Package 'genset'

September 16, 2020

Type Package

Title Generates Data Sets for Class Demonstrations

Version 0.1.0

Author Lori Murray [aut, cre], John Wilson [aut, cre]

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

Description 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

LazyData true

Suggests knitr, rmarkdown, testthat

VignetteBuilder knitr

RoxygenNote 7.1.1

NeedsCompilation no

Repository CRAN

Date/Publication 2020-09-16 09:40:02 UTC

R topics documented:

genset *Generate Data Sets for Class Demonstrations*

Description

Generate data sets to demonstrate the importance of multiple regression. 'genset' generates a data set from an initial data set to have the same summary statistics (mean, median, and standard deviation) but opposing regression results. 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, option, n, decrease, output)
```

Arguments

y	a vector containing the response variable (continuous),
x1	a vector containing the first predictor variable (continuous)
x2	a vector containing the 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 be statistically significant in the new data set ("x1" (default), "x2" or "both")
n	maximum number of iterations
decrease	decreases the signficance level when TRUE, default is FALSE
output	print each interation when TRUE, default is FALSE

Details

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

Value

Returns an object of class "data.frame" containing the generated data set: (in order) the response variable, first predictor variable and second predictor variable.

Author(s)

Lori Murray & John Wilson

References

Murray, L. and Wilson, J. (2020). The Need for Regression: Generating Multiple Data Sets with Identical Summary Statistics but Differing Conclusions. *Decision Sciences Journal of Innovative Education.* Accepted for publication.

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