Package 'dcmdata'

August 20, 2025

Title Data Sets for Diagnostic Classification Modeling

Version 0.1.0

Description Access data sets for demonstrating or testing diagnostic classification models. Simulated data sets can be used to compare estimated model output to true data-generating values. Real data sets can be used to demonstrate real-world applications of diagnostic models.

License MIT + file LICENSE

URL https://dcmdata.r-dcm.org, https://github.com/r-dcm/dcmdata

BugReports https://github.com/r-dcm/dcmdata/issues

Depends R (>= 2.10)

Imports cli, rlang (>= 1.1.0), tibble

Config/testthat/edition 3

Config/Needs/website r-dcm/rdcmtemplate

Config/Needs/documentation openpharma/roxylint

Config/roxylint list(linters = roxylint::tidy)

Encoding UTF-8

Language en-US

LazyData true

RoxygenNote 7.3.2

Suggests spelling, testthat (>= 3.0.0)

NeedsCompilation no

Author W. Jake Thompson [aut, cre] (ORCID:
<https://orcid.org/0000-0001-7339-0300>),
University of Kansas [cph],
Institute of Education Sciences [fnd]

Maintainer W. Jake Thompson <wjakethompson@gmail.com>

Repository CRAN

Date/Publication 2025-08-20 17:30:02 UTC

Contents

`dtmr_data` *Diagnosing teachers' multiplicative reasoning (DTMR)*

Description

This is a simulated data set modeled after the DTMR study described by Bradshaw et al. (2014) and Izsák et al. (2019). The data was simulated from the loglinear cognitive diagnostic model (LCDM), which is the model that was used to analyze the data in the referenced articles. The data set consists of 990 responses to the 27 items included in the final version of the DTMR data, matching the sample that was collected by the original authors. Each respondent was randomly assigned a mastery profile using the profile proportions reported in Figure 10 of Izsák et al. (2019). Item responses were then generated for each respondent using their assigned mastery profile and the item parameters reported in Table 1 of Bradshaw et al. (2014). Reproducible code for generating the simulated data is available in the GitHub repository for this package.

Usage

```
dtmr_data

dtmr_qmatrix

dtmr_true_structural

dtmr_true_profiles

dtmr_true_items
```

Format

`dtmr_data` is a tibble containing simulated DTMR response data with 990 rows and 28 variables.

- `id`: Respondent identifier.
- `1-22`: Simulated dichotomous item responses to the 27 DTMR items.

`dtmr_qmatrix` is a tibble that identifies which skills are measured by each DTMR item, as reported in Bradshaw et al. (2014). The DTMR assessment contains 27 items measuring 4 skills. The `dtmr_qmatrix` correspondingly is made up of 27 rows and 5 variables.

- `item`: Item identifier, corresponds to `1-22` in `dtmr_data`.

- `referent_units`, `partitioning_iterating`, `appropriateness`, and `multiplicative_comparison`: Dichotomous indicator for whether or not the skill is measured by each item. A value of `1` indicates the skill is measured by the item and a value of `0` indicates the skill is not measured by the item.

Simulation values:

In addition to the simulated data sets, the true values used to simulate the data are included for reference. This may be useful if, for example, you want to estimate a model and then check how well the estimated parameters match values that were used to create the data.

To simulate the data, we first need `dtmr_true_structural`. This is a tibble that contains the structural parameters reported in Figure 10 of Izsák et al. (2019). The structural parameters define the probability of observing each possible profile in the population of respondents. Each row represents one possible mastery profile. Therefore, there are 16 rows and 5 variables.

- `referent_units`, `partitioning_iterating`, `appropriateness`, `multiplicative_comparison`: Integer values indicating whether each attribute has been mastered by respondents with the given profile.
- `class_probability`: The proportion of respondents estimated to demonstrate the given pattern of mastery.

Using the `dtmr_true_structural` values, we randomly sampled a mastery profile for each of the 990 respondents. The true profiles for each respondent are available in `dtmr_true_profiles`. There are a total of 990 rows and 5 variables.

- `id`: Respondent identifier, corresponds to `id` in `dtmr_data`.
- `referent_units`, `partitioning_iterating`, `appropriateness`, `multiplicative_comparison`: Integer values indicating whether each attribute has been mastered by the respondent.

We use the `dtmr_true_profiles` and the `dtmr_qmatrix` to identify whether each respondent possess the attributes required by each item. Based on which attributes are required and possessed, we use the `dtmr_true_items` to calculate the log odds of each respondent providing a correct response to each item. `dtmr_true_items` contains the estimated item parameters reported in Table 1 of Bradshaw et al. (2014). This a tibble with 27 rows and 7 columns.

- `item`: Item identifier, corresponds to `1-22` in `dtmr_data`.
- `intercept`: The LCDM intercept parameter for each item.
- `referent_units`: The LCDM main effect parameter for items measuring the referent units attribute.
- `partitioning_iterating`: The LCDM main effect parameter for items measuring the partitioning and iterating attribute.
- `appropriateness`: The LCDM main effect parameter for items measuring the appropriateness attribute.
- `multiplicative_comparison`: The LCDM main effect parameter for items measuring the multiplicative comparisons attribute.
- `referent_units__partitioning_iterating`: The LCDM interaction parameter for items measuring both referent units and partitioning and iterating attributes.

Finally, we convert the log odds values to probabilities and draw a random Bernoulli variable using the probabilities of a correct response. The drawn Bernoulli values are the simulated item scores that make up the `dtmr_data`.

Details

The skills correspond to knowledge of:

1. Referent units
2. Partitioning and iterating
3. Appropriateness
4. Multiplicative comparisons

References

Bradshaw, L., Izsák, A., Templin, J., & Jacobson, E. (2014). Diagnosing teachers' understandings of rational numbers: Building a multidimensional test within the diagnostic classification framework. *Educational Measurement: Issues and Practice, 33*(1), 2-14. doi:10.1111/emip.12020

Izsák, A., Jacobson, E., & Bradshaw, L. (2019). Surveying middle-grades teachers' reasoning about fraction arithmetic in terms of measured quantities. *Journal for Research in Mathematics Education, 50*(2), 156-209. doi:10.5951/jresematheduc.50.2.0156

`ecpe_data` *Examination for the certificate of proficiency in English (ECPE)*

Description

This is data from the grammar section of the ECPE, administered annually by the English Language Institute at the University of Michigan. This data contains responses to 28 questions from 2,922 respondents, which ask respondents to complete a sentence with the correct word. This data set has been used by Templin & Hoffman (2013) and Templin & Bradshaw (2014) for demonstrating the log-linear cognitive diagnosis model (LCDM) and the hierarchical diagnostic classification model (HDCM), respectively.

Usage

```
ecpe_data

ecpe_qmatrix
```

Format

`ecpe_data` is a tibble containing ECPE response data with 2,922 rows and 29 variables.

- `resp_id`: Respondent identifier.
- `E1-E28`: Dichotomous item responses to the 28 ECPE items.

`ecpe_qmatrix` is a tibble that identifies which skills are measured by each ECPE item. This section of the ECPE contains 28 items measuring 3 skills. The `ecpe_qmatrix` correspondingly is made up of 28 rows and 4 variables.

- `item_id`: Item identifier, corresponds to `E1-E28` in `ecpe_data`.

- `morphosyntactic`, `cohesive`, and `lexical`: Dichotomous indicator for whether or not the skill is measured by each item. A value of `1` indicates the skill is measured by the item and a value of `0` indicates the skill is not measured by the item.

Details

The skills correspond to knowledge of:

1. Morphosyntactic rules
2. Cohesive rules
3. Lexical rules

For more details, see Buck & Tatsuoka (1998) and Henson & Templin (2007).

References

Buck, G., & Tatsuoka, K. K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing, 15*(2), 119-157. doi:10.1177/026553229801500201

Henson, R., & Templin, J. (2007, April). *Large-scale language assessment using cognitive diagnosis models.* Paper presented at the Annual meeting of the National Council on Measurement in Education, Chicago, IL.

Templin, J., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement: Issues and Practice, 32*(2), 37-50. doi:10.1111/emip.12010

Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika, 79*(2), 317-339. doi:10.1007/s1133601393620

`generate_ids` *Generate unique identifiers*

Description

Create unique alphanumeric identifiers with a specified character length and proportions of alpha and numeric characters.

Usage

```
generate_ids(n, characters, prop_numeric = 1, n_attempt = n * 3)
```

Arguments

n	The number of unique identifiers to generate.
characters	The number of characters to be included in each identifier.
prop_numeric	The proportion of characters that should be numeric. The default is 1 (i.e., all numbers). If less than 1, identifiers will also include lowercase and uppercase letters.
n_attempt	The number of allowed attempts for generating the requested number of identifiers. See details for more information.

Details

When identifiers are long (e.g., characters >= 10), it is slow and computationally intensive to find all possible permutations of the specified number of alpha and numeric characters. Therefore, identifiers are generated one at a time by sampling the required number of characters. This greatly increases efficiency, as we don't waste time generating multiple millions of identifiers when we might only need a few hundred. However, this means that it is possible we could generate duplicate identifiers. The n_attempt argument allows us to control how many identifiers we can generate in order to achieve our desired n unique identifiers. If we fail to find n unique identifiers after n_attempt, the function will error. For example, consider a request for 1,000 identifiers, each with 2 characters and only using numbers. With the number 0-9, there are only 100 possible two-character permutations. Thus, after n_attempt, the function will fail as 1,000 unique identifiers cannot be found.

Value

A factor vector of length n.

Examples

```
generate_ids(n = 10, characters = 5)
generate_ids(n = 100, characters = 10, prop_numeric = 0.5)
```

---

log_odds *Log-odds transformation*

Description

These functions implement the log-odds (or logit) transformation. This is a common transformation for psychometric models that is used to put probabilities on a continuous scale.

Usage

```
logit(x)

inv_logit(x)
```

Arguments

x	A number to be transformed.

Value

A transformed double.

Examples

```
logit(0.6)
logit(0.5)

inv_logit(3.5)
inv_logit(0)
```

---

mdm_data *MacReady & Dayton multiplication data (MDM)*

Description

This is a small data set of multiplication item responses. This data contains responses to 4 items from 142 respondents, which ask respondents to complete an integer multiplication problem.

Usage

```
mdm_data

mdm_qmatrix
```

Format

`mdm_data` is a tibble containing responses to multiplication items, as described in MacReady and Dayton (1977). There are 142 rows and 5 variables.

- `respondent`: Respondent identifier.
- `mdm1-mdm4`: Dichotomous item responses to the 4 multiplication items.

`mdm_qmatrix` is a tibble that identifies which skills are measured by each MDM item. This MDM data contains 4 items, all of which measure the skill of multiplication. The `mdm_qmatrix` correspondingly is made up of 4 rows and 2 variables.

- `item`: Item identifier, corresponds to `mdm1-mdm4` in `mdm_data`.
- `multiplication`: Dichotomous indicator for whether or not the multiplication skill is measured by each item. A value of `1` indicates the skill is measured by the item and a value of `0` indicates the skill is not measured by the item.

References

MacReady, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 2*(2), 99-120. doi:10.2307/1164802

Index