Package 'catdap'

March 11, 2020

Version 1.3.5

Title Categorical Data Analysis Program Package

Author The Institute of Statistical Mathematics

Maintainer Masami Saga <msaga@mtb.biglobe.ne.jp>

Depends R (>= 3.2.0)

Suggests utils, datasets, methods

Imports graphics, grDevices

Description Categorical data analysis by AIC. The methodology is described in Sakamoto (1992) <ISBN 978-0-7923-1429-5>.

License GPL (>= 2)

MailingList Please send bug reports to ismrp@jasp.ism.ac.jp.

NeedsCompilation yes

R topics documented:

`catdap-package` *Categorical Data Analysis Program Package*

Description

R functions for categorical data analysis

Details

This package provides functions for analyzing multivariate data. Dependencies of the distribution of specified variable (response variable) to other variables (explanatory variables) are derived and evaluated by AIC (Akaike Information Criterion).

Functions `catdap1` and `catdap1c` are for the analysis of categorical data. Every variable is specified as the response variable in turn and the goodness of other variables as the explanatory variables to the specified variable is evaluated by AIC.

Function `catdap2` can be applied to data where categorical variable and numerical variable are mixed. Specifying one variable as the response variable, the dependencies of its distribution on sets of other variables are investigated. If the response variable is categorical, contingency table analysis method is employed. If the response variable is numerical, categorizing the response variable by pooling, the problem is reduced to the categorical response variable case. This method eventually finds the dependency of the histogram of numerical response variable on sets of explanatory variables.

The Fortran source program codes for above functions were published in Sakamoto, Ishiguro and Kitagawa (1983), and Frontiers of Times Series Modeling 3 : Modeling Seasonality & Periodicity ; ISM (2002), respectively.

References

Y.Sakamoto and H.Akaike (1978) *Analysis of Cross-Classified Data by AIC.* Ann. Inst. Statist. Math., 30, pp.185-197.

K.Katsura and Y.Sakamoto (1980) *Computer Science Monograph, No.14, CATDAP, A Categorical Data Analysis Program Package.* The Institute of Statistical Mathematics.

Y.Sakamoto, M.Ishiguro and G.Kitagawa (1983) *Information Statistics* Kyoritsu Shuppan Co., Ltd., Tokyo. (in Japanese)

Y.Sakamoto (1985) *Model Analysis of Categorical Data.* Kyoritsu Shuppan Co., Ltd., Tokyo. (in Japanese)

Y.Sakamoto (1985) *Categorical Data Analysis by AIC.* Kluwer Academic publishers.

An AIC-based Tool for Data Visualization (2015), NTT DATA Mathematical Systems Inc. (in Japanese)

Barplot2WayTable *Bar Plots for Two-Way Tables*

Description

Create bar plots for output two-way tables of catdap1() or catdap2().

Usage

```
Barplot2WayTable(vname, resvar, exvar = NULL, tway.table, interval = NULL)
```

Arguments

vname	variable names.
resvar	names of the response variables.
exvar	names of the explanatory variables. Default is all variables except resvar.
tway.table	output tway.table of catdap1 or catdap2.
interval	output interval of catdap2.

Details

For continuous variables, we assume that $b_1, b_2, \ldots, b_{m+1}$ are boundary values of m bins. Output value ranges r_i $(1 \leq i \leq m)$ are defined as follows :

$$r_i = [\ b_i,\ b_{i+1}\ )\ \text{ for } 1 \leq i < m,$$

$$r_m = [\ b_m,\ b_{m+1}\ ].$$

Examples

```
# catdap1c (Titanic data)
resvar <- "Survived"
z1 <- catdap1c(Titanic, resvar)

vname <- names(dimnames(Titanic))
Barplot2WayTable(vname, resvar, , z1$tway.table)

# catdap2 (Edgar Anderson's Iris Data)
# "Sepal.Length" "Sepal.Width" "Petal.Length" "Petal.Width" "Species"
data(iris)
resvar <- "Petal.Width"
vname <- names(iris)
z2 <- catdap2(iris, c(0, 0, 0, -7, 2), resvar, c(0.1, 0.1, 0.1, 0.1, 0))

exvar <- c("Sepal.Length", "Petal.Length")
Barplot2WayTable(vname, resvar, exvar, z2$tway.table, z2$interval)
```

catdap1 *Categorical Data Analysis Program Package 01*

Description

Calculates the degree of association between all the possible pairs of categorical variables.

Usage

```
catdap1(cdata, response.names = NULL, plot = 1, ask = TRUE)
catdap1c(ctable, response.names = NULL, plot = 1, ask = TRUE)
```

Arguments

`cdata`	categorical data matrix with variable names on the first row.
`ctable`	cross-tabulation table with a list of variable names.
`response.names`	variable names of response variables. If `NULL` (default), all variables are regarded as response variables.
`plot`	split directions for each level of the mosaic: **0 :** no plot, **1 :** horizontal (default), **2 :** alternating directions, beginning with a vertical split.
`ask`	logical; if `TRUE` (default), the user is asked to confirm before a new page is started. if `FALSE`, each new plot create a new page.

Details

This function is an R-function style clone of Sakamoto's CATDAP-01 program for categorical data analysis. CATDAP-01 calculates the degree of association between all the possible pairs of categorical variables.

The degree of association is evaluated by AIC value. See help(catdap2) for details about AIC.

`catdap2` should be used when the best subset and categorization of explanatory variables are sought for. Continuous explanatory variables could be explanatory variables in case of catdap2.

Value

`tway.table`	two-way tables and ratio.
`total`	total number of data with corresponding code of variables.
`aic`	AIC's of explanatory variables for each response variable.
`aic.order`	list of explanatory variable numbers arranged in ascending order of AIC.

References

Y.Sakamoto and H.Akaike (1978) *Analysis of Cross-Classified Data by AIC*. Ann. Inst. Statist. Math., 30, pp.185-197.

K.Katsura and Y.Sakamoto (1980) *Computer Science Monograph, No.14, CATDAP, A Categorical Data Analysis Program Package*. The Institute of Statistical Mathematics.

Y.Sakamoto, M.Ishiguro and G.Kitagawa (1983) *Information Statistics* Kyoritsu Shuppan Co., Ltd., Tokyo. (in Japanese)

Y.Sakamoto (1985) *Categorical Data Analysis by AIC.* Kluwer Academic publishers.

Examples

```
## example 1 (The Japanese National Character)
data(JNcharacter)
response <- c("born.again", "difficult", "pleasure", "women.job", "money")
catdap1(JNcharacter, response)

# or, simply
data(JNcharacter)
catdap1(JNcharacter)

## example 2 (Titanic data)
# A data set with 2201 observations on 4 variables (Class, Sex, Age and Survived)
# cross-tabulating data
catdap1c(Titanic, "Survived")

# individual data
x <- data.frame(Titanic)
y <- data.matrix(x)
n <- dim(y)[1]
nc <- dim(y)[2]
z <- array(, dim = c(nc-1, sum(y[, 5])))
k <- 1
for (i in 1:n)
  if (y[i, nc] != 0) {
    np <- y[i, nc]
    for (j in 1:(nc-1))
      z[j, k:(k+np-1)] <- dimnames(Titanic)[[j]][[y[i, j]]]
    k <- k + np
  }
data <- data.frame(aperm(array(z, dim = c(4,2201)), c(2,1)),
                   stringsAsFactors = TRUE)
names(data) <- names(dimnames(Titanic))
catdap1(data, "Survived")
```

Description

Search for the best single explanatory variable and detect the best subset of explanatory variables.

Usage

```
catdap2(data, pool = NULL, response.name, accuracy = NULL, nvar = NULL,
       additional.output = NULL, missingmark = NULL, pa1 = 1, pa2 = 4, pa3 = 10,
        print.level = 0, plot = 1)
```

Arguments

`data`	data matrix with variable names on the first row.
`pool`	the ways of pooling to categorize each variable must be specified by integer parameters: **(-m) < 0 :** m-bin histogram is employed to describe the distribution of continuous response variable (this option is valid only for the response variable), **0 :** equally spaced pooling via a top-down algorithm, **1 :** unequally spaced pooling via a bottom-up algorithm (default), **2 :** no pooling for discrete variables.
`response.name`	variable name of the response variable.
`accuracy`	minimum width for the discretization for each variable.
`nvar`	number of variables to be retained for the analysis of multidimensional tables. Default is the number of variables in `data`.
`additional.output`	list of sets of explanatory variable names for additional output.
`missingmark`	positive number for handling missing value. See 'Details'.
`pa1, pa2, pa3`	control parameter for size of the working area. If error message is output, please change the value of parameter according to it.
`print.level`	this argument determines the level of output printing. The default value of '`0`' means that lists of "AIC's of the models with k explanatory variables (k=1,2,...)" are printed. A value of '1' means that those lists are not printed and "Summary of subsets of explanatory variables" within the top 30 is listed.
`plot`	split directions for each level of the mosaic: **0 :** no plot, **1 :** horizontal (default), **2 :** alternating directions, beginning with a vertical split.

Details

This function is an R-function style clone of Sakamoto's CATDAP-02 program for categorical data analysis. CATDAP-02 can be used to search for the best subset of explanatory variables which have the most effective information on a specified response variable. Continuous explanatory variables could be explanatory variables. In that case CATDAP-02 searches for optimal categorization of continuous values.

The basic statistic adopted is obtained by the application of the statistic AIC to the models.

E denotes the response variable and F denotes candidate explanatory variable, and their cell frequencies by $n_E(i)(i \in (E))$ and $n_F(j)(j \in (F))$. The cross frequency is denoted by $n_F, F(i,j)$ $(i,j \in (E,F))$. To measure the strength of dependence of a specific set of response variables E on the explanatory variable F, we use the following statistic:

$$AIC(E;F) = -2 \sum_{i,j \in (E,F)} n_E, F(i,j)\ ln\{n_E, F(i,j)/(n_F(j)\} + 2(C_E - 1)C_F, \quad (1)$$

where C_E and C_F denote the total number of categories of the corresponding sets of variables, respectively.

The selection of the best subset of explanatory variables is realized by the search for F which gives the minimum $AIC(E;F)$.

In case of $F = \phi$, the formula (1) reduces to

$$AIC(E;\phi) = -2 \sum_{i \in (E)} n_E(i)\ ln\{n_E(i)/n\} + 2(C_E - 1).$$

Here it is assumed that $C_\phi = 1$ and $n_\phi(1) = n$.

Sakamoto's original CATDAP outputs $AIC(E;F) - AIC(E;\phi)$ as the AIC value instead of $AIC(E;F)$. By this way the positive value of AIC indicates that the variable F is judged to be useless as the explanatory variable of the E.

On the other hand, this policy make impossible to compare the goodness of the CATDAP model with other models, logit models for example.

Considering the convenience of users, present "R version CATDAP" provides not only $AIC = AIC(E;F) - AIC(E;\phi)$, but $AIC(E;\phi)$, either. The latter value is given as base_AIC in the output.

Users could recover $AIC(E;F)$ by adding AIC and base_AIC.

`missingmark` enables missing value handling. When a positive values, say 1000, is set here, any value, say x, greater than or equal to 1000 is treated as a missing value. If $1000 \leq x < 2000$, x is treated as a missing value of the 1st type. If $2000 \leq x < 3000$, x is treated as a missing value of the 2nd type, and so on. Generally speaking, any x that $1000k \leq x < 1000(k+1)$ is treated as the k-th type missing value. Users are referred to the reference for the technical details of the missing value handling procedure.

For continuous variables, we assume that $b_1, b_2, \ldots, b_{m+1}$ are boundary values of m bins. Output value ranges r_i $(1 \leq i \leq m)$ are defined as follows :

$$r_i = [\ b_i,\ b_{i+1}\ )\ \text{ for } 1 \leq i < m,$$
$$r_m = [\ b_m,\ b_{m+1}\ ].$$

Specifically, for continuous response variable V,

$$r_i = [\ x_{min} + (i-1) * s,\ x_{min} + i * s\ )\ \text{ for } 1 \leq i < m,$$
$$r_m = [\ x_{min} + (m-1) * s,\ x_{max}\ ],$$

where x_{min} and x_{max} are the minimum and the maximums of variable V respectively and $s = (x_{max} - x_{min})/m$.

Value

`tway.table`	two-way tables.
`total`	total number of data with corresponding code of variables.
`interval`	class interval for continuous and discrete explanatory variables.
`base.aic`	base_AIC.
`aic`	AIC's of single explanatory variables.
`aic.order`	list of explanatory variable numbers arranged in ascending order of AIC.
`nsub`	number of subsets of explanatory variables.
`subset`	list of subsets of explanatory variables in ascending order of AIC with the following components: **nv:** number of explanatory variables, **ncc:** number of categories, **aic:** AIC's, **exv:** explanatory variables, **vname:** explanatory variable names.
`ctable`	contingency table constructed by the best subset and additional subsets if any variables is specified by `additional.output`.
`ctable.interval`	class interval for continuous and discrete explanatory variables in contingency table.
`caic`	AIC of subset of explanatory variables in contingency table.
`missing`	number of types of the missing values for each variable.

References

K.Katsura and Y.Sakamoto (1980) *Computer Science Monograph, No.14, CATDAP, A Categorical Data Analysis Program Package.* The Institute of Statistical Mathematics.

Y.Sakamoto (1985) *Model Analysis of Categorical Data.* Kyoritsu Shuppan Co., Ltd., Tokyo. (in Japanese)

Y.Sakamoto (1985) *Categorical Data Analysis by AIC.* Kluwer Academic publishers.

An AIC-based Tool for Data Visualization (2015), NTT DATA Mathematical Systems Inc. (in Japanese)

Examples

```
# Example 1 (medical data "HealthData")
# as additional output, contingency tables for explanatory variable sets
# c("aortic.wav","min.press") and c("ecg","age") are obtained.

data(HealthData)
catdap2(HealthData, c(2, 2, 2, 0, 0, 0, 0, 2), "symptoms",
        c(0., 0., 0., 1., 1., 1., 0.1, 0.), ,
        list(c("aortic.wav", "min.press"), c("ecg", "age")))
```

```
# Example 2 (Edgar Anderson's Iris Data)
# continuous response variable handling and the usage of Barplot2WayTable
# function to visualize the result in shape of stacked histogram.

data(iris)
resvar <- "Petal.Width"
z <- catdap2(iris, c(0, 0, 0, -7, 2), resvar, c(0.1, 0.1, 0.1, 0.1, 0))
z

vname <- names(iris)
exvar <- c("Sepal.Length", "Petal.Length")
Barplot2WayTable(vname, resvar, exvar, z$tway.table, z$interval)

# Example 3  (in the case of a large number of variables)
data(HelloGoodbye)
pool <- rep(2, 56)

## using the default values of parameters pa1, pa2, pa3
## catdap2(HelloGoodbye, pool, "Isay", nvar = 10, print.level = 1, plot = 0)
## Error : Working area for contingency table is too short, try pa1 = 12.

### According to the error message, set the parameter p1 at 12, then ..
catdap2(HelloGoodbye, pool, "Isay", nvar = 10, pa1 = 12, print.level = 1,
        plot = 0)

# Example 4 (HealthData with missing values)
data(MissingHealthData)
catdap2(MissingHealthData, c(2, 2, 2, 0, 0, 0, 0, 2), "symptoms",
        c(0., 0., 0., 1., 1., 1., 0.1, 0.), missingmark = 300)
```

HealthData *Health Data*

Description

Medical data containing both continuous and categorical explanatory variables.

Usage

```
data(HealthData)
```

Format

A data frame with 52 observations on the following 8 variables.

A part of the source data was recoded according to an input example of original program CATDAP-02. In addition, we converted 1 into 'A' and 2 into 'B' of symptoms data, and converted cholesterol data less than 198 into 'low' and the others into 'high'.

[, 1] opthalmo. 1, 2

[, 2]	ecg	1, 2
[, 3]	symptoms	A, B
[, 4]	age	49-59
[, 5]	max.press	98-216
[, 6]	min.press	56-120
[, 7]	aortic.wav	6.3-10.2
[, 8]	cholesterol	low, high

Source

Y.Sakamoto, M.Ishiguro and G.Kitagawa (1980) *Computer Science Monograph, No.14, CATDAP, A CATEGORICAL DATA ANALYSIS PROGRAM PACKAGE, DATA No.2.* The Institute of Statistical Mathematics.

Y.Sakamoto (1985) *Categorical Data Analysis by AIC, p. 74.* Kluwer Academic publishers.

HelloGoodbye *Anonymous Binary Data*

Description

Real data contributed from an anonymous organization. We borrowed the wording of a famous song to hide the true nature of the data.

Usage

```
data(HelloGoodbye)
```

Format

A data frame of with 13954 observations (rows) and 56 variables (columns).

Source

An anonymous organization.

JNcharacter *The Japanese National Character*

Description

A part of the Survey on the Japanese National Character.

Usage

```
data(JNcharacter)
```

Format

A data frame with 85 observations on the following 10 variables.

A part of the source data was deleted and recoded according to an input example of original program CATDAP-01.

[, 1]	sex	1, 2
[, 2]	age	1, 2, 3, 4
[, 3]	pol.party	1, 2, 3, 4
[, 4]	education	1, 2, 3
[, 5]	occupation	1, 2
[, 6]	born.again	1, 2
[, 7]	difficult	1, 2
[, 8]	pleasure	1, 2
[, 9]	women.job	1, 2, 3
[, 10]	money	1, 2, 3

Source

K.Katsura and Y.Sakamoto (1980) *Computer Science Monograph, No.14, CATDAP, A Categorical Data Analysis Program Package, DATA No.1.* The Institute of Statistical Mathematics.

Y.Sakamoto, M.Ishiguro and G.Kitagawa (1983) *Information Statistics*, III-2, DATA No. 9, Kyoritsu Shuppan Co., Ltd., Tokyo. (in Japanese)

`MissingHealthData` *Health Data with Missing Values*

Description

Medical data containing both categorical variables and continuous variables, the latter include two variables with missing values.

Usage

```
data(MissingHealthData)
```

Format

A data frame with 52 observations on the following 8 variables.

A part of the source data was recoded according to an input example of original program CATDAP-02. In addition, we converted 1 into 'A' and 2 into 'B' of symptoms data, and converted cholesterol data less than 198 into 'low' and the others into 'high'.

[, 1]	opthalmo.	1, 2
[, 2]	ecg	1, 2
[, 3]	symptoms	A, B
[, 4]	age	49-59

[, 5]	max.press	98-216, 300 (missing value)
[, 6]	min.press	56-120, 300 (missing value)
[, 7]	aortic.wav	6.3-10.2
[, 8]	cholesterol	low, high

Source

Y.Sakamoto, M.Ishiguro and G.Kitagawa (1980) *Computer Science Monograph, No.14, CATDAP, A CATEGORICAL DATA ANALYSIS PROGRAM PACKAGE, DATA No.2.* The Institute of Statistical Mathematics.

Y.Sakamoto (1985) *Categorical Data Analysis by AIC, p. 74.* Kluwer Academic publishers.

Index