurrence in
```


#streams against four independent variables describing
#catchment characteristics (from Walsh et al. 2004)
data(amphipod)
env1 <- amphipod[,2:5]
all.regs(amphipod$australis, env1, fam = "binomial",
gof = "logLik", print.vars = TRUE)

---

**combos**

*All combinations of a hierarchy of models of n variables*

**Description**

Lists a matrix of combinations of 1 to n variables in ascending order

**Usage**

`combos(n)`

**Arguments**

- `n` an integer: the number of variables

**Details**

Lists hierarchy of all possible combinations of n variables in ascending order, starting with 1 variable, then all combinations of 2 variables, and so on until the one combination with all n variables. This function is used by `all.regs` to structure the models required for hierarchical partitioning. This function requires the gtools package in the gregmisc bundle.

**Value**

a list containing

- `ragged` a matrix with zeroes in empty elements and 1 in column 1, 2 in column 2 ... n in column n for full elements
- `binary` a matrix as for `ragged`, but with 1 in all full elements

**Author(s)**

Chris Walsh (Chris.Walsh@sci.monash.edu.au)

**See Also**

`all.regs`
hier.part  

**Goodness of fit calculation and hierarchical partitioning**

**Description**
Partitions variance in a multivariate dataset

**Usage**

```r
hier.part(y, xcan, family = "gaussian", gof = "RMSPE", barplot = TRUE)
```

**Arguments**

- `y` a vector containing the dependent variables
- `xcan` a dataframe containing the n independent variables
- `family` family argument of `glm`
- `gof` Goodness-of-fit measure. Currently "RMSPE", Root-mean-square ’prediction’ error, "logLik", Log-Likelihood or "Rsqu", R-squared
- `barplot` If TRUE, a barplot of I and J for each variable is plotted expressed as percentage of total explained variance.

**Details**
This function calculates goodness of fit measures for the entire hierarchy of models using all combinations of N independent variables using the function `all.regs`. It takes the list of goodness of fit measures and, using the `partition` function, applies the hierarchical partitioning algorithm of Chevan and Sutherland (1991) to return a simple table listing each variable, its independent contribution (I) and its conjoint contribution with all other variables (J).

Note earlier versions of hier.part (<1.0) produced a matrix and barplot of percentage distribution of effects as a percentage of the sum of all Is and Js, as portrayed in Hatt et al. (2004) and Walsh et al. (2004). This version plots the percentage distribution of independent effects only. The sum of Is equals the goodness of fit measure of the full model minus the goodness of fit measure of the null model.

The distribution of joint effects shows the relative contribution of each variable to shared variability in the full model. Negative joint effects are possible for variables that act as suppressors of other variables (Chevan and Sutherland 1991).

At this stage, the partition routine will not run for more than 12 independent variables. This function requires the gtools package in the gregmisc bundle

**Value**

- a list containing
  - `gfs` a data frame or vector listing all combinations of independent variables in the first column in ascending order, and the corresponding goodness of fit measure for the model using those variables
  - `IJ` a data frame of I, the independent and J the joint contribution for each independent variable
  - `I.perc` a data frame of I as a percentage of total explained variance
Note

The function produces a minor rounding error for analyses with more than 9 independent variables. To check if this error affects the inference from an analysis, run the analysis several times with the variables entered in a different order. There are no known problems for analyses with 9 or fewer variables.

Author(s)

Chris Walsh (Chris.Walsh@sci.monash.edu.au) using c and fortran code written by Ralph Mac Nally (Ralph.MacNally@sci.monash.edu.au).

References


See Also

all.regs, partition, rand.hp

Examples

# linear regression of log(electrical conductivity) in streams
# against seven independent variables describing catchment
# characteristics (from Hatt et al. 2004)
data(urbanwq)
env <- urbanwq[,2:8]
hier.part(urbanwq$lec, env, fam = "gaussian", gof = "Rsqu")

# logistic regression of an amphipod species occurrence in
# streams against four independent variables describing
# catchment characteristics (from Walsh et al. 2004)
data(amphipod)
env1 <- amphipod[,2:5]
hier.part(amphipod$australis, env1, fam = "binomial", gof = "logLik")
Example data sets for hier.part package.

Usage

\begin{verbatim}
data(urbanwq)
data(amphipod)
data(chevan)
\end{verbatim}

Details

urbanwq.txt

Seven catchment variables (fimp, catchment imperviousness\(^{0.25}\); sconn, drainage connection\(^{0.5}\); sdensep, septic tank density\(^{0.5}\); unsealden, unsealed road density; farea, catchment area\(^{0.25}\); selev, elevation\(^{0.5}\); amgeast, longitude) and median baseflow concentrations of three water quality variables in sixteen streams draining 16 independent subcatchments (ldoc, log(dissolved organic carbon); lec, log(electrical conductivity); lnox, log(nitrate/nitrite). Data from Hatt et al. (2004).

amphipod.txt

Presence-absence data for a stream-dwelling amphipod (australis) in 58 sites, with four catchment variables (fimp, imperviousness\(^{0.25}\); fconn, drainage connection\(^{0.25}\); densep, density of septic tanks; unseal, unsealed road density). Data from Walsh et al. (2004).

chevan.txt

Chi-squared (chisq) and R-squared (rsqu) for logistic and linear regression example from Chevan and Sutherland (1991)

Author(s)

Chris Walsh (Chris.Walsh@sci.monash.edu.au)

References


hier.part.internal  Internal functions for hier.part

Description

Internal functions for hier.part

partition

Hierarchical partitioning from a list of goodness of fit measures

Description

Partitions variance in a multivariate dataset from a list of goodness of fit measures

Usage

partition(gfs, pcan, var.names = NULL)

Arguments

gfs  an array as outputted by the function all.regs or a vector of goodness of fit measures from a hierarchy of regressions based on pcan variables in ascending order (as produced by function combos, but also including the null model as the first element)

pcan  the number of variables from which the hierarchy was constructed (maximum = 12)

var.names  an array of pcan variable names, if required

Details

This function applies the hierarchical partitioning algorithm of Chevan and Sutherland (1991) to return a simple table listing of each variable, its independent contribution (I) and its conjoint contribution with all other variables (J). The output is identical to the function hier.part, which takes the dependent and independent variable data as its input.

Note earlier versions of partition (hier.part<1.0) produced a matrix and barplot of percentage distribution of effects as a percentage of the sum of all Is and Js, as portrayed in Hatt et al. (2004) and Walsh et al. (2004). This version plots the percentage distribution of independent effects only. The sum of Is equals the goodness of fit measure of the full model minus the goodness of fit measure of the null model.

The distribution of joint effects shows the relative contribution of each variable to shared variability in the full model. Negative joint effects are possible for variables that act as suppressors of other variables (Chevan and Sutherland 1991).

At this stage, the partition routine will not run for more than 12 independent variables. This function requires the gtools package in the gregmisc bundle.
Value

a list containing

- **gfs**: a data frame listing all combinations of independent variables in the first column in ascending order, and the corresponding goodness of fit measure for the model using those variables
- **IJ**: a data frame of I, the independent and J the joint contribution for each independent variable
- **I.perc**: a data frame of I as a percentage of total explained variance
- **J.perc**: a data frame of J as a percentage of sum of all Js

Note

The function produces a minor rounding error for hierarchies constructed from more than 9 variables. To check if this error affects the inference from an analysis, run the analysis several times with the variables entered in a different order. There are no known problems for hierarchies with 9 or fewer variables.

Author(s)

Chris Walsh (Chris.Walsh@sci.monash.edu.au) using c and fortran code written by Ralph MacNally (Ralph.MacNally@sci.monash.edu.au).

References


See Also

- `all.regs`
- `partition`
- `rand.hp`

Examples

```r
#linear regression of log(electrical conductivity) in streams
#against seven independent variables describing catchment
#characteristics (from Hatt et al. 2004)
data(urbanwq)
env <- urbanwq[,2:8]
gofs <- all.regs(urbanwq$lec, env, fam = "gaussian",
goF = "Rsq", print.vars = TRUE)
partition(gofs, pcan = 7, var.names = names(urbanwq[,2,8]))

#hierarchical partitioning of logistic and linear regression
#goodness of fit measures from Chevan and Sutherland (1991)
data(chevan)
partition(chevan$chiSq, pcan = 4)
partition(chevan$rsqu, pcan = 4)
```
Randomization test for hierarchical partitioning

Description

Randomizes elements in each column in xcan and recalculates hier.part num.reps times

Usage

rand.hp(y, xcan, family = "gaussian", gof = "RMSPE", num.reps = 100)

Arguments

y a vector containing the dependent variables
xcan a dataframe containing the n independent variables
family family argument of glm
gof Goodness-of-fit measure. Currently "RMSPE", Root-mean-square 'Prediction' error, "NLL", Negative log Likelihood or "Rsqu", R-squared
num.reps Number of repeated randomizations

Details

This function is a randomization routine for the hier.part function which returns a matrix of I values (the independent contribution towards explained variance in a multivariate dataset) for all values from “num.reps” randomizations. For each randomization, the values in each variable (i.e each column of xcan) are randomized independently, and hier.part is run on the randomized xcan. As well as the randomized I matrix, the function returns a summary table listing the observed I values, the 95th and 99th percentile values of I for the randomized dataset.

Value

a list containing

Irands matrix of num.reps + 1 rows of I values for each dependent variable. The first row contains the observed I values and the remaining num.reps rows contains the I values returned for each randomization.
Iprobs data.frame of observed I values for each variable, Z-scores for the generated distribution of randomized Is and an indication of statistical significance. Z-scores are calculated as (observed - mean(randomizations))/sd(randomizations), and statistical significance (*) is based on upper 0.95 confidence limit (Z >= 1.65).

Author(s)

Chris Walsh (Chris.Walsh@sci.monash.edu.au).
References


See Also

`hier.part.partition`

Examples

```r
#linear regression of log(electrical conductivity) in streams
#against four independent variables describing catchment
#characteristics (from Hatt et al. 2004)
data(urbanwq)
env <- urbanwq[,2:5]
rand.hp(urbanwq$lec, env, fam = "gaussian", gof = "Rsqu")$Iprobs

#logistic regression of an amphipod species occurrence in
#streams against four independent variables describing
#catchment characteristics (from Walsh et al. 2004)
data(amphipod)
env1 <- amphipod[,2:5]
rand.hp(amphipod$australis, env1, fam = "binomial",
gof = "logLik")$Iprobs
```
Index

*Topic datasets
  hier.part.data, 5

*Topic internal
  hier.part.internal, 6

all.regs, 1, 3, 5, 8
amphipod(hier.part.data), 5
chevan(hier.part.data), 5
combos, 3
combos1(hier.part.internal), 6
current.model
  (hier.part.internal), 6
hier.part, 2, 3, 10
hier.part.data, 5
hier.part.internal, 6
partition, 2, 5, 7, 8, 10
rand.hp, 2, 5, 8, 8
urbanwq(hier.part.data), 5