Package 'MissingDataGUI'

December 27, 2015

Type Package

Title A GUI for Missing Data Exploration

Version 0.2-4

Date 2015-12-25

Author Xiaoyue Cheng, Dianne Cook, Heike Hofmann

Maintainer Xiaoyue Cheng <xycheng@unomaha.edu>

Description Provides numeric and graphical summaries for the missing values from both categorical and quantitative variables. A variety of imputation methods are applied, including the univariate imputations like fixed or random values, multivariate imputations like the nearest neighbors and multiple imputations, and imputations conditioned on a categorical variable.

Depends gWidgetsRGtk2, ggplot2

Imports GGally, cairoDevice (>= 2.23), grid, reshape

Suggests Hmisc, norm, mice, mi (>= 1.0)

License GPL (>= 2.0)

LazyData true

Collate 'MissingDataGUI-package.r' 'MissingDataGUI.r' 'WatchMissingValues.r' 'imputation.r' 'utils.r'

RoxygenNote 5.0.1

NeedsCompilation no

Repository CRAN

Date/Publication 2015-12-27 16:00:20

R topics documented:

`MissingDataGUI-package` *A Graphical User Interface for Exploring Missing Values in Data*

Description

This package was designed mainly for the exploration of missing values structure, and results of imputation, using static graphics and numerical summaries. A graphical user interface (GUI) makes it accessible to novice users.

References

Xiaoyue Cheng, Dianne Cook, Heike Hofmann (2015). Visually Exploring Missing Values in Multivariable Data Using a Graphical User Interface. Journal of Statistical Software, 68(6), 1-23. doi:10.18637/jss.v068.i06

Examples

```
if (interactive()) {
    MissingDataGUI()
}
```

`brfss` *The Behavioral Risk Factor Surveillance System (BRFSS) Survey Data, 2009.*

Description

The data is a subset of the 2009 survey from BRFSS, an ongoing data collection program designed to measure behavioral risk factors for the adult population (18 years of age or older) living in households.

Usage

```
data(brfss)
```

Details

Also see the codebook: http://ftp.cdc.gov/pub/data/brfss/codebook_09.rtf

Format: a data frame with 245 observations on the following 34 variables.

`STATE` A factor with 52 levels. The labels and states corresponding to the labels are as follows. 1:Alabama, 2:Alaska, 4:Arizona, 5:Arkansas, 6:California, 8:Colorado, 9:Connecticut, 10:Delaware, 11:District of Columbia, 12:Florida, 13:Georgia, 15:Hawaii, 16:Idaho, 17:Illinois, 18:Indiana, 19:Iowa, 20:Kansas, 21:Kentucky, 22:Louisiana, 23:Maine, 24:Maryland, 25:Massachusetts, 26:Michigan, 27:Minnesota, 28:Mississippi, 29:Missouri, 30:Montana, 31:Nebraska, 32:Nevada, 33:New Hampshire, 34:New Jersey, 35:New Mexico, 36:New York, 37:North Carolina, 38:North Dakota, 39:Ohio, 40:Oklahoma, 41:Oregon, 42:Pennsylvania, 44:Rhode Island, 45:South Carolina, 46:South Dakota, 47:Tennessee, 48:Texas, 49:Utah, 50:Vermont, 51:Virginia, 53:Washington, 54:West Virginia, 55:Wisconsin, 56:Wyoming, 66:Guam, 72:Puerto Rico, 78:Virgin Islands

`SEX` A factor with levels `Male` `Female`.

`AGE` A numeric vector from 7 to 97.

`HISPANC2` A factor with levels `Yes` `No` corresponding to the question: are you Hispanic or Latino?

`VETERAN2` A factor with levels `1` `2` `3` `4` `5`. The question for this variable is: Have you ever served on active duty in the United States Armed Forces, either in the regular military or in a National Guard or military reserve unit? Active duty does not include training for the Reserves or National Guard, but DOES include activation, for example, for the Persian Gulf War. And the labels are meaning: 1: Yes, now on active duty; 2: Yes, on active duty during the last 12 months, but not now; 3: Yes, on active duty in the past, but not during the last 12 months; 4: No, training for Reserves or National Guard only; 5: No, never served in the military.

`MARITAL` A factor with levels `Married` `Divorced` `Widowed` `Separated` `NeverMarried` `UnmarriedCouple`.

`CHILDREN` A numeric vector giving the number of children less than 18 years of age in household.

`EDUCA` A factor with the education levels `1` `2` `3` `4` `5` `6` as 1: Never attended school or only kindergarten; 2: Grades 1 through 8 (Elementary); 3: Grades 9 through 11 (Some high school); 4: Grade 12 or GED (High school graduate); 5: College 1 year to 3 years (Some college or technical school); 6: College 4 years or more (College graduate).

`EMPLOY` A factor showing the employment status with levels `1` `2` `3` `4` `5` `7` `8`. The labels mean – 1: Employed for wages; 2: Self-employed; 3: Out of work for more than 1 year; 4: Out of work for less that 1 year; 5: A homemaker; 6: A student; 7: Retired; 8: Unable to work.

`INCOME2` The annual household income from all sources with levels `<10k` `10-15k` `15-20k` `20-25k` `25-35k` `35-50k` `50-75k` `>75k` `Dontknow` `Refused`.

`WEIGHT2` The weight without shoes in pounds.

`HEIGHT3` The weight without shoes in inches.

`PREGNANT` Whether pregnant now with two levels `Yes` and `No`.

`GENHLTH` The answer to the question "in general your health is" with levels `Excellent` `VeryGood` `Good` `Fair` `Poor` `Refused`.

`PHYSHLTH` The number of days during the last 30 days that the respondent's physical health was not good. -7 is for "Don't know/Not sure", and -9 is for "Refused".

`MENTHLTH` The number of days during the last 30 days that the respondent's mental health was not good. -7 is for "Don't know/Not sure", and -9 is for "Refused".

`POORHLTH` The number of days during the last 30 days that poor physical or mental health keep the respondent from doing usual activities, such as self-care, work, or recreation. -7 is for "Don't know/Not sure", and -9 is for "Refused".

`HLTHPLAN` Whether having any kind of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare. The answer has two levels: `Yes` and `No`.

`CAREGIVE` Whether providing any such care or assistance to a friend or family member during the past month, with levels `Yes` and `No`.

`QLACTLM2` Whether being limited in any way in any activities because of physical, mental, or emotional problems, with levels `Yes` and `No`.

`DRNKANY4` Whether having had at least one drink of any alcoholic beverage such as beer, wine, a malt beverage or liquor during the past 30 days, with levels `Yes` and `No`.

`ALCDAY4` The number of days during the past 30 days that the respondent had at least one drink of any alcoholic beverage. -7 is for "Don't know/Not sure", and -9 is for "Refused".

`AVEDRNK2` The number of drinks on the average the respondent had on the days when he/she drank, during the past 30 days. -7 is for "Don't know/Not sure", and -9 is for "Refused".

`SMOKE100` Whether having smoked at least 100 cigarettes in the entire life, with levels `Yes` and `No`.

`SMOKDAY2` The frequency of days now smoking, with levels `Everyday` `Somedays` and `NotAtAll`(not at all).

`STOPSMK2` Whether having stopped smoking for one day or longer during the past 12 months because the respondent was trying to quit smoking, with levels `Yes` and `No`.

`LASTSMK1` A factor with levels `3` `4` `5` `6` `7` `8` corresponding to the question: how long has it been since last smokeing cigarettes regularly? The labels mean: 3: Within the past 6 months (3 months but less than 6 months ago); 4: Within the past year (6 months but less than 1 year ago); 5: Within the past 5 years (1 year but less than 5 years ago); 6: Within the past 10 years (5 years but less than 10 years ago); 7: 10 years or more; 8: Never smoked regularly.

`FRUIT` The number of fruit the respondent eat every year, not counting juice. -7 is for "Don't know/Not sure", and -9 is for "Refused".

`GREENSAL` The number of servings of green salad the respondent eat every year. -7 is for "Don't know/Not sure", and -9 is for "Refused".

`POTATOES` The number of servings of potatoes, not including french fries, fried potatoes, or potato chips, that the respondent eat every year. -7 is for "Don't know/Not sure", and -9 is for "Refused".

`CARROTS` The number of carrots the respondent eat every year. -7 is for "Don't know/Not sure", and -9 is for "Refused".

`VEGETABL` The number of servings of vegetables the respondent eat every year, not counting carrots, potatoes, or salad. -7 is for "Don't know/Not sure", and -9 is for "Refused".

`FRUITJUI` The number of fruit juices such as orange, grapefruit, or tomato that the respondent drink every year. -7 is for "Don't know/Not sure", and -9 is for "Refused".

`BMI4` Body Mass Index (BMI). Computed by WEIGHT in Kilograms/(HEIGHT in Meters * HEIGHT3 in Meters). Missing if any of WEIGHT2 or HEIGHT3 is missing.

Source

http://www.cdc.gov/brfss/data_documentation/index.htm

Examples

```
if (interactive()) {
    data(brfss)
    MissingDataGUI(brfss)
}
```

`compute_missing_pct` *Compute the numeric summary of the missingness*

Description

Compute the numeric summary of the missingness

Usage

```
compute_missing_pct(dat)
```

Arguments

`dat`	A data frame.

Value

A list including (1) a data frame 'missingsummary' that provides a table of missingness; (2) the total missing percentage; (3) the percent of variables that contain missing values; (4) the ratio of observations that have missings.

Author(s)

Xiaoyue Cheng <<xycheng@unomaha.edu>>

Examples

```
data(tao)
compute_missing_pct(tao)
```

`imputation` *Impute the missing data with the method selected under the condition.*

Description

This function provides eight methods for imputation with categorical varaibles as conditions.

Usage

```
imputation(origdata, method, vartype = NULL, missingpct = NULL, condition = NULL,
    knn = 5, mi.n = 3, mi.seed = 1234567, row_var = NULL)
```

Arguments

`origdata`	A data frame whose missing values need to be imputed. This data frame should be selected from the missing data GUI.
`method`	The imputation method selected from the missing data GUI. Must be one of 'Below 10 'MI:areg','MI:norm','MI:mice','MI:mi'. If method='MI:mice', then the methods of the variables containing NA's must be attached with argument method. If not, then default methods are used.
`vartype`	A vector of the classes of origdata. The length is the same as the number of columns of origdata. The value should be from "integer", "numeric", "logical", "character", "factor", and "ordered".
`missingpct`	A vector of the percentage of missings of the variables in origdata. The length is the same as the number of columns of origdata. The values should be between 0 and 1.
`condition`	A vector of categorical variables. The dataset will be partitioned based on those variables, and then the imputation is implemented in each group. There are no missing values in those variables. If it is null, then there is no division. The imputation is based on the whole dataset.
`knn`	number of the neighbors.
`mi.n`	number of the imputation sets for multiple imputation
`mi.seed`	random number seed for multiple imputation
`row_var`	A column name (character) that defines the ID of rows.

Details

The imputation methods: This list displays all the imputation methods. Users can make one selection. (1) 'Below 10 NA's of one variable will be replaced by the value which equals to the minimum of the variable minus 10 For categorical variables, NA's are treated as a new category. Under this status the selected conditioning variables are ignored. If the data are already imputed, then this item will show the imputed result. (2) 'Simple' will create three tabs: Median, Mean, and Random Value. 'Median' means NA's will be replaced by the median of this variable (omit NA's). 'Mean' means NA's will be replaced by the mean of the variable (omit NA's). The median does not apply to the nominal variable, neither does the mean to the categorical variable. In these cases the

mode (omit NA's) is provided. 'Random Value' means NA's will be replaced by any values of this variable (omit NA's) which are randomly selected. (3) 'Neighbor' contains two methods: 'Average Neighbor' and 'Random Neighbor'. 'Average Neighbor' will replace the NA's by the mean of the nearest neighbors. 'Random Neighbor' substitutes the missing for a random sample of the k nearest neighbors. The number of neighbors is default to 5, and can be changed by argument knn. The Neighbor methods require at lease one case to be complete, at least two variables to be selected, and no factor/character variables. The ordered factors are treated as integers. The method will return the overall mean or a global random sample value if the observation only contains NA's. (4) 'MI:areg' uses function `aregImpute` from package **Hmisc**. It requires at lease one case to be complete, and at least two variables to be selected. (5) 'MI:norm' uses function `imp.norm` from package **norm**. It requires all selected variables to be numeric(at least integer), and at least two variables to be selected. Sometimes it cannot converge, then the programme will leave NA's without imputation. (6) 'MI:mice' uses the **mice** package. The methods of the variables containing NA's must be attached with argument method. If not, then default methods are used. (7) 'MI:mi' employes the **mi** package.

Value

The imputed data frame with the last column being the row number from the original dataset. During the procedure of the function, rows may be exchanged, thus a column of row number could keep track of the original row number and then help to find the shadow matrix.

Author(s)

Xiaoyue Cheng <<xycheng@unomaha.edu>>

`MissingDataGUI` *The Starting of Missing Data GUI.*

Description

This function starts an open-files GUI, allowing 1) selecting one or more data files; 2)opening the main missing-data GUI for one data file. The missing data GUI consists of two tabs. In the summary tab, there are a list of all variables, a list of variables having missing values to color by, two radios for imputation methods and graph types respectively, a checkbox group for the conditional variables, four buttons and a graphics device. In the help tab, the layout is the same as the summary tab. But when the users move their mouse on those widgets, or click any of those radios or buttons, the functions of all widgets will be described at the place of the graphics device. The attributes of the variables can be changed. If the user double clicks on any variables in the top left table of missing-data GUI, an attribute window will pop up. Then the name could be edited, and the class could be changed to one of the four classes: integer, numeric, factor, and character. When a numeric variable is changed to a categorical variable, the condtions in the bottom left checkbox group will be updated. If the list of the color by variables is very long, the selector allows text entry to find the variable when this widget is active.

Usage

```
MissingDataGUI(data = NULL, width = 1000, height = 750)
```

Arguments

data	A data frame which is shown in the main missing-data GUI. If it is null, then the open-files GUI opens.
width	the width of window. Default to be 1000, and the minimal is 800.
height	the height of window. Default to be 750, and the minimal is 600.

Details

If more than one files are listed in the window but no file is focused when clicking the "Watch Missing Values", then the first file is selected for the main missing-data GUI. If more than one files are focused, then the first file of the focused files is selected for the main GUI.

Author(s)

Xiaoyue Cheng <<xycheng@unomaha.edu>>

Examples

```
if (interactive()) {
    MissingDataGUI()

    data(tao)
    MissingDataGUI(tao)

    data(brfss)
    MissingDataGUI(brfss)
}
```

---

scale_colour_discrete *Change the discrete color scale for the plots generated by ggplot2*

Description

Change the discrete color scale for the plots generated by ggplot2

Usage

```
scale_colour_discrete(...)
```

Arguments

...	parameters passed into the function

scale_fill_discrete *Change the discrete fill scale for the plots generated by ggplot2*

Description

Change the discrete fill scale for the plots generated by ggplot2

Usage

```
scale_fill_discrete(...)
```

Arguments

...	parameters passed into the function

tao *West Pacific Tropical Atmosphere Ocean Data, 1993 & 1997.*

Description

Real-time data from moored ocean buoys for improved detection, understanding and prediction of El Ni'o and La Ni'a.

Usage

```
data(tao)
```

Details

The data is collected by the Tropical Atmosphere Ocean project (http://www.pmel.noaa.gov/tao/index.shtml).

Format: a data frame with 736 observations on the following 8 variables.

`year` A factor with levels `1993` `1997`.

`latitude` A factor with levels `-5` `-2` `0`.

`longitude` A factor with levels `-110` `-95`.

`sea.surface.temp` Sea surface temperature(degree Celsius), measured by the TAO buoys at one meter below the surface.

`air.temp` Air temperature(degree Celsius), measured by the TAO buoys three meters above the sea surface.

`humidity` Relative humidity(buoys 3 meters above the sea surface.

`uwind` The East-West wind vector components(M/s). TAO buoys measure the wind speed and direction four meters above the sea surface. If it is positive, the East-West component of the wind is blowing towards the East. If it is negative, this component is blowing towards the West.

vwind The North-South wind vector components(M/s). TAO buoys measure the wind speed and direction four meters above the sea surface. If it is positive, the North-South component of the wind is blowing towards the North. If it is negative, this component is blowing towards the South.

Source

http://www.pmel.noaa.gov/tao/data_deliv/deliv.html

Examples

```
if (interactive()) {
    data(tao)
    MissingDataGUI(tao)
}
```

WatchMissingValues *The Main Window of Missing Data GUI.*

Description

This function is to open the missing data GUI. The widgets shown in the GUI include: a table of all variables in the dataset, a checkbox group of categorical variables to condition on, a table of variables which have missing values to coloy by, a radio of imputation methods, a radio of graph types, three command buttons, and a graphics device. In this GUI the user can: 1)change the name and class of the selected variable; 2)look at the numeric summary for the missing values in the selected variables; 3)look at the plot of imputed data, under one of the imputation methods and one of the graph types and one color-by variable, with or without the conditions; 4)export the imputed data as well as the missing shadow matrix, and save them to a data file(csv).

Usage

```
WatchMissingValues(h, data = NULL, gt = NULL, size.width = 1000, size.height = 750,
    ...)
```

Arguments

h	A list with components obj referring to the button "Watch Missing Values" in MissingDataGUI.
data	A data frame which is shown in the missing-data GUI. If it is null, then parameter gt must not be null.
gt	A widget created by gtable(). It should be passed from MissingDataGUI.
size.width	the width of window. Default to be 1000, and the minimal is 800.
size.height	the height of window. Default to be 750, and the minimal is 600.
...	Other parameters to be passed to this function.

Details

The missing data GUI consists of two tabs. In the summary tab, there are a list of all variables, a list of variables having missing values to color by, two radios for imputation methods and graph types respectively, a checkbox group for the conditional variables, four buttons and a graphics device. In the help tab, the layout is the same as the summary tab. But when the users move their mouse on those widgets, or click any of those radios or buttons, the functions of all widgets will be described at the place of the graphics device. The attributes of the variables can be changed. If the user double clicks on any variables in the top left table of missing-data GUI, an attribute window will pop up. Then the name could be edited, and the class could be changed to one of the four classes: integer, numeric, factor, and character. When a numeric variable is changed to a categorical variable, the condtions in the bottom left checkbox group will be updated. If the list of the color by variables is very long, the selector allows text entry to find the variable when this widget is active.

Author(s)

Xiaoyue Cheng <<xycheng@unomaha.edu>>

Examples

```
if (interactive()) {
    data(tao)
    WatchMissingValues(data = tao)
    data(brfss)
    WatchMissingValues(data = brfss)
}
```

Index