Package 'CopyDetect'

October 8, 2018

Version 1.3

Date 2018-09-25

Title Computing Response Similarity Indices for Multiple-Choice Tests

Author Cengiz Zopluoglu

Maintainer Cengiz Zopluoglu <c.zopluoglu@miami.edu>

Depends mirt

Description Contains several IRT and non-IRT based response similarity indices proposed in the literature for multiple-choice examinations such as the Omega index, Wollack (1997) <doi:10.1177/01466216970214002>; Generalized Binomial Test, van der Linden & Sotaridona (2006) <doi:10.3102/10769986031003283>; K index, K1 and K2 indices, Sotaridona & Meijer (2002) <doi:10.1111/j.1745-3984.2002.tb01138.x>; and S1 and S2 indices, Sotaridona & Meijer (2003) <doi:10.1111/j.1745-3984.2003.tb01096.x>.

License GPL-3 | file LICENSE

URL http://sites.education.miami.edu/zopluoglu/

NeedsCompilation no

Repository CRAN

Date/Publication 2018-10-08 18:30:09 UTC

R topics documented:

form1 *Form 1 of a real credentialing dataset*

Description

This is the Form 1 of a real credentialing dataset provided by Cizek and Wollack (2017). It has 1,636 test takers and their response to 170 items. The dataset also includes the unique individual IDs in the first column and the unique center IDs in the second column. Column 3 to Column 172 includes the dichotomous item responses to 170 items.

Usage

```
data(form1)
```

Format

A data frame with 1,636 rows and 172 columns.

References

Cizek, G. J., & Wollack, J. A. (Eds.). (2017). Handbook of quantitative methods for detecting cheating on tests. New York, NY: Routledge.

form2 *Form 2 of a real credentialing dataset*

Description

This is the Form 2 of a real credentialing dataset provided by Cizek and Wollack (2017). It has 1,607 test takers and their response to 170 items. The dataset also includes the unique individual IDs in the first column and the unique center IDs in the second column. Column 3 to Column 172 includes the nominal item responses to 170 items.

Usage

```
data(form2)
```

Format

A data frame with 1,607 rows and 172 columns.

References

Cizek, G. J., & Wollack, J. A. (Eds.). (2017). Handbook of quantitative methods for detecting cheating on tests. New York, NY: Routledge.

similarity1 *Response Similarity for Dcihotomously Scored Items*

Description

Computes the response similarity indices such as the Omega index (Wollack, 1996), Generalized Binomial Test ([GBT], van der Linden & Sotaridona (2006), K index (Holland, 1996), K1 and K2 indices (Sotaridona & Meijer, 2002), and S1 and S2 indices (Sotaridona & Meijer, 2003), and the M4 index (Maynes, 2014).

Usage

```
similarity1(data, item.par=NULL, model="1PL", prior = TRUE, person.id, center.id=NULL,
item.loc, single.pair=NULL, many.pairs=NULL, centers=NULL)
```

Arguments

data	a data frame with *N* rows and *n* columns. The data must include at least one column for unique individual IDs and item responses. All items should be scored dichotomously, with 0 indicating an incorrect response and 1 indicating a correct response. All columns must be "numeric". Missing values (NA) are allowed. Please see the details below for the treatment of missing data in the analysis.
item.par	a data matrix with *n* rows and three columns, where *n* denotes the number of items. The first, second, and third columns representitem discrimination, item difficulty, and item guessing parameters, respectively. If item parameters are not provided by user, the **mirt** package is internally called to estimate the parameeters of a chosen IRT model. The rows in the item parameter matrix must be in the same order as the columns in the response data.
model	IRT model to be used for computing IRT-based indices (omega, GBT, and M4). The available options are "1PL", "2PL", and "3PL". Default is "1PL".
prior	a logical argument if the model is equal 3PL. If TRUE, then a prior distribution is specified for the guessing parameter when fitting the 3PL model. Otherwise, this argument is ignored.
person.id	column label in the dataset for the variable indicating unique person ids.
center.id	column label in the dataset for the variable indicating center ids. This is used if the user asks computing the indices for all pairs in some centers. For a single pair or multiple pair calculation, this argument is ignored.
item.loc	a numeric vector indicating the location of item response in the dataset.
single.pair	a character vector of length 2 for the suspected pair of examinees. The first element of the vector indicates the unique ID of the suspected copier examinee, and the second element of the vector indicates the unique ID of the suspected source examinee.

many.pairs	a matrix with two columns. Users can request to compute the indices simultaneously as many pairs as they wish. Each row of this matrix represents a pair of examinees. The first column of each row indicates the unique ID of the suspected copier examinee, and the second column of each row indicates the unique ID of the suspected source examinee.
centers	a character vector including unique center ids. Users can request to compute the indices for all pairs in the centers listed using this argument.

Details

Test fraud has recently been receiving increased attention in the field of educational testing. The current R package provides a set of useful statistical indices recently proposed in the literature for detecting a specific type of test fraud - answer copying from a nearby examinee on multiple-choice examinations. The information obtained from these procedures may provide additional statistical evidence of answer copying, but they should be used cautiously. These statistical procedures should not be used as sole evidence of answer copying, especially when used for general screening purposes.

There are more than twenty different statistical procedures recommended in the literature for detecting answer copying on multiple-choice examinations. However, the CopyDetect package includes the indices that have been shown as effective and reliable based on the simulation studies in the literature (Sotaridona & Meijer, 2002, 2003; van der Linden & Sotaridona, 2006; Wollack, 1996, 2003, 2006; Wollack & Cohen, 1998; Maynes, 2014). Among these indices, ω, GBT, and M4 use IRT models, and K and K variants are the non-IRT counterparts.

Since `similarity1` uses dichotomous responses as input, any (0,0) response combination between two response vectors is counted as an "identical incorrect response", and any (1,1) response combination between two response vectors is counted as an "identical correct response". `similarity1` also counts any (NA,NA) response combination between two response vectors as an "identical incorrect response". Other response combinations such as (0,1),(1,0),(0,NA),(1,NA) between two response vectors are not counted as identical responses. When computing the number-correct/number-incorrect scores or estimating the IRT ability parameters, missing values (NA) are counted as an incorrect response.

Value

If a single-pair is requested, `similarity1()` returns a list containing the following components.

data	original data file provided by user
W.index	statistics for the W index
GBT.index	statistics for the GBT index
K.index	statistics for the K index
K.variants	statistics for the K1, K2, S1, and S2 indices
M4.index	statistics for the M4 index
item.loc	columns in the dataset that stores the item responses

item.par	estimated item parameter matrix for a chosen IRT model
center.id	column label for the unique center IDs
person.id	column label for the unique person IDs
single.pair	Unique IDs for a pair of examinee requested
single.pair2	corresponding row numbers for the pair of examinee requested
thetas	maximum likelihood estimates for IRT ability parameter for each individual

If a multiple pairs are requested in the matrix form, similarity1() returns a list containing the following components.

data	original data file provided by user
output.manypairs	a matrix including the IDs for each pair requested and corresponding indices computed for these pairs
item.loc	columns in the dataset that stores the item responses
item.par	estimated item parameter matrix for a chosen IRT model
center.id	column label for the unique center IDs
person.id	column label for the unique person IDs
thetas	maximum likelihood estimates for IRT ability parameter for each individual

If all possible pairs in certain test centers are requested in the matrix form, similarity1() returns a list containing the following components.

data	original data file provided by user
output.centers	a matrix including the IDs for all pairs requested and corresponding indices computed for these pairs
item.loc	columns in the dataset that stores the item responses
item.par	estimated item parameter matrix for a chosen IRT model
center.id	column label for the unique center IDs
person.id	column label for the unique person IDs
thetas	maximum likelihood estimates for IRT ability parameter for each individual

Note

* A recursive algorithm to compute the compound binomial probability distribution required for the GBT index is partially adapted from an S-plus code provided by Dr. Leonardo Sotaridona. The author acknowledges his contribution and permission.

* The indices in the package rely on a sample of sufficient size to estimate CTT- or IRT-based parameters with enough precision for computational procedures. Users should be careful when using these indices with small samples such as those containing fewer than 100 examinees.

Author(s)

Cengiz Zopluoglu

References

Sotaridona, L.S., & Meijer, R.R.(2002). Statistical properties of the K-index for detecting answer copying. *Journal of Educational Measurement, 39*, 115-132.

Sotaridona, L.S., & Meijer, R.R.(2003). Two new statistics to detect answer copying. *Journal of Educational Measurement, 40*, 53-69.

van der Linden, W.J., & Sotaridona, L.S.(2006). Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioral Statistics, 31*, 283-304.

Wollack, J.A.(1996). Detection of answer copying using item response theory. *Dissertation Abstracts International, 57/05*, 2015.

Wollack, J.A.(2003). Comparison of answer copying indices with real data. *Journal of Educational Measurement, 40*, 189-205.

Wollack, J.A.(2006). Simultaneous use of multiple answer copying indexes to improve detection rates. *Applied Measurement in Education, 19*, 265-288.

Wollack, J.A., & Cohen, A.S.(1998). Detection of answer copying with unknown item and trait parameters. *Applied Psychological Measurement, 22*, 144-152.

Examples

```
data(form1)
dim(form1)
head(form1)

  # the first column of this dataset is unique individual IDs
  # the second column of this dataset is unique center IDs
  # From Column 3 to Column 172, dichotomous item responses


# For the sake of reducing the computational time,
# I will analyze a subset of this dataset (first 20 items)

 subset <- form1[1:1000,1:22]

 dim(subset)
 head(subset)


# Computing similarity for a single pair

a <- similarity1(data       = subset,
```

```
                  model      = "1PL",
                  person.id  = "EID",
                  item.loc   = 3:22,
                  single.pair= c("e100287","e100869"))

print(a)



# Computing for multiple pairs

pairs <- matrix(as.character(sample(subset$EID,20)),nrow=10,ncol=2)

a <- similarity1(data       = subset,
                  model      = "1PL",
                  person.id  = "EID",
                  item.loc   = 3:20,
                  many.pairs = pairs)

print(a)


# Computing all possible pairs in the requested centers

a <- similarity1(data       = subset,
                  model      = "1PL",
                  person.id  = "EID",
                  center.id  = "cent_id",
                  item.loc   = 3:20,
                  centers    = c(1802,5130,67,9056))

print(a)
```

`similarity2` *Response Similarity Indices for Nominal Response Items*

Description

Computes the response similarity indices such as the Omega index (Wollack, 1996), Generalized Binomial Test ([GBT], van der Linden & Sotaridona (2006), K index (Holland, 1996), K1 and K2 indices (Sotaridona & Meijer, 2002), and S1 and S2 indices (Sotaridona & Meijer, 2003), and the M4 index (Maynes, 2014).

Usage

```
similarity2(data,resp.options,key, person.id, center.id=NULL, item.loc,
single.pair=NULL, many.pairs=NULL, centers=NULL)
```

Arguments

`data`	a data frame with N rows and n columns. The data must include at least one column for unique individual IDs and item responses. All items should be scored dichotomously, with 0 indicating an incorrect response and 1 indicating a correct response. All columns must be "numeric". Missing values (NA) are allowed. Please see the details below for the treatment of missing data in the analysis.
`resp.options`	a vector of labels for the nominal response options in the dataset.
`key`	a vector of key response options for the items in the dataset. The order of key responses should be the same as the order of columns that stores item responses given in the item.loc argument.
`person.id`	column label in the dataset for the variable indicating unique person ids.
`center.id`	column label in the dataset for the variable indicating center ids. This is used if the user asks computing the indices for all pairs in some centers. For a single pair or multiple pair calculation, this argument is ignored.
`item.loc`	a numeric vector indicating the location of item response in the dataset.
`single.pair`	a character vector of length 2 for the suspected pair of examinees. The first element of the vector indicates the unique ID of the suspected copier examinee, and the second element of the vector indicates the unique ID of the suspected source examinee.
`many.pairs`	a matrix with two columns. Users can request to compute the indices simultaneously as many pairs as they wish. Each row of this matrix represents a pair of examinees. The first column of each row indicates the unique ID of the suspected copier examinee, and the second column of each row indicates the unique ID of the suspected source examinee.
`centers`	a character vector including unique center ids. Users can request to compute the indices for all pairs in the centers listed using this argument.

Details

`similarity2` uses nominally scored items. Therefore, the definition of "identical incorrect response" and "identical correct response" is slightly different from `similarity1`. For example, let A, B, C, and D be the response alternatives for items in a multiple-choice test, and let A be the key response for an item. There are 10 possible response combinations between two response vectors: (A,A), (A,B), (A,C), (A,D), (B,B), (B,C), (B,D), (C,C), (C,D), and (D,D). `similarity2` counts the (A,A) response combination as an "identical correct response", and any of the (B,B), (C,C), and (D,D) response combinations as an "identical incorrect response". Similar to `similarity1`, the (NA,NA) response combination is counted as an "identical incorrect response". All other response combinations (A,B), (A,C), (A,D), (B,C), (B,D), (C,D), (A,NA), (B,NA), (C,NA), and (D,NA) are counted as non-identical responses. When computing the number-correct/number-incorrect scores or estimating the IRT ability parameters, missing values (NA) in a response vector are counted as an incorrect response.

Value

If a single-pair is requested, `similarity1()` returns a list containing the following components.

data	original data file provided by user
scored.data	scored item response matrix based on the key response vector provided
W.index	statistics for the W index
GBT.index	statistics for the GBT index
K.index	statistics for the K index
K.variants	statistics for the K1, K2, S1, and S2 indices
M4.index	statistics for the M4 index
item.loc	columns in the dataset that stores the item responses
item.par	estimated item parameter matrix for the Bock's Nominal Response IRT Model
center.id	column label for the unique center IDs
person.id	column label for the unique person IDs
single.pair	Unique IDs for a pair of examinee requested
single.pair2	corresponding row numbers for the pair of examinee requested
thetas	maximum likelihood estimates for IRT ability parameter for each individual

If a multiple pairs are requested in the matrix form, `similarity1()` returns a list containing the following components.

data	original data file provided by user
output.manypairs	a matrix including the IDs for each pair requested and corresponding indices computed for these pairs
item.loc	columns in the dataset that stores the item responses
item.par	estimated item parameter matrix for a chosen IRT model
center.id	column label for the unique center IDs
person.id	column label for the unique person IDs
thetas	maximum likelihood estimates for IRT ability parameter for each individual

If all possible pairs in certain test centers are requested in the matrix form, `similarity1()` returns a list containing the following components.

data	original data file provided by user
output.centers	a matrix including the IDs for all pairs requested and corresponding indices computed for these pairs
item.loc	columns in the dataset that stores the item responses
item.par	estimated item parameter matrix for a chosen IRT model
center.id	column label for the unique center IDs
person.id	column label for the unique person IDs
thetas	maximum likelihood estimates for IRT ability parameter for each individual

Author(s)

Cengiz Zopluoglu

References

Sotaridona, L.S., & Meijer, R.R.(2002). Statistical properties of the K-index for detecting answer copying. *Journal of Educational Measurement, 39*, 115-132.

Sotaridona, L.S., & Meijer, R.R.(2003). Two new statistics to detect answer copying. *Journal of Educational Measurement, 40*, 53-69.

van der Linden, W.J., & Sotaridona, L.S.(2006). Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioral Statistics, 31*, 283-304.

Wollack, J.A.(1996). Detection of answer copying using item response theory. *Dissertation Abstracts International, 57/05*, 2015.

Wollack, J.A.(2003). Comparison of answer copying indices with real data. *Journal of Educational Measurement, 40*, 189-205.

Wollack, J.A.(2006). Simultaneous use of multiple answer copying indexes to improve detection rates. *Applied Measurement in Education, 19*, 265-288.

Wollack, J.A., & Cohen, A.S.(1998). Detection of answer copying with unknown item and trait parameters. *Applied Psychological Measurement, 22*, 144-152.

Examples

```
data(form2)
dim(form2)
head(form2)

  # the first column of this dataset is unique individual IDs
  # the second column of this dataset is unique center IDs
 # From Column 3 to Column 172, nominal item responses. 1, 2, 3, and 4 represent different
  # nominal response options.


# For the sake of reducing the computational time,
# I will analyze a subset of this dataset (first 10 items)

 subset <- form2[1:500,1:12]

 dim(subset)
 head(subset)


# Computing similarity for a single pair
```

```
key.resp <- c(2,3,1,4,1,2,2,1,1,1)

a <- similarity2(data         = subset,
                 resp.options = c(1,2,3,4),
                 key          = key.resp,
                 person.id    = "EID",
                 item.loc     = 3:12,
                 single.pair  = c("e200287","e200169"))

print(a)



# Computing for multiple pairs

pairs <- matrix(as.character(sample(subset$EID,20)),nrow=10,ncol=2)

a <- similarity2(data         = subset,
                 resp.options = c(1,2,3,4),
                 key          = key.resp,
                 person.id    = "EID",
                 item.loc     = 3:12,
                 many.pairs   = pairs)

print(a)


# Computing all possible pairs in the requested centers

a <- similarity2(data         = subset,
                 resp.options = c(1,2,3,4),
                 key          = key.resp,
                 person.id    = "EID",
                 center.id    = "cent_id",
                 item.loc     = 3:12,
                 centers      = c(42,45,4114))

print(a)


# Key response vector for all 170 items for future reference

key.resp <- c(2,3,1,4,1,2,2,1,1,1,4,1,3,1,3,3,1,2,1,3,3,4,1,
              3,3,2,3,2,2,3,1,4,1,2,3,3,2,3,4,1,2,1,1,4,3,3,
              1,1,4,2,2,1,4,1,2,3,3,1,2,4,1,4,2,4,1,1,2,3,4,
              4,1,4,2,1,2,2,2,2,4,4,3,2,1,3,2,3,2,2,1,2,4,3,
              2,1,2,1,2,3,1,1,4,3,4,3,4,3,1,3,3,4,2,1,1,4,3,
              2,4,4,1,1,1,2,2,1,3,1,2,3,3,3,4,4,1,4,4,3,4,2,
              3,1,4,1,4,1,3,2,2,4,4,4,1,2,2,3,4,1,2,1,4,4,4,
              1,3,1,2,1,2,3,2,2)
```

Index