Package 'genset'

July 22, 2025

Type Package

Title Generates Multiple Data Sets

Version 0.1.1

Maintainer Lori Murray <lori.murray@uwo.ca>

Description
Generate multiple data sets for educational purposes to demonstrate the importance of multiple regression. The genset function generates a data set from an initial data set to have the same summary statistics (mean, median, and standard deviation) but opposing regression results.

License GPL-2

Encoding UTF-8

Suggests knitr, rmarkdown

VignetteBuilder knitr

RoxygenNote 7.2.3

NeedsCompilation no

Author Lori Murray [aut, cre]

Repository CRAN

Date/Publication 2025-03-29 18:20:02 UTC

Contents

genset *Generate Mutliple Data Sets for Hands-on Learning*

Description

Generate mutliple data sets to demonstrate the importance of multiple regression. Data sets are generated from an initial data set input to have the same summary statistics (mean, median, and standard deviation) but opposing regression results (significance in the predictor variables). The initial data set will have one response variable (continuous) and two predictor variables (continous or one continuous and one categorical with 2 levels) that are statistically significant in a linear regression model.

Usage

```
genset(y, x1, x2, method=c(1,2), option=c("x1","x2","both"), n, decrease, output)
```

Arguments

`y`	response variable (continuous).
`x1`	first predictor variable (continuous).
`x2`	second predictor variable (continuous or categorical with 2 levels). If variable is categorical then argument is `factor(x2)`.
`method`	the method `1` or `2` to be used to generate the data set. `1` (default) rearranges the values within each variable, and `2` is a perturbation method that makes subtle changes to the values of the variables.
`option`	the variable(s) that will not statistically significant in the new data set (`"x1"`, `"x2"` or `"both"`).
`n`	the number of iterations. Default is 2000 iterations.
`decrease`	indicates an increase or decrease in level of significance. `FALSE` is the default.
`output`	shows the iterations. `FALSE` is the default.

Details

The summary statistics are within a (predetermined) tolerance level, and when rounded, will be the same as the original data set. The standard convention of 0.05 is used as the significance level threshold. Less than `n=2000` iterations may or may not be sufficient and is dependent on the initial data set.

Author(s)

Lori Murray and John Wilson

References

Murray, L.L. & Wilson, J.G. (2021). Generating data sets for teaching the importance of regression analysis. *Decision Sciences Journal of Innovative Education (DSJIE)*, Vol 19 (2), 157-166.

Examples

```
## Choose variables of interest
y <- mtcars$mpg
x1 <- mtcars$hp
x2 <- mtcars$wt
## Create a dataframe
set1 <- data.frame(y, x1, x2)
## Check summary statistics
multi.fun <- function(x) {
  c(mean = mean(x), media=median(x), sd=sd(x))
}
round(multi.fun(set1$y), 0)
round(multi.fun(set1$x1), 1)
round(multi.fun(set1$x2), 1)
## Fit linear regression model
## to verify regressors are statistically
## significant (p-value < 0.05)
summary(lm(y ~ x1, x2, data=set1))

## Set seed to reproduce same data set
set.seed(101)
set2 <- genset(y, x1, x2, method=1, option="x1", n=1000)
## Verify summary statistics match set 1
round(multi.fun(set2$y), 0)
round(multi.fun(set2$x1), 1)
round(multi.fun(set2$x2), 1)
## Fit linear regression model
## to verify x1 is not statistically
## significant (p-value > 0.05)
summary(lm(y ~ x1 + x2, data=set2))
```

Index