

Package ‘nearfar’

January 23, 2024

Type Package

Title Near-Far Matching

Version 1.3

Date 2024-01-22

Author Joseph Rigdon <jrigdon@wakehealth.edu>

Maintainer Joseph Rigdon <jrigdon@wakehealth.edu>

Imports GenSA, MASS, car, stats

Description Near-far matching is a study design technique for preprocessing observational data to mimic a pair-randomized trial. Individuals are matched to be near on measured confounders and far on levels of an instrumental variable. Methods outlined in further detail in Rigdon, Baiocchi, and Basu (2018) <[doi:10.18637/jss.v086.c05](https://doi.org/10.18637/jss.v086.c05)>.

License GPL-3

Depends nbpMatching

NeedsCompilation no

Repository CRAN

Date/Publication 2024-01-23 13:00:02 UTC

R topics documented:

| | |
|---------------------------|----|
| nearfar-package | 2 |
| angrist | 2 |
| calipers | 3 |
| eff_ratio | 4 |
| matches | 5 |
| opt_nearfar | 6 |
| smahal | 8 |
| summary.nf | 9 |
| summ_matches | 10 |

| | |
|--------------|-----------|
| Index | 11 |
|--------------|-----------|

| | |
|-----------------|--------------------------|
| nearfar-package | <i>Near-Far Matching</i> |
|-----------------|--------------------------|

Description

Near-far matching is a study design technique for preprocessing observational data to mimic a pair-randomized trial. Individuals are matched to be near on measured confounders and far on levels of an instrumental variable.

Details

| | |
|----------|------------|
| Package: | nearfar |
| Type: | Package |
| Version: | 1.3 |
| Date: | 2024-01-15 |
| License: | GPL-3 |

Author(s)

Joseph Rigdon <jrigdon@wakehealth.edu>

References

Rigdon J, Baiocchi M, Basu S (2018). Near-far matching in R: The nearfar package. *Journal of Statistical Software*, 86(5), 1-21.

Baiocchi M, Small D, Lorch S, Rosenbaum P (2010). Building a stronger instrument in an observational study of perinatal care for premature infants. *Journal of the American Statistical Association*, 105(492), 1285-1296.

Baiocchi M, Small D, Yang L, Polsky D, Groeneveld P (2012). Near-far matching: a study design approach to instrumental variables. *Health Services and Outcomes Research Methodology*, 12(4), 237-253.

| | |
|---------|-------------------------------------------------|
| angrist | <i>Angrist data set for education and wages</i> |
|---------|-------------------------------------------------|

Description

A random sample of 1000 observations from the data set used by Angrist and Krueger in their investigation of the impact of education on future wages.

Format

A data frame with 1000 observations on the following 7 variables.

wage a numeric vector

educ a numeric vector

qob a numeric vector

IV a numeric vector

age a numeric vector

married a numeric vector

race a numeric vector

Details

This data set is a random sample of 1000 observations from the URL listed below.

Source

<https://economics.mit.edu/people/faculty/josh-angrist/angrist-data-archive>

References

Angrist JD, Krueger AB (1991). Does Compulsory School Attendance Affect Schooling and Earnings? The Quarterly Journal of Economics, 106(4), 979-1014.

Examples

```
library(nearfar)
str(angrist)
## maybe str(angrist) ; plot(angrist) ...
```

calipers

Matching priority function

Description

Updates given distance matrix to prioritize specified measured confounders in a pair match. Used in consort with [matches](#) function to prioritize specific measured confounders in a near-far match in the [opt_nearfar](#) function.

Usage

```
calipers(distmat, variable, tolerance = 0.2)
```

Arguments

| | |
|-----------|-------------------------------------------------------------------------------------|
| distmat | An object of class distance matrix |
| variable | Named variable from list of measured confounders |
| tolerance | Penalty to apply to mismatched observations; values near 0 penalize mismatches more |

Value

Returns an updated distance matrix

See Also

[matches](#), [opt_nearfar](#)

Examples

```
dd = mtcars[1:4, 2:3]
cc = calipers(distmat=smahal(dd), variable=dd$cyl, tolerance=0.2)
cc
```

eff_ratio

Inference for effect ratio

Description

Conducts inference on effect ratio as described in Section 3.3 of Baiocchi (2010), resulting in an estimate and a permutation based confidence interval for the effect ratio.

Usage

```
eff_ratio(dta, match, outc, trt, alpha)
```

Arguments

| | |
|-------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| dta | The name of the data frame object |
| match | Data frame where first column contains indices for those individuals encouraged into treatment by instrumental variable and second column contains indices for those individuals discouraged from treatment by instrumental variable; returned by both opt_nearfar and matches |
| outc | The name of the outcome variable in quotes, e.g., “wages” |
| trt | The name of the treatment variable, e.g., “educ” |
| alpha | Level of confidence interval |

Value

| | |
|---------|---------------------------------------------------------------|
| est.emp | Empirical estimate of effect ratio |
| est.HL | Hodges-Lehmann type estimate of effect ratio |
| lower | Lower limit to 1-alpha/2 confidence interval for effect ratio |
| upper | Upper limit to 1-alpha/2 confidence interval for effect ratio |

Author(s)

Joseph Rigdon <jrigdon@wakehealth.edu>

References

Baiocchi M, Small D, Lorch S, Rosenbaum P (2010). Building a stronger instrument in an observational study of perinatal care for premature infants. *Journal of the American Statistical Association*, 105(492), 1285-1296.

Examples

```
k2 = matches(dta=mtcars, covs=c("cyl", "disp"), sinks=0.2, iv="carb",
            cutpoint=2, imp.var=c("cyl"), tol.var=0.03)

eff_ratio(dta=mtcars, match=k2, outc="wt", trt="gear", alpha=0.05)
```

| | |
|---------|-------------------------------------------------------------------------------------------------------------------------------------------|
| matches | <i>Function to find pair matches using a distance matrix. Called by opt_nearfar to discover optimal near-far matches.</i> |
|---------|-------------------------------------------------------------------------------------------------------------------------------------------|

Description

Given values of percent sinks and cutpoint, this function will find the corresponding near-far match

Usage

```
matches(dta, covs, iv = NA, imp.var = NA, tol.var = NA, sinks = 0,
        cutpoint = NA)
```

Arguments

| | |
|---------|--------------------------------------------------------------------------------------------------------------|
| dta | The name of the data frame on which to do the matching |
| covs | A vector of the names of the covariates to make “near”, e.g., covs=c("age", "sex", "race") |
| iv | The name of the instrumental variable, e.g., iv="QOB" |
| imp.var | A list of (up to 5) named variables to prioritize in the “near” matching |
| tol.var | A list of (up to 5) tolerances attached to the prioritized variables where 0 is highest penalty for mismatch |

| | |
|----------|-------------------------------------------------------------------------------------------------------|
| sinks | Percentage of the data to match to sinks (and thus remove) if desired; default is 0 |
| cutpoint | Value below which individuals are too similar on iv; increase to make individuals more “far” in match |

Details

Default settings yield a "near" match on only observed confounders in X; add IV, sinks, and cutpoint to get near-far match.

Value

A two-column matrix of row indices of paired matches

Author(s)

Joseph Rigdon <jrigdon@wakehealth.edu>

References

Lu B, Greevy R, Xu X, Beck C (2011). Optimal nonbipartite matching and its statistical applications. *The American Statistician*, 65(1), 21-30.

See Also

[opt_nearfar](#)

Examples

```
k2 = matches(dta=mtcars, covs=c("cyl", "disp"), sinks=0.2, iv="carb",
             cutpoint=2, imp.var=c("cyl"), tol.var=0.03)
k2[1:5, ]
```

opt_nearfar

Finds optimal near-far match

Description

Discovers optimal near-far matches using the partial F statistic (for continuous treatments) or partial deviance (for binary and treatments)

Usage

```
opt_nearfar(dta, trt, covs, iv, trt.type = "cont", imp.var = NA,
            tol.var = NA, adjust.IV = TRUE, sink.range = c(0, 0.5), cutp.range = NA,
            max.time.seconds = 300)
```

Arguments

| | |
|------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| dta | The name of the data frame on which matching was performed |
| trt | The name of the treatment variable, e.g., "educ" |
| iv | The name of the instrumental variable, e.g., iv="QOB" |
| covs | A vector of the names of the covariates to make "near", e.g., covs=c("age", "sex", "race") |
| trt.type | Treatment variable type: "cont" for continuous, or "bin" for binary |
| imp.var | A list of (up to 5) named variables to prioritize in the "near" matching |
| tol.var | A list of (up to 5) tolerances attached to the prioritized variables where 0 is highest penalty for mismatch |
| adjust.IV | if TRUE, include measured confounders in treatment~IV model that is optimized; if FALSE, exclude |
| sink.range | A two element vector of (min, max) for range of sinks over which to optimize in the near-far match; default (0, 0.5) such that maximally 50% of observations can be removed |
| cutp.range | a two element vector of (min, max) for range of cutpoints (how far apart the IV will become) over which to optimize in the near-far match; default is (one SD of IV, range of IV) |
| max.time.seconds | How long to let the optimization algorithm run; default is 300 seconds = 5 minutes |

Value

| | |
|------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| n.calls | Number of calls made to the objective function |
| sink.range | A two element vector of (min, max) for range of sinks over which to optimize in the near-far match; default (0, 0.5) such that maximally 50% of observations can be removed |
| cutp.range | a two element vector of (min, max) for range of cutpoints (how far apart the IV will become) over which to optimize in the near-far match; default is (one SD of IV, range of IV) |
| pct.sink | Optimal percent sinks |
| cutp | Optimal cutpoint |
| maxF | Highest value of partial F-statistic (continuous treatment) or residual deviance (binary treatment) found by simulated annealing optimizer |
| match | A two column matrix where the first column is the index of an "encouraged" individual and the second column is the index of the corresponding "discouraged" individual from the pair matching |
| summ | A table of mean variable values for both the "encouraged" and "discouraged" groups across all variables plus absolute standardized differences for each variable |

Author(s)

Joseph Rigdon <jrigdon@wakehealth.edu>

References

Lu B, Greevy R, Xu X, Beck C (2011). Optimal nonbipartite matching and its statistical applications. *The American Statistician*, 65(1), 21-30.

Xiang Y, Gubian S, Suomela B, Hoeng J (2013). Generalized Simulated Annealing for Efficient Global Optimization: the GenSA Package for R. *The R Journal*, 5(1). URL <http://journal.r-project.org/>.

Examples

```
k = opt_nearfar(dta=mtcars, trt="drat", covs=c("cyl", "disp"),
  trt.type="cont", iv="carb", imp.var=NA, tol.var=NA, adjust.IV=TRUE,
  max.time.seconds=2)
summary(k)
```

smahal

Compute rank-based Mahalanobis distance matrix between each pair

Description

This function computes the rank-based Mahalanobis distance matrix between each pair of observations in the data set. Called by `matches` (and ultimately `opt_nearfar`) function to set up a distance matrix used to create pair matches.

Usage

```
smahal(X)
```

Arguments

X A matrix of observed confounders with n rows (observations) and p columns (variables)

Value

Returns the rank-based Mahalanobis distance matrix between every pair of observations

Examples

```
smahal(mtcars[1:4, 2:3])
```

| | |
|------------|------------------------------------------------|
| summary.nf | <i>Summary method for object of class "nf"</i> |
|------------|------------------------------------------------|

Description

Displays key information, e.g., number of matches tried, and post-match balance, for [opt_nearfar](#) function

Usage

```
## S3 method for class 'nf'  
summary(object, ...)
```

Arguments

| | |
|--------|--------------------------------------------------------------|
| object | Object of class "nf" returned by opt_nearfar |
| ... | additional arguments affecting the summary produced |

Value

Returns a summary of results from [opt_nearfar](#) function

Author(s)

Joseph Rigdon <jrigdon@wakehealth.edu>

See Also

[opt_nearfar](#)

Examples

```
k = opt_nearfar(dta=mtcars, trt="drat", covs=c("cyl", "disp"),  
               trt.type="cont", iv="carb", imp.var=NA, tol.var=NA, adjust.IV=TRUE,  
               max.time.seconds=1)  
summary(k)
```

| | |
|--------------|------------------------------------------------------------|
| summ_matches | <i>Computes table of absolute standardized differences</i> |
|--------------|------------------------------------------------------------|

Description

Computes absolute standardized differences for both continuous and binary variables. Called by [opt_nearfar](#) to summarize results of near-far match.

Usage

```
summ_matches(dta, iv, covs, match)
```

Arguments

| | |
|-------|--------------------------------------------------------------------------------------------|
| dta | The name of the data frame on which matching was performed |
| iv | The name of the instrumental variable, e.g., iv="QOB" |
| covs | A vector of the names of the covariates to make "near", e.g., covs=c("age", "sex", "race") |
| match | A two-column matrix of row indices of paired matches |

Value

A table of mean variable values for both the "encouraged" and "discouraged" groups across all variables plus absolute standardized differences for each variable

Author(s)

Joseph Rigdon <jrigdon@wakehealth.edu>

See Also

[opt_nearfar](#)

Examples

```
k2 = matches(dta=mtcars, covs=c("cyl", "disp"), sinks=0.2, iv="carb",
             cutpoint=2, imp.var=c("cyl"), tol.var=0.03)
summ_matches(dta=mtcars, iv="carb", covs=c("cyl", "disp"), match=k2)
```

Index

* **datasets**

 angrist, [2](#)

angrist, [2](#)

calipers, [3](#)

eff_ratio, [4](#)

matches, [3](#), [4](#), [5](#), [8](#)

nearfar-package, [2](#)

opt_nearfar, [3-6](#), [6](#), [8-10](#)

smahal, [8](#)

summ_matches, [10](#)

summary.nf, [9](#)