

Package ‘GoodmanKruskal’

March 18, 2020

Title Association Analysis for Categorical Variables

Version 0.0.3

Description Association analysis between categorical variables using the Goodman and Kruskal tau measure. This asymmetric association measure allows the detection of asymmetric relations between categorical variables (e.g., one variable obtained by re-grouping another).

Depends R (>= 3.2.4)

Imports classInt, corrplot

License MIT + file LICENSE

LazyData true

Suggests knitr, MASS, testthat, car, rmarkdown

VignetteBuilder knitr

RoxygenNote 7.0.2

NeedsCompilation no

Author Ron Pearson [aut, cre]

Maintainer Ron Pearson <ronald.k.pearson@gmail.com>

Repository CRAN

Date/Publication 2020-03-18 16:40:06 UTC

R topics documented:

GKtau	2
GKtauDataframe	3
GoodmanKruskal	4
GroupNumeric	4
plot.GKtauMatrix	5
Index	7

`GKtau`*Compute Goodman and Kruskal tau measure of association.*

Description

GKtau returns forward and backward Goodman and Kruskal tau measures between categorical variables.

Usage

```
GKtau(x, y, dgts = 3, includeNA = "ifany")
```

Arguments

<code>x</code>	A categorical vector (factor).
<code>y</code>	A categorical vector (factor).
<code>dgts</code>	Integer, number of digits for results; optional (default = 3).
<code>includeNA</code>	Character, passed to <code>useNA</code> parameter for table; default is "ifany"; other valid options are "no" and "always"

Details

The Goodman and Kruskal tau measure is an asymmetric association measure between two categorical variables, based on the extent to which variation in one variable can be explained by the other. This function returns a dataframe with both forward and backward associations.

Value

A one-row dataframe with the following columns:

- the names of the x and y variables,
- the numbers of distinct values N_x and N_y for each variable, and
- the forward and backward associations, $\tau(x,y)$ and $\tau(y,x)$.

Author(s)

Ron Pearson

Examples

```
x <- rep(c("a", "b", "c", "d"), each = 3)
y <- rep(c("a", "b", "c", "d"), times = 3)
z <- rep(c("a", "b", "a", "c"), each = 3)
GKtau(x, y)
GKtau(x, z)
GKtau(y, z)
```

GKtauDataframe *Compute Goodman and Kruskal's tau for a dataframe.*

Description

GKtauDataframe returns the square matrix of Goodman and Kruskal measures computed between each pair of columns in a dataframe. Numeric variables in the dataframe are treated as factors.

Usage

```
GKtauDataframe(df, dgts = 3, includeNA = "ifany")
```

Arguments

df	Dataframe from which to compute association measures.
dgts	Integer, number of digits for results; optional (default = 3).
includeNA	Character, passed to useNA parameter for table; default is "ifany"; other valid options are "no" and "always"

Details

The Goodman and Kruskal tau measure is an asymmetric association measure between two categorical variables, based on the extent to which variation in one variable can be explained by the other. This function returns an S3 object of class 'GKtauMatrix' that gives the number of levels for each variable on the diagonal of the matrix and the association between variables in the off-diagonal elements. Note that this matrix is generally NOT symmetric, in contrast to standard correlation matrices.

Value

An S3 object of class 'GKtauMatrix' consisting of a square matrix with one row and column for each column of the dataframe df. The structure of this matrix is:

- row and column names are the names of the variables in the dataframe.
- the diagonal matrix element contains the number of unique levels for the corresponding variable.
- off-diagonal matrix elements contain the forward Goodman-Kruskal tau association from the variable listed in the row names to the variable listed in the column names.

Author(s)

Ron Pearson

GoodmanKruskal	<i>GoodmanKruskal: The Goodman and Kruskal tau measure</i>
----------------	--

Description

Association analysis between categorical variables using the Goodman and Kruskal tau measure

GroupNumeric	<i>Group a numerical variable into an n-level factor</i>
--------------	--

Description

GroupNumeric converts a numerical variable x into a factor variable with n groups for categorical association analysis.

Usage

```
GroupNumeric(
  x,
  n = NULL,
  groupNames = NULL,
  orderedFactor = FALSE,
  style = "quantile",
  ...
)
```

Arguments

x	Numeric vector to be grouped.
n	Integer number of groups; if NULL (the default), the number of groups will be inferred from the groupNames parameter.
groupNames	Character vector of names for the levels of the factor variable created; if NULL (the default), the default names from R's cut function will be used.
orderedFactor	Logical, specifying whether the factor returned is ordered or not; default is FALSE.
style	Character string, passed to the classIntervals function from the classInt package as its style parameter (see help file for the classInterval function for details); default is "quantile".
...	Optional parameters passed to the classIntervals function from the classInt package.

Details

This function uses the classIntervals function from the classInt package to compute the breakpoints that define the groups. The style parameter is passed to the classIntervals function to specify the grouping method. Note that some methods may return a different number of groups than that requested via the n parameter. If groupNames is specified consistently with n, this different number of returned groups will cause an error. The recommended approach in this case is to either change the style parameter or to re-run without groupNames specified.

Value

Factor variable with n distinct levels, named according to groupNames (if specified).

Author(s)

Ron Pearson

plot.GKtauMatrix *Plot method for DataRobot S3 objects of class GKtauMatrix*

Description

Method for R's generic plot function for DataRobot S3 objects of class GKtauMatrix. This function generates an array of Goodman-Kruskal tau association measures as described under Details. Note that, in general, this matrix is asymmetric.

Usage

```
## S3 method for class 'GKtauMatrix'
plot(
  x,
  y,
  colorPlot = TRUE,
  corrColors = NULL,
  backgroundColor = "gray",
  diagColor = "black",
  diagSize = 1,
  ...
)
```

Arguments

x	S3 object of class GKtauMatrix to be plotted.
y	Not used; included for conformance with plot() generic function parameter requirements.
colorPlot	Logical variable indicating whether to generate a color plot (the default, for colorPlot = TRUE) or a black-and-white plot.

<code>corrColors</code>	Character vector giving the color names for the correlation values printed on the plot; default value is <code>NULL</code> , causing <code>rainbow(n)</code> to be used, where <code>n</code> is the number of rows and columns in the matrix <code>x</code> .
<code>backgroundColor</code>	Character variable naming the background color used for the correlation ellipses in the plot.
<code>diagColor</code>	Character variable naming the color of the text used to display the number of levels per variable along the diagonal of the correlation matrix plot.
<code>diagSize</code>	Numeric scale factor to adjust the text size for the number of levels displayed on the diagonal of the plot array.
<code>...</code>	Not used; included for conformance with <code>plot()</code> generic function parameter requirements.

Details

This function calls the `corrplot` function from the `corrplot` package to generate an array of correlation plots from the matrix of Goodman-Kruskal tau measures returned by the `GKtauDataframe` function. The off-diagonal elements of this array contain ellipses whose shape is determined by the corresponding tau value, and the diagonal elements display the number of unique levels for each variable in the plot.

This plot may be rendered either in color (the default, obtained by specifying `colorPlot` as `TRUE`) or black-and-white. In color plots, the color of the text for the correlation values is set by the `corrColors` parameter. The default value for this parameter is `NULL`, which causes the function to use the color vector `rainbow(n)` where `n` is the number of rows and columns in the `GKtauMatrix` object `x`. The background color used to fill in each ellipse is specified by the `backgroundColor` parameter, and the text for the diagonal entries is determined by the `diagColor` parameter. In cases where the default choices make the correlation values difficult to read, a useful alternative is to specify `corrColors = "blue"`.

Value

None. This function is called for its side-effect of generating a plot.

Author(s)

Ron Pearson

Index

GKtau, [2](#)

GKtauDataframe, [3](#)

GoodmanKruskal, [4](#)

GroupNumeric, [4](#)

plot.GKtauMatrix, [5](#)