Package ‘corlink’

October 12, 2022

Title  Record Linkage, Incorporating Imputation for Missing Agreement Patterns, and Modeling Correlation Patterns Between Fields

Version  1.0.0

Description  A matrix of agreement patterns and counts for record pairs is the input for the procedure. An EM algorithm is used to impute plausible values for missing record pairs. A second EM algorithm, incorporating possible correlations between per-field agreement, is used to estimate posterior probabilities that each pair is a true match - i.e. constitutes the same individual.

Depends  R (>= 3.2.4)

License  CC0

Encoding  UTF-8

LazyData  true

RoxygenNote  5.0.1

NeedsCompilation  no

Author  John Ferguson [aut, cre]

Maintainer  John Ferguson <john.ferguson@nuigalway.ie>

Repository  CRAN

Date/Publication  2016-10-20 23:17:42

R topics documented:

  corlink ................................................................. 2
  linkd ................................................................. 2

Index  4
corlink: Record linkage, Incorporating Imputation for Missing Agreement Patterns, and Modeling correlation patterns between fields

Description

A matrix of agreement patterns and counts for record pairs is the input for the procedure. An EM algorithm is used to impute plausible values for missing record pairs. A second EM algorithm, incorporating possible correlations between per-field agreement, is used to estimate posterior probabilities that each pair is a true match - i.e. constitutes the same individual.

corlink functions

linkd

linkd Function to impute missing agreement patterns and then to link data

Description

Function to impute missing agreement patterns and then to link data

Usage

linkd(d, initial_m = NULL, initial_u = NULL, p_init = 0.5,
fixed_col = NULL, alg = "m")

Arguments

d Matrix of agreement patterns with final column counting the number of times that pattern was observed. See Details

initial_m starting probabilities for per-field agreement in record pairs, both records being generated from the same individual. Defaults to NULL

initial_u starting probabilities for per-field agreement in record pairs, with the two records being generated from differing individuals. Defaults to NULL

p_init starting probability that both records for a randomly selected record pair is associated with the same individual

fixed_col vector indicating columns that are not to be updated in initial EM algorithm. Useful if good prior estimates of the mis-match probabilities. See details

alg character; see Details
**Details**

$d$ is a numeric matrix with $N$ rows corresponding to $N$ record pairs, and $L+1$ columns the first $L$ of which show the field agreement patterns observed over the record pairs, and the last column the total number of times that pattern was observed in the database. The code 0 is used for a field that differs for two record, 1 for a field that agrees, and 2 for a missing field. $\text{fixed_col}$ indicates the components of the $u$ vector (per field probabilities of agreement for 2 records from differing individuals) that are not to be updated when applying the EM algorithm to estimate components of the Feligi Sunter model. $\text{alg}$ has four possible values. The default 'm' fits a log-linear model for the agreement counts only within the record pairs that corresponds to the same individual, 'b' fits differing log-linear models for the 2 clusters, 'i' corresponds to the original Feligi Sunter algorithm, with probabilities estimated via the EM algorithm, 'a' fits all the previously listed models.

**Value**

A list, the first component is a matrix - the posterior probabilities of being a true match is the last column, the second component are the fitted models used to generate the predicted probabilities.

**Examples**

```r
# Simulate data
m_probs <- rep(0.8,6)
u_probs <- rep(0.2,6)
means_match <- -1*qnorm(1-m_probs)
means_mismatch <- -1*qnorm(1-u_probs)
missingprobs <- rep(.2,6)
thedata <- do_sim(cor_match=0.2,cor_mismatch=0,nsample=10^4,pi_match=.5,m_probs=rep(0.8,5),u_probs=rep(0.2,5),missingprobs=rep(0.4,5))
colnames(thedata) <- c(paste("V",1:5,sep="_"),"count")
output <- linkd(thedata)
output$fitted_probs
```
Index

* EM
  linkd, 2
* Feligi/Sunter,
  linkd, 2
* algorithm,
  linkd, 2
* class,
  linkd, 2
* correlation
  linkd, 2
* latent
  linkd, 2
* linkage,
  linkd, 2
* probabilistic
  linkd, 2

corlink, 2
corlink-package (corlink), 2

linkd, 2