Package 'sitepickR'

December 5, 2022

Type Package

Title Two-Level Sample Selection with Optimal Site Replacement

Version 0.0.1

Date 2022-11-29

Description Carries out a two-level sample selection where the possibility of an initially selected site not wanting to participate is anticipated, and the site is optimally replaced. The procedure aims to reduce bias (and/or loss of external validity) with respect to the target population. In selecting units and sub-units, 'sitepickR' uses the cube method developed by 'Deville & Tillé', (2004) <http://www.math.helsinki.fi/msm/banocoss/Deville_Tille_2004.pdf> and described in Tillé (2011) <https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2011002/article/11609-eng.pdf?st=5-sx8Q8n>. The cube method is a probability sampling method that is designed to satisfy criteria for balance between the sample and the population. Recent research has shown that this method performs well in simulations for studies of educational programs (see Fay & Olsen (2021, under review). To implement the cube method, 'sitepickR' uses the sampling R package <https://cran.r-project.org/package=sampling>. To implement statistical matching, 'sitepickR' uses the 'MatchIt' R package <https://cran.r-project.org/package=MatchIt>.

Imports MatchIt, sampling, dplyr, ggplot2, reshape2, data.table, stats, stringr, tidyr, magrittr, tidyselect, scales

License GPL (>= 3)

Encoding UTF-8

LazyData true

RoxygenNote 7.2.2

Depends R (>= 2.10)

Suggests knitr, rmarkdown, devtools

VignetteBuilder knitr

NeedsCompilation no

Author Elena Badillo-Goicoechea [aut, cre], Robert Olsen [aut], Elizabeth Stuart [aut]

Maintainer Elena Badillo-Goicoechea <egoicoe1@jhu.edu>

Repository CRAN

Date/Publication 2022-12-05 11:00:02 UTC

R topics documented:

`getSummary` *Summary tables*

Description

Build summary tables, with unit/match/sub-unit balance between initially selected units and a target population, for each covariate of interest

Usage

```
getSummary(smOut, diagnostic)
```

Arguments

`smOut`	list; selectMatch() output
`diagnostic`	numeric; balance Diagnostic: "unitBal" = original unit balance, "matchBal" = match balance, "matchFreq" = sucessful match frequency, "matchCount" = match success count by replacement group, "subunitBal" =sub-unit balance

Value

ggplot object

Examples

```
##############################################################################
############## Balance Diagnostics [sitepickR Package] #######################
######### Robert Olsen, Elizabeth A. Stuart & Elena Badillo-Goicoechea (2022) ##
##############################################################################

# Basic usage of getSummary()

rawCCD <- sitepickR::rawCCD

uSampVarsCCD <- c("w.pct.frlunch", "w.pct.black", "w.pct.hisp", "w.pct.female")
suSampVarsCCD <- c("sch.pct.frlunch", "sch.pct.black", "sch.pct.hisp", "sch.pct.female")

dfCCD <- prepDF(rawCCD,
                unitID="LEAID", subunitID="NCESSCH")
dfCCD <- dplyr::filter(dfCCD, unitID %in% unique(dfCCD$unitID)[1:80])

smOut <- selectMatch(df = dfCCD, # user dataset
                     unitID = "LEAID", # column name of unit ID in user dataset
                     subunitID = "NCESSCH", # column name of sub-unit ID in user dataset
                     unitVars = uSampVarsCCD, # name of unit level covariate columns
                  subunitSampVars = suSampVarsCCD, # name of sub-unit level covariate columns
                     nUnitSamp = 30,
                     nRepUnits = 5,
                     nsubUnits = 2
)

getSummary(smOut, diagnostic="unitBal")
```

`matchBalance` *Match balance*

Description

Balance between initially sampled units and their K matches, for each covariate of interest

Usage

```
matchBalance(
  smOut,
  title = "Standardized Mean Difference:",
  subtitle = "Replacement Unit Groups (1...K) vs. Originally Selected Units"
)
```

Arguments

`smOut`	list; selectMatch() output
`title`	character; user-specified figure title
`subtitle`	character; user-specified figure title

Value

ggplot object

Examples

```
################################################################################
############## Balance Diagnostics [sitepickR Package] ########################
######### Robert Olsen, Elizabeth A. Stuart & Elena Badillo-Goicoechea (2022) ##
################################################################################

# Basic usage of matchBalance()

rawCCD <- sitepickR::rawCCD

uSampVarsCCD <- c("w.pct.frlunch", "w.pct.black", "w.pct.hisp", "w.pct.female")
suSampVarsCCD <- c("sch.pct.frlunch", "sch.pct.black", "sch.pct.hisp", "sch.pct.female")

dfCCD <- prepDF(rawCCD,
                unitID="LEAID", subunitID="NCESSCH")
dfCCD <- dplyr::filter(dfCCD, unitID %in% unique(dfCCD$unitID)[1:80])

smOut <- selectMatch(df = dfCCD, # user dataset
                     unitID = "LEAID", # column name of unit ID in user dataset
                     subunitID = "NCESSCH", # column name of sub-unit ID in user dataset
                     unitVars = uSampVarsCCD, # name of unit level covariate columns
                   subunitSampVars = suSampVarsCCD, # name of sub-unit level covariate columns
                     nUnitSamp = 30,
                     nRepUnits = 5,
                     nsubUnits = 2
)
matchBalance(smOut)
```

`matchCount` *Successful matches for each replacement group*

Description

Percentage of successful matches in each unit replacement group, 1...K

Usage

```
matchCount(smOut, title = "Percentage of Successful Matches per Unit Group")
```

Arguments

`smOut`	list; selectMatch() output
`title`	character; user-specified figure title

Value

ggplot object

Examples

```
################################################################################
############## Balance Diagnostics [sitepickR Package] ########################
######### Robert Olsen, Elizabeth A. Stuart & Elena Badillo-Goicoechea (2022) ##
################################################################################

# Basic usage of matchCount()

rawCCD <- sitepickR::rawCCD

uSampVarsCCD <- c("w.pct.frlunch", "w.pct.black", "w.pct.hisp", "w.pct.female")
suSampVarsCCD <- c("sch.pct.frlunch", "sch.pct.black", "sch.pct.hisp", "sch.pct.female")

dfCCD <- prepDF(rawCCD,
                unitID="LEAID", subunitID="NCESSCH")
dfCCD <- dplyr::filter(dfCCD, unitID %in% unique(dfCCD$unitID)[1:80])

smOut <- selectMatch(df = dfCCD, # user dataset
                     unitID = "LEAID", # column name of unit ID in user dataset
                     subunitID = "NCESSCH", # column name of sub-unit ID in user dataset
                     unitVars = uSampVarsCCD, # name of unit level covariate columns
                  subunitSampVars = suSampVarsCCD, # name of sub-unit level covariate columns
                     nUnitSamp = 30,
                     nRepUnits = 5,
                     nsubUnits = 2
)
matchCount(smOut)
```

`matchFreq` *Match frequency*

Description

Distribution of successful matches among original units

Usage

```
matchFreq(smOut, title = "Match Frequency per Original Unit")
```

Arguments

`smOut`	list; selectMatch() output
`title`	character; user-specified figure title

Value

ggplot object

Examples

```
################################################################################
############## Balance Diagnostics [sitepickR Package] ########################
######### Robert Olsen, Elizabeth A. Stuart & Elena Badillo-Goicoechea (2022) ##
################################################################################

# Basic usage of matchFreq()

rawCCD <- sitepickR::rawCCD

uSampVarsCCD <- c("w.pct.frlunch", "w.pct.black", "w.pct.hisp", "w.pct.female")
suSampVarsCCD <- c("sch.pct.frlunch", "sch.pct.black", "sch.pct.hisp", "sch.pct.female")

dfCCD <- prepDF(rawCCD,
                unitID="LEAID", subunitID="NCESSCH")
dfCCD <- dplyr::filter(dfCCD, unitID %in% unique(dfCCD$unitID)[1:80])

smOut <- selectMatch(df = dfCCD, # user dataset
                     unitID = "LEAID", # column name of unit ID in user dataset
                     subunitID = "NCESSCH", # column name of sub-unit ID in user dataset
                     unitVars = uSampVarsCCD, # name of unit level covariate columns
                  subunitSampVars = suSampVarsCCD, # name of sub-unit level covariate columns
                     nUnitSamp = 30,
                     nRepUnits = 5,
                     nsubUnits = 2
)
matchFreq(smOut)
```

`prepDF` *Prepare nested dataset*

Description

Prepare nested dataset

Usage

```
prepDF(df, unitID, subunitID)
```

Arguments

`df`	dataframe
`unitID`	character; unit column name in original dataset
`subunitID`	character; sub-unit column name in original dataset

Value

processed dataframe

Examples

```
################################################################################
############## Prepare dataframe [sitepickR Package] ###########################
######### Robert Olsen, Elizabeth A. Stuart & Elena Badillo-Goicoechea (2022) ##

# Basic usage of prepDF()

rawCCD <- sitepickR::rawCCD

uSampVarsCCD <- c("w.pct.frlunch", "w.pct.black", "w.pct.hisp", "w.pct.female")
suSampVarsCCD <- c("sch.pct.frlunch", "sch.pct.black", "sch.pct.hisp", "sch.pct.female")

dfCCD <- prepDF(rawCCD,
                unitID="LEAID", subunitID="NCESSCH")
```

---

rawCCD *Common Core of Data (CCD) data for California schools (2017-18).*

Description

A pre-processed dataset containing key variables from administrative data compiled by the CCD, aggregated at the district and school level for public schools in California for the 2017 and 2018 school years.

Usage

```
data(rawCCD)
```

Format

A data frame with 1890 rows and 11 variables.

LEAID school district unique identifier

NCESSCH school unique identifier

w.pct.frlunch percentage of students in the school district who are under free/reduced price lunch program; weighted by school size.

w.pct.black percentage of students in the school district who are Black; weighted by school size.

w.pct.hisp percentage of students in the school district who are Hispanic; weighted by school size.

w.pct.female percentage of students in the school district who are female; weighted by school size.

sch.pct.frlunch percentage of students in the school who are under free/reduced price lunch program.

sch.pct.black percentage of students in the school who are Black.

sch.pct.hisp percentage of students in the school who are Hispanic.

sch.pct.female percentage of students in the school who are female.

distr.type school district type (constructed for illustration purposes; (values="A", "B", "C", "D")).

dtrct_size number of schools in the district

Source

https://nces.ed.gov/ccd/files.asp#FileNameId:15,VersionId:10,FileSchoolYearId:33,Page:1

`selectMatch` *Two-level sample selection*

Description

Carries out a two-level sample selection where the possibility of an initially selected site not wanting to participate is anticipated, and the site is optimally replaced. The procedure aims to reduce the bias (and/or loss of generalizability) with respect to the target population.

Usage

```
selectMatch(
  df,
  unitID,
  subunitID,
  subunitSampVars,
  unitVars,
  nUnitSamp,
  nRepUnits,
  nsubUnits,
  exactMatchVars = NULL,
  calipMatchVars = NULL,
  calipValue = 0.2,
  seedN = NA,
  matchDistance = "mahalanobis",
  sizeFlag = TRUE,
  repFlag = TRUE,
  writeOut = FALSE,
  replacementUnitsFilename = "replacementUnits.csv",
  subUnitTableFilename = "subUnitTable.csv"
)
```

Arguments

df	dataframe; sub-unit level dataframe with both sub-unit and unit level variables
unitID	character; name of unit ID column
subunitID	character; name of sub-unit ID column
subunitSampVars	vector; column names of unit level variables to sample units on
unitVars	vector; column names of unit level variables to match units on
nUnitSamp	numeric; number of units to be initially randomly selected
nRepUnits	numeric; number of replacement units to find for each selected unit
nsubUnits	numeric; number of sub-units to be randomly selected for each unit
exactMatchVars	vector; column names of categorical variables on which units must be matched exactly. Must be present in 'unitVars'; default = NULL
calipMatchVars	vector; column names of continuous variables on which units must be matched within a specified caliper. Must be present in 'unitVars'; default = NULL
calipValue	numeric; number of standard deviations to be used as caliper for matching units on calipMatchVars
seedN	numeric; seed number to be used for sampling. If NA, calls set.seed(); default = NA
matchDistance	character; MatchIt distance parameter to obtain optimal matches (nearest neigboors); default = "mahalanois"
sizeFlag	logical; if TRUE, sampling is made proportional to unit size; default = TRUE
repFlag	logical; if TRUE, pick unit matches with/without repetition; default = TRUE
writeOut	logical; if TRUE, writes a .csv file for each output table; default = FALSE
replacementUnitsFilename	character; csv filename for saving unit:replacement directory when writeOut == TRUE; default = "replacementUnits.csv"
subUnitTableFilename	character; csv filename for saving unit:replacement directory when writeOut == TRUE; default = "subUnitTable.csv"

Value

list with: 1) table of the form: selected unit i: (unit i replacements), 2) table of the form: potential unit i:(unit i sub-units), 3) balance diagnostics.

Examples

```
##############################################################################
############## Prepare dataframe [sitepickR Package] ########################
######### Robert Olsen, Elizabeth A. Stuart & Elena Badillo-Goicoechea (2022) ##

# Basic usage of selectMatch()

rawCCD <- sitepickR::rawCCD
```

```
uSampVarsCCD <- c("w.pct.frlunch", "w.pct.black", "w.pct.hisp", "w.pct.female")
suSampVarsCCD <- c("sch.pct.frlunch", "sch.pct.black", "sch.pct.hisp", "sch.pct.female")

dfCCD <- prepDF(rawCCD,
                unitID="LEAID", subunitID="NCESSCH")
dfCCD <- dplyr::filter(dfCCD, unitID %in% unique(dfCCD$unitID)[1:80])

smOut <- selectMatch(df = dfCCD, # user dataset
                     unitID = "LEAID", # column name of unit ID in user dataset
                     subunitID = "NCESSCH", # column name of sub-unit ID in user dataset
                     unitVars = uSampVarsCCD, # name of unit level covariate columns
                  subunitSampVars = suSampVarsCCD, # name of sub-unit level covariate columns
                     nUnitSamp = 30,
                     nRepUnits = 5,
                     nsubUnits = 2
)
```

`subUnitBalance` *Sub-unit balance*

Description

Sub-unit balance between initially selected units and all units in population, for each covariate of interest

Usage

```
subUnitBalance(
  smOut,
  title = "Subunits from Original and Replacement Unit Groups vs. Population (SMD)"
)
```

Arguments

`smOut`	list; selectMatch() output
`title`	character; user-specified figure title

Value

ggplot object

Examples

```
################################################################################
############## Balance Diagnostics [sitepickR Package] ########################
######### Robert Olsen, Elizabeth A. Stuart & Elena Badillo-Goicoechea (2022) ##
################################################################################
```

```
# Basic usage of subUnitBalance()
rawCCD <- sitepickR::rawCCD

uSampVarsCCD <- c("w.pct.frlunch", "w.pct.black", "w.pct.hisp", "w.pct.female")
suSampVarsCCD <- c("sch.pct.frlunch", "sch.pct.black", "sch.pct.hisp", "sch.pct.female")

dfCCD <- prepDF(rawCCD,
                unitID="LEAID", subunitID="NCESSCH")
dfCCD <- dplyr::filter(dfCCD, unitID %in% unique(dfCCD$unitID)[1:80])

smOut <- selectMatch(df = dfCCD, # user dataset
                     unitID = "LEAID", # column name of unit ID in user dataset
                     subunitID = "NCESSCH", # column name of sub-unit ID in user dataset
                     unitVars = uSampVarsCCD, # name of unit level covariate columns
                subunitSampVars = suSampVarsCCD, # name of sub-unit level covariate columns
                     nUnitSamp = 30,
                     nRepUnits = 5,
                     nsubUnits = 2
)
subUnitBalance(smOut =smOut,
               title="Standardized Mean Difference:
               Sub-units from Original + Replacement Unit Groups vs. Population")
```

`unitLovePlot` *Original units balance*

Description

Balance between initially sampled units and all units in the population

Usage

```
unitLovePlot(
  smOut,
  title = "Standardized Mean Difference",
  subtitle = "Initially Selected Units vs. Population"
)
```

Arguments

`smOut`	list; selectMatch() output
`title`	character; user-specified figure title
`subtitle`	character; user-specified figure subtitle

Value

ggplot object

Examples

```
################################################################################
############## Balance Diagnostics [sitepickR Package] ########################
######### Robert Olsen, Elizabeth A. Stuart & Elena Badillo-Goicoechea (2022) ##
################################################################################

# Basic usage of unitLovePlot()
rawCCD <- sitepickR::rawCCD

uSampVarsCCD <- c("w.pct.frlunch", "w.pct.black", "w.pct.hisp", "w.pct.female")
suSampVarsCCD <- c("sch.pct.frlunch", "sch.pct.black", "sch.pct.hisp", "sch.pct.female")

dfCCD <- prepDF(rawCCD,
                unitID="LEAID", subunitID="NCESSCH")
dfCCD <- dplyr::filter(dfCCD, unitID %in% unique(dfCCD$unitID)[1:80])

smOut <- selectMatch(df = dfCCD, # user dataset
                     unitID = "LEAID", # column name of unit ID in user dataset
                     subunitID = "NCESSCH", # column name of sub-unit ID in user dataset
                     unitVars = uSampVarsCCD, # name of unit level covariate columns
                   subunitSampVars = suSampVarsCCD, # name of sub-unit level covariate columns
                     nUnitSamp = 30,
                     nRepUnits = 5,
                     nsubUnits = 2
)
unitLovePlot(smOut)
```

Index