Package ‘DisImpact’

September 20, 2020

Title Calculates Disproportionate Impact When Binary Success Data are Disaggregated by Subgroups

Version 0.0.14

Description Implements methods for calculating disproportionate impact: the percentage point gap, proportionality index, and the 80% index.

Depends R (>= 3.4.0)
Imports dplyr (>= 0.8.5), rlang, tidyselect, purrr, tidyr
License GPL-3

URL https://github.com/vinhdizzo/DisImpact

BugReports https://github.com/vinhdizzo/DisImpact/issues

Encoding UTF-8

LazyData true

RoxygenNote 7.1.0

Suggests knitr, rmarkdown, ggplot2, forcats, scales

VignetteBuilder knitr

NeedsCompilation no

Author Vinh Nguyen [aut, cre]

Maintainer Vinh Nguyen <nguyenvq714@gmail.com>

Repository CRAN

Date/Publication 2020-09-19 22:20:30 UTC
R topics documented:

- di_80_index .......................................................... 2
- di_iterate ............................................................ 4
- di_ppg ............................................................... 7
- di_ppg_iterate ........................................................ 10
- di_prop_index .......................................................... 11
- ppg_moe .............................................................. 13
- student_equity ......................................................... 14

Index 15

di_80_index Calculate disproportionate impact per the 80% index

Description

Calculate disproportionate impact per the 80% index method.

Usage

di_80_index(
  success,
  group,
  cohort,  # Optional
  weight,  # Optional
  data,
  di_80_index_cutoff = 0.8,
  reference_group = "hpg",
  check_valid_reference = TRUE
)

Arguments

- **success** A vector of success indicators (1/0 or TRUE/FALSE) or an unquoted reference (name) to a column in data if it is specified. It could also be a vector of counts, in which case weight should also be specified (group size).

- **group** A vector of group names of the same length as success or an unquoted reference (name) to a column in data if it is specified.

- **cohort** (Optional) A vector of cohort names of the same length as success or an unquoted reference (name) to a column in data if it is specified. Disproportionate impact is calculated for every group within each cohort. When cohort is not specified, then the analysis assumes a single cohort.

- **weight** (Optional) A vector of case weights of the same length as success or an unquoted reference (name) to a column in data if it is specified. If success consists of counts instead of success indicators (1/0), then weight should also be specified to indicate the group size.
data (Optional) A data frame containing the variables of interest. If data is specified, then success, group, and cohort will be searched within it.

di_80_index_cutoff
A numeric value between 0 and 1 that is used to determine disproportionate impact if the index comparing the success rate of the current group to the reference group falls below this threshold; defaults to 0.80.

reference_group
The reference group value in group that each group should be compared to in order to determine disproportionate impact. By default (‘hpg’), the group with the highest success rate is used as reference.

check_valid_reference
Check whether reference_group is a valid value; defaults to TRUE. This argument exists to be used in di_iterate as when iterating DI calculations, there may be some scenarios where a specified reference group does not contain any students.

Details
This function determines disproportionate impact based on the 80% index method, as described in this reference from the California Community Colleges Chancellor’s Office. It assumes that a higher rate is good ("success"). For rates that are deemed negative (eg, rate of drop-outs, high is bad), then consider looking at the converse of the non-success (eg, non drop-outs, high is good) instead in order to leverage this function properly.

Value
A data frame consisting of:

• cohort (if used),
• group,
• n (sample size),
• success (number of successes for the cohort-group),
• pct (proportion of successes for the cohort-group),
• reference_group (the reference group used to compare and determine disproportionate impact),
• reference (the reference rate used for comparison, corresponding to reference_group),
• di_80_index (ratio of pct to the reference),
• di_indicator (1 if di_80_index < di_80_index_cutoff),
• success_needed_not_di (the number of additional successes needed in order to no longer be considered disproportionately impacted as compared to the reference), and
• success_needed_full_parity (the number of additional successes needed in order to achieve full parity with the reference).

References
Examples

```r
library(dplyr)
data(student_equity)
di_80_index(success=Transfer, group=Ethnicity, data=student_equity) %>%
as.data.frame
```

**di_iterate**

Iteratively calculate disproportionate impact using multiple method for many variables.

Description

Iteratively calculate disproportionate impact via the percentage point gap (PPG), proportionality index, and 80% index methods for many success variables, disaggregation variables, and scenarios.

Usage

```r
di_iterate(
data,
success_vars,
group_vars,
cohort_vars = NULL,
scenario_repeat_by_vars = NULL,
exclude_scenario_df = NULL,
weight_var = NULL,
include_non_disagg_results = TRUE,
ppg_reference_groups = "overall",
min_moe = 0.03,
use_prop_in_moe = FALSE,
prop_sub_0 = 0.5,
prop_sub_1 = 0.5,
di_prop_index_cutoff = 0.8,
di_80_index_cutoff = 0.8,
di_80_index_reference_groups = "hpg",
check_valid_reference = TRUE
)
```

Arguments

- **data**: A data frame for which to iterate DI calculations for a set of variables.
- **success_vars**: A character vector of success variable names to iterate across.
- **group_vars**: A character vector of group (disaggregation) variable names to iterate across.
- **cohort_vars**: (Optional) A character vector of the same length as success_vars to indicate the cohort variable to be used for each variable specified in success_vars. A vector of length 1 could be specified, in which case the same cohort variable is used for each success variable. If not specified, then a single cohort is assumed for all success variables.
**di_iterate**

scenario_repeat_by_vars

(Optional) A character vector of variables to repeat DI calculations for across all combination of these variables. For example, the following variables could be specified:

- Ed Goal: Degree/Transfer, Shot-term Career, Non-credit
- First time college student: Yes, No
- Full-time status: Yes, No

Each combination of these variables (e.g., full time, first time college students with an ed goal of degree/transfer as one combination) would constitute an iteration/sample for which to calculate disproportionate impact for outcomes listed in success_vars and for the disaggregation variables listed in group_vars. The overall rate of success for full time, first time college students with an ed goal of degree/transfer would just include these students and not others. Each variable specified is also collapsed to an '- All' group so that the combinations also reflect all students of a particular category. The total number of combinations for the previous example would be \((+1\text{ representing the all category}): (3 + 1) \times (2 + 1) \times (2 + 1) = 36\).

exclude_scenario_df

(Optional) A data frame with variables that match scenario_repeat_by_vars for specifying the combinations to exclude from DI calculations. Following the example specified above, one could choose to exclude part-time non-credit students from consideration.

weight_var

(Optional) A character variable specifying the weight variable if the input data set is summarized (i.e., the success variables specified in success_vars contain count of successes). Weight here corresponds to the denominator when calculating the success rate. Defaults to NULL for an input data set where each row describes each individual.

include_non_disagg_results

A logical variable specifying whether or not the non-disaggregated results should be returned; defaults to TRUE. When TRUE, a new variable `~None` is added to the data set with a single data value `~All` and this variable is added group_vars as a disaggregation/group variable. The user would want these results returned to review non-disaggregated results.

ppg_reference_groups

Either 'overall', 'hpg', 'all but current', or a character vector of the same length as group_vars that indicates the reference group value for each group variable in group_vars when determining disproportionate impact using the percentage point gap method.

min_moe

The minimum margin of error to be used in the PPG calculation, passed to di_ppg.

use_prop_in_moe

Whether the estimated proportions should be used in the margin of error calculation by the PPG, passed to di_ppg.

prop_sub_0

passed to di_ppg; defaults to 0.50.

prop_sub_1

passed to di_ppg; defaults to 0.50.
di_prop_index_cutoff
  Threshold used for determining disproportionate impact using the proportionality index; passed to `di_prop_index`; defaults to 0.80.

di_80_index_cutoff
  Threshold used for determining disproportionate impact using the 80% index; passed to `di_80_index`; defaults to 0.80.

di_80_index_reference_groups
  A character vector of the same length as `groupVars` that indicates the reference group value for each group variable in `groupVars` when determining disproportionate impact using the 80% index; defaults to `'hpg'` (highest performing group as reference).

check_valid_reference
  Check whether `ppg_reference_groups` and `di_80_index_reference_groups` contain valid values; defaults to TRUE.

Details
Iteratively calculate disproportionate impact via the percentage point gap (PPG), proportionality index, and 80% index methods for all combinations of `success_vars`, `group_vars`, and `cohort_vars`, for each combination of subgroups specified by `scenario_repeat_by_vars`.

Value
A summarized data set (data frame) consisting of:

- `success_variable` (elements of `success_vars`),
- `disaggregation` (elements of `group_vars`),
- `cohort` (values corresponding to the variables specified in `cohort_vars`),
- `di_indicator_ppg` (1 if there is disproportionate impact per the percentage point gap method, 0 otherwise),
- `di_indicator_prop_index` (1 if there is disproportionate impact per the proportionality index, 0 otherwise),
- `di_indicator_80_index` (1 if there is disproportionate impact per the 80% index, 0 otherwise), and
- other relevant fields returned from `di_ppg`, `di_prop_index`, and `di_80_index`.

Examples
```r
library(dplyr)
data(student_equity)
# Multiple group variables
di_iterate(data=student_equity, success_vars=c('Transfer'),
  group_vars=c('Ethnicity', 'Gender'), cohort_vars=c('Cohort'),
  ppg_reference_groups='overall')
```
**di_ppg**

Calculate disproportionate impact per the percentage point gap (PPG) method.

---

**Description**

Calculate disproportionate impact per the percentage point gap (PPG) method.

**Usage**

```r
di_ppg(
  success,
  group,
  cohort,
  weight,
  reference = c("overall", "hpg", "all but current", unique(group)),
  data,
  min_moe = 0.03,
  use_prop_in_moe = FALSE,
  prop_sub_0 = 0.5,
  prop_sub_1 = 0.5,
  check_valid_reference = TRUE
)
```

**Arguments**

- **success**: A vector of success indicators (1/0 or TRUE/FALSE) or an unquoted reference (name) to a column in `data` if it is specified. It could also be a vector of counts, in which case `weight` (group size) should also be specified.

- **group**: A vector of group names of the same length as `success` or an unquoted reference (name) to a column in `data` if it is specified.

- **cohort** (Optional): A vector of cohort names of the same length as `success` or an unquoted reference (name) to a column in `data` if it is specified. Disproportionate impact is calculated for every group within each cohort. When `cohort` is not specified, then the analysis assumes a single cohort.

- **weight** (Optional): A vector of case weights of the same length as `success` or an unquoted reference (name) to a column in `data` if it is specified. If `success` consists of counts instead of success indicators (1/0), then `weight` should also be specified to indicate the group size.

- **reference**: Either 'overall' (default), 'hpg' (highest performing group), 'all but current' (success rate of everyone excluding the comparison group; also known as 'ppg minus 1'), a value from `group` (specifying a reference group), a single proportion (e.g., 0.50), or a vector of proportions (one for each cohort). Reference is used as a point of comparison for disproportionate impact for each group. When `cohort` is specified:
di_pgg

- 'overall' will use the overall success rate of each cohort group as the reference;
- 'hpg' will use the highest performing group in each cohort as reference;
- 'all but current' will use the calculated success rate of each cohort group excluding the comparison group
- the success rate of the specified reference group from group in each cohort will be used;
- the specified proportion will be used for all cohorts;
- the specified vector of proportions will refer to the reference point for each cohort in alphabetical order (so the number of proportions should equal to the number of unique cohorts).

data (Optional) A data frame containing the variables of interest. If data is specified, then success, group, and cohort will be searched within it.

min_moe The minimum margin of error (MOE) to be used in the calculation of disproportionate impact and is passed to ppg_moe. Defaults to 0.03.

use_prop_in_moe A logical value indicating whether or not the MOE formula should use the observed success rates (TRUE). Defaults to FALSE, which uses 0.50 as the proportion in the MOE formula. If TRUE, the success rates are passed to the proportion argument of ppg_moe.

prop_sub_0 For cases where proportion is 0, substitute with prop_sub_0 (defaults to 0.5) to account for the zero MOE. This is relevant only when use_prop_in_moe=TRUE.

prop_sub_1 For cases where proportion is 1, substitute with prop_sub_1 (defaults to 0.5) to account for the zero MOE. This is relevant only when use_prop_in_moe=TRUE.

check_valid_reference Check whether reference is a valid value; defaults to TRUE. This argument exists to be used in di_iterate as when iterating DI calculations, there may be some scenarios where a specified reference group does not contain any students.

Details

This function determines disproportionate impact based on the percentage point gap (PPG) method, as described in this reference from the California Community Colleges Chancellor’s Office. It assumes that a higher rate is good ("success"). For rates that are deemed negative (eg, rate of drop-outs, high is bad), then consider looking at the converse of the non-success (eg, non drop-outs, high is good) instead in order to leverage this function properly. Note that the margin of error (MOE) is calculated using using $1.96 \times \sqrt{0.25^2/n}$, with a min_moe used as the minimum by default.

Value

A data frame consisting of:
- cohort (if used),
- group,
- n (sample size),
- success (number of successes for the cohort-group),
• **pct** (proportion of successes for the cohort-group),
• **reference_group** (reference group used in DI calculation),
• **reference** (reference value used in DI calculation),
• **moe** (margin of error),
• **pct_lo** (lower 95% confidence limit for pct),
• **pct_hi** (upper 95% confidence limit for pct),
• **di_indicator** (1 if there is disproportionate impact, i.e., when \( \text{pct}_\text{hi} \leq \text{reference} \)),
• **success_needed_not_di** (the number of additional successes needed in order to no longer be considered disproportionately impacted as compared to the reference), and
• **success_needed_full_parity** (the number of additional successes needed in order to achieve full parity with the reference).

**References**


**Examples**

```r
library(dplyr)
data(student_equity)
# Vector
di_ppg(success=student.equity$Transfer , group=student.equity$Ethnicity) %>% as.data.frame
# Tidy and column reference
di_ppg(success=Transfer, group=Ethnicity, data=student_equity) %>%
as.data.frame
# Cohort
di_ppg(success=Transfer, group=Ethnicity, cohort=Cohort , data=student_equity) %>%
as.data.frame
# With custom reference (single)
di_ppg(success=Transfer, group=Ethnicity, reference=0.54 , data=student_equity) %>%
as.data.frame
# With custom reference (multiple)
di_ppg(success=Transfer, group=Ethnicity, cohort=Cohort , reference=c(0.5, 0.55), data=student_equity) %>%
as.data.frame
# min_moe
di_ppg(success=Transfer, group=Ethnicity, data=student_equity , min_moe=0.02) %>%
as.data.frame
# use_prop_in_moe
di_ppg(success=Transfer, group=Ethnicity, data=student_equity , min_moe=0.02 , use_prop_in_moe=TRUE) %>%
as.data.frame
```
**di_ppg_iterate**

Iteratively calculate disproportionate impact via the percentage point gap (PPG) method for many variables.

**Description**

Iteratively calculate disproportionate impact via the percentage point gap (PPG) method for many disaggregation variables.

**Usage**

```r
di_ppg_iterate(
  data,
  success_vars,
  group_vars,
  cohort_vars,
  reference_groups,
  repeat_by_vars = NULL,
  weight_var = NULL,
  min_moe = 0.03,
  use_prop_in_moe = FALSE,
  prop_sub_0 = 0.5,
  prop_sub_1 = 0.5
)
```

**Arguments**

- `data` A data frame for which to iterate DI calculation for a set of variables.
- `success_vars` A character vector of success variable names to iterate across.
- `group_vars` A character vector of group (disaggregation) variable names to iterate across.
- `cohort_vars` A character vector of cohort variable names to iterate across.
- `reference_groups` Either 'overall', 'hpg', or a character vector of the same length as 'group_vars' that indicates the reference group value for each group variable in 'group_vars'.
- `repeat_by_vars` A character vector of variables to repeat DI calculations for across all combination of these variables, including '- All' as a group for each variable. The reference rate used for DI comparison differs for every combination of the variables listed here.
- `weight_var` A character scalar specifying the weight variable if the input data set is summarized (ie, the success variables specified in 'success_vars' contain count of successes). Weight here corresponds to the denominator when calculating the success rate. Defaults to 'NULL' for an input data set where each row describes each individual.
- `min_moe` The minimum margin of error to be used in the PPG calculation, passed to 'di_ppg'.
**di_prop_index**

---

**use_prop_in_moe**

Whether the estimated proportions should be used in the margin of error calculation by the PPG, passed to `di_ppg`.

**prop_sub_0**

Passed to `di_ppg`.

**prop_sub_1**

Passed to `di_ppg`.

---

**Details**

Iteratively calculate disproportionate impact via the percentage point gap (PPG) method for all combinations of `success_vars`, `group_vars`, and `cohort_vars`, for each combination of subgroups specified by `repeat_by_vars`.

---

**Value**

A data frame with all relevant returned fields from `di_ppg` plus `success_variable` (elements of `success_vars`), `disaggregation` (elements of `group_vars`), and `reference_group` (elements of `reference_groups`).

---

**Examples**

```r
library(dplyr)
data(student_equity)
# Multiple group variables
di_ppg_iterate(data=student_equity, success_vars=c('Transfer'),
               group_vars=c('Ethnicity', 'Gender'), cohort_vars=c('Cohort'),
               reference_groups='overall')
```

---

**di_prop_index**

Calculate disproportionate impact per the proportionality index (PI) method.

---

**Description**

Calculate disproportionate impact per the proportionality index (PI) method.

**Usage**

```r
di_prop_index(success, group, cohort, weight, data, di_prop_index_cutoff = 0.8)
```

**Arguments**

- **success**
  
  A vector of success indicators (1/0 or TRUE/FALSE) or an unquoted reference (name) to a column in data if it is specified. It could also be a vector of counts, in which case weight should also be specified (group size).

- **group**
  
  A vector of group names of the same length as success or an unquoted reference (name) to a column in data if it is specified.
cohort (Optional) A vector of cohort names of the same length as success or an unquoted reference (name) to a column in data if it is specified. disproportionate impact is calculated for every group within each cohort. When cohort is not specified, then the analysis assumes a single cohort.

weight (Optional) A vector of case weights of the same length as success or an unquoted reference (name) to a column in data if it is specified. If success consists of counts instead of success indicators (1/0), then weight should also be specified to indicate the group size.

data (Optional) A data frame containing the variables of interest. If data is specified, then success, group, and cohort will be searched within it.

di_prop_index_cutoff A numeric value between 0 and 1 that is used to determine disproportionate impact if the proportionality index falls below this threshold; defaults to 0.80.

Details
This function determines disproportionate impact based on the proportionality index (PI) method, as described in this reference from the California Community Colleges Chancellor’s Office. It assumes that a higher rate is good (“success”). For rates that are deemed negative (eg, rate of drop-outs, high is bad), then consider looking at the converse of the non-success (eg, non drop-outs, high is good) instead in order to leverage this function properly.

Value
A data frame consisting of:
• cohort (if used),
• group,
• n (sample size),
• success (number of successes for the cohort-group),
• pct_success (proportion of successes attributed to the group within the cohort),
• pct_group (proportion of sample attributed to the group within the cohort),
• di_prop_index (ratio of pct_success to pct_group),
• di_indicator (1 if di_prop_index < di_prop_index_cutoff), and
• success_needed_not_di (the number of additional successes needed in order to no longer be considered disproportionately impacted as compared to the reference), and
• success_needed_full_parity (the number of additional successes needed in order to achieve full parity with the reference).

When di_prop_index < 1, then there are signs of disproportionate impact.

References
Examples

```r
library(dplyr)
data(student_equity)
di_prop_index(success=Transfer, group=Ethnicity, data=student_equity) %>%
as.data.frame
```

ppg_moe

Margin of error for the PPG

Description

Calculate the margin of error (MOE) for the percentage point gap (PPG) method.

Usage

```r
ppg_moe(n, proportion, min_moe = 0.03, prop_sub_0 = 0.5, prop_sub_1 = 0.5)
```

Arguments

- `n`: Sample size for the group of interest.
- `proportion`: (Optional) The proportion of successes for the group of interest. If specified, then the proportion is used in the MOE formula. Otherwise, a default proportion of 0.50 is used (conservative and yields the maximum MOE).
- `min_moe`: The minimum MOE returned even if the sample size is large. Defaults to 0.03. This equates to a minimum threshold gap for declaring disproportionate impact.
- `prop_sub_0`: For cases where 'proportion' is 0, substitute with prop_sub_0 (defaults to 0.5) to account for the zero MOE.
- `prop_sub_1`: For cases where 'proportion' is 1, substitute with prop_sub_1 (defaults to 0.5) to account for the zero MOE.

Value

The margin of error for the PPG given the specified sample size.

References


Examples

```r
ppg_moe(n=800)
ppg_moe(n=c(200, 800, 1000, 2000))
ppg_moe(n=800, proportion=0.20)
ppg_moe(n=800, proportion=0.20, min_moe=0)
ppg_moe(n=c(200, 800, 1000, 2000), min_moe=0.01)
```
Fake data on student equity

Description
Data randomly generated to illustrate the use of the package.

Usage
data(student_equity)

Format
A data frame with 20,000 rows:

- **Ethnicity**  ethnicity (one of: Asian, Black, Hispanic, Multi-Ethnicity, Native American, White).
- **Gender**  gender (one of: Male, Female, Other).
- **Cohort**  year student first enrolled in any credit course at the institution (one of: 2017, 2018).
- **Transfer**  1 or 0 indicating whether or not a student transferred within 2 years of first enrollment (Cohort).
- **Cohort_Math**  year student first enrolled in a math course at the institution; could be NA if the student have not attempted math.
- **Math**  1 or 0 indicating whether or not a student completed transfer-level math within 1 year of their first math attempt (Cohort_Math); could be NA if the student have not attempted math.
- **Cohort_English**  year student first enrolled in a math course at the institution; could be NA if the student have not attempted math.
- **English**  1 or 0 indicating whether or not a student completed transfer-level English within 1 year of their first math attempt (Cohort_English); could be NA if the student have not attempted English.
- **Ed_Goal**  student’s educational goal (one of: Deg/Transfer, Other).
- **College_Status**  student’s educational status (one of: First-time College, Other).
- **Student_ID**  student’s unique identifier.

Examples
data(student_equity)
Index

* datasets
  student_equity, 14
  di_80_index, 2, 6
  di_iterate, 3, 4, 8
  di_ppg, 5, 6, 7
  di_ppg_iterate, 10
  di_prop_index, 6, 11
  ppg_moe, 8, 13
  student_equity, 14