Package 'mutualinf'

October 28, 2021

Type Package

Title Computation and Decomposition of the Mutual Information Index

Version 1.1.2

Description The Mutual Information Index (M) introduced to social science literature by Theil and Finizza (1971) <doi:10.1080/0022250X.1971.9989795> is a multigroup segregation measure that is highly decomposable and that according to Frankel and Volij (2011) <doi:10.1016/j.jet.2010.10.008> and Mora and Ruiz-Castillo (2011) <doi:10.1111/j.1467-9531.2011.01237.x> satisfies the Strong Unit Decomposability and Strong Group Decomposability properties. This package allows computing and decomposing the total index value into its ``between" and ``within" terms. These last terms can also be decomposed into their contributions, either by group or unit characteristics. The factors that produce each ``within" term can also be displayed at the user's request. The results can be computed considering a variable or sets of variables that define separate clusters.

License GPL-3

Imports data.table, parallel, runner, stats

Encoding UTF-8

LazyData true

URL https://github.com/RafaelFuentealbaC/mutualinf

BugReports https://github.com/RafaelFuentealbaC/mutualinf/issues

Depends R (>= 2.10)

RoxygenNote 7.1.1

Collate 'Data_Source.R' 'get_general_contribution.R' 'get_proportion.R' 'get_internal_data.R' 'M_within.R' 'mutual.R' 'get_contribution.R' 'M_value.R' 'M.R' 'globals.R' 'prepare_data.R'

NeedsCompilation no

Author Rafael Fuentealba-Chaura [aut, cre],
Ricardo Mora [aut],
Julio Rojas-Mora [aut],

FONDECYT/ANID Project 11170583 [fnd],
MCIN/AEI/10.13039/501100011033 (Project no. PID2019-108576RB-I00) [fnd],
UCT VIP Project FEQUIP2019-INRN-03 [fnd]

Maintainer Rafael Fuentealba-Chaura <rfuentealba@inf.uct.cl>

Repository CRAN

Date/Publication 2021-10-28 18:00:01 UTC

R topics documented:

mutualinf-package 2
DF_Seg_Chile 3
DT_Seg_Chile 4
DT_test 5
mutual 6
prepare_data 8

Index **10**

`mutualinf-package` *An R package to compute and decompose the Mutual Information Index (M).*

Description

The Mutual Information Index (M) introduced to the social sciences by Theil and Finizza (1971). The M index is a multigroup segregation measure that is highly decomposable, satisfiying both the Strong Unit Decomposability (SUD) and the Strong Group Decomposability (SGD) properties (Frankel and Volij, 2011; Mora and Ruiz-Castillo, 2011).

The package allows for:

- The computation of the M index, either overall or over subsamples defined by the user.
- The decomposition of the M index into a "between" and a "within" term.
- The identification of the "exclusive contributions" of segregation sources defined either by group or unit characteristics.
- The computation of all the elements that conform the "within" term in the decomposition.
- Fast computation employing more than one CPU core in Mac, Linux, Unix, and BSD systems. This option uses the data.table and parallel libraries (which Windows does not permit to run with more than one CPU core).

Author(s)

Rafael Fuentealba-Chaura <rfuentealba@inf.uct.cl>
Ricardo Mora <ricmora@eco.uc3m.es>
Julio Rojas-Mora <jrojas@inf.uct.cl>

References

Frankel, D. and Volij, O. (2011). Measuring school segregation. *Journal of Economic Theory, 146*(1):1-38. doi: 10.1016/j.jet.2010.10.008.

Guinea-Martin, D., Mora, R., & Ruiz-Castillo, J. (2018). The evolution of gender segregation over the life course. *American Sociological Review, 83*(5), 983-1019. doi: 10.1177/0003122418794503.

Mora, R. and Guinea-Martin, D. (2021). Computing decomposable multigroup indexes of segregation. *UC3M Working papers, Economics 31803*, Universidad Carlos III de Madrid. Departamento de Economía.

Mora, R. and Ruiz-Castillo, J. (2011). Entropy-based segregation indices. *Sociological Methodology, 41*(1):159-194. doi: 10.1111/j.14679531.2011.01237.x.

Theil, H. and Finizza, A. J. (1971). A note on the measurement of racial integration of schools by means of informational concepts. *The Journal of Mathematical Sociology, 1*(2):187-193. doi: 10.1080/0022250X.1971.9989795.

`DF_Seg_Chile` *Segregation data in southern Chile*

Description

The data set included in this package was build using two data sets. The first one is the student enrollment reported by the Ministry of Education (MINEDUC, `https://datosabiertos.mineduc.cl/`) for students of primary education (first eight years of formal education) who attended establishments officially recognized by the State. The second one is the Quality and Context of Education Questionnaire for Parents and Guardians, and the Student Questionnaire, both applied by the Education Quality Agency (`https://www.agenciaeducacion.cl/`) to all students in grades 4 and 8 of primary education. Both sources are limited to the period 2016-2018. Contains information related to students and educational system characteristics in southern Chile (Biobio, La Araucania and Los Rios regions).

Usage

```
DF_Seg_Chile
```

Format

A `data.frame` with 191495 observations and 11 variables:

year Student enrollment year. From 2016 to 2018.

school School ID (RBD, Rol de Base de Datos).

district Administrative district where the school is located.

csep Preferential Scholar Subsidy Category (from the SpanishCategoría de Sub-vención Escolar Preferencial). Students belong to either the non-subsidized, the partially-subsidized, or the subsidized group acording to the Act 20.248 of Preferencial Scholar Subsidy (SEP).

ethnicity Self-reported Mapuche ethnicity. Students belong to Mapuche ethnicity or not.

rural School with multiage classrooms. The school is located in a urban zone or not.

region Administrative region where the school is located. Schools can belong either Biobio region, La Araucania region or Los Rios region.

sch_type Whether the school is public, charter, or private.

gender Student gender code. Students can either be female or male.

grade Student grade. Students can either belong to the 4th (4) or 8th (8) grade of basic school.

nobs Number of students in a cell or combination of variables.

Source

Ministry of Education (MINEDUC): https://datosabiertos.mineduc.cl/

Education Quality Agency: https://www.agenciaeducacion.cl/

`DT_Seg_Chile` *Segregation data in southern Chile*

Description

The data set included in this package was build using two data sets. The first one is the student enrollment reported by the Ministry of Education (MINEDUC, https://datosabiertos.mineduc.cl/) for students of primary education (first eight years of formal education) who attended establishments officially recognized by the State. The second one is the Quality and Context of Education Questionnaire for Parents and Guardians, and the Student Questionnaire, both applied by the Education Quality Agency (https://www.agenciaeducacion.cl/) to all students in grades 4 and 8 of primary education. Both sources are limited to the period 2016-2018. Contains information related to students and educational system characteristics in southern Chile (Biobio, La Araucania and Los Rios regions).

Usage

```
DT_Seg_Chile
```

Format

A `data.table` with 55960 observations and 11 variables:

year Student enrollment year. From 2016 to 2018.

school School ID (RBD, Rol de Base de Datos).

district Administrative district where the school is located.

csep Preferential Scholar Subsidy Category (from the SpanishCategoría de Sub-vención Escolar Preferencial). Students belong to either the non-subsidized, the partially-subsidized, or the subsidized group acording to the Act 20.248 of Preferencial Scholar Subsidy (SEP).

ethnicity Self-reported Mapuche ethnicity. Students belong to Mapuche ethnicity or not.

rural School with multiage classrooms. The school is located in a urban zone or not.

region Administrative region where the school is located. Schools can belong either Biobio region, La Araucania region or Los Rios region.

sch_type Whether the school is public, charter, or private.

gender Student gender code. Students can either be female or male.

grade Student grade. Students can either belong to the 4th (4) or 8th (8) grade of basic school.

nobs Number of students in a cell or combination of variables.

Source

Ministry of Education (MINEDUC): https://datosabiertos.mineduc.cl/

Education Quality Agency: https://www.agenciaeducacion.cl/

`DT_test` *Segregation data in southern Chile*

Description

The data set included in this package was build using two data sets. The first one is the student enrollment reported by the Ministry of Education (MINEDUC, https://datosabiertos.mineduc.cl/) for students of primary education (first eight years of formal education) who attended establishments officially recognized by the State. The second one is the Quality and Context of Education Questionnaire for Parents and Guardians, and the Student Questionnaire, both applied by the Education Quality Agency (https://www.agenciaeducacion.cl/) to all students in grades 4 and 8 of primary education. Both sources are limited to 2018. Contains information related to students and educational system characteristics in southern Chile (Biobio, La Araucania and Los Rios regions).

Usage

```
DT_test
```

Format

A `data.table` with 6703 observations and 5 variables, only for testing pourposes:

school School ID (RBD, Rol de Base de Datos).

csep Preferential Scholar Subsidy Category (from the SpanishCategoría de Sub-vención Escolar Preferencial). Students belong to either the non-subsidized, the partially-subsidized, or the subsidized group acording to the Act 20.248 of Preferencial Scholar Subsidy (SEP).

ethnicity Self-reported Mapuche ethnicity. Students belong to Mapuche ethnicity or not.

region Administrative region where the school is located. Schools can belong either Biobio region, La Araucania region or Los Rios region.

fw Number of students in a cell or combination of variables.

Source

Ministry of Education (MINEDUC): https://datosabiertos.mineduc.cl/

Education Quality Agency: https://www.agenciaeducacion.cl/

`mutual` *Computes and decomposes the Mutual Information index*

Description

Computes and decomposes the Mutual Information index into "between" and "within" terms. The "within" terms can also be decomposed into "exclusive contributions" of segregation sources defined either by group or unit characteristics. The mathematical components required to compute each "within" term can also be displayed at the user's request. The results can be computed over subsamples defined by the user.

Usage

```
mutual(
  data,
  group,
  unit,
  within = NULL,
  by = NULL,
  contribution.from = NULL,
  components = FALSE,
  cores = NULL
)
```

Arguments

`data`	An object from the "data.table" and "mutual.data" classes.
`group`	A categorical variable name or vector of categorical variables names contained in `data`, or also, a column number or vector of column numbers of `data`. Defines the first dimension over which segregation is computed.
`unit`	A categorical variable name or vector of categorical variables names contained in `data`, or also, a column number or vector of column numbers of `data`. Defines the second dimension over which segregation is computed.
`within`	A categorical variable name or vector of categorical variables names contained in `data`, or also, a column number or vector of column numbers of `data`. Defines the partitions to compute the between and within decompositions. By default is `NULL`.
`by`	A categorical variable name or vector of categorical variables names contained in `data`, or also, a column number or vector of column numbers of `data`. Defines the subsamples over which indexes are computed. By default is `NULL`.

`contribution.from`
: A variable of character type that can be 'group_vars' or 'unit_vars', or also, a categorical variable name or vector of categorical variables names contained in the `group` parameter or `unit` parameter, or also, a column number or vector of column numbers in the `group` parameter or the `unit` parameter. Defines the segregation sources whose exclusive contributions to the "within" terms and the overall index are computed. By default is `NULL`.

`components`
: A boolean value. If TRUE and the `within` option is not `NULL` and the by option is `NULL`, then it returns a list where the first element is a `data.table` that contains a summary of the index total value and its decompositions, while the second element is a `data.table` with more detailed information of the decomposition of the "within" term (the mathematical components required to compute the within terms). If the `within` and by options are not `NULL`, then the function returns a list of lists where each first element is a `data.table` that contains the summary of the index total value and decompositions in a given subsample, while each second element is a `data.table` with more detailed information of the decomposition of the within term displayed in each first element in the same subsample. By default is FALSE.

`cores`
: A positive integer. Defines the amount of CPU cores to use in parallelization tasks. If `NULL`, then the computation is carried out in only one core. This option is available to Mac, Linux, Unix, and BSD systems but is not available to Windows sytems. By default is `NULL`.

Details

Mixing `group` variables with `unit` variables in `contribution.from` will produce an error.

Value

A `data.table` if the `components` option is `FALSE`; a list if the `components` option is `TRUE`, the `within` option is not `NULL` and the by option is `NULL`; or a list of lists if the `components` option is `TRUE`, and both `within` and by options are not `NULL`.

References

Frankel, D. and Volij, O. (2011). Measuring school segregation. *Journal of Economic Theory, 146*(1):1-38. doi: 10.1016/j.jet.2010.10.008.

Guinea-Martin, D., Mora, R., & Ruiz-Castillo, J. (2018). The evolution of gender segregation over the life course. *American Sociological Review, 83*(5), 983-1019. doi: 10.1177/0003122418794503.

Mora, R. and Guinea-Martin, D. (2021). Computing decomposable multigroup indexes of segregation. *UC3M Working papers, Economics 31803*. Universidad Carlos III de Madrid. Departamento de Economía.

Mora, R. and Ruiz-Castillo, J. (2011). Entropy-based segregation indices. *Sociological Methodology, 41*(1):159-194. doi: 10.1111/j.14679531.2011.01237.x.

Theil, H. and Finizza, A. J. (1971). A note on the measurement of racial integration of schools by means of informational concepts. *The Journal of Mathematical Sociology, 1*(2):187-193. doi: 10.1080/0022250X.1971.9989795.

Examples

```
# To compute the overall measure of school segregation by socioeconomic and ethnic status.
mutual(data = DT_test, group = c("csep", "ethnicity"), unit = "school")

# Computation of the exclusive effect of specific segregation sources on the overall measure, e.g.,
# socioeconomic and ethnic contributions, and the contribution that cannot be attributed to any of
# them (the "interaction" term).
mutual(data = DT_test, group = c("csep", "ethnicity"), unit = "school", by = "region",
contribution.from = "group_vars")

# For more information on the package, refer to the manual and the README file.
```

`prepare_data` *Prepares the data to be used by the* `mutual` *function*

Description

Receives the data that is later used in the `mutual` function. Generates a `data.table` with the entry variables.

Usage

```
prepare_data(data, vars, fw = NULL, col.order = NULL)
```

Arguments

`data`	A tabular format object (`data.frame`, `data.table`, `tibble`). The data expected is microdata or frequency weight data for each combination of variables. The variables must be of "factor" class.
`vars`	A vector of variable names or vector of columns numbers contained in `data`. Also can be used "all_vars" to select all variables contained in `data`.
`fw`	Variable name or column number contained in `data` that contains frecuency weight for each combination of variables of the dataset. If this variable exists then the function will change its original name to `fw`. If this variable does not exist or is `NULL`, then the function will compute the frecuency weight given the combination of variables of `vars` and will create a new variable called `fw`. By default is `NULL`.
`col.order`	A variable name or vector of variables names contained in `vars`, or a column number or vector of column numbers contained in `vars`. Selects the columns to sort the dataset. By default is `NULL`.

Value

Returns a `data.table` of class "data.table" "data.frame" "mutual.data".

Examples

```
## Not run:
# Using some variable names in 'data' with explicit 'fw'.
my_data <- prepare_data(data = DF_Seg_Chile, vars = c("csep", "ethnicity", "school", "district"),
fw = "nobs")

# Using some column numbers in 'data' and explicit 'fw' as another column number.
my_data <- prepare_data(data = DF_Seg_Chile, vars = c(4, 5, 2, 3), fw = 11)

# Using all variables of 'data' with explicit 'fw'.
my_data <- prepare_data(data = DF_Seg_Chile, vars = "all_vars", fw = "nobs")

# Using some variable names in 'data' and 'fw' does not exist (in this case, the new 'fw' will
# be equal to 1 for all variable combinations as 'data' already has a frequency weights variable)
my_data <- prepare_data(data = DF_Seg_Chile, vars = c("csep", "ethnicity", "school", "district"))

# Using the 'col.order' option to sort data according to the 'csep' column.
my_data <- prepare_data(data = DF_Seg_Chile, vars = c("csep", "ethnicity", "school", "district"),
fw = "nobs", col.order = "csep")

# The class of the resulting object in all cases must be "data.table", "data.frame" and
# "mutual.data".
class(my_data)

## End(Not run)
```

Index

∗ **datasets**
    DF_Seg_Chile, 3
    DT_Seg_Chile, 4
    DT_test, 5
∗ **package**
    mutualinf-package, 2

DF_Seg_Chile, 3
DT_Seg_Chile, 4
DT_test, 5

mutual, 6
mutualinf-package, 2

prepare_data, 8