Package 'fakeR'

May 26, 2016

Type Package

Title Simulates Data from a Data Frame of Different Variable Types

Version 1.0

Date 2016-05-25

Description Generates fake data from a dataset of different variable types.
The package contains the functions simulate_dataset and simulate_dataset_ts
to simulate time-independent and time-dependent data. It randomly samples
character and factor variables from contingency tables and numeric and
ordered factors from a multivariate normal distribution. It currently supports the
simulation of stationary and zero-inflated count time series.

License CC0

Imports mvtnorm, polycor, pscl, VGAM, stats

Suggests knitr, rmarkdown, testthat

VignetteBuilder knitr

NeedsCompilation no

Author Lily Zhang [aut, cre],
Dustin Tingley [aut]

Maintainer Lily Zhang <lilyhzhang1029@gmail.com>

Repository CRAN

Date/Publication 2016-05-26 17:54:23

R topics documented:

fakeR-package *Simulates Data from a Data Frame of Different Variable Types*

Description

Generates fake data from a dataset of different variable types. The package contains the functions simulate_dataset and simulate_dataset_ts to simulate time-independent and time-dependent data. It randomly samples character and factor variables from contingency tables and numeric and ordered factors from a multivariate normal distribution. It currently supports the simulation of stationary and zero-inflated count time series.

Details

The DESCRIPTION file:

Package:	fakeR
Type:	Package
Title:	Simulates Data from a Data Frame of Different Variable Types
Version:	1.0
Date:	2016-05-25
Authors@R:	c(person("Lily", "Zhang", email = "lilyhzhang1029@gmail.com", role = c("aut", "cre")), person("Dustin","'
Description:	Generates fake data from a dataset of different variable types. The package contains the functions simulate
License:	CC0
Imports:	mvtnorm, polycor, pscl, VGAM, stats
Suggests:	knitr, rmarkdown, testthat
VignetteBuilder:	knitr
Author:	Lily Zhang [aut, cre], Dustin Tingley [aut]
Maintainer:	Lily Zhang <lilyhzhang1029@gmail.com>

Index of help topics:

This package is used to simulate datasets of different variable types. The package contains the functions simulate_dataset and simulate_dataset_ts to simulate time-independent and time-dependent data.

Author(s)

NA

Maintainer: NA

References

~~ Literature or other references for background information ~~

Examples

```
## time-independent data frame of multiple types
# single column of an unordered, string factor
state_df <- data.frame(division=state.division)
# character variable
state_df$division <- as.character(state_df$division)
# numeric variable
state_df$area <- state.area
# factor variable
state_df$region <- state.region
state_sim <- simulate_dataset(state_df)

## time-independent data frame with missingness
df <- mtcars
# change one of the variable types to an unordered factor
df$carb <- as.factor(df$carb)
# change another variable type to an ordered factor
df$gear <- as.ordered(as.factor(df$gear))
df[2,] <- NA
sim_df <- simulate_dataset(df, stealth.level=2)

## time series dataframe
tree_ring <- data.frame(treering)
tree_ring$year <- c(1: nrow(tree_ring))
sim_tree_ring <- simulate_dataset_ts(tree_ring,
                                     cluster="treering",
                                     time.variable="year")
par(mfrow = c(2, 1), mar = c(3, 3, 4, 2), mgp = 0.9 * 2:0)
plot (tree_ring$year, tree_ring$treering, type='l',
      main=paste("Original","Normalized ring width"),
      ylab="Ring width", xlab="Year index")
plot (tree_ring$year, tree_ring$treering, type='l',
      main=paste("Simulated","Normalized ring width"),
      ylab="Ring width", xlab="Year index")
```

`simulate_dataset` *Simulate from a data frame of time-independent data.*

Description

This function takes as argument an existing dataset in the form of a data frame and outputs a randomized version of all its columns. The function accepts the following types: character variables, numeric variables, and ordered and unordered factor variables.

Usage

```
simulate_dataset(dataset, digits=2, n=NA,
                              use.levels=TRUE, use.miss=TRUE,
                              mvt.method="eigen", het.ML=FALSE,
                              het.suppress=TRUE, stealth.level=1,
                              level3.noise=FALSE, ignore=NA)
```

Arguments

dataset	the data frame from which to generate a randomized version
digits	the number of digits after the decimal point to include in the new values
n	number of rows in the new data frame. Equal to the number of rows in the original if set to NA, the default.
use.levels	when set to true, gives the simulated factor variables the same number of levels as the original.
use.miss	when set to TRUE, inserts the missing data like is present in the original (i.e. based on the distribution of missingness in the original data).
mvt.method	specifies the matrix decomposition to be used in sampling from the multivariate normal.
het.ML	as per the hetcor function, if TRUE, compute maximum-likelihood estimates;if FALSE, compute quick two-step estimates in computing the heterogeneous correlation matrix.
het.suppress	when set to TRUE, suppresses stops from the het.corr function.
stealth.level	when set to 1 (deafult), takes into account the covariances between all the unordered factors and the covariances between the numeric and ordered factors. When set to 2, simulates each variable independently. When set to 3, does not take into account any covariances and instead randomly samples from a uniform distribution ranging from the min to the max of the data for each variable.
level3.noise	when set to TRUE, add Gaussian noise to the min and max parameter for the uniform distribution in stealth.level 3. The noise term has a variance of one fourth of the range of the data for any particular variable.
ignore	specifies which columns to ignore (i.e. to leave as is instead of simulate). Takes in a list of column names as input.

Details

This function does not account for clustered time series data (see simulate_dataset_ts).

This function randomly samples each each character and factor variable from the population distribution given in the original dataset. It simulates numeric and ordered factors from a multivariate normal distribution. When both numeric and ordered factors are included, a heterogeneous correlation matrix is used, coercing the means of the ordered factor variables to be 0.

The function only accounts for between-column correlations for numeric and ordered factor variables. Each unordered factor and character column is treated as independent.

The order of the columns in the simulated dataset may differ from the order of the original dataset since the function puts the numeric and ordered factor data in the front and the character and unordered factor data afterwards. The column names stay consistent, however.

Value

Returns a data frame with the same number of columns and same type for each.

Author(s)

Lily Zhang Dustin Tingley

References

Inspired by the fakeR function originally created by Ryne Estabrook.

Examples

```
# single column of an unordered, string factor
state_df <- data.frame(division=state.division)
# character variable
state_df$division <- as.character(state_df$division)
# numeric variable
state_df$area <- state.area
# factor variable
state_df$region <- state.region
state_sim <- simulate_dataset(state_df)
```

---

simulate_dataset_ts *Simulate a dataframe of time series data*

Description

This function simulates clustered numeric time series data from an ARIMA model fit or, if specified, a zero-inflated Poission regression model fit, with each column variable regressed on the first lag.

Usage

```
simulate_dataset_ts(dataset, digits=2, n=NA, cluster=NA, time.variable=NA,
                    date.index=FALSE, complete.panel=FALSE, zero.inflate=FALSE,
                  stealth.level=2, level3.noise=FALSE, use.miss=TRUE, ignore=NA)
```

Arguments

dataset	the data frame from which to generate a randomized version
digits	the number of digits after the decimal point to include in the new values
n	number of rows in the new data frame. Equal to the number of rows in the original if set to NA, the default.
cluster	the column names of the time series variables. Argument should be in the form of a list if multiple values.

time.variable	the column name(s) of the time variables corresponding to each time series variable. Should be the same length as cluster, even if that means including the same time variable multiple times.
date.index	whether the time variable is a date and should be treated as a Date object.
complete.panel	when set to TRUE, indicates a preprocessing step needed to complete the time series columns. Specifically, inserts all missing dates and zero values at each of those time points.
zero.inflate	when set to TRUE, indicates that time series variables require a zero-inflated Poisson regression model fit. When set to FALSE (default), series all fit to an ARIMA model to simulate from.
stealth.level	when set to 2 (default), simulates independent time series observations. When set to 3, does not take into account any covariances between time points and instead randomly samples from a uniform distribution ranging from the min to the max of the data for each variable. No option 1.
level3.noise	when set to TRUE, add Gaussian noise to the min and max parameter for the uniform distribution in stealth.level 3. The noise term has a variance of one fourth of the range of the data for any particular variable.
use.miss	when set to TRUE, inserts the missing data like is present in the original.
ignore	specifies which columns to ignore (i.e. to leave as is instead of simulate). Takes in a list of column names as input.

Details

Note that this function is specific to two types of numeric time series, stationary ones and zero-inflated count ones. For modeling clustered numeric data assuming a multivariate normal distribution, look at simulate_dataset. Note that this function only accepts numeric observation types.

The function assumes each time series process is independent of the others, and allows for a different time variable to be associated with each series. Thus, there is no stealth level of 1 for simulate_dataset_ts(), as this function does not simulate multivariate time series. Columns not part of time series values or time indices are ignored and not simulated.

Value

A data frame. Columns alternating time variable and cluster variable, with the cluster/time variable pairs in the order inputted into the function arguments.

Author(s)

Lily Zhang Dustin Tingley

Examples

```
# An example using the treering dataset from the R datasets package
tree_ring <- data.frame(treering)
tree_ring$year <- c(1: nrow(tree_ring))
sim_tree_ring <- simulate_dataset_ts(tree_ring,
                                     cluster="treering",
```

```
                                        time.variable="year")
par(mfrow = c(2, 1), mar = c(3, 3, 4, 2), mgp = 0.9 * 2:0)
plot (tree_ring$year, tree_ring$treering, type='l',
      main=paste("Original","Normalized ring width"),
      ylab="Ring width", xlab="Year index")
plot (tree_ring$year, tree_ring$treering, type='l',
      main=paste("Simulated","Normalized ring width"),
      ylab="Ring width", xlab="Year index")
```

Index