Package 'nncc'

May 9, 2026

Title Nearest Neighbors Matching of Case-Control Data

Version 2.0.0

Description Provides nearest-neighbors matching and analysis of case-control data. Cui, Z., Marder, E. P., Click, E. S., Hoekstra, R. M., & Bruce, B. B. (2022) <doi:10.1097/EDE.0000000000001504>.

Depends R (>= 3.3.2)

License GPL (>= 3)

Encoding UTF-8

LazyData true

Imports dplyr, furrr, tidyr, igraph, ggplot2, cluster, rlang, stats

RoxygenNote 7.2.3

Suggests rmarkdown, knitr, future, future.batchtools, logistf, mice, survival

VignetteBuilder knitr

NeedsCompilation no

Author Beau Bruce [aut, cre], Zhaohui Cui [aut]

Maintainer Beau Bruce <lue7@cdc.gov>

Repository CRAN

Date/Publication 2024-01-11 14:10:02 UTC

Contents

`anifood` *case-control data*

Description

A toy dataset containing 7-day exposure history of 250 cases and 250 controls

Usage

```
anifood
```

Format

A data frame with 500 rows and 11 variables:

case case status, 1 = case, 0 = control

exp01 whether exposed to exp01, 1 = yes, 0 = no

exp09 whether exposed to exp09, 1 = yes, 0 = no

exp20 whether exposed to exp20, 1 = yes, 0 = no

exp24 whether exposed to exp24, 1 = yes, 0 = no

exp27 whether exposed to exp27, 1 = yes, 0 = no

exp43 whether exposed to exp43, 1 = yes, 0 = no

exp45 whether exposed to exp45, 1 = yes, 0 = no

exp50 whether exposed to exp50, 1 = yes, 0 = no

exp52 whether exposed to exp52, 1 = yes, 0 = no

exp57 whether exposed to exp57, 1 = yes, 0 = no

cacheit *Function to cache long operations*

Description

Save results from code that takes a long time to execute to a .rds file if that file does not exist in the cache directory. If the file exists in the cache directory, that file will be loaded to memory without evaluating the code.

Usage

```
cacheit(name, code, dir, createdir = FALSE, clearcache = FALSE)
```

Arguments

name	Name of the file to create without extension
code	Expression of the code to execute and cache
dir	Name of cache directory which should be placed in the working directory
createdir	Logical about whether to create the directory if it does not exist
clearcache	Logical about whether to recalculate the cached .rds file for this object

Details

For more information, please refer to the vignette using `browseVignettes("nncc")`.

Value

Output of code, either freshly executed if the file does not exist or or clearcache is TRUE otherwise returns result from the cache file

calc_strata_or *Calculate the pooled strata OR*

Description

Each case and matched controls form a stratum in the data set. This function is to calculate the pooled OR for the data set.

Usage

```
calc_strata_or(dfs, filter = TRUE, filterdata = NULL)
```

Arguments

dfs	A named list of dataframes created by package functions
filter	Filter statement to apply
filterdata	Extra data to left join to the dfs for filtering

Details

Uses the M-H method unless there is only one strata for which the fisher.test is used. For more information, please refer to the vignette using browseVignettes("nncc").

---

distance_density_plot *Distance density plots comparing closest to random choices*

Description

Distance density plots comparing closest to random choices

Usage

```
distance_density_plot(threshold_results)
```

Arguments

threshold_results	See get_threshold

Value

The ggplot showing the distances of cases matched to their nearest neighbor vs. a random control

---

excl_vars *Variables excluded from matching*

Description

A dataset lists variables that are excluded from matching for each exposure. This dataset is supplied to the rmvars argument of the function make_knn_strata. The two columns must be named with "exp_var" and "rm_vars".

Usage

```
excl_vars
```

Format

A data frame with two variables:

exp_var exposures of interest

rm_vars variables to be excluded from matching for a given exposure

---

`finalize_data` *Final cleaning of the matched dataset(s)*

Description

Ensures that a control retained in a data frame is used once and remove strata without any case or any control. In this process, priority is first given to the smallest strata then smallest distance if a control is matched to multiple cases (i.e., that control exists in multiple strata).

Usage

```
finalize_data(dfs, filter = TRUE, filterdata = NULL)
```

Arguments

`dfs`	A list of data frames generated by `make_analysis_sets`
`filter`	Filter statement to apply
`filterdata`	Extra data to left join to the `dfs` for filtering

Details

For more information, please refer to the vignette using `browseVignettes("nncc")`.

Value

A list of data frames

`fix_df` *Fix the strata so they all have at least one case and control*

Description

Fix the strata so they all have at least one case and control

Usage

```
fix_df(d)
```

Arguments

`d`	A stratified dataset

get_paf *Calculate population attributable fraction using odds ratio*

Description

Calculate population attributable fraction using odds ratio

Usage

```
get_paf(df_or, which_or, exp_var, exp_level, df_matched)
```

Arguments

df_or	A data frame that stores odds ratios for all exposure of interest
which_or	An unquoted name of the name of the column that stores odds ratio, or its lower or upper confidence limit in df_or.
exp_var	An unquoted name of the column that stores the name of exposures in df_or
exp_level	An unquoted name of the column that stores the level of the exposure variable in df_or
df_matched	The list of data frames used to calculate odds ratios

Details

Use odds ratio, its upper confidence limit, and its lower confidence limit to calculate population attributable fraction, its upper confidence limit, and its lower confidence limit, respectively.

For more information, please refer to the vignette using `browseVignettes("nncc")`.

Value

A data frame.

get_threshold *Identify the right threshold*

Description

To find a threshold for distance to define controls that are qualified to be matched with a case.

Usage

```
get_threshold(data, vars, case_var = "case", p_threshold = 0.5, seed = 1600)
```

Arguments

data	The dataset
vars	The variables to use for calculating distance
case_var	The name of the case identifier variable
p_threshold	The probability that the closest matching approach produces the closer matching relative to the random matching approach. The greater p_threshold, the smaller the threshold.
seed	A random seed.

Details

This function uses logistic regression to predict by the distance whether a control is the closest (unique) match for each case vs. a random selection and by default returns the 50

For more information, please refer to the vignette using browseVignettes("nncc").

Value

A list with items:

threshold	The numeric threshold chosen
modeldata	The data used to fit the logistic regression model
strata	The strata made by make_knn_strata
model	The fit logisitic regression model

---

make_analysis_set *Make analysis set*

Description

Set a maximum number of controls that are allowed to be matched to a case; ensure that matched case-control pairs have a distance closer than the predefined threshold; merge strata sharing same controls.

Usage

```
make_analysis_set(
  var,
  stratified_data,
  data,
  maxdist = 0,
  maxcontrols = 20,
  silent = FALSE
)
```

Arguments

var	Character of current exposure variable in make_analysis_sets
stratified_data	Stratified dataset, see make_knn_strata
data	Original case control data
maxdist	Reject any controls more than maxdist from their case
maxcontrols	Maximum number of controls to keep per strata
silent	Suppress exposure info useful for *apply/loop implementations

Details

For more information, please refer to the vignette using browseVignettes("nncc").

Value

A list of data frames with the length of number of exposures.

---

make_analysis_sets *Make analysis datasets*

Description

This helper function facilitates the implement the make_analysis_set() to each exposure.

Usage

```
make_analysis_sets(stratified_data, expvars, data, threshold)
```

Arguments

stratified_data	List of stratified data sets, see make_knn_strata
expvars	Character vector of exposure variable for each set in stratifed_data
data	Original case control data
threshold	Maximum distance threshold for cases and controls created by get_threshold

Details

For more information, please refer to the vignette using browseVignettes("nncc").

Value

A list of data frames with the length of number of exposures

make_knn_strata *Make case-control strata using k nearest neighbors (knn)*

Description

Select a pre-defined number of controls for each case based on calculated distances between cases and controls.

Usage

```
make_knn_strata(
  expvar,
  matchvars,
  df,
 rmvars = data.frame(exp_var = character(), rm_vars = character(), stringsAsFactors =
   FALSE),
  casevar = "case",
  ncntls = 250,
  metric = "gower",
  silent = FALSE
)
```

Arguments

expvar	A character - the name of the exposure variable in df.
matchvars	Character vector - what are the variables to match on. Note that the function automatically excludes the the exposure variable.
df	A dataframe that contains the case-control data.
rmvars	A data frame that lists variables to be excluded from matching for each exposure. For details, please see the vignette of this package.
casevar	A character - what is the name of the variable indicating case status (1 = case, 0 = control)
ncntls	An integer to specify number of controls to find for each case (k in knn).
metric	A character to specify a metric for measuring distance between a case and a control. See daisy.
silent	Suppress exposure info useful for *apply/loop implementations?

Details

For more information, please refer to the vignette using browseVignettes("nncc").

Value

A list of data frames with a length of number of exposures of interest.

nncc *nncc: nearest-neighbors matching for case-control data*

Description

The nncc package implements an approach to match cases with their nearest controls defined by Gower distance. This approach may achieve better confounding control than conventional analytic approaches such as (conditional) logistic regression when you have a relatively large number of exposures of interest. To learn more about nncc, start with the vignettes: `browseVignettes("nncc")`.

Authors(s)

Maintainer: Beau B. Bruce <lue7@cdc.gov>

Coauthor: Zhaohui Cui

Functions

- get_threshold
- distance_density_plot
- threshold_model_plot
- original_compare_plot
- make_knn_strata
- make_analysis_sets
- finalize_data
- test_mh
- get_paf

original_compare_plot *Compare the original strata's distances to the knn version*

Description

Compare the original strata's distances to the knn version

Usage

```
original_compare_plot(data, casevar, stratavar, threshold_results)
```

Arguments

data	The original data
casevar	The variable that defines cases vs. controls
stratavar	The variable that defines the strata
threshold_results	See get_threshold

Value

An list with items:

`plot_density` The ggplot displayed

`prop_distance_gt_threshold`
A table showing proportion of pairs exceeding numeric threshold chosen

---

`plot_results` *Plot the OR results*

Description

Plot the OR results

Usage

```
plot_results(csvfilename, filter = TRUE)
```

Arguments

`csvfilename`	CSV results file, see `write_strata_or_output`
`filter`	How to filter the results

Details

For more information, please refer to the vignette using `browseVignettes("nncc")`.

Value

Returns `csvfilename` to allow chaining

`sex2` *Urinary Tract Infection in American College Students*

Description

This data set deals with urinary tract infection in sexually active college women, along with covariate information on age and contraceptive use. The variables are all binary and coded in 1 (condition is present) and 0 (condition is absent).

Usage

```
sex2
```

Format

sex2: a data.frame containing 239 observations

case urinary tract infection, the study outcome variable

age >= 24 years

dia use of diaphragm

oc use of oral contraceptive

vic use of condom

vicl use of lubricated condom

vis use of spermicide

Source

<https://www.cytel.com/>

References

Cytel Inc., (2010) LogXact 9 user manual, Cambridge, MA:Cytel Inc

---

`sexagg` *Urinary Tract Infection in American College Students*

Description

This data set deals with urinary tract infection in sexually active college women, along with covariate information on age an contraceptive use. The variables are all binary and coded in 1 (condition is present) and 0 (condition is absent): case (urinary tract infection, the study outcome variable), age (>= 24 years), dia (use of diaphragm), oc (use of oral contraceptive), vic (use of condom), vicl (use of lubricated condom), and vis (use of spermicide).

Usage

```
sexagg
```

Format

sexagg: an aggregated data.frame containing 31 observations with case weights (COUNT).

case urinary tract infection, the study outcome variable

age >= 24 years

dia use of diaphragm

oc use of oral contraceptive

vic use of condom

vicl use of lubricated condom

vis use of spermicide

Source

<https://www.cytel.com/>

References

Cytel Inc., (2010) LogXact 9 user manual, Cambridge, MA:Cytel Inc

---

`test_mh` *Calculate odds ratios*

Description

Calculate odds ratios using the M-H method when the matched dataset has more than 1 stratum, and using the Fisher's exact test when the matched dataset has only one stratum.

Usage

```
test_mh(case, exp, strata)
```

Arguments

`case`	The case statuses
`exp`	The exposure statuses
`strata`	The strata identifiers

Details

For more information, please refer to the vignette using `browseVignettes("nncc")`.

Value

The list of statistical results

`threshold_model_plot` *Show the prediction of the logistic regression model*

Description

Show the prediction of the logistic regression model

Usage

```
threshold_model_plot(threshold_results, p_threshold = 0.5)
```

Arguments

threshold_results	See get_threshold
p_threshold	The probability that the closest matching approach produces the closer matching relative to the random matching approach. The greater p_threshold, the smaller the threshold.

Value

The ggplot showing the threshold logistic regression model

---

unique_controls *Ensures controls are unique to avoid possible pseudoreplication issues*

Description

Ensures controls are unique to avoid possible pseudoreplication issues

Usage

```
unique_controls(stratifieddata)
```

Arguments

stratifieddata	See make_knn_strata and make_analysis_set.

Value

A tibble after it has been examined and filtered for duplicate controls

---

write_strata_or_output *Format strata output into CSV*

Description

Format strata output into CSV

Usage

```
write_strata_or_output(results, varnames, filename)
```

Arguments

`results`	Output of `test_mh`
`varnames`	Vector of exposure variable names
`filename`	String of the filename to output to

Value

Returns the filename to allow chaining

Index