Package: durhamSLR (via r-universe)

September 7, 2024

Type Package

Title The durhamSLR package

Version 0.2.0

Author Sarah Heaps

Maintainer Sarah.Heaps <sarah.e.heaps@durham.ac.uk>

Imports plyr

Description Data for Statistical Learning modules at Durham University.

License GPL-3 | GPL-2

Encoding UTF-8

LazyData true

RoxygenNote 7.2.1

Repository https://nseg4.r-universe.dev

RemoteUrl https://github.com/nseg4/durhamSLR

RemoteRef HEAD

RemoteSha 8e74cb7fae87e46339d5611d7b9b641aedd8db23

Contents

`admission` *MBA admissions data.*

Description

Data on applicants to the Masters of Business Administration (MBA) programme of a US business graduate school.

Usage

```
data(admission)
```

Value

A data frame with 85 rows and 3 variables. The data frame contains the following columns:

GPA The applicant's grade point average (GPA) on a 0.0 - 4.0 scale.

GMAT The applicant's graduate management admission test (GMAT) score on a 200 - 800 scale.

decision A factor with three levels, `admit`, `border` and `notadmit`, which refer to the category to which the student was assigned by admissions tutors (admit, borderline or do not admit).

Source

The data were taken from Johnson and Wichern (2008).

References

Johnson, R.A and Wichern, D.W. (2008) *Applied Multivariate Statistical Analysis, Sixth Edition.* Pearson.

Examples

```
data(admission)
head(admission)
```

`airpollution` *Air pollution data.*

Description

A data comprising of observations from 80 US cities for the year 1960 on 11 variables. These include a number of measures of air pollution, specifically concentrations of sulphate and suspended particulate, as well as a number of demographic variables.

Usage

```
data(airpollution)
```

Value

A data frame with 80 rows and 11 variables. The data frame contains the following columns:

SMIN Smallest biweekly sulphate reading in micrograms per cubic metre (x 10).

SMEAN Arithmetic mean of biweekly sulphate reading in micrograms per cubic metre (x 10).

SMAX Largest biweekly sulphate reading in micrograms per cubic metre (x 10).

PMIN Smallest biweekly suspended particulate reading in micrograms per cubic metre (x 10).

PMEAN Arithmetic mean of biweekly suspended particulate reading in micrograms per cubic metre (x 10).

PMAX Largest biweekly suspended particulate reading in micrograms per cubic metre (x 10).

PM2 Population density per square mile (x 0.1).

PERWH Percent of population who are white.

NONPOOR Percent of families with income above the poverty level.

GE65 Percent of population who are at least 65 (x 10).

LPOP Logarithm (base 10) of population (x 10).

Source

The complete data set is described in Gibbons *et al.* (1987).

References

D.I. Gibbons and G.C. McDonald and R.F. Gunst (1987), The complementary use of regression diagnostics and robust estimators. *Naval Research Logistics*, **34**, 109–131.

Examples

```
data(airpollution)
head(airpollution)
```

banknotes *Banknote authentication data.*

Description

Data extracted from images taken from genuine and forged banknotes, digitized into 400 x 400 arrays of pixels, and then summarised into four continuously valued summary statistics. For each banknote, the data set records whether the banknote was genuine or forged, along with the four numerical summaries of the image.

Usage

```
data(banknotes)
```

Value

A data frame with 1372 rows and 5 variables. The data frame contains the following columns:

variance Variance of Wavelet Transformed image.

skewness Skewness of Wavelet Transformed image.

kurtosis Kurtosis of Wavelet Transformed image.

entropy Entropy of image.

class A factor with two levels, `0` and `1`, which refer to whether the banknote was a forgery or real.

Source

The data were taken from the UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/banknote+authentication.

Examples

```
data(banknotes)
head(banknotes)
```

Boston *Housing values in suburbs of Boston.*

Description

The Boston data frame has 506 rows and 14 columns.

Usage

```
data(Boston, package="durhamSLR")
```

Value

This data frame contains the following columns:

lcrim Natural logarithm of the per capita crime rate by town.

zn Proportion of residential land zoned for lots over 25,000 sq.ft.

indus Proportion of non-retail business acres per town.

chas Charles River dummy variable (=1 if tract bounds river; =0 otherwise).

nox Nitrogen oxides concentration (parts per 10 million).

rm Average number of rooms per dwelling.

age Proportion of owner-occupied units built prior to 1940.

disf A numerical vector representing an ordered categorical variable with four levels depending on the weighted mean of the distances to five Boston employment centres (=1 if distance < 2.5, =2 if 2.5 <= distance < 5, =3 if 5 <= distance < 7.5, =4 if distance >= 7.5).

rad Index of accessibility to radial highways.

tax Full-value property-tax rate per $10,000.

pratio Pupil-teacher ratio by town.

black $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town.

lstat Lower status of the population (percent).

medv Median value of owner-occupied homes in $1000s.

Source

Harrison, D. and Rubinfeld, D.L. (1978) Hedonic prices and the demand for clean air. *J. Environ. Economics and Management* **5**, 81–102.

Belsley D.A., Kuh, E. and Welsch, R.E. (1980) *Regression Diagnostics. Identifying Influential Data and Sources of Collinearity.* New York: Wiley.

Examples

```
data(Boston, package="durhamSLR")
head(Boston)
```

`centipedes` *Counts of centipedes.*

Description

A data set containing the counts of *Lithobius forficatus*, more commonly known as the brown or stone centipede, at each of 30 sites in microhabitats of rotting wood. For each site, a number of soil and habitat variables are recorded in addition to their altitude and geographical coordinates.

Usage

```
data(centipedes)
```

Value

A data frame with 30 rows and 10 variables. The data frame contains the following columns:

site The abbreviated site name.

count The number of centipedes found at the site.

offset The area sampled at the site in square metres.

type A factor with two levels, `Synanthropic` and `Deciduous`, which refer to the habitat in which the site was located; either deciduous woods or "synanthropic" areas associated with human habitation, e.g. parks and gardens.

lorg The natural loagarithm of the percentage of organic matter in the soil.

lalt The natural loarithm of the altitude of the site in metres.

airt The air temperature in degrees Celcius.

soilt The soil temperature in degrees Celcius.

east The Easting of the site in tenths of a kilometre.

north The Northing of the site in tenths of a kilometre.

Source

The complete data set, which involved more species of centipede and more microhabitats, is described in Blackburn *et al.* (2002).

References

J. Blackburn and M. Farrow and W. Arthur (2002), Factors influencing the distribution, abundance and diversity of geophilomorph and lithobiomorph centipedes. *Journal of Zoology*, **256**, 221–232.

Examples

```
data(centipedes)
head(centipedes)
```

---

chapman *Chapman data.*

Description

Data from a study on heart disease by Dr. John M. Chapman in the mid-twentieth century. The data were taken from the Los Angeles Heart Study and comprise measurements from 200 male patients.

Usage

```
data(chapman)
```

Value

A data frame with 200 rows and 7 variables. The data frame contains the following columns:

age Patient's age; a numeric vector.

highbp Patient'systolic blood pressure; a numeric vector.

lowbp Patient's diastolic blood pressure; a numeric vector.

chol Patient's cholesterol; a numeric vector.

height Patient's height; a numeric vector.

weight Patient's weight; a numeric vector.

y A binary numeric vector which takes the value 1 if the patient experienced a coronory incident in the preceeding 10 years and 0 otherwise.

Source

The data were taken from the StatLib Datasets Archive at Carnegie Mellon University: https://lib.stat.cmu.edu/datasets/christensen-llm.

Examples

```
data(chapman)
head(chapman)
```

---

diabetes *Blood and other measurements in diabetics.*

Description

Data collected in a study concerning patients with diabetes. The response variable of interest was disease progression one year after taking baseline measurements on various clinical variables. For each of n=442 patients, the data comprise a quantitative measure of disease progression (`dis`) and measurements on p=10 baseline (explanatory) variables: age (`age`), sex (`sex`), body mass index (`bmi`), average blood pressure (`map`) and six blood serum measurements (`tc`, `ldl`, `hdl`, `tch`, `ltg`, `glu`). The explanatory variables have been transformed to have mean 0, with sum of squares equal to 1.

Usage

```
data(diabetes)
```

Value

A data frame with 442 rows and 11 variables. The data frame contains the following columns:

age Age.

sex Gender.

bmi Body mass index.

map Average blood pressure.

tc Blood serum measurement 1.

ldl Blood serum measurement 2.

hdl Blood serum measurement 3.

tch Blood serum measurement 4.

ltg Blood serum measurement 5.

glu Blood serum measurement 6.

dis Quantitative measure of disease progression.

Source

http://www-stat.stanford.edu/~hastie/Papers/LARS/LeastAngle_2002.ps.

References

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004), Least Angle Regression (with discussion). *Annals of Statistics*, **32**, 407–499.

Examples

```
data(diabetes)
head(diabetes)
```

---

diagnostics *Graphical diagnostics for arrays of MCMC output.*

Description

This function generates graphical diagnostics for an array of MCMC output. such For every parameter, a row of three plots is generated: a trace plot, an ACF plot and a kernel density plot. If there is output from more than one chain, the default behaviour is to overlay the plots for different chains in different colours.

Usage

```
diagnostics(mcmc, rows = 3, lag.max = 50, pool = FALSE, colours = NULL)
```

Arguments

mcmc	A matrix with dimensions: iterations, parameters; or a three dimensional array with dimensions: iterations, chains, parameters. The final (i.e. parameter) component of the dimnames attribute of the matrix or array should contain the parameter names.
rows	A number indicating the number of parameters to plot per page on the graphics device.
lag.max	A number indicating the maximum lag for the ACF plots.
pool	A logical. If TRUE the samples are pooled across chains before generating the plots.
colours	A vector indicating the colours to use to represent each chain. Colours can be specified using any of the three kinds of R colour specifications, i.e. a colour name (as listed by colors()), a hexadecimal string of the form "#rrggbb" or "#rrggbbaa" or a positive integer i meaning palette()[i].

Value

NULL

Examples

```
srs = array(rnorm(8000), c(1000, 2, 4)) # Example for illustration only!
dimnames(srs) = list(NULL, NULL, paste("theta[",1:4,"]",sep=""))
diagnostics(srs)
```

---

durhamSLR *The durhamSLR package*

Description

Data for Statistical Learning modules at Durham University.

Author(s)

Sarah Heaps <sarah.e.heaps@durham.ac.uk>

heptathlon *Heptathlon data.*

Description

Results for the heptathlon at the 2012 Olympic Games in London for the 29 athletes who completed all events and were not disqualified.

Usage

```
data(heptathlon)
```

Value

A data frame with 29 rows and 7 variables. The data frame contains the following columns:

H100M 100 metres hurdles (seconds).

HJ High jump (metres).

SP Shot put (metres).

R200M 200 metres (seconds).

LJ Long jump (metres).

JT Javelin throw (metres).

R800M 800 metres (seconds).

Source

The data were taken from Wikipedia at https://en.wikipedia.org/wiki/Athletics_at_the_2012_Summer_Olympics_%E2%80%93_Women%27s_heptathlon.

Examples

```
data(heptathlon)
head(heptathlon)
```

heptathlon_points *Heptathlon points data.*

Description

Final scores for the heptathlon at the 2012 Olympic Games in London for the 29 athletes who completed all events and were not disqualified.

Usage

```
data(heptathlon_points)
```

Value

A data frame with 29 rows and 1 variable in the column Points.

Source

The data were taken from Wikipedia at https://en.wikipedia.org/wiki/Athletics_at_the_2012_Summer_Olympics_%E2%80%93_Women%27s_heptathlon.

Examples

```
data(heptathlon_points)
head(heptathlon_points)
```

---

plasma *Blood screening data.*

Description

The erythrocyte sedimentation rate (ESR) and measurements of two plasma proteins (fibrinogen and globulin).

Usage

```
data(plasma)
```

Details

The erythrocyte sedimentation rate (ESR) is the rate at which red blood cells (erythrocytes) settle out of suspension in blood plasma, when measured under standard conditions. If the ESR increases when the level of certain proteins in the blood plasma rise in association with conditions such as rheumatic diseases, chronic infections and malignant diseases, its determination might be useful in screening blood samples taken form people suspected to being suffering from one of the conditions mentioned. The absolute value of the ESR is not of great importance rather it is whether it is less than 20mm/hr since lower values indicate a healthy individual.

The question of interest is whether there is any association between the probability of an ESR reading greater than 20mm/hr and the levels of the two plasma proteins. If there is not then the determination of ESR would not be useful for diagnostic purposes.

Value

A data frame with 32 observations on the following 3 variables:

fibrinogen The fibrinogen level in the blood.

globulin The globulin level in the blood.

ESR A factor with two levels representing the erythrocyte sedimentation rate, either less or greater 20 mm / hour.

Source

D. Collett and A. A. Jemain (1985), Residuals, outliers and influential observations in regression analysis. *Sains Malaysiana*, **4**, 493–511.

Examples

```
data(plasma)
layout(matrix(1:2, ncol = 2))
boxplot(fibrinogen ~ ESR, data = plasma, varwidth = TRUE)
boxplot(globulin ~ ESR, data = plasma, varwidth = TRUE)
```

USArrests *Violent crime rates by US state with region.*

Description

This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas and the Census Bureau-designated region

Usage

```
data(USArrests, package="durhamSLR")
```

Value

A data frame with 50 observations on 5 variables.:

Murder Murder arrests (per 100,000); a numeric vector.

Assault Assault arrests (per 100,000); a numeric vector.

UrbanPop Percent urban population; a numeric vector.

Rape Rape arrests (per 100,000); a numeric vector.

Region A factor with four levels indicating the Census Bureau-designated region.

Note

USArrests contains the data as in McNeil's monograph. For the UrbanPop percentages, a review of the table (No. 21) in the Statistical Abstracts 1975 reveals a transcription error for Maryland (and that McNeil used the same "round to even" rule that R's round() uses), as found by Daniel S Coven (Arizona).

Source

World Almanac and Book of facts 1975. (Crime rates).

Statistical Abstracts of the United States 1975, p.20, (Urban rates), possibly available as https://books.google.ch/books?id=zl9qAAAAMAAJ&pg=PA20.

References

McNeil, D. R. (1977) *Interactive Data Analysis*. New York: Wiley.

Examples

```
data(USArrests, package="durhamSLR")
head(USArrests)
```

Index