Package: modgo (via r-universe)

September 13, 2024

Type Package

Title MOck Data GeneratiOn

Version 1.0.1

Date 2022-03-09

Maintainer George Koliopanos <george.koliopanos@cardio-care.ch>

Description Generation of mock data from a real dataset using rank normal inverse transformation.

Author Andreas Ziegler [aut], Francisco Miguel Echevarria [aut, ctb], George Koliopanos [aut, cre]

Encoding UTF-8

RoxygenNote 7.2.3

Suggests knitr, rmarkdown, survival

VignetteBuilder knitr

License GPL (>= 3) + file LICENSE

Depends R (>= 4.1)

Imports ggplot2 (>= 3.4.0), patchwork (>= 1.1.2), wesanderson (>= 0.3.6.9000), Matrix (>= 1.6.1.1), ggcorrplot (>= 0.1.4.1), gridExtra (>= 2.3), psych (>= 2.2.9), GLDEX (>= 2.0.0.9.2), MASS (>= 7.3), gp (>= 1.0), stats, utils

URL https://github.com/GeorgeKoliopanos/modgo

BugReports https://github.com/GeorgeKoliopanos/modgo/issues

Repository https://georgekoliopanos.r-universe.dev

RemoteUrl https://github.com/georgekoliopanos/modgo

RemoteRef HEAD

RemoteSha 52273957ea132eef06448d009a9e9c08f7bf6fd1

Contents

`checkArguments` *Check Arguments*

Description

This function is used internally by modgo to check the correctness of the arguments passed to it.

Usage

```
checkArguments(
  data = NULL,
  ties_method = "max",
  variables = colnames(data),
  bin_variables = NULL,
  categ_variables = NULL,
  count_variables = NULL,
  n_samples = nrow(data),
  sigma = NULL,
  nrep = 100,
  noise_mu = FALSE,
  pertr_vec = NULL,
  change_cov = NULL,
  change_amount = 0,
  seed = 1,
  thresh_var = NULL,
  thresh_force = FALSE,
  var_prop = NULL,
  var_infl = NULL,
  infl_cov_stable = FALSE,
  tol = 1e-06,
  stop_sim = FALSE,
  new_mean_sd = NULL,
  multi_sugg_prop = NULL,
  generalized_mode = FALSE,
  generalized_mode_model = NULL,
  generalized_mode_lmbds = NULL
)
```

Arguments

data
A data frame containing the data whose characteristics are to be mimicked during the data simulation.

ties_method
Method used to deal with ties during rank transformation. Allowed input: "max","average" or "min". This parameter is passed by `rbi_normal_transform` to the parameter `ties.method` of `rank`.

variables
A character vector indicating the columns in `data` to be used. Default: `colnames(data)`.

bin_variables
A character vector listing those entries in `variables` to be treated as binary variables.

categ_variables
A character vector listing those entries in `variables` to be treated as ordinal categorical variables, with more than two categories. See Details.

count_variables
A character vector listing those entries `categ_variables` to be treated as count variables. Relevant only when `generalized_mode = TRUE`.

n_samples
Number of rows of each simulated dataset. Default is the number of rows of `data`.

sigma
A covariance matrix of NxN (N= number of variables) provided by the user to bypass the covariance matrix calculations

nrep
Number of simulated datasets to be generated.

noise_mu
Logical. Should noise be added to the mean vector of the multivariate normal distribution used to draw the simulated values? Default: FALSE.

pertr_vec
A named vector. Vector's names are the continuous variables that the user want to perturb. Variance of simulated dataset mimic original data's variance.

change_cov
Change the covariance of a specific pair of variables.

change_amount
the amount of change in the covariance of a specific pair of variables.

seed
A numeric value specifying the random seed. If `seed = NA`, no random seed is set.

thresh_var
A data frame that contains the thresholds(left and right) of specified variables (1st column: variable names, 2nd column: Left thresholds, 3rd column: Right thresholds)

thresh_force
A logical value indicating if you want to force threshold in case the proportion of samples that can surpass the threshold are less than 10%

`var_prop` A named vector that provides a proportion of value=1 for a specific binary variable (=name of the vector) that will be the proportion of this value in the simulated datasets.[this may increase execution time drastically]

`var_infl` A named vector. Vector's names are the continuous variables that the user want to perturb and increase their variance

`infl_cov_stable`
Logical value. If TRUE,perturbation is applied to original dataset and simulations values mimic the perturbed original dataset. Covariance matrix used for simulation = original data's correlations. If FALSE, perturbation is applied to the simulated datasets.

`tol` A numeric value that set up tolerance(relative to largest variance) for numerical lack of positive-definiteness in Sigma

`stop_sim` A logical value indicating if the analysis should stop before simulation and produce only the correlation matrix

`new_mean_sd` A matrix that contains two columns named "Mean" and "SD" that the user specifies desired Means and Standard Deviations in the simulated datasets for specific continues variables. The variables must be declared as ROWNAMES in the matrix.

`multi_sugg_prop`
A named vector that provides a proportion of value=1 for specific binary variables (=name of the vector) that will be the close to the proportion of this value in the simulated datasets.

`generalized_mode`
A logical value indicating if generalized lambda/Poisson distributions or set up thresholds will be used to generate the simulated values

`generalized_mode_model`
A matrix that contains two columns named "Variable" and "Model". This matrix can be used only if a generalized_mode_model argument is provided. It specifies what model should be used for each Variable. Model values should be "rmfmkl", "rprs", "star" or a combination of them, e.g. "rmfmkl-rprs" or "star-star", in case the user wants a bimodal simulation. The user can select Generalised Poisson model for Poisson variables, but this model cannot be included in bimodal simulation

`generalized_mode_lmbds`
A matrix that contains lambdas values for each of the variables of the dataset to be used for either Generalized Lambda Distribution Generalized Poisson Distribution or setting up thresholds

Details

All variables passed to `modgo` should be of class double or integer. This includes the variables passed to the parameter `categ_variables`. The character vector `variables`, indicating the variables in `data` to be used in the simulation, should contain at least two variables. The variables in `variables` not present in `bin_variables` nor `categ_variables` will be treated as continuous variables.

Author(s)

Francisco M. Ojeda, George Koliopanos

Cleveland — *Cleveland Dataset (‘Cleveland‘)*

Description

Rows: samples (303) x Columns: Variables (11)

Usage

```
data("Cleveland")
```

Format

A data frame

Details

Selected 11 variables from Cleveland Clinic Heart Disease Dataset (Detrano et al. (1989)). The dataset was dowloaded from the University of California in Irvine machine learning data repository (Dua et al. (2019)).

Missing values were imputed. For each continuous variable values were drawn from a normal distribution using the sample mean and standard deviation computed on the complete observations. For categorical variables values were drawn from the empirical distribution of the complete observations.

References

Detrano, R. et al. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology*, 64(5), 304-310.

Dua, D. and Graff C (2019). UCI machine learning repository. Irvine: University of California, School of Information and Computer Science. http://archive.ics.uci.edu/ml. Accessed March 20th 2023.

Examples

```
data("Cleveland", package="modgo")
```

corr_plots *Plots correlation matrix of original and simulated data*

Description

Produces a graphical display of the Pearson correlation matrix of the original dataset, a single simulated dataset and also of the average of the correlation matrices across all simulations for an object returned by `modgo`.

Usage

```
corr_plots(
  Modgo_obj,
  sim_dataset = 1,
  variables = colnames(Modgo_obj[["simulated_data"]][[1]])
)
```

Arguments

`Modgo_obj`	An object returned by `modgo`.
`sim_dataset`	Number indicating the simulated dataset in `Modgo_obj` to be used in plots.
`variables`	A character vector indicating the columns in the data to be used in plots.

Value

A patchwork object created by `patchwork::wrap_plots`.

Author(s)

Francisco M. Ojeda, George Koliopanos

Examples

```
data("Cleveland",package="modgo")

test_modgo <- modgo(data = Cleveland,
     bin_variables = c("CAD","HighFastBloodSugar","Sex","ExInducedAngina"),
     categ_variables =c("Chestpaintype"))

corr_plots(test_modgo)
```

distr_plots *Plots distribution of original and simulated data*

Description

Produces a graphical display of the distribution of the variables in the original and a single simulated dataset for an object returned by modgo.

Usage

```
distr_plots(
  Modgo_obj,
  variables = colnames(Modgo_obj[["original_data"]]),
  sim_dataset = 1,
  wespalette = "Cavalcanti1",
  text_size = 12
)
```

Arguments

Modgo_obj	An object returned by modgo.
variables	A character vector indicating the columns in the data to be used in plots.
sim_dataset	Number indicating the simulated dataset in Modgo_obj to be used in plots.
wespalette	Name of selected Wes Anderson color palette. Passed to wesanderson::wes_palette.
text_size	Text size in plot for legend, tick mark and axes labels. Passed to ggplot2::element_text.

Details

Box-and-whisker plots and bar charts are produced for continuous and categorical variables, respectively.

Value

a gtable object from package gtable.

Author(s)

Andreas Ziegler, Francisco M. Ojeda, George Koliopanos

Examples

```
data("Cleveland",package="modgo")
test_modgo <- modgo(data = Cleveland,
     bin_variables = c("CAD","HighFastBloodSugar","Sex","ExInducedAngina"),
     categ_variables =c("Chestpaintype"))

distr_plots(test_modgo)
```

generalizedMatrix *Generalized Lambda and Poisson preparation*

Description

Prepare the four moments matrix for GLD and GPD

Usage

```
generalizedMatrix(
  data,
  variables = colnames(data),
  bin_variables = NULL,
  generalized_mode_model = NULL,
  multi_sugg_prop = NULL
)
```

Arguments

`data` A data frame with original variables.

`variables` A vector of which variables you want to transform. Default:colnames(data)

`bin_variables` A character vector listing the binary variables.

`generalized_mode_model`
A matrix that contains two columns named "Variables" and "Model". This matrix can be used only if a generalized_mode_model argument is provided. It specifies what model should be used for each Variable. Model values should be "RMFMKL", "RPRS", "STAR" or a combination of them, e.g. "RMFMKL-RPRS" or "STAR-STAR", in case the use wants a bimodal simulation. The user can select Generalized Poisson model for Poisson variables, but this model cannot be included in bimodal simulation

`multi_sugg_prop`
A named vector that provides a proportion of value=1 for specific binary variables(=name of the vector) that will be the close to the proportion of this value in the simulated datasets.

Value

A numeric matrix

Author(s)

Francisco M. Ojeda, George Koliopanos

Examples

```
data("Cleveland",package="modgo")
Variables <- c("Age","STDepression")
Model <- c("rprs", "star-rmfmkl")
model_matrix <- cbind(Variables,
                      Model)
test_modgo <- generalizedMatrix(data = Cleveland,
     generalized_mode_model = model_matrix,
     bin_variables = c("CAD","HighFastBloodSugar","Sex","ExInducedAngina"))
```

`general_transform_inv` *Inverse gldex transformation*

Description

Inverse transforms z values of a vector to simulated values driven by the original dataset using Generalized Lambda and Generalized Poisson percentile functions.

Usage

```
general_transform_inv(x, data = NULL, n_samples, lmbds)
```

Arguments

`x`	A vector of z values.
`data`	A data frame with original variables.
`n_samples`	Number of samples you need to produce.
`lmbds`	A vector with generalized lambdas values

Value

A numeric vector.

Author(s)

Andreas Ziegler, Francisco M. Ojeda, George Koliopanos

Examples

```
data("Cleveland",package="modgo")
test_rank <- rbi_normal_transform(Cleveland[,1])
test_generalized_lmbds <- generalizedMatrix(Cleveland,
                  bin_variables = c("Sex", "HighFastBloodSugar", "CAD"))
test_inv_rank <- general_transform_inv(x = test_rank,
                  data = Cleveland[,1],
                  n_samples = 100,
                  lmbds = test_generalized_lmbds[,1])
```

generate_simulated_data *Generate new dataset by using previous correlation matrix*

Description

This function is used internally by modgo. It conducts the computation of the correlation matrix of the transformed variables, which are assumed to follow a multivariate normal distribution.

Usage

```
generate_simulated_data(
  data,
  df_sim,
  variables,
  bin_variables,
  categ_variables,
  count_variables,
  n_samples,
  generalized_mode,
  generalized_mode_lmbds,
  multi_sugg_prop,
  pertr_vec,
  var_infl,
  infl_cov_stable
)
```

Arguments

`data` a data frame with original variables.

`df_sim` a data frame with simulated values.

`variables` variables a character vector indicating which columns of `data` should be used.

`bin_variables` a character vector listing the binary variables.

`categ_variables` a character vector listing the ordinal categorical variables.

`count_variables` a character vector listing the count as a sub sub category of categorical variables. Count variables should be part of categorical variables vector. Count variables are treated differently when using gldex to simulate them.

`n_samples` Number of rows of each simulated dataset. Default is the number of rows of `data`.

`generalized_mode`
A logical value indicating if generalized lambda/Poisson distributions or set up thresholds will be used to generate the simulated values

`generalized_mode_lmbds`
A matrix that contains lmbds values for each of the variables of the dataset to be used for either Generalized Lambda Distribution Generalized Poisson Distribution or setting up thresholds

`multi_sugg_prop`
A named vector that provides a proportion of value=1 for specific binary variables(=name of the vector) that will be the close to the proportion of this value in the simulated datasets.

`pertr_vec` A named vector.Vector's names are the continuous variables that the user want to perturb. Variance of simulated dataset mimic original data's variance.

`var_infl` A named vector.Vector's names are the continuous variables that the user want to perturb and increase their variance

`infl_cov_stable`
Logical value. If TRUE,perturbation is applied to original dataset and simulations values mimic the perturbed original data set.Covariance matrix used for simulation = original data's correlations. If FALSE, perturbation is applied to the simulated datasets.

Value

Simulation Data Frame

Author(s)

Francisco M. Ojeda, George Koliopanos

`Inverse_transformation_variables` *Inverse transform variables*

Description

This function is used internally by `modgo`. It transforms all variables to their original scale.

Usage

```
Inverse_transformation_variables(
  data,
  df_sim,
  variables,
  bin_variables,
  categ_variables,
  count_variables,
  n_samples,
  generalized_mode,
  generalized_mode_lmbds
)
```

Arguments

`data` a data frame with original variables.

`df_sim` data frame with transformed variables.

`variables` variables a character vector indicating which columns of `data` should be used.

`bin_variables` a character vector listing the binary variables.

`categ_variables` a character vector listing the ordinal categorical variables.

`count_variables` a character vector listing the count as a sub sub category of categorical variables. Count variables should be part of categorical variables vector. Count variables are treated differently when using gldex to simulate them.

`n_samples` Number of rows of each simulated dataset. Default is the number of rows of `data`.

`generalized_mode` A logical value indicating if generalized lambda/Poisson distributions or set up thresholds will be used to generate the simulated values

`generalized_mode_lmbds` A matrix that contains lambdas values for each of the variables of the dataset to be used for either Generalized Lambda Distribution Generalized Poisson Distribution or setting up thresholds

Value

A correlation matrix.

Author(s)

Francisco M. Ojeda, George Koliopanos

modgo *MOck Data GeneratiOn*

Description

Creates synthetic dataset based on real data by means of the rank-based inverse normal transformation. Data with perturbed characteristics can be generated.

Usage

```
modgo(
  data,
  ties_method = ”max”,
  variables = colnames(data),
  bin_variables = NULL,
  categ_variables = NULL,
  count_variables = NULL,
  n_samples = nrow(data),
  sigma = NULL,
  nrep = 100,
  noise_mu = FALSE,
  pertr_vec = NULL,
  change_cov = NULL,
  change_amount = 0,
  seed = 1,
  thresh_var = NULL,
  thresh_force = FALSE,
  var_prop = NULL,
  var_infl = NULL,
  infl_cov_stable = FALSE,
  tol = 1e-06,
  stop_sim = FALSE,
  new_mean_sd = NULL,
  multi_sugg_prop = NULL,
  generalized_mode = FALSE,
  generalized_mode_model = NULL,
  generalized_mode_lmbds = NULL
)
```

Arguments

`data` A data frame containing the data whose characteristics are to be mimicked during the data simulation.

`ties_method` Method used to deal with ties during rank transformation. Allowed input: "max","average" or "min". This parameter is passed by `rbi_normal_transform` to the parameter `ties.method` of `rank`.

`variables` A character vector indicating the columns in `data` to be used. Default: `colnames(data)`.

`bin_variables` A character vector listing those entries in `variables` to be treated as binary variables.

`categ_variables`
A character vector listing those entries in `variables` to be treated as ordinal categorical variables, with more than two categories. See Details.

`count_variables`
A character vector listing those entries `categ_variables` to be treated as count variables. Relevant only when `generalized_mode` = `TRUE`.

`n_samples` Number of rows of each simulated dataset. Default is the number of rows of `data`.

`sigma` A covariance matrix of NxN (N= number of variables) provided by the user to bypass the covariance matrix calculations

`nrep` Number of simulated datasets to be generated.

`noise_mu` Logical. Should noise be added to the mean vector of the multivariate normal distribution used to draw the simulated values? Default: FALSE.

`pertr_vec` A named vector. Vector's names are the continuous variables that the user want to perturb. Variance of simulated dataset mimic original data's variance.

`change_cov` Change the covariance of a specific pair of variables.

`change_amount` the amount of change in the covariance of a specific pair of variables.

`seed` A numeric value specifying the random seed. If `seed = NA`, no random seed is set.

`thresh_var` A data frame that contains the thresholds(left and right) of specified variables (1st column: variable names, 2nd column: Left thresholds, 3rd column: Right thresholds)

`thresh_force` A logical value indicating if you want to force threshold in case the proportion of samples that can surpass the threshold are less than 10%

`var_prop` A named vector that provides a proportion of value=1 for a specific binary variable (=name of the vector) that will be the proportion of this value in the simulated datasets.[this may increase execution time drastically]

`var_infl` A named vector. Vector's names are the continuous variables that the user want to perturb and increase their variance

`infl_cov_stable`
Logical value. If TRUE,perturbation is applied to original dataset and simulations values mimic the perturbed original dataset. Covariance matrix used for simulation = original data's correlations. If FALSE, perturbation is applied to the simulated datasets.

`tol` A numeric value that set up tolerance(relative to largest variance) for numerical lack of positive-definiteness in Sigma

`stop_sim` A logical value indicating if the analysis should stop before simulation and produce only the correlation matrix

`new_mean_sd` A matrix that contains two columns named "Mean" and "SD" that the user specifies desired Means and Standard Deviations in the simulated datasets for specific continues variables. The variables must be declared as ROWNAMES in the matrix.

`multi_sugg_prop`
A named vector that provides a proportion of value=1 for specific binary variables (=name of the vector) that will be the close to the proportion of this value in the simulated datasets.

`generalized_mode`
A logical value indicating if generalized lambda/Poisson distributions or set up thresholds will be used to generate the simulated values

`generalized_mode_model`
: A matrix that contains two columns named "Variable" and "Model". This matrix can be used only if a generalized_mode_model argument is provided. It specifies what model should be used for each Variable. Model values should be "rmfmkl", "rprs", "star" or a combination of them, e.g. "rmfmkl-rprs" or "star-star", in case the user wants a bimodal simulation. The user can select Generalised Poisson model for Poisson variables, but this model cannot be included in bimodal simulation

`generalized_mode_lmbds`
: A matrix that contains lambdas values for each of the variables of the dataset to be used for either Generalized Lambda Distribution Generalized Poisson Distribution or setting up thresholds

Details

Simulated data is generated based on available data. The algorithm used is based on the ranked-based inverse normal transformation (Koliopanos et al. (2023)) and attempts to mimic the characteristics of the original data.

All variables passed to `modgo` should be of class double or integer. This includes the variables passed to the parameter `categ_variables`. The character vector `variables`, indicating the variables in `data` to be used in the simulation, should contain at least two variables. The variables in `variables` not present in `bin_variables` nor `categ_variables` will be treated as continuous variables.

Value

A list with the following components:

`simulated_data` A list of data frames containing the simulated data.

`original_data` A data frame with the input data.

`correlations` A list of correlation matrices. The ith element is the correlation matrix for the ith simulated dataset. The `(repn + 1)`the (last) element of the list is the average of the correlation matrices.

`bin_variables` A character vector listing the binary variables

`categ_variables`
: A character vector listing the ordinal categorical variables

`covariance_matrix`
: Covariance matrix used when generating observations from a multivariate normal distribution.

`seed` Random seed used.

`samples_produced`
: Number of rows of each simulated dataset.

`sim_dataset_number`
: Number of simulated datasets produced.

Author(s)

Francisco M. Ojeda, George Koliopanos

References

Koliopanos, G., Ojeda, F. and Ziegler A. (2023). A simple-to-use R package for mimicking study data by simulations. *Methods Inf Med*, 62(03/04), 119-129.

Examples

```
data("Cleveland",package="modgo")
test_modgo <- modgo(data = Cleveland,
     bin_variables = c("CAD","HighFastBloodSugar","Sex","ExInducedAngina"),
     categ_variables =c("Chestpaintype"))
```

modgo_survival *MOck Data GeneratiOn*

Description

`modgo_survival` Create mock dataset from a real one by using Generalized Lambdas Distributions and by separating the dataset in 2 based in the event status.

Usage

```
modgo_survival(
  data,
  event_variable = NULL,
  time_variable = NULL,
  surv_method = 1,
  ties_method = "max",
  variables = colnames(data),
  bin_variables = NULL,
  categ_variables = NULL,
  count_variables = NULL,
  n_samples = nrow(data),
  sigma = NULL,
  nrep = 100,
  noise_mu = FALSE,
  pertr_vec = NULL,
  change_cov = NULL,
  change_amount = 0,
  seed = 1,
  thresh_var = NULL,
  thresh_force = FALSE,
  var_prop = NULL,
  var_infl = NULL,
  infl_cov_stable = FALSE,
  tol = 1e-06,
  stop_sim = FALSE,
  new_mean_sd = NULL,
  multi_sugg_prop = NULL,
  generalized_mode = TRUE,
  generalized_mode_model = NULL,
  generalized_mode_model_event = "rprs",
  generalized_mode_model_no_event = "rprs",
  generalized_mode_lmbds = NULL
)
```

Arguments

`data` A data frame containing the data whose characteristics are to be mimicked during the data simulation.

`event_variable` a character string listing the event variable.

`time_variable` a character string listing the time variable.

`surv_method` A numeric value that indicates which one of the 2 survival methods will be used. First method (surv_method = 1): Event and no event datasets are using different covariance matrices for the simulation. Second method(surv_method = 2): Event and no event datasets are using the same covariance matrix for the simulation

`ties_method` Method on how to deal with equal values during rank transformation. Acceptable input:"max","average","min". This parameter is passed by `rbi_normal_transform` to the parameter `ties.method` of `rank`.

`variables` a vector of which variables you want to transform. Default:colnames(data)

`bin_variables` a character vector listing the binary variables.

`categ_variables` a character vector listing the ordinal categorical variables.

`count_variables` a character vector listing the count as a sub sub category of categorical variables. Count variables should be part of categorical variables vector. Count variables are treated differently when using gldex to simulate them.

`n_samples` Number of rows of each simulated dataset. Default is the number of rows of `data`.

`sigma` a covariance matrix of NxN (N= number of variables) provided by the user to bypass the covariance matrix calculations

`nrep` number of repetitions.

`noise_mu` Logical value if you want to apply noise to multivariate mean. Default: FALSE

`pertr_vec` A named vector.Vector's names are the continuous variables that the user want to perturb. Variance of simulated dataset mimic original data's variance.

`change_cov` change the covariance of a specific pair of variables.

`change_amount` the amount of change in the covariance of a specific pair of variables.

`seed` A numeric value specifying the random seed. If `seed = NA`, no random seed is set.

`thresh_var` A data frame that contains the thresholds(left and right) of specified variables (1st column: variable names, 2nd column: Left thresholds, 3rd column: Right thresholds) `thresh_force` A logical value indicating if you want to force threshold in case the proportion of samples that can surpass the threshold are less than 10%

`var_prop` A named vector that provides a proportion of value=1 for a specific binary variable(=name of the vector) that will be the proportion of this value in the simulated datasets.[this may increase execution time drastically]

`var_infl` A named vector.Vector's names are the continuous variables that the user want to perturb and increase their variance

`infl_cov_stable`
Logical value. If TRUE,perturbation is applied to original dataset and simulations values mimic the perturbed original data set.Covariance matrix used for simulation = original data's correlations. If FALSE, perturbation is applied to the simulated datasets.

`tol` A numeric value that set up tolerance(relative to largest variance) for numerical lack of positive-definiteness in Sigma

`stop_sim` A logical value indicating if the analysis should stop before simulation and produce only the correlation matrix

`new_mean_sd` A matrix that contains two columns named "Mean" and "SD" that the user specifies desired Means and Standard Deviations in the simulated datasets for specific continues variables. The variables must be declared as ROWNAMES in the matrix.

`multi_sugg_prop`
A named vector that provides a proportion of value=1 for specific binary variables(=name of the vector) that will be the close to the proportion of this value in the simulated datasets.

`generalized_mode`
A logical value indicating if generalized lambda/Poisson distributions or set up thresholds will be used to generate the simulated values

`generalized_mode_model`
A matrix that contains two columns named "Variable" and "Model". This matrix can be used only if a generalized_mode_model argument is provided. It specifies what model should be used for each Variable. Model values should be "rmfmkl", "rprs", "star" or a combination of them, e.g. "rmfmkl-rprs" or "star-star", in case the use wants a bimodal simulation. The user can select Generalised Poisson model for Poisson variables, but this model cannot be included in bimodal simulation

`generalized_mode_model_event`
A matrix that contains two columns named "Variable" and "Model" and it is to be used for the event dataset (event = 1). This matrix can be used only if a generalized_mode_model argument is provided. It specifies what model should be used for each Variable. Model values should be "rmfmkl", "rprs", "star" or a combination of them, e.g. "rmfmkl-rprs" or "star-star", in case the use wants a bimodal simulation. The user can select Generalised Poisson model for Poisson variables, but this model cannot be included in bimodal simulation.

`generalized_mode_model_no_event`
A matrix that contains two columns named "Variable" and "Model" and it is to be used for the non-event dataset (event = 0). This matrix can be used only if a

generalized_mode_model argument is provided. It specifies what model should be used for each Variable. Model values should be "rmfmkl", "rprs", "star" or a combination of them, e.g. "rmfmkl-rprs" or "star-star", in case the use wants a bimodal simulation. The user can select Generalised Poisson model for Poisson variables, but this model cannot be included in bimodal simulation

`generalized_mode_lmbds`
A matrix that contains lambdas values for each of the variables of the dataset to be used for either Generalized Lambda Distribution Generalized Poisson Distribution or setting up thresholds

Details

Simulated data is generated based on available data. The simulated data mimics the characteristics of the original data. The algorithm used is based on the ranked based inverse normal transformation (Koliopanos et al. (2023)).

Value

A list with the following components:

`simulated_data` A list of data frames containing the simulated data.

`original_data` A data frame with the input data.

`correlations` a list of correlation matrices. The ith element is the correlation matrix for the ith simulated dataset. The `(repn + 1)`the (last) element of the list is the average of the correlation matrices.

`bin_variables` character vector listing the binary variables

`categ_variables`
a character vector listing the ordinal categorical variables

`covariance_matrix`
Covariance matrix used when generating observations from a multivariate normal distribution.

`seed` Random seed used.

`samples_produced`
Number of rows of each simulated dataset.

`sim_dataset_number`
Number of simulated datasets produced.

Author(s)

Francisco M. Ojeda, George Koliopanos

`multicenter_comb` *Modgo multi-studies*

Description

Combines modgo objects from a multiple studies to a single one in order to calculate new correlations and visualise the data

Usage

```
multicenter_comb(modgo_1, ...)
```

Arguments

`modgo_1` a list modgo object.

`...` multiple modgo object names.

Value

A modgo object/list.

Author(s)

Francisco M. Ojeda, George Koliopanos

`rbi_normal_transform` *Rank-based inverse normal transformation*

Description

Applies the rank-based inverse normal transformation to a numeric vector.

Usage

```
rbi_normal_transform(x, ties_method = c("max", "min", "average"))
```

Arguments

`x` a numeric vector.

`ties_method` character string indicating how to handle ties when computing sample ranks. Can be any of "max","average" or "min". This is passed to the parameter `ties.method` of `rank`.

Details

The rank-based inverse normal transformation (Beasley et al. (2009)), transforms values of a vector of length n to ranks/(n + 1) and then applies the quantile function of the standard normal distribution.

Value

A numeric vector.

Author(s)

Andreas Ziegler, Francisco M. Ojeda, George Koliopanos

References

Beasley, T.M. and Erickson S. and Allison D.B. (2009). Rank-based inverse normal transformations are increasingly used, but are they merited? *Behavior genetics*, 39, 580-595.

Examples

```
data("Cleveland",package="modgo")
test_rank <- rbi_normal_transform(Cleveland[,1])
```

---

rbi_normal_transform_inv

Inverse of rank-based inverse normal transformation

Description

Transforms a vector `x` using the inverse of a rank-based inverse normal transformation associated with a given vector `x_original`. This inverse is defined as $F_n^{-1}\Phi(x)$, where F_n^{-1} is the inverse empirical cumulative distribution function of `x_original` and Φ is the cumulative distribution function of a standard normal random variable.

Usage

```
rbi_normal_transform_inv(x, x_original)
```

Arguments

`x`	a numeric vector to which the inverse of a rank-based inverse normal transformation associated with `x_original` will be applied.
`x_original`	a numeric vector.

Value

A numeric vector.

Author(s)

Andreas Ziegler, Francisco M. Ojeda, George Koliopanos

Examples

```
data("Cleveland",package="modgo")
test_rank <- rbi_normal_transform(Cleveland[,1])
test_inv_rank <- rbi_normal_transform_inv(x = test_rank,
                                         x_original = Cleveland[,1])
```

`Sigma_calculation` *Calculate Sigma with the help of polychoric and polyserial functions*

Description

This function is used internally by `modgo`. It conducts the computation of the correlation matrix of the transformed variables, which are assumed to follow a multivariate normal distribution.

Usage

```
Sigma_calculation(data, variables, bin_variables, categ_variables, ties_method)
```

Arguments

`data` a data frame with original variables.

`variables` variables a character vector indicating which columns of `data` should be used.

`bin_variables` a character vector listing the binary variables.

`categ_variables` a character vector listing the ordinal categorical variables.

`ties_method` Method on how to deal with equal values during rank transformation. Acceptable input:"max","average","min". This parameter is passed by `rbi_normal_transform` to the parameter `ties.method` of `rank`.

Value

A correlation matrix.

Author(s)

Francisco M. Ojeda, George Koliopanos

`Sigma_transformation` *Correlation of transformed variables*

Description

This function is used internally by `modgo`. It finishes the computation of the correlation matrix of the transformed variables, which are assumed to follow a multivariate normal distribution. It computes the correlations involving at least one categorical variable. For this purpose the biserial, tetrachoric, polyserial and polychoric correlations are used.

Usage

```
Sigma_transformation(
  data,
  data_z,
  Sigma,
  variables,
  bin_variables = c(),
  categ_variables = c()
)
```

Arguments

`data` a data frame with original variables.

`data_z` data frame with transformed variables.

`Sigma` A numeric square matrix.

`variables` variables a character vector indicating which columns of `data` should be used.

`bin_variables` a character vector listing the binary variables.

`categ_variables` a character vector listing the ordinal categorical variables.

Value

A correlation matrix.

Author(s)

Francisco M. Ojeda, George Koliopanos

Index