Package: SDCNway (via r-universe)

September 6, 2024

Depends R (>= 4.1.0)

Imports methods, plyr (>= 1.8.5), dplyr (>= 0.8.4), ggplot2 (>= 3.2.1), MASS (>= 3.6.0), Rdpack

RdMacros Rdpack

Title Tools to evaluate disclosure risk

Version 1.1.0

Description A package for calculating disclosure risk measures. This includes record-level measures primarily using exhaustive tabulation, as well as file-level measures using a loglinear model.

License GPL-2

Encoding UTF-8

LazyData true

RoxygenNote 7.2.0

Repository https://dataprotectiontoolkit.r-universe.dev

RemoteUrl https://github.com/dataprotectiontoolkit/sdcnway

RemoteRef HEAD

RemoteSha c2064517f42466090507f7067a54e29611e24aa2

Contents

`exampledata` *A subset of the 1992 National Adult Literacy Study (NALS) prison study public-use microdata file.*

Description

A subset of the 1992 National Adult Literacy Study (NALS) prison study public-use microdata file. It has 20 variables and 182 records.

Usage

```
data(exampledata)
```

Format

An object of class `"data.frame"`;

`sdc_extabs` *Calculate risk measures through exhaustive tabulations, Mu-Argus, and other methods.*

Description

This function primarily uses the exhaustive tabulation method to quantify disclosure risk. It tabulates cell counts for different combinations of variables provided by the user. Using these counts, this function identifies variable categories and records which are considered high risk for disclosure. File-level re-identification risk measures are also provided, e.g., Mu-Argus (Polettini 2003) and the risk metrics promosed in El Emam (2011).

Usage

```
sdc_extabs(
  data,
  ID = NULL,
  weight = NULL,
  varpool = names(data),
  forcelist = character(0),
  forcenum = 1,
  missingdef = list(),
  mindim = 1,
  maxdim = 2,
  threshold = NULL,
  wgtthreshold = NULL,
  condition = NULL,
  output_filename = NULL,
```

```
  tau1 = 0.2,
  tau2 = 0.2,
  include_mu_argus = TRUE
)

## S3 method for class 'sdc_extabs'
print(x, cutoff = 50, summary_outfile = NULL, ...)

## S3 method for class 'sdc_extabs'
plot(x, plotpath = NULL, plotvar1 = character(0), plotvar2 = character(0), ...)
```

Arguments

data	Data frame containing the data for which we are to measure disclosure risk. Unexpected behavior may result if any column name begins with a period.
ID	Name of column which identifies records. If NULL (default), an ID column named .ROW_NUMBER is created and used in reports.
weight	Column name for sampling weights. NULL or empty if none.
varpool	Vector of column names over which to form tables.
forcelist	Vector of variable names. Some are included in all tabulations. Optional.
forcenum	Number of variables in forcelist that are mandatory for all tabulations. That is, all tabulations will have a number of variables from forcelist exactly equal to forcenum.
missingdef	A named list specifying missing values. The names correspond to column names in data.
mindim	Integer specifying the minimum number of varpool variables (including forcelist variables) that can be used to form tables.
maxdim	Integer specifying the maximum number of varpool variables (including]code-forcelist variables) that can be used to form tables.
threshold	Threshold to determine the number of violations in terms of cell counts. If the number of cases in a cell is less than threshold, the cell is flagged as a violation. If threshold is NULL and wgthreshold is not NULL, then only a weighted threshold will be used. If both are NULL, threshold will be set to 3 and the weighted threshold will not be used.
wgtthreshold	Threshold to determine violations in terms of weighted cell counts. If NULL, a weighted threshold will not be used.
condition	Character string describing how weighted and unweighted thresholds are combined when both are used. If used, it must be "and" or "or" (case insensitive). This parameter is ignored if weight is NULL.
output_filename	Name of the csv file to save the data set with violation counts and Mu-Argus scores attached. NULL if no output file is to be saved.
tau1	A threshold to compute the risk measure, pRa. See User Manual for more details.

tau2	A threshold to compute the risk measure, jRa. This parameter is ignored if weight is NULL. See User Manual for more details.
include_mu_argus	Flag indicating whether Mu-Argus and El-Emam metrics should be calculated.
x	An object of class sdc_extabs, as returned by the sdc_extabs function.
cutoff	The number of variable categories with the highest percentage of cell violations for each table dimension. Default is 50.
summary_outfile	Name of summary output .txt file. If not NULL, console output is copied to the file. Default is NULL (no logging of output). Errors and warnings are not diverted (consider running in batch mode if logging of errors and warnings is needed).
...	Currently unused. For NextMethod compatibility.
plotpath	Directory to save plots. Plots are saved as *jpeg* files (quality = 100%). If the directory does not exist, it is first created. If plotpath is NULL (default), plots are not saved.
plotvar1	A vector of names of discrete variables for boxplots. If none, boxplots are not produced.
plotvar2	A vector of names of continuous variables for scatterplots. If none, scatterplots are not produced.

Details

If a specified missing value contains only whitespace, it will match any element with only whitespace. NA values in data are treated as missing regardless of missingdef. If you do not want NA values to be treated as missing, please recode them before passing the data to this function.

Note that if a weight variable is not provided, the number of statistics and plots that are produced is significantly reduced.

Value

An object of type sdc_extabs. Internally, a named list of statistics.

tabulation Cell counts and violation flags. Represented as a list with each element corresponding to a varpool combination.

data_with_statistics The original data with new columns showing statistics such as violation counts and Mu-Argus score for each record.

recoded_data_with_statistics Same as data_with_statistics but with missing value recodes.

mu_argus_summary Summary table of Mu-Argus by cell count. For this summary, all variables in varpool are used to define a cell. If weight is NULL, then this summary is omitted.

el_emam_measures List of file-level re-identification risk measures.

percent_violations_by_var_and_level Table with percent of records that are in violation for each variable/category.

percent_violations_by_dim_var_and_level Table with percent of cells that are in violation for each dimension/variable/category.

options Options provided to sdc_extabs by the user, such as missingdef, mindim, etc.

Methods (by generic)

- `print`: S3 print method for `sdc_extabs` objects
  Prints a nicely formatted version of the percent record violations by variable/category and percent cell violations by dimension/variable/category
- `plot`: S3 plot method for `sdc_extabs` objects
  Produces boxplots and scatterplots of violation counts and mu-argus scores.

References

El Emam K (2011). "Methods for the de-identification of electronic health records for genomic research." *Genome medicine*, **3**(4), 25.

Polettini S (2003). "Some remarks on the individual risk methodology." *Joint ECE/EUROSTAT Work Session on Data Confidentiality, Luxembourg*.

Examples

```
data(exampledata)
vars <- c("BIB1201", "BIC0501", "BID0101", "BIE0601", "BORNUSA", "CENREG",
          "DAGE3", "DRACE3", "EDUC3", "GENDER")
results <- sdc_extabs(exampledata,
                      ID="CASEID",
                      weight="WEIGHT",
                      varpool=vars,
                      mindim=2,
                      maxdim=3,
                      missingdef=list(BIE0601=5),
                      wgtthreshold=3000,
                      condition="or")
print(results, cutoff=15)
plot(results, plotvar1="BORNUSA", plotvar2="WEIGHT")
```

`sdc_loglinear` *sdc_loglinear*

Description

Calculates file-level risk measures using a loglinear model.

Usage

```
sdc_loglinear(
  data,
  weight,
  varpool,
  degree = 2,
  numiter = 40,
  epsilon = 0.001,
```

```
  blanks_as_missing = TRUE,
  output_filename = NULL
)

## S3 method for class 'sdc_loglinear'
print(x, summary_outfile = NULL, ...)

## S3 method for class 'sdc_loglinear'
plot(x, plotpath = NULL, plotvar1 = character(0), plotvar2 = character(0), ...)
```

Arguments

data	Data frame containing the data to be evaluated.
weight	Column name for sampling weights.
varpool	Vector of column names to be used in model.
degree	Highest degree of interaction terms to be used in the model.
numiter	Maximum number of iterations to run iterative proportional fitting for the log-linear model.
epsilon	Maximum deviation allowed between observed and fitted margins.
blanks_as_missing	If TRUE, character and factor variables that are blank or pure whitespace are treated as missing values.
output_filename	Name of the csv file to save the data set with record-level risk measures, .tau1_rec and .tau2_rec, attached. NULL if no output file is to be saved.
x	Object of class sdc_loglinear, as returned by sdc_loglinear.
summary_outfile	Name of summary output .txt file. If not NULL, console output is copied to the file. Default is NULL (no logging of output). Errors and warnings are not diverted (consider running in batch mode if logging is needed).
...	Currently unused. For NextMethod compatibility.
plotpath	Directory to save plots. Plots are saved as *jpeg* files (quality = 100%). If the directory does not exist, it is first created. If plotpath is NULL (default), plots are not saved.
plotvar1	A vector of names of discrete variables for boxplots. If none, boxplots are not produced.
plotvar2	A vector of names of continuous variables for scatterplots. If none, scatterplots are not produced.

Details

The data should not contain any missing values among varpool variables or the weight variable.

Value

An object of type sdc_loglinear containing calculated risk measures.

Methods (by generic)

- `print`: S3 print method for sdc_loglinear objects
  Prints tables of file-level reidentification risk measures.
- `plot`: S3 plot method for `sdc_loglinear` objects
  Produces boxplots and scatterplots of record-level risk measures, tau1 and tau2.

Examples

```
data(exampledata)
vars <- c("BORNUSA", "CENREG", "DAGE3", "DRACE3", "EDUC3", "GENDER")
wgt <- "WEIGHT"

results <- sdc_loglinear(exampledata, wgt, vars, degree=3)
print(results)
plot(results, plotvar1="BORNUSA", plotvar2="WEIGHT")
```

---

sdc_loglinear_iter *sdc_loglinear_iter*

Description

Calculates file-level risk measures using a loglinear model with forward stepwise variable selection for interaction terms.

Usage

```
sdc_loglinear_iter(
  data,
  weight,
  varpool,
  numiter = 40,
  epsilon = 0.01,
  fixed_pi = TRUE,
  intermediate_fname = "__loglin_intermediate__.rds",
  restart = FALSE,
  delta = NULL,
  verbose = TRUE,
  blanks_as_missing = TRUE,
  output_filename = NULL
)

## S3 method for class 'sdc_loglinear_iter'
print(x, summary_outfile = NULL, ...)

## S3 method for class 'sdc_loglinear_iter'
plot(x, plotpath = NULL, plotvar1 = character(0), plotvar2 = character(0), ...)
```

Arguments

`data`	Data frame containing the data to be evaluated.
`weight`	Column name for sampling weights.
`varpool`	Vector of column names to be used in model.
`numiter`	Maximum number of iterations to run iterative proportional fitting for the log-linear model.
`epsilon`	Maximum deviation allowed between observed and fitted margins.
`fixed_pi`	If TRUE, sampling rate assumed to be the same across cells.
`intermediate_fname`	Name of intermediate rds file. At each iteration of variable selection, the results so far are saved to this file. This file allows for the process to be restarted if interrupted.
`restart`	If TRUE, restart an interrupted run.
`delta`	Stopping condition for variable selection. If the relative change in all risk measures is smaller than delta, stop iteration. If NULL iteration continues till all variables are used or goodness of fit measures are all negative.
`verbose`	If TRUE, print updates to console at each iteration of the variable selection process.
`blanks_as_missing`	If TRUE, character and factor variables that are blank or pure whitespace are treated as missing values.
`output_filename`	Name of the csv file to save the data set with record-level risk measures, .tau1_rec and .tau2_rec, attached. NULL if no output file is to be saved.
`x`	Object of class sdc_loglinear_iter, as returned by sdc_loglinear.
`summary_outfile`	Name of summary output .txt file. If not NULL, console output is copied to the file. Default is NULL (no logging of output). Errors and warnings are not diverted (consider running in batch mode if logging is needed).
`...`	Currently unused. For NextMethod compatibility.
`plotpath`	Directory to save plots. Plots are saved as *jpeg* files (quality = 100%). If the directory does not exist, it is first created. If `plotpath` is NULL (default), plots are not saved.
`plotvar1`	A vector of names of discrete variables for boxplots. If none, boxplots are not produced.
`plotvar2`	A vector of names of continuous variables for scatterplots. If none, scatterplots are not produced.

Value

An object of type `sdc_loglinear_iter` containing calculated risk measures.

Methods (by generic)

- `print`: S3 print method for sdc_loglinear_iter objects
  Prints summary of iterative loglinear fit.
- `plot`: S3 plot method for `sdc_loglinear_iter` objects
  Produces boxplots and scatterplots of record-level risk measures, tau1 and tau2.

Examples

```
data(exampledata)
vars <- c("BORNUSA", "CENREG", "DAGE3", "DRACE3", "EDUC3", "GENDER")
wgt <- "WEIGHT"

results <- sdc_loglinear_iter(exampledata, wgt, vars)
print(results)
```

Index