Package 'sda1'

January 22, 2023

Title R-paket för kursen Statistik och dataanalys I vid SU

Version 0.0.1

Description Funktioner för grundläggande statistik, inkl regression.

License MIT

Encoding UTF-8

Roxygen list(markdown = TRUE)

RoxygenNote 7.2.3

Imports glmnet,
mvtnorm,
manipulate,
RColorBrewer,
SUdatasets,
ggplot2,
cowplot

Remotes StatisticsSU/SUdatasets

Depends R (>= 2.10)

LazyData true

Suggests rmarkdown,
knitr

VignetteBuilder knitr

R topics documented:

bike — *Number of daily rides for a bike share company in Washington D.C.*

Description

A dataset containing the number of rides per day and other attributes over the course of 2 years

Usage

```
bike
```

Format

A data frame with 731 rows and 12 variables:

dteday date in YYYY-MM-DD format

season categorical variable (1="winter", 2 = "spring", 3 = "summer", 4 = "fall")

yr year (0="2011", 1 = "2012")

mnth month from 1-12 where 1 = "January"

holiday binary variable for public holidays

weekday day of the week 0-6, 0 ="Sunday"

workingday binary variable for working days (=1)

weathersit categorical variable (1="clear", 2 = "mist", 3 = "light snow")

temp continuous temperature variable, normalized between 0,1

hum continuous humidity variable, normalized between 0,1

windspeed continuous windspeed variable, normalized between 0,1 ...

Source

https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset

corr_matrix — *Compute pair-wise correlations and hypothesis test*

Description

Computes pair-wise correlations between variables in a dataframe df Uses p-values to test:

H0: rho = 0
H1: rho != 0

Usage

```
corr_matrix(df)
```

Arguments

df	dataframe

Value

list with two tables: corrs (correlations), pvals (p-values)

Examples

```
library(sda1)
corr_matrix(mtcars[,c("mpg","hp","drat","wt")])
```

---

reg_crossval *K-fold cross-validation of regression models estimated with lm()*

Description

K-fold cross-validation of regression models estimated with lm()

Usage

```
reg_crossval(formula, data, nfolds, obs_order = "random")
```

Arguments

formula	an object of class "formula": a symbolic description of the model to be fitted.
data	a data frame with the data used for fitting the models.
nfolds	the number of folds in the cross-validation.
obs_order	order of the observations when splitting the data. obs_order = "random" gives a random order.

Value

RMSE Root mean squared prediction error on test data

Examples

```
library(sda1)
RMSE_CV = reg_crossval(mpg ~ hp, data = mtcars, nfolds = 4, obs_order = 1:32)
print(RMSE_CV)
```

reg_predict *Plot confidence and prediction intervals for simple linear regression*

Description

Plot confidence and prediction intervals for simple linear regression

Usage

```
reg_predict(formula, data, level = 0.95, conf_int_line = T, pred_interval = T)
```

Arguments

formula	an object of class "formula": a symbolic description of the model to be fitted.
data	a data frame with the data.
level	confidence level, default is level = 0.95
conf_int_line	if TRUE, then conf intervals for regression line are plotted.
pred_interval	if TRUE, then prediction intervals are plotted.

Value

plot of data with overlayed intervals

Examples

```
library(sda1)
reg_predict(mpg ~ hp, data = mtcars)
```

reg_simulate *Simulate from a linear regression model*

Description

Simulates a dataset with n observation from the linear regression model

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon, \epsilon \sim N(0, \sigma_\epsilon^2)$$

with covariates (x) simulated from a normal distribution with the same correlation rho_x between all pairs of covariates. Covariate x_j has standard deviation sigma_x[j]. Alternatively the covariate can follow a uniform distribution.

Usage

```
reg_simulate(
  n,
  betavect,
  sigma_eps,
  intercept = TRUE,
  covdist = "normal",
  rho_x = 0,
  sigma_x = rep(1, length(betavect) - intercept)
)
```

Arguments

n	the number of observations in the simulated dataset.
betavect	a vector with regression coefficients c(beta_0,beta_1,...beta_k). First element is intercept if `intercept = TRUE`
sigma_eps	standard deviation of the error terms, epsilon.
intercept	if `TRUE` an intercept is added to the model.
covdist	distribution of the covariates. Options: `'normal'` or `'uniform'`.
rho_x	correlation among the covariates. Same for all covariate pairs.
sigma_x	vector with standard deviation of the covariates.

Value

dataframe with simulated data (y, X1, X2, ..., XK) (no intercept included).

Examples

```
library(sda1)
simdata <- reg_simulate(n = 500, betavect = c(1, -2, 1, 0), sigma_eps = 2)
lmfit <- lm(y ~ X1 + X2 + X3, data = simdata)
reg_summary(lmfit, anova = F)
```

---

reg_summary *Summarize the results from a regression analysis*

Description

Alternative to `summary.lm` to summarize a regression from `lm`. Prints a table similar to the one generated by SAS and Minitab.

Usage

```
reg_summary(
  lmobject,
  anova = T,
  fit_measures = T,
  param = T,
  conf_intervals = F,
  vif_factors = F
)
```

Arguments

lmobject	a fitted regression model from lm.
anova	TRUE if an ANOVA table is computed.
fit_measures	TRUE if measures of fit (R^2 etc) is computed.
param	TRUE if parameter estimates, standard errors etc is computed.
conf_intervals	TRUE if confidence intervals for parameters.
vif_factors	TRUE if variance inflation factors are to be printed.

Value

list with three tables: param, anova and fit_measures

Examples

```
library(sda1)
lmfit = lm(nRides ~ temp + hum + windspeed, data = bike)
regsumm = reg_summary(lmfit, anova = T, conf_intervals = T, vif_factors = T)
regsumm$param
regsumm$anova
regsumm$fit_measures
```

---

residuals4in1 *Residual analysis mimicing the 4-in-1 plots from Minitab*

Description

Plots:

1. Normal QQ-plot
2. Residuals vs fitted values
3. Histogram and normal density fit
4. Residuals vs order.

Usage

```
residuals4in1(lm_object)
```

Arguments

`lm_object`	a fitted regression model from `lm`.

Examples

```
library(sda1)
fit = lm(mpg ~ hp, data = mtcars)
residuals4in1(fit)
```

---

titanic *Survival of passengers on the Titanic*

Description

This data set provides information on the fate of passengers on the fatal maiden voyage of the ocean liner 'Titanic', summarized according to economic status (class), sex, age and survival.

NOTE: this is not the same as the dataset Titanic (note capital T) which has more observations, but also missing values.

Usage

```
titanic
```

Format

A data frame with 887 rows and 8 variables:

name passenger name

survived 0 = no, 1 = yes

sex male/female

age age of passenger

fare ticket cost

firstclass first class ticket ...

Details

The sinking of the Titanic is a famous event, and new books are still being published about it. Many well-known facts—from the proportions of first-class passengers to the 'women and children first' policy, and the fact that that policy was not entirely successful in saving the women and children in the third class—are reflected in the survival rates for various classes of passenger.

These data were originally collected by the British Board of Trade in their investigation of the sinking. Note that there is not complete agreement among primary sources as to the exact numbers on board, rescued, or lost.

Due in particular to the very successful film 'Titanic', the last years saw a rise in public interest in the Titanic. Very detailed data about the passengers is now available on the Internet, at sites such as Encyclopedia Titanica (https://www.encyclopedia-titanica.org/).

Source

Dawson, Robert J. MacG. (1995), The 'Unusual Episode' Data Revisited. Journal of Statistics Education, 3. doi: 10.1080/10691898.1995.11910499.

Index