Package ‘sdcLog’

March 19, 2022

**Title**  Tools for Statistical Disclosure Control in Research Data Centers

**Version**  0.5.0

**Description**  Tools for researchers to explicitly show that their results comply to rules for statistical disclosure control imposed by research data centers. These tools help in checking descriptive statistics and models and in calculating extreme values that are not individual data. Also included is a simple function to create log files. The methods used here are described in the `Guidelines for the checking of output based on microdata research` by Bond, Brandt, and de Wolf (2015) <https://ec.europa.eu/eurostat/cros/system/files/dwb_standalone-document_output-checking-guidelines.pdf>.

**License**  GPL-3

**URL**  https://github.com/matthiasgomonka/sdcLog

**BugReports**  https://github.com/matthiasgomonka/sdcLog/issues

**Depends**  R (>= 3.5)

**Imports**  broom (>= 0.5.5),
checkmate (>= 2.0.0),
cli,
data.table (>= 1.12.8),
mathjaxr,
stats,
utils

**Suggests**  cfrr,
knitr,
lfe,
rmarkdown,
skimr,
spelling,
testthat (>= 3.0.0),
tibble

**VignetteBuilder**  knitr

**RdMacros**  mathjaxr
common_arguments

Description

Arguments

data data.frame from which the descriptive statistics are calculated.
id_var character The name of the id variable. Defaults to getOption("sdc.id_var") so that you can provide options(sdc.id_var = "my_id_var") at the top of your script.
val_var character vector of value variables on which descriptive statistics are computed.
by character vector of grouping variables.
zero_as_NA logical If TRUE, zeros in 'val_var' are treated as NA.
fill_id_var logical Only for very specific use cases. For example:

- id_var contains NA values which represent missing values in the sense that there actually exist values identifying the entity but are unknown (or deleted for privacy reasons).
• id_var contains NA values which result from the fact that an observation features more than one confidential identifier and not all of these identifiers are present in each observation. Examples for such identifiers are the role of a broker in a security transaction or the role of a collateral giver in a credit relationship.

If TRUE, NA values within id_var will internally be filled with <filled_[i]>, assuming that all NA values of id_var can be treated as different small entities for statistical disclosure control purposes. Thus, set TRUE only if this is a reasonable assumption.

Defaults to FALSE.

model The estimated model object. Can be a model type like lm, glm and various others (anything which can be handled by broom::augment()).

min_obs integer The minimum number of observations used to calculate the minimum and maximum. Defaults to getOption("sdc.n_ids",5L). This is not the number of distinct entities.

max_obs integer The maximum number of observations used to calculate the minimum and maximum. Defaults to nrow(data). This is not the number of distinct entities.

---

print.sdc_distinct_ids

Print methods for SDC objects

**Description**

These methods print SDC objects. Tables containing information are only printed when relevant.

**Usage**

```r
## S3 method for class 'sdc_distinct_ids'
print(x, ...)

## S3 method for class 'sdc_dominance'
print(x, ...)

## S3 method for class 'sdc_options'
print(x, ...)

## S3 method for class 'sdc_settings'
print(x, ...)

## S3 method for class 'sdc_descriptives'
print(x, ...)

## S3 method for class 'sdc_model'
```
print(x, ...)
## S3 method for class 'sdct_min_max'
print(x, ...)

Arguments

x The object to be printed
...
Ignored.

sdc_descriptives Disclosure control for descriptive statistics

Description

Checks the number of distinct entities and the (n, k) dominance rule for your descriptive statistics.
That means that sdc_descriptives() checks if there are at least 5 distinct entities and if the largest
2 entities account for 85% or more of val_var. The parameters can be changed using options. For
details see vignette("options",package = "sdcLog").

Usage

sdc_descriptives(
data,
id_var = getOption("sdc.id_var"),
val_var = NULL,
by = NULL,
zero_as_NA = NULL,
fill_id_var = FALSE
)

Arguments
data data.frame from which the descriptive statistics are calculated.
id_var character The name of the id variable. Defaults to getOption("sdc.id_var")
so that you can provide options(sdc.id_var = "my_id_var") at the top of
your script.
val_var character vector of value variables on which descriptive statistics are computed.
by character vector of grouping variables.
zero_as_NA logical If TRUE, zeros in 'val_var' are treated as NA.
fill_id_var logical Only for very specific use cases. For example:
• id_var contains NA values which represent missing values in the sense that
  there actually exist values identifying the entity but are unknown (or deleted
  for privacy reasons).
id_var contains NA values which result from the fact that an observation features more than one confidential identifier and not all of these identifiers are present in each observation. Examples for such identifiers are the role of a broker in a security transaction or the role of a collateral giver in a credit relationship.

If TRUE, NA values within id_var will internally be filled with <filled_[i]>, assuming that all NA values of id_var can be treated as different small entities for statistical disclosure control purposes. Thus, set TRUE only if this is a reasonable assumption.

Defaults to FALSE.

Details

The general form of the \((n, k)\) dominance rule can be formulated as:

\[
\sum_{i=1}^{n} x_i > \frac{k}{100} \sum_{i=1}^{N} x_i
\]

where \(x_1 \geq x_2 \geq \cdots \geq x_N\). \(n\) denotes the number of largest contributions to be considered, \(x_n\) the \(n\)-th largest contribution, \(k\) the maximal percentage these \(n\) contributions may account for, and \(N\) is the total number of observations.

If the statement above is true, the \((n, k)\) dominance rule is violated.

Value

A list of class sdc_descriptives with detailed information about options, settings, and compliance with the criteria distinct entities and dominance.

Examples

```r
sdc_descriptives(
  data = sdc_descriptives_DT,
  id_var = "id",
  val_var = "val_1"
)

sdc_descriptives(
  data = sdc_descriptives_DT,
  id_var = "id",
  val_var = "val_1",
  by = "sector"
)

sdc_descriptives(
  data = sdc_descriptives_DT,
  id_var = "id",
  val_var = "val_1",
  by = c("sector", "year")
)
```
sdc_descriptives(
  data = sdc_descriptives_DT,
  id_var = "id",
  val_var = "val_2",
  by = c("sector", "year")
)

sdc_descriptives(
  data = sdc_descriptives_DT,
  id_var = "id",
  val_var = "val_2",
  by = c("sector", "year"),
  zero_as_NA = FALSE
)

---

sdc_descriptives_DT  Example data for sdc_descriptives()

**Description**

Utilized in the vignette.

**Usage**

```r
data("sdc_descriptives_DT")
```

**Format**

A data.table with 20 rows and 5 columns.

**Details**

The data.table contains the following columns:

- `id` factor random identifier
- `sector` factor economic sector
- `year` integer time variable
- `val_1, val_2` numeric value variables
sdc_log

Create Stata-like log files from R Scripts

Description

This function creates Stata-like log files from R Scripts. It can handle several files (in a character vector) at once.

Usage

sdc_log(r_script, destination, replace = FALSE, append = FALSE, local = FALSE)

Arguments

r_script character Path of the R script to be run with logging.
destination One of:

• character Path of the log file to be used.
• file connection to which the log should be written. This is especially useful, when you have nested calls to sdc_log() and want to write everything into the same log file. Then, create a single file connection and provide this connection to all calls to sdc_log() (and close it afterwards).

replace logical Indicates whether to replace an existing log file.
append logical Indicates whether to append an existing log file.
local One of:

• logical Indicates whether to evaluate within the global environment (FALSE) or the calling environment (TRUE).
• environment A specific evaluation environment. Determines the evaluation environment. Useful whenever sdc_log() is called from within a function, or for nested sdc_log() calls. By default (FALSE) evaluation occurs in the global environment. See also source.

Value

character vector holding the path(s) of the written log file(s).
sdc_min_max

Calculate RDC rule-compliant extreme values

**Description**

Checks if calculation of extreme values comply to RDC rules. If so, function returns average min and max values according to RDC rules.

**Usage**

```r
sdc_min_max(
  data,
  id_var = getOption("sdc.id_var"),
  val_var,
  by = NULL,
  max_obs = nrow(data),
  fill_id_var = FALSE
)
```

**Arguments**

- `data`  
  *data.frame* from which the descriptive statistics are calculated.
- `id_var`  
  *character* The name of the id variable. Defaults to `getOption("sdc.id_var")` so that you can provide `options(sdc.id_var = "my_id_var")` at the top of your script.
- `val_var`  
  *character* vector of value variables on which descriptive statistics are computed.
- `by`  
  *character* vector of grouping variables.
- `max_obs`  
  *integer* The maximum number of observations used to calculate the minimum and maximum. Defaults to `nrow(data)`. *This is not the number of distinct entities.*
- `fill_id_var`  
  *logical* Only for very specific use cases. For example:
  - `id_var` contains NA values which represent missing values in the sense that there actually exist values identifying the entity but are unknown (or deleted for privacy reasons).
  - `id_var` contains NA values which result from the fact that an observation features more than one confidential identifier and not all of these identifiers are present in each observation. Examples for such identifiers are the role of a broker in a security transaction or the role of a collateral giver in a credit relationship.

If TRUE, NA values within `id_var` will internally be filled with `<filled_[i]>`, assuming that all NA values of `id_var` can be treated as different small entities for statistical disclosure control purposes. Thus, set TRUE only if this is a reasonable assumption.

Defaults to FALSE.
Value

A list \texttt{list} of class \texttt{sdc_min_max} with detailed information about options, settings and the calculated extreme values (if possible).

Examples

\begin{verbatim}
sdc_min_max(sdc_min_max_DT, id_var = "id", val_var = "val_1")
sdc_min_max(sdc_min_max_DT, id_var = "id", val_var = "val_2")
sdc_min_max(sdc_min_max_DT, id_var = "id", val_var = "val_3", max_obs = 10)
sdc_min_max(sdc_min_max_DT, id_var = "id", val_var = "val_1", by = "year")
sdc_min_max(
    sdc_min_max_DT, id_var = "id", val_var = "val_1", by = c("sector", "year")
)
\end{verbatim}

\begin{longtable}{ll}
\textbf{sdc_min_max_DT} & \textit{Example data for sdc_min_max()} \\
\end{longtable}

Description

Utilized in the vignette

Usage

\begin{verbatim}
data("sdc_min_max_DT")
\end{verbatim}

Format

A data.table with 20 rows and 6 columns.

Details

The data.table contains the following columns:

- \textbf{id} \texttt{factor} random identifier
- \textbf{sector} \texttt{factor} economic sector
- \textbf{year} \texttt{integer} time variable
- \textbf{val_1} - \textbf{val_3} \texttt{numeric} value variables
**sdc_model**  
*Disclosure control for models*

**Description**
Checks if your model complies to RDC rules. Checks for overall number of entities and number of entities for each level of dummy variables.

**Usage**

```r
sdc_model(data, model, id_var = getOption("sdc.id_var"), fill_id_var = FALSE)
```

**Arguments**
- `data`  
  data.frame which was used to build the model.
- `model`  
  The estimated model object. Can be a model type like `lm`, `glm` and various others (anything which can be handled by `broom::augment()`).
- `id_var`  
  character The name of the id variable. Defaults to `getOption("sdc.id_var")` so that you can provide options(`sdc.id_var = "my_id_var"`) at the top of your script.
- `fill_id_var`  
  logical Only for very specific use cases. For example:
  - `id_var` contains NA values which represent missing values in the sense that there actually exist values identifying the entity but are unknown (or deleted for privacy reasons).
  - `id_var` contains NA values which result from the fact that an observation features more than one confidential identifier and not all of these identifiers are present in each observation. Examples for such identifiers are the role of a broker in a security transaction or the role of a collateral giver in a credit relationship.

  If TRUE, NA values within `id_var` will internally be filled with `<filled_[i]>`, assuming that all NA values of `id_var` can be treated as different small entities for statistical disclosure control purposes. Thus, set TRUE only if this is a reasonable assumption.

  Defaults to FALSE.

**Value**
A list of class `sdc_model` with detailed information about options, settings, and compliance with the distinct entities criterion.

**Examples**

```r
# Check simple models
model_1 <- lm(y ~ x_1 + x_2, data = sdc_model_DT)
sdc_model(data = sdc_model_DT, model = model_1, id_var = "id")
```
model_2 <- lm(y ~ x_1 + x_2 + x_3, data = sdc_model_DT)
sdc_model(data = sdc_model_DT, model = model_2, id_var = "id")

model_3 <- lm(y ~ x_1 + x_2 + dummy_3, data = sdc_model_DT)
sdc_model(data = sdc_model_DT, model = model_3, id_var = "id")

---

**sdc_model_DT**

*Example data for sdc_model()*

### Description
Utilized in the vignette

### Usage

```r
data("sdc_model_DT")
```

### Format
A data.table with 80 rows and 9 columns.

### Details
The data.table contains the following columns:

- id factor random identifier
- y, x_4 numeric value variables
- dummy_1 - dummy_3 factor dummy variables
Index

* datasets
  sdc_descriptives_DT, 6
  sdc_min_max_DT, 9
  sdc_model_DT, 10

broom::augment(), 3, 9
character, 2, 4, 6–9
common_arguments, 2
data.frame, 2, 4, 8, 9
environment, 7
factor, 6, 9, 11
file, 7
glm, 3, 9
integer, 3, 6, 8, 9
list, 5, 8, 10
lm, 3, 9
logical, 2, 4, 7, 8, 10
numeric, 6, 9, 11

print.sdc_descriptives
  (print.sdc_distinct_ids), 3
print.sdc_distinct_ids, 3
print.sdc_dominance
  (print.sdc_distinct_ids), 3
print.sdc_min_max
  (print.sdc_distinct_ids), 3
print.sdc_model
  (print.sdc_distinct_ids), 3
print.sdc_options
  (print.sdc_distinct_ids), 3
print.sdc_settings
  (print.sdc_distinct_ids), 3

sdc_descriptives, 4
sdc_descriptives_DT, 6
sdc_log, 6
sdc_min_max, 7
sdc_min_max_DT, 9
sdc_model, 9
sdc_model_DT, 10
source, 7