Package 'sdcLog'

November 20, 2020

Title Tools for Statistical Disclosure Control in Research Data Centers

Version 0.1.0

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/matthiasgomolka/sdcLog

BugReports https://github.com/matthiasgomolka/sdcLog/issues

Depends R (>= 3.5)

Imports broom (>= 0.5.5), checkmate (>= 2.0.0), crayon (>= 1.3.4), data.table (>= 1.12.8), methods

Suggests callr, covr (>= 3.5.0), devtools, here, knitr, rmarkdown, skimr, spelling, testthat (>= 3.0.0)

VignetteBuilder knitr

Config/testthat/edition 3

Encoding UTF-8

Language en-US

LazyData true

RoxygenNote 7.1.1

NeedsCompilation no

Author Matthias Gomolka [aut, cre], Tim Becker [aut]

Maintainer Matthias Gomolka <matthias.gomolka@posteo.de>

Repository CRAN

Date/Publication 2020-11-20 08:20:02 UTC

R topics documented:

`check_distinct_ids` *Internal function which creates cross-tables with number of distinct id's*

Description

Internal function which creates cross-tables with number of distinct id's

Usage

```
check_distinct_ids(data, id_var, val_var, by = NULL)
```

Arguments

`data`	data.frame from which the descriptive statistics are calculated.
`id_var`	character The name of the id variable.
`val_var`	character vector of value variables on which descriptive statistics are computed.
`by`	Grouping variables (or expression) as in data.table's by.

check_dominance *Internal function which creates cross-tables with number of distinct id's*

Description

Internal function which creates cross-tables with number of distinct id's

Usage

```
check_dominance(data, id_var, val_var, by = NULL)
```

Arguments

data	data.frame from which the descriptive statistics are calculated.
id_var	character The name of the id variable.
val_var	character vector of value variables on which descriptive statistics are computed.
by	Grouping variables (or expression) as in data.table's by.

common_arguments *arguments*

Description

arguments

Arguments

data	data.frame from which the descriptive statistics are calculated.
id_var	character The name of the id variable.
val_var	character vector of value variables on which descriptive statistics are computed.
by	Grouping variables (or expression) as in data.table's by.
zero_as_NA	logical If TRUE, zeros in 'val_var' are treated as NA.
model	The estimated model object. Can be a model type like lm, glm and various others (anything which can be handled by `broom::augment()`).
n_min	integer The number of values used to calculate the minimum, by default 5.
n_max	integer The number of values used to calculate the maximum, by default 5.

generate_log *Source a single R script and generate a log file*

Description

Source a single R script and generate a log file

Usage

```
generate_log(r_script, log_file)
```

Arguments

r_script	R script to be run and logged
log_file	destination file of the log file to be generated

sdc_descriptives *Disclosure control for descriptive statistics*

Description

Checks if your descriptive statistics comply to statistical disclosure control. Checks for number of distinct entities and dominance.

Usage

```
sdc_descriptives(data, id_var, val_var, by = NULL, zero_as_NA = NULL)
```

Arguments

data	data.frame from which the descriptive statistics are calculated.
id_var	character The name of the id variable.
val_var	character vector of value variables on which descriptive statistics are computed.
by	Grouping variables (or expression) as in data.table's by.
zero_as_NA	logical If TRUE, zeros in 'val_var' are treated as NA.

Value

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

Examples

```
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

```
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_DT` *Example datasets used in the vignette*

Description

Recreated after the example data from corresponding Stata function.

Usage

```
data("sdc_DT")
```

Format

A data.table with 100 rows and 8 columns.

Details

The data.table contains the following columns:

- id integer random identifier
- time integer random time variable
- V1 - V3 numeric random variables
- D1 - D3 logical non-random dummy variables

sdc_extreme *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

```
sdc_extreme(
  data,
  id_var,
  val_var,
  by = NULL,
  n_min = getOption("sdc.n_ids", 5L),
  n_max = n_min
)
```

Arguments

data	data.frame from which the descriptive statistics are calculated.
id_var	character The name of the id variable.
val_var	character vector of value variables on which descriptive statistics are computed.
by	Grouping variables (or expression) as in data.table's by.
n_min	integer The number of values used to calculate the minimum, by default 5.
n_max	integer The number of values used to calculate the maximum, by default 5.

Value

A list list of class sdc_extreme with detailed information about options, settings and the calculated extreme values (if possible).

Examples

```
sdc_extreme(data = sdc_extreme_DT, id_var = "id", val_var = "val_1")
sdc_extreme(data = sdc_extreme_DT, id_var = "id", val_var = "val_2")
sdc_extreme(data = sdc_extreme_DT, id_var = "id", val_var = "val_2",
            n_min = 7)
sdc_extreme(data = sdc_extreme_DT, id_var = "id", val_var = "val_3",
            n_min = 10, n_max = 10)
sdc_extreme(data = sdc_extreme_DT, id_var = "id", val_var = "val_3",
            n_min = 8, n_max = 8)
sdc_extreme(data = sdc_extreme_DT, id_var = "id", val_var = "val_1",
            by = year)
sdc_extreme(data = sdc_extreme_DT, id_var = "id", val_var = "val_1",
            by = c("sector", "year"))
```

sdc_extreme_DT *Example data for* sdc_extreme()

Description

Utilized in the vignette

Usage

```
data("sdc_extreme_DT")
```

Format

A data.table with 20 rows and 6 columns.

Details

The data.table contains the following columns:

- id factor random identifier
- sector factor economic sector
- year integer time variable
- val_1 - val_3 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_scripts, log_files, replace = FALSE)
```

Arguments

r_scripts	character vector containing the path(s) of the R script(s) which should be run with logging.
log_files	character vector containing the path(s) of the text file(s) where the log(s) should be stored.
replace	logical Indicates whether to replace existing log files.

Value

Invisible NULL.

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

```
sdc_model(data, model, id_var)
```

Arguments

data	data.frame from which the descriptive statistics are calculated.
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.

Value

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

Examples

```
# 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

```
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