Package: faraway (via r-universe)

March 9, 2025

Version 1.0.9

Date 2025-02-07

Title Datasets and Functions for Books by Julian Faraway

Description Books are ``Linear Models with R'' published 1st Ed. August 2004, 2nd Ed. July 2014, 3rd Ed. February 2025 by CRC press, ISBN 9781439887332, and ``Extending the Linear Model with R'' published by CRC press in 1st Ed. December 2005 and 2nd Ed. March 2016, ISBN 9781584884248 and ``Practical Regression and ANOVA in R'' contributed documentation on CRAN (now very dated).

Depends R (>= 3.5.0)

License GPL

URL https://github.com/julianfaraway/faraway

LazyData yes

Imports methods, lme4, nlme

Suggests leaps

Encoding UTF-8

RoxygenNote 7.3.2

NeedsCompilation no

Author Julian Faraway [aut, cre]

Maintainer Julian Faraway <jjf23@bath.ac.uk>

Repository CRAN

Date/Publication 2025-02-07 13:00:02 UTC

Config/pak/sysreqs cmake make

Contents

`aatemp` *Annual mean temperatures in Ann Arbor, Michigan*

Description

The data comes from the U.S. Historical Climatology Network.

Format

A data frame with 115 observations on the following 2 variables.

year year from 1854 to 2000

temp annual mean temperatures in degrees F in Ann Arbor

Source

United States Historical Climatology Network: https://www.ncei.noaa.gov/products/land-based-station/us-historical-climatology-network

---

`abrasion` *Wear on materials according to type, run and position*

Description

The `abrasion` data frame has 16 rows and 4 columns. Four materials were fed into a wear testing machine and the amount of wear recorded. Four samples could be processed at the same time and the position of these samples may be important. A Latin square design was used.

Format

This data frame contains the following columns:

run The run number 1-4

position The position number 1-4

material The material A-D

wear The wear measured loss of weight in 0.1mm over testing period

Source

The Design and Analysis of Industrial Experiments by O. Davies, 1954, published by Wiley

aflatoxin *aflatoxin dosage and liver cancer in lab animals*

Description

Aflatoxin B1 was fed to lab animals at vary doses and the number responding with liver cancer recorded.

Format

A data frame with 6 observations on the following 3 variables.

dose dose in ppb

total number of test animals

tumor number with liver cancer

Source

Gaylor DW (1987) "Linear nonparametric upper limits for low dose extrapolation" ASA Proceedings of the Biopharmaceutical Section.

Examples

```
data(aflatoxin)
```

africa *miltary coups and politics in sub-Saharan Africa*

Description

Data is a subset of a larger study on factors affecting regime stability in Sub-Saharan Africa

Format

A data frame with 47 observations on the following 9 variables.

miltcoup number of successful military coups from independence to 1989

oligarchy number years country ruled by military oligarchy from independence to 1989

pollib Political liberalization - 0 = no civil rights for political expression, 1 = limited civil rights for expression but right to form political parties, 2 = full civil rights

parties Number of legal political parties in 1993

pctvote Percent voting in last election

popn Population in millions in 1989

size Area in 1000 square km

numelec Total number of legislative and presidential elections

numregim Number of regime types

Source

Bratton, Michael, and Nicholas Van De Walle. 1997. “Political Regimes and Regime Transitions in Africa, 1910-1994.” *Study Number I06996.* Ann Arbor: Inter-University Consortium for Political and Social Research.

References

"Bayesian Methods: A Social and Behavioral Sciences Approach" by Jeff Gill 2002.

`airpass` *Airline passengers*

Description

Monthly totals of airline passengers from 1949 to 1951

Format

A data frame with 144 observations on the following 2 variables.

pass number of passengers in thousands

year the date as a decimal

Details

Well known time series example dataset

Source

Brown, R.G.(1962) Smoothing, Forecasting and Prediction of Discrete Time Series. Englewood Cliffs, N.J.: Prentice-Hall.

References

Box, G.E.P., Jenkins, G.M. and Reinsel, G.C. (1994) Time Series Analysis, Forecasting and Control, 3rd edn. Englewood Cliffs, N.J.: Prentice-Hall.

Examples

```
data(airpass)
## maybe str(airpass) ; plot(airpass) ...
```

alfalfa *Effects of seed inoculum, irrigation and shade on alfalfa yield*

Description

The alfalfa data frame has 25 rows and 4 columns. Data comes from an experiment to test the effects of seed inoculum, irrigation and shade on alfalfa yield. A latin square design has been used.

Format

This data frame contains the following columns:

shade Distance of location from tree line divided into 5 shade areas

irrigation Irrigation effect divided into 5 levels

inoculum Four types of seed incolum, A-D with E as control.

yield Dry matter yield of alfalfa

Source

Petersen, R.G. 1994. Agricultural Field Experiments, Design and Analysis. Marcel Dekker, Inc., New York. Pages 70-74. 1994

amlxray *Match pair study for AML and Xray link*

Description

A matched case control study carried out to investigate the connection between X-ray usage and acute myeloid leukemia in childhood. The pairs are matched by age, race and county of residence.

Format

A data frame with 238 observations on the following 11 variables.

ID a factor denoting the matched pairs

disease 0=control, 1=case

Sex F or M

downs Presence of Downs syndrome: no or yes

age Age in years

Mray Did the mother ever have an Xray: no or yes

MupRay Did the mother have an Xray of the upper body during pregnancy: no or yes

MlowRay Did the mother have an Xray of the lower body during pregnancy: no or yes

Fray Did the father ever have an Xray: no or yes

Cray Did the child ever have an Xray: no or yes

CnRay Total number of Xrays of the child 1=none < 2=1 or 2 < 3=3 or 4 < 4= 5 or more

Source

Chap T. Le (1998) "Applied Categorical Data Analysis" Wiley.

`anaesthetic` *Time in minutes to eye opening after reversal of anaesthetic.*

Description

A doctor at major London hospital compared the effects of 4 anaesthetics used in major operations. 80 patients were divided into groups of 20.

Format

A data frame with 80 observations on the following 2 variables.

breath time in minutes to start breathing unassisted

tgrp Four treatment groups `A` `B` `C` `D`

Source

Chatfield C. (1995) Problem Solving: A Statistician's Guide, 2ed Chapman Hall.

Examples

```
data(anaesthetic)
## maybe str(anaesthetic) ; plot(anaesthetic) ...
```

`babyfood` *Respiratory disease rates of babies fed in different ways*

Description

Study on infant respiratory disease, namely the proportions of children developing bronchitis or pneumonia in their first year of life by type of feeding and sex.

Format

A data frame with 6 observations on the following 4 variables.

disease number with disease

nondisease number without disease

sex a factor with levels `Boy` `Girl`

food a factor with levels `Bottle` `Breast` `Suppl`

Source

Payne, C. (1987). The GLIM System Release 3.77 Manual (2 ed.). Oxford: Nu- merical Algorithms Group.

Examples

```
data(babyfood)
## maybe str(babyfood) ; plot(babyfood) ...
```

`beetle` *Beetles exposed to fumigant*

Description

Grain beetles were exposed to ethylene oxide

Format

A data frame with 10 observations on the following 3 variables.

conc concentration of ethylene oxide in mg/l

affected number affected

exposed number exposed

Source

Busvine (1938)

References

Collet D. "Modelling Binary Data"

Examples

```
data(beetle)
## maybe str(beetle) ; plot(beetle) ...
```

bliss *Insect mortality due to insecticide*

Description

An experiment measuring death rates for insects, with 30 insects at each of five treatment levels.

Format

A data frame with 5 observations on the following 3 variables.

dead number dead

alive number alive

conc concentration of insecticide

Source

Bliss (1935). The calculation of the dosage-mortality curve. Annals of Applied Biology 22, 134-167.

Examples

```
data(bliss)
## maybe str(bliss) ; plot(bliss) ...
```

breaking *Breaking strength of materials*

Description

An experiment was conducted to select the supplier of raw materials for production of a component. The breaking strength of the component was the objective of interest. Four suppliers were considered. The four operators can only produce one component each per day. A Latin square design was used.

Format

A data frame with 16 observations on the following 4 variables.

y The breaking strength of the component

operator the operator - a factor with levels `op1` `op2` `op3` `op4`

day the day of production - a factor with levels `day1` `day2` `day3` `day4`

supplier the supplier of the raw material - a factor with levels `A` `B` `C` `D`

Source

Lentner M. and Bishop T. (1986) Experimental Design and Analysis, Valley Book Company

broccoli *Broccoli weight variation*

Description

A number of growers supply broccoli to a food processing plant. The plant instructs the growers to pack the broccoli into standard size boxes. There should be 18 clusters of broccoli per box and each cluster should weigh between 1.33 and 1.5 pounds. Because the growers use different varieties, methods of cultivation etc, there is some variation in the cluster weights. The plant manager selected 3 growers at random and then 4 boxes at random supplied by these growers. 3 clusters were selected from each box.

Format

A data frame with 36 observations on the following 4 variables.

wt weight of broccoli

grower the grower - a factor with levels `1` `2` `3`

box the box - a factor with levels `1` `2` `3` `4`

cluster the cluster - a factor with levels `1` `2` `3`

Source

Lentner M. and Bishop T. (1986) Experimental Design and Analysis, Valley Book Company

butterfat *Butterfat content of milk by breed*

Description

Average butterfat content (percentages) of milk for random samples of twenty cows (ten two-year old and ten mature (greater than four years old)) from each of five breeds. The data are from Canadian records of pure-bred dairy cattle.

Format

A data frame with 100 observations on the following 3 variables.

Butterfat butter fat content by percentage

Breed a factor with levels `Ayrshire` `Canadian` `Guernsey` `Holstein-Fresian` `Jersey`

Age a factor with levels `2year` `Mature`

Source

Sokal, R. R. and Rohlf, F. J. (1994) Biometry. W. H. Freeman, New York, third edition.

Examples

```
data(butterfat)
## maybe str(butterfat) ; plot(butterfat) ...
```

---

cathedral *Cathedral nave heights and lengths in England*

Description

Example Dataset from "Practical Regression and Anova"

Format

A dataset with 25 cases

style of the cathedral - romanesque or gothic

height in feet

width in feet

Source

Weisberg, S. (2005). Applied Linear Regression, 3rd edition. New York: Wiley

References

Reference details may be found in "Practical Regression and Anova" by Julian Faraway

---

cheddar *Taste of Cheddar cheese*

Description

In a study of cheddar cheese from the LaTrobe Valley of Victoria, Australia, samples of cheese were analyzed for their chemical composition and were subjected to taste tests. Overall taste scores were obtained by combining the scores from several tasters.

Format

A data frame with 30 observations on the following 4 variables.

taste a subjective taste score

Acetic concentration of acetic acid (log scale)

H2S concentration of hydrogen sulfide (log scale)

Lactic concentration of lactic acid

Source

David S. Moore and George P. McCabe (1993) Introduction to the Practice of Statistics, W. H. Freeman and company, second edition.

Examples

```
data(cheddar)
## maybe str(cheddar) ; plot(cheddar) ...
```

`chicago` *Chicago insurance redlining*

Description

Data from a 1970's study on the relationship between insurance redlining in Chicago and racial composition, fire and theft rates, age of housing and income in 47 zip codes.

Format

This dataframe contains the following columns

race racial composition in percent minority

fire fires per 100 housing units

theft theft per 1000 population

age percent of housing units built before 1939

involact new FAIR plan policies and renewals per 100 housing units

income median family income in thousands of dollars

side North or South side of Chicago

Source

Adapted from "Data : A Collection of Problems from Many Fields for the Student and Research Worker" by D. Andrews and A. Herzberg published by Springer-Verlag, in 1985

chiczip *Chicago zip codes north-south*

Description

Complements the chicago and chmiss datasets by dividing the zip codes into north and south

Format

chiczip takes the values "n" (north) and "s" south

References

Reference details may be found in "Practical Regression and Anova" by Julian Faraway

See Also

chicago

chmiss *Chicago insurance redlining*

Description

Data from a 1970's study on the relationship between insurance redlining in Chicago and racial composition, fire and theft rates, age of housing and income in 47 zip codes. Missing values have been randomly added.

Format

This dataframe contains the following columns

race racial composition in percent minority

fire fires per 100 housing units

theft theft per 1000 population

age percent of housing units built before 1939

involact new FAIR plan policies and renewals per 100 housing units

income median family income in thousands of dollars

side North or South side of Chicago

Source

Adapted from "Data : A Collection of Problems from Many Fields for the Student and Research Worker" by D. Andrews and A. Herzberg published by Springer-Verlag, in 1985

choccake *Chocolate cake experiment with split plot design*

Description

An experiment was conducted to determine the effect of recipe and baking temperature on chocolate cake quality. 15 batches of cake mix for each recipe were prepared. Each batch was sufficient for six cakes. Each of the six cakes was baked at a different temperature which was randomly assigned. Several measures of cake quality were recorded of which breaking angle was just one.

Format

A data frame with 270 observations on the following 4 variables.

recipe Chocolate for recipe 1 was added at 40C, Chocolate for recipe 2 was added at 60C and recipe 3 had extra sugar

batch batch number from 1 to 15

temp temperature at which cake was baked: `175C` `185C` `195C` `205C` `215C` `225C`

breakang the breaking angle of the cake

Source

Cochran W. and Cox G. (1992) Experimental Designs, 2nd Edition Wiley

chredlin *Chicago insurance redlining*

Description

Data from a 1970's study on the relationship between insurance redlining in Chicago and racial composition, fire and theft rates, age of housing and income in 47 zip codes

Format

This dataframe contains the following columns

race racial composition in percent minority

fire fires per 100 housing units

theft theft per 1000 population

age percent of housing units built before 1939

involact new FAIR plan policies and renewals per 100 housing units

income median family income in thousands of dollars

side North or South side of Chicago

Source

Adapted from "Data : A Collection of Problems from Many Fields for the Student and Research Worker" by D. Andrews and A. Herzberg published by Springer-Verlag, in 1985

`clot` *Blood clotting times*

Description

The clotting times of blood for plasma diluted with nine different percentage concentrations with prothrombin-free plasma

Format

This data frame contains the following columns:

time time in seconds to clot

conc concentration in percent

lot lot number - either one or two

Source

Hurn et al (1945)

References

Nelder & McCullagh (1989) Generalized Linear Models (2ed)

`cmob` *Social class mobility from 1971 to 1981 in the UK*

Description

Social class mobility from 1971 to 1981 for 42425 men from the United Kingdom census. Subjects were aged 45-64.

Format

A data frame with 36 observations on the following 3 variables.

y Frequency of observation

class71 social class in 1971 - a factor with levels `I`, professionals, `II` semi-professionals, `IIIN` skilled non-manual, `IIIM` skilled manual, `IV` semi-skilled, `V` unskilled

class81 social class in 1971 - a factor with levels `I` `II` `IIIN` `IIIM` `IV` `V` with same classification

Source

D. Blane and S. Harding and M. Rosato (1999) "Does social mobility affect the size of the socioeconomic mortality differential?: Evidence from the Office for National Statistics Longitudinal Study" JRSS-A, 162 59-70.

`cns` *Malformations of the central nervous system*

Description

Frequencies of various malformations of the central nervous system recorded on live births in South Wales, UK. Study was designed to determine the effect of water hardness on the incidence of such malformations.

Format

A data frame with 16 observations on the following 7 variables.

Area a factor with levels `Cardiff` `GlamorganC` `GlamorganE` `GlamorganW` `MonmouthOther` `MonmouthV` `Newport` `Swansea` being areas of South Wales

NoCNS count of births with no CNS problem

An count of Anencephalus births

Sp count of Spina Bifida births

Other count of other CNS births

Water water hardeness

Work a factor with levels `Manual` `NonManual` being the type of work done by the parents

Source

C. Lowe and C. Roberts and S. Lloyd, (1971) Malformations of the central nervous system and softness of local water supplies, British Medical Journal, 15,357-361.

References

P. McCullagh and J. Nelder (1989), Generalized Linear Models, Chapman and Hall, 2nd Ed.

coagulation *Blood coagulation times by diet*

Description

Dataset comes from a study of blood coagulation times. 24 animals were randomly assigned to four different diets and the samples were taken in a random order.

Format

This dataframe contains the following columns

coag coagulation time in seconds

diet diet type - A,B,C or D

Source

"Statistics for Experimenters" by G. P. Box, W. G. Hunter and J. S. Hunter, Wiley, 1978

composite *Strength of a thermoplastic composite depending on two factors*

Description

The composite data frame has 9 rows and 3 columns. Data comes from an experiment to test the strength of a thermoplastic composite depending on the power of a laser and speed of a tape.

Format

This data frame contains the following columns:

strength interply bond strength of the composite

laser laser power at 40, 50 or 60W

tape tape speed, slow=6.42 m/s, medium=13m/s and fast=27m/s

Source

Mazumdar, S and Hoa S (1995) "Application of a Taguchi Method for Process enhancement of an online consolidation technique" Composites 26, 669-673

cornnit *Corn yields from nitrogen application*

Description

The relationship between corn yield (bushels per acre) and nitrogen (pounds per acre) fertilizer application were studied in Wisconsin.

Format

A data frame with 44 observations on the following 2 variables.

yield corn yield in bushels per acre

nitrogen pounds per acre

Source

Unknown

corrosion *Corrosion loss in Cu-Ni alloys*

Description

Data consist of thirteen specimens of 90/10 Cu-Ni alloys with varying iron content in percent. The specimens were submerged in sea water for 60 days and the weight loss due to corrosion was recorded in units of milligrams per square decimeter per day.

Format

This dataframe contains the following columns

Fe Iron content in percent

loss Weight loss in mg per square decimeter per day

Source

"Applied Regression Analysis" by N. Draper and H. Smith, Wiley, 1998

cpd *Projected and actual sales of 20 consumer products*

Description

Projected and actual sales of 20 consumer products. Data have been disguised from original form.

Format

A data frame with 20 observations on the following 2 variables.

projected projected sales in dollars
actual actual sales in dollars

Source

G. Whitmore (1986) "Inverse Gaussian Ratio Estimation" Applied Statistics, 35, 8-15.

Cpplot *Cp plot*

Description

Makes a Cp plot

Usage

```
Cpplot(cp)
```

Arguments

cp	A leaps object returned from leaps()

Details

Requires leaps package

Value

none

Author(s)

Julian Faraway

See Also

leaps()

crawl *Crawling babies by month*

Description

A study investigated whether babies take longer to learn to crawl in cold months when they are often bundled in clothes that restrict their movement, than in warmer months. The study sought an association between babies' first crawling age and the average temperature during the month they first try to crawl (about 6 months after birth). Parents brought their babies into the University of Denver Infant Study Center between 1988-1991 for the study. The parents reported the birth month and age at which their child was first able to creep or crawl a distance of four feet in one minute. Data were collected on 208 boys and 206 girls (40 pairs of which were twins)

Format

A data frame with 12 observations on the following 4 variables.

crawling average crawling age in weeks

SD standard deviation of crawling age

n sample size

temperature average temperature(F) six months after birth

Source

Benson, Janette. (1993). Infant Behavior and Development

Examples

```
data(crawl)
## maybe str(crawl) ; plot(crawl) ...
```

ctsib *Effects of surface and vision on balance.*

Description

An experiment was conducted to study the effects of surface and vision on balance. The balance of subjects were observed for two different surfaces and for restricted and unrestricted vision. Balance was assessed qualitatively on an ordinal four-point scale based on observation by the experimenter. Forty subjects were studied, twenty males and twenty females ranging in age from 18 to 38, with heights given in cm and weights in kg. The subjects were tested while standing on foam or a normal surface and with their eyes closed or open or with a dome placed over their head. Each subject was tested twice in each of the surface and eye combinations for a total of 12 measures per subject.

Format

A data frame with 480 observations on the following 8 variables.

Subject an indicator

Sex a factor with levels `female` `male`

Age in years

Height in cm

Weight in kg

Surface a factor with levels `foam` `norm`

Vision a factor with levels `closed` `dome` `open`

CTSIB a four point scale measuring balance

Source

Steele, R. (1998). Effect of surface and vision on balance. Ph. D. thesis, Depart- ment of Physiotherapy, University of Queensland.

References

OzDasl

Examples

```
data(ctsib)
## maybe str(ctsib) ; plot(ctsib) ...
```

`death` *Death penalty in Florida 1977*

Description

Data on 326 defendents in homicide indictments in 20 Florida counties during 1976-77.

Format

A data frame with 8 observations on the following 4 variables.

y a numeric vector

penalty Did the subject recieve the death penalty? `no` or `yes`

victim Was the victim black or white?

defend Was the defendent black or white?

Source

Radelet M. (1981) Racial characteristics and the imposition of the death penalty. Amer. Sociol. Rev. **46** 918-927.

References

Agresti A. (1990) Categorical Data Analysis, Wiley.

`debt` *psychology of debt*

Description

The data arise from a large postal survey on the psychology of debt.

Format

A data frame with 464 observations on the following 13 variables.

incomegp income group (1=lowest, 5=highest)

house security of housing tenure (1=rent, 2=mortgage, 3=owned outright)

children number of children in household

singpar is the respondent a single parent?

agegp age group (1=youngest)

bankacc does the respondent have a bank account?

bsocacc does the respondent have a building society account?

manage self-rating of money management skill (high values=high skill)

ccarduse how often did s/he use credit cards (1=never... 3=regularly)

cigbuy does s/he buy cigarettes?

xmasbuy does s/he buy Christmas presents for children?

locintrn score on a locus of control scale (high values=internal)

prodebt score on a scale of attitudes to debt (high values=favourable to debt

Details

All yes/no questions are coded 0=no, 1=yes. Locus of control is a personality measure introduced by Rotter, which claims to differentiate people according to how much they feel things that happen to them are as a result of processes within themselves (internal locus of control) or outside events (external locus of control).

Source

Lea, Webley & Walker, 1995, Journal of Economic Psychology, 16, 181-201 Data obtained from http://au.exeter.ac.uk/SEGLea/.

denim *Denim wastage by supplier*

Description

Five suppliers cut denim material for a jeans manufacturer. An algorithm is used to estimate how much material will be wasted given the dimensions of the material supplied. Typically, a supplier wastes more material than the target based on the algorithm although occasionally they waste less. The percentage of waste relative to target was collected weekly for the 5 suppliers. In all, 95 observations were recorded.

Format

A data frame with 95 observations on the following 2 variables.

waste percentage wastage

supplier a factor with levels 1 2 3 4 5

Source

Unknown

Examples

```
data(denim)
## maybe str(denim) ; plot(denim) ...
```

diabetes *Diabetes and obesity, cardiovascular risk factors*

Description

403 African Americans were interviewed in a study to understand the prevalence of obesity, diabetes, and other cardiovascular risk factors in central Virginia.

Format

A data frame with 403 observations on the following 19 variables.

id Subject ID

chol Total Cholesterol

stab.glu Stabilized Glucose

hdl High Density Lipoprotein

ratio Cholesterol/HDL Ratio

glyhb Glycosolated Hemoglobin

location County - a factor with levels `Buckingham` `Louisa`

age age in years

gender a factor with levels `male` `female`

height height in inches

weight weight in pounds

frame a factor with levels `small` `medium` `large`

bp.1s First Systolic Blood Pressure

bp.1d First Diastolic Blood Pressure

bp.2s Second Systolic Blood Pressure

bp.2d Second Diastolic Blood Pressure

waist waist in inches

hip hip in inches

time.ppn Postprandial Time (in minutes) when Labs were Drawn

Details

Glycosolated hemoglobin greater than 7.0 is usually taken as a positive diagnosis of diabetes

Source

Willems JP, Saunders JT, DE Hunt, JB Schorling: Prevalence of coronary heart disease risk factors among rural blacks: A community-based study. Southern Medical Journal 90:814-820; 1997

References

Schorling JB, Roach J, Siegel M, Baturka N, Hunt DE, Guterbock TM, Stewart HL: A trial of church-based smoking cessation interventions for rural African Americans. Preventive Medicine 26:92-101; 1997

`dicentric` *Radiation dose effects on chromosomal abnormality*

Description

An experiment was conducted to determine the effect of gamma radiation on the numbers of chromosomal abnormalities observed

Format

A data frame with 27 observations on the following 4 variables.

cells Number of cells in hundreds

ca Number of chromosomal abnormalities

doseamt amount of dose in Grays

doserate rate of dose in Grays/hour

Source

Purott R. and Reeder E. (1976) The effect of changes in dose rate on the yield of chromosome aberrations in human lymphocytes exposed to gamma radiation. Mutation Research. 35, 437-444.

References

Frome E. and DuFrain R. (1986) Maximum Likelihood Estimation for Cytogenic Dose-Response Curves. Biometrics. 42, 73-84.

---

divusa *Divorce in the USA 1920-1996*

Description

Divorce rates in the USA from 1920-1996

Format

A data frame with 77 observations on the following 7 variables.

year the year from 1920-1996

divorce divorce per 1000 women aged 15 or more

unemployed unemployment rate

femlab percent female participation in labor force aged 16+

marriage marriages per 1000 unmarried women aged 16+

birth births per 1000 women aged 15-44

military military personnel per 1000 population

Source

Unknown

---

drugpsy *Choice of drug treatment for psychiatry patients*

Description

A sample of psychiatry patients were cross-classified by their diagnosis and whether a drug treatment was prescribed.

Format

A data frame with 10 observations on the following 3 variables.

y the number of patients

diagnosis a factor with levels `Affective.Disorder` `Neurosis` `Personality.Disorder` `Schizophrenia` `Special.Symptoms`

drug a factor with levels `no` `yes`

Source

Helmes E. and Fekken G. (1986) Effects of psychotropic drugs and psychiatric illness on vocational aptitude and interest assessment. J. Clin. Psychol. **42** 569-576

References

Agresti A. (1990) "Categorical Data Analysis" Wiley

`dvisits` *Doctor visits in Australia*

Description

The data come from the Australian Health Survey of 1977-78 and consist of 5190 single adults where young and old have been oversampled.

Format

A data frame with 5190 observations on the following 19 variables.

sex 1 if female, 0 if male

age Age in years divided by 100 (measured as mid-point of 10 age groups from 15-19 years to 65-69 with 70 or more coded treated as 72)

agesq age squared

income Annual income in Australian dollars divided by 1000 (measured as mid-point of coded ranges Nil, less than 200, 200-1000, 1001-, 2001-, 3001-, 4001-, 5001-, 6001-, 7001-, 8001-10000, 10001-12000, 12001-14000, with 14001- treated as 15000

levyplus 1 if covered by private health insurance fund for private patient in public hospital (with doctor of choice), 0 otherwise

freepoor 1 if covered by government because low income, recent immigrant, unemployed, 0 otherwise

freerepa 1 if covered free by government because of old-age or disability pension, or because invalid veteran or family of deceased veteran, 0 otherwise

illness Number of illnesses in past 2 weeks with 5 or more coded as 5

actdays Number of days of reduced activity in past two weeks due to illness or injury

hscore General health questionnaire score using Goldberg's method. High score indicates bad health

chcond1 1 if chronic condition(s) but not limited in activity, 0 otherwise

chcond2 1 if chronic condition(s) and limited in activity, 0 otherwise

doctorco Number of consultations with a doctor or specian the past 2 weeks

nondocco Number of consultations with non-doctor health professionals (chemist, optician, physiotherapist, social worker, district community nurse, chiropodist or chiropractor) in the past 2 weeks

hospadmi Number of admissions to a hospital, psychiatric hospital, nursing or convalescent home in the past 12 months (up to 5 or more admissions which is coded as 5)

hospdays Number of nights in a hospital, etc. during most recent admission: taken, where appropriate, as the mid-point of the intervals 1, 2, 3, 4, 5, 6, 7, 8-14, 15-30, 31-60, 61-79 with 80 or more admissions coded as 80. If no admission in past 12 months then equals zero

medicine Total number of prescribed and nonprescribed medications used in past 2 days

prescrib Total number of prescribed medications used in past 2 days

nonpresc Total number of nonprescribed medications used in past 2 days

Source

Cameron A, Trivedi P, Milne F and Piggot J (1988) A Microeconometric model of the demand for health care and health insurance in Australia, Review of Economic Studies 55, 85-106

---

eco *Ecological regression example*

Description

Relationship between 1998 per capita income dollars from all sources and the proportion of legal state residents born in the United States in 1990 for each of the 50 states plus the District of Columbia

Format

This dataframe contains the following columns

usborn Percentage of population born in the United States

income Per capita annual income in dollars

home Percentage born in state

pop Population of state

Source

US Bureau of the Census

eggprod *Treatment and block effects on egg production*

Description

The eggprod data frame has 12 rows and 3 columns. Six pullets were placed into each of 12 pens. Four blocks were formed from groups of 3 pens based on location. Three treatments were applied. The number of eggs produced was recorded

Format

This data frame contains the following columns:

treat Three treatments: O, E or F

block Four blocks labeled 1-4

eggs Number of eggs produced

Source

Mead, R., R.N. Curnow, and A.M. Hasted. 1993. Statistical Methods in Agriculture and Experimental Biology. Chapman and Hall, London, p. 64. 1993

eggs *Nested data on lab testing of eggs*

Description

Consistency between laboratory tests is important and yet the results may depend on who did the test and where the test was performed. In an experiment to test levels of consistency, a large jar of dried egg powder was divided up into a number of samples. Because the powder was homogenized, the fat content of the samples is the same, but this fact is withheld from the laboratories. Four samples were sent to each of six laboratories. Two of the samples were labeled as G and two as H, although in fact they were identical. The laboratories were instructed to give two samples to two different technicians. The technicians were then instructed to divide their samples into two parts and measure the fat content of each. So each laboratory reported eight measures, each technician four measures, that is, two replicated measures on each of two samples.

Format

A data frame with 48 observations on the following 4 variables.

Fat a numeric vector

Lab a factor with levels I II III IV V VI

Technician a factor with levels one two

Sample a factor with levels G H

Source

Bliss, C. I. (1967). Statistics in Biology. New York: McGraw Hill.

Examples

```
data(eggs)
## maybe str(eggs) ; plot(eggs) ...
```

`epilepsy` *Epileptic seizures in clinical trial of drug*

Description

Data from a clinical trial of 59 epileptics. For a baseline, patients were observed for 8 weeks and the number of seizures recorded. The patients were then randomized to treatment by the drug Progabide (31 patients) or to the placebo group (28 patients). They were observed for four 2-week periods and the number of seizures recorded.

Format

A data frame with 295 observations on the following 6 variables.

seizures number of seizures

id identifying number

treat 1=treated, 0=not

expind 0=baseline period, 1=treatment period

timeadj weeks of period

age in years

Source

Thall, P. F. and S. C. Vail (1990). Some covariance models for longitudinal count data with overdispersion. Biometrics 46, 657-671.

References

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 88, 9-25. Diggle, P. J., P. Heagerty, K. Y. Liang, and S. L. Zeger (2002). Analysis of Longitudinal Data (2 ed.). Oxford: Oxford University Press.

Examples

```
data(epilepsy)
## maybe str(epilepsy) ; plot(epilepsy) ...
```

esdcomp *Complaints about emergency room doctors*

Description

Data was recorded on 44 doctors working in an emergency service at a hospital to study the factors affecting the number of complaints received.

Format

A data frame with 44 observations on the following 6 variables.

visits the number of patient visits

complaints the number of complaints

residency is the doctor in residency training `N` or `Y`

gender gender of doctor `F` or `M`

revenue dollars per hour earned by the doctor

hours total number of hours worked

Source

Chap T. Le (1998) "Applied Categorical Data Analysis" Wiley

exa *Simulated non-parametric regression data*

Description

True function is f(x)=sin^3(2pi x^3).

Format

A data frame with 256 observations on the following 3 variables.

x input

y response

m true value

Source

Haerdle, W. (1991). Smoothing Techniques with Implementation in S. New York:Springer.

Examples

```
data(exa)
## maybe str(exa) ; plot(exa) ...
```

exb *Simulated non-parametric regression data*

Description

True function is f(x)=0

Format

A data frame with 256 observations on the following 3 variables.

x input

y response

m true value

Source

Haerdle, W. (1991). Smoothing Techniques with Implementation in S. New York:Springer.

Examples

```
data(exa)
## maybe str(exa) ; plot(exa) ...
```

eyegrade *grading of eye pairs for distance vision*

Description

A sample of women are rated for the performance of distance vision in each eye.

Format

A data frame with 16 observations on the following 3 variables.

y the observed count

right rated vision in the right eye - a factor with levels `best` `second` `third` `worst`

left rated vision in the left eye - a factor with levels `best` `second` `third` `worst`

Source

A. Stuart (1955) A test for homogeneity of the marginal distributions in a two-way classification, Biometrika, 42, 412-416.

`fat` *Percentage of Body Fat and Body Measurements in Men*

Description

Age, weight, height, and 10 body circumference measurements are recorded for 252 men. Each man's percentage of body fat was accurately estimated by an underwater weighing technique.

Format

A data frame with 252 observations on the following 18 variables.

brozek Percent body fat using Brozek's equation, 457/Density - 414.2

siri Percent body fat using Siri's equation, 495/Density - 450

density Density (gm/cm^3)

age Age (yrs)

weight Weight (lbs)

height Height (inches)

adipos Adiposity index = Weight/Height^2 (kg/m^2)

free Fat Free Weight = (1 - fraction of body fat) * Weight, using Brozek's formula (lbs)

neck Neck circumference (cm)

chest Chest circumference (cm)

abdom Abdomen circumference (cm) at the umbilicus and level with the iliac crest

hip Hip circumference (cm)

thigh Thigh circumference (cm)

knee Knee circumference (cm)

ankle Ankle circumference (cm)

biceps Extended biceps circumference (cm)

forearm Forearm circumference (cm)

wrist Wrist circumference (cm) distal to the styloid processes

Source

Johnson R. Journal of Statistics Education v.4, n.1 (1996)

femsmoke *Mortality due to smoking according age group in women*

Description

In 1972-74, a survey of one in six residents of Whickham, near Newcastle, England was made. Twenty years later, this data recorded in a follow-up study. Only women who are current smokers or who have never smoked are included.

Format

A data frame with 28 observations on the following 4 variables.

y observed count for given combination

smoker a factor with levels `yes` `no`

dead a factor with levels `yes` `no`

age a factor with agegroup levels `18-24` `25-34` `35-44` `45-54` `55-64` `65-74` `75+`

Source

D. Appleton, J. French, M. Vanderpump (1996) "Ignoring a Covariate: An Example of Simpson's Paradox" American Statistician, 50, 340-341

fortune *Billionaires' wealth and age*

Description

Fortune magazine publishes a f the world's billionaires each year. The 1992 list includes 233 individuals. Their wealth, age, and geographic location (Asia, Europe, Middle East, United States, and Other) are reported.

Format

A data frame with 232 observations on the following 3 variables.

wealth Billions of dollars

age age in years

region a factor with levels `A`, Asia, `E`, Europe, `M`, Middle East, `O` Other, `U` USA

Source

Fortune magazine

Examples

```
data(fortune)
## maybe str(fortune) ; plot(fortune) ...
```

---

fpe *1981 French Presidential Election*

Description

Elections for the French presidency proceed in two rounds. In 1981, there were 10 candidates in the first round. The top two candidates then went on to the second round, which was won by Francois Mitterand over Valery Giscard-d'Estaing. The losers in the first round can gain political favors by urging their supporters to vote for one of the two fina Since voting is private, we cannot know how these votes were transferred, we might hope to infer from the published vote totals how this might have happened. Data is given for vote totals in every fourth department of France:

Format

This dataframe contains the following columns (vote totals are in thousands)

list("EI Electeur Inscrits (registered voters)

A Voters for Mitterand in the first round

B Voters for Giscard in the first round

C Voters for Chirac in the first round

D Voters for Communists in the first round

E Voters for Ecology party in the first round

F Voters for party F in the first round

G Voters for party G in the first round

H Voters for party H in the first round

I Voters for party I in the first round

J Voters for party J in the first round

K Voters for party K in the first round

A2 Voters for Mitterand in the second round

B2 Voters for party Giscard in the second round

N Difference between the number of voters in the second round and in the first round

Source

"The Teaching of Practical Statistics" by C.W. Anderson and R.M. Loynes, Wiley,1987

`fround` *Formating the Rounding of Numbers*

Description

`fround` rounds the values in its first argument to the specified number of decimal places with surrounding quotes.

Usage

```
fround(x, digits)
```

Arguments

`x`	a numeric vector.
`digits`	integer indicating the precision to be used.

Details

`pfround` rounds the values in its first argument to the specified number of decimal places without surrounding quotes.

Author(s)

Andrew Gelman; Yu-Sung Su

References

Copied from the `arm` package

See Also

`round`

Examples

```
x <- 3.1415926
fround(x, digits=2)
pfround(x, digits=2)
```

fruitfly *Longevity of fruitflies depending on sexual activity and thorax length*

Description

The `fruitfly` data frame has 9 rows and 3 columns. 125 fruitflies were divided randomly into 5 groups of 25 each. The response was the longevity of the fruitfly in days. One group was kept solitary, while another was kept individually with a virgin female each day. Another group was given 8 virgin females per day. As an additional control the fourth and fifth groups were kept with one or eight pregnant females per day. Pregnant fruitflies will not mate. The thorax length of each male was measured as this was known to affect longevity. One observation in the many group has been lost.

Format

This data frame contains the following columns:

thorax Thorax length

longevity Lifetime in days

activity The group: isolated = fly kept solitary, one = fly kept with one pregnant fruitfly, many = fly kept with eight pregnant fruitflies, low= fly kept with one virgin fruitfly, high = fly kept with eight virgin fruitflies.

Source

"Sexual Activity and the Lifespan of Male Fruitflies" by L. Partridge and M. Farquhar, Nature, 1981, 580-581

gala *Species diversity on the Galapagos Islands*

Description

There are 30 Galapagos islands and 7 variables in the dataset. The relationship between the number of plant species and several geographic variables is of interest. The original dataset contained several missing values which have been filled for convenience. See the `galamiss` dataset for the original version.

Format

The dataset contains the following variables

Species the number of plant species found on the island

Endemics the number of endemic species

Area the area of the island (km^2)

Elevation the highest elevation of the island (m)

Nearest the distance from the nearest island (km)

Scruz the distance from Santa Cruz island (km)

Adjacent the area of the adjacent island (square km)

Source

M. P. Johnson and P. H. Raven (1973) "Species number and endemism: The Galapagos Archipelago revisited" Science, 179, 893-895

`galamiss` *Species diversity on the Galapagos Islands*

Description

There are 30 Galapagos islands and 7 variables in the dataset. The relationship between the number of plant species and several geographic variables is of interest. This is the original version of the dataset containing missing values.

Format

The dataset contains the following variables

Species the number of plant species found on the island

Endemics the number of endemic species

Area the area of the island (km^2)

Elevation the highest elevation of the island (m)

Nearest the distance from the nearest island (km)

Scruz the distance from Santa Cruz island (km)

Adjacent the area of the adjacent island (square km)

Source

M. P. Johnson and P. H. Raven (1973) "Species number and endemism: The Galapagos Archipelago revisited" Science, 179, 893-895

gammaray *Xray decay from a gamma ray burst*

Description

The X-ray decay light curve of Gamma ray burst 050525a obtained with the X-Ray Telescope (XRT) on board the Swift satellite. The dataset has 63 brightness measurements in the 0.4-4.5 keV spectral band at times ranging from 2 minutes to 5 days after the burst.

Format

A data frame with 63 observations on the following 3 variables.

time in seconds since burst

flux X-ray flux in units of 10^-11 erg/cm2/s, 2-10 keV

error measurement error of the flux based on detector signal-to-noise values

Source

A. J. Blustin and 64 coauthors, Astrophys. J. 637, 901-913 2006. Available at http://arxiv.org/abs/astro-ph/0507515.

Examples

```
data(gammaray)
## maybe str(gammaray) ; plot(gammaray) ...
```

gavote *Undercounted votes in Georgia in 2000 presidential election*

Description

The data comes from the US presidential election in the state of Georgia. The undercount is the difference between the number of ballots cast and votes recorded. Voters may have chosen not to vote for president, voted for more than one candidate (disqualified) or the equipment may have failed to register their choice.

Format

A data frame with 159 observations on the following 10 variables. Each case represents a county in Georgia.

equip The voting equipment used: `LEVER`, `OS-CC` (optical, central count), `OS-PC` (optical, precinct count) `PAPER`, `PUNCH`

econ economic status of county: `middle` `poor` `rich`

perAA percent of African Americans in county

rural indicator of whether county is `rural` or `urban`

atlanta indicator of whether county is in `Atlanta` or not: `notAtlanta`

gore number of votes for Gore

bush number of votes for Bush

other number of votes for other candidates

votes number of votes

ballots number of ballots

Source

Meyer M. (2002) Uncounted Votes: Does Voting Equipment Matter? Chance, 15(4), 33-38

`globwarm` *Northern Hemisphere temperatures and climate proxies in the last millenia*

Description

Average Northen Hemisphere Temperature from 1856-2000 and eight climate proxies from 1000-2000AD. Data can be used to predict temperatures prior to 1856.

Format

A data frame with 1001 observations on the following 10 variables.

nhtemp Northern hemisphere average temperature (C) provided by the UK Met Office (known as HadCRUT2)

wusa Tree ring proxy information from the Western USA.

jasper Tree ring proxy information from Canada.

westgreen Ice core proxy information from west Greenland

chesapeake Sea shell proxy information from Chesapeake Bay

tornetrask Tree ring proxy information from Sweden

urals Tree ring proxy information from the Urals

mongolia Tree ring proxy information from Mongolia

tasman Tree ring proxy information from Tasmania

year Year 1000-2000AD

Details

See the source and references below for the original data. Only some proxies have been included here. Some missing values have been imputed. The proxy data have been smoothed. This version of the data is intended only for demonstration purposes. If you are specifically interested in the subject matter, use the original data.

Source

P.D. Jones and M.E. Mann (2004) "Climate Over Past Millennia" Reviews of Geophysics, Vol. 42, No. 2, RG2002, doi:10.1029/2003RG000143

References

www.ncdc.noaa.gov/paleo/pubs/jones2004/jones2004.html

Examples

```
data(globwarm)
## maybe str(globwarm) ; plot(globwarm) ...
```

---

haireye *Hair and eye color*

Description

Data collected from 592 students in an introductory statistics class

Format

A data frame with 16 observations on the following 3 variables.

y count of the number of student with given hair/eye combination

eye a factor with levels `green` `hazel` `blue` `brown`

hair a factor with levels `BLACK` `BROWN` `RED` `BLOND`

Source

Snee R. (1974) Graphical display of two-way contingency tables. American Statistician, 28, 9-12

---

halfnorm *Half Normal Plot*

Description

Makes a half-normal plot

Usage

```
halfnorm(
  x,
  nlab = 2,
  labs = as.character(1:length(x)),
  ylab = "Sorted Data",
  ...
)
```

Arguments

x	a numeric vector
nlab	number of points to label
labs	labels for points
ylab	label for Y-axis
...	arguments passed to plot()

Value

none

Author(s)

Julian Faraway

See Also

qqnorm

Examples

```
halfnorm(runif(10))
```

---

happy *MBA students experience love, sex, work and happiness*

Description

Data were collected from 39 students in a University of Chicago MBA class

Usage

```
mba
```

Format

A data frame with 39 observations on the following 5 variables.

happy Happiness on a 10 point scale where 10 is most happy

money family income in thousands of dollars

sex 1 = satisfactory sexual activity, 0 = not

love 1 = lonely, 2 = secure relationships, 3 = deep feeling of belonging and caring

work 5 point scale where 1 = no job, 3 = OK job, 5 = great job

An object of class `data.frame` with 39 rows and 5 columns.

Source

George and McCulloch (1993) "Variable Selection via Gibbs Sampling" JASA, 88, 881-889

---

hemoglobin *Treatment of insulin dependent diabetic children*

Description

16 insulin-dependent diabetic children were enrolled in a study involving a new treatment. 8 children received the new treatment(N) while the other 8 received the standard treatment(S). The age and sex of the child was recorded along with the measured value of gycosolated hemoglobin both before and after treatment.

Format

A data frame with 16 observations on the following 5 variables.

age age in years

sex a factor with levels `F M`

treatment a factor with levels `N S`

pre measured value of hemoglobin before treatment

post measured value of hemoglobin after treatment

Source

Unknown

Examples

```
data(hemoglobin)
## maybe str(hemoglobin) ; plot(hemoglobin) ...
```

hips *Ankylosing Spondylitis*

Description

Data from Royal Mineral Hospital in Bath. AS is a chronic form of arthritis. A study conducted to determine whether daily stretching of the hip tissues would improve mobility. 39 "typical" AS patients were randomly allocated to control (standard treatment) group or the treatment group in a 1:2 ratio. Responses were flexion and rotation angles at the hip measured in degrees. Larger numbers indicate more flexibility.

Format

A data frame with 78 observations on the following 7 variables.

fbef flexion angle before

faft flexion angle after

rbef rotation angle before

raft rotation angle after

grp treatment group - a factor with levels `control` `treat`

side side of the body - a factor with levels `right` `left`

person id for the individual

Source

Chatfield C. (1995) Problem Solving: A Statistician's Guide, 2ed Chapman Hall.

Examples

```
data(hips)
## maybe str(hips) ; plot(hips) ...
```

hormone *Hormone concentrations in gay and straight men*

Description

Urinary androsterone (androgen) and etiocholanolone (estrogen) values were recorded from 26 healthy males.

Format

A data frame with 26 observations on the following 3 variables.

androgen concentration

estrogen concentration

orientation sexual orientation with levels `g` `s`

Source

Margolese, M. (1970). Homosexuality: A new endocrine correlate. Hormones and Behavior 1, 151-155.

References

Hand, D. (1981). Discrimination and Classification. Chichester, UK: Wiley.

Examples

```
data(hormone)
## maybe str(hormone) ; plot(hormone) ...
```

---

hprice *Housing prices in US cities 86-94*

Description

Data on housing prices in 36 US metropolitan statistical areas (MSAs) over 9 years from 1986-1994 were collected.

Format

A data frame with 324 observations on the following 8 variables.

narsp natural log average sale price in thousands of dollars

ypc average per capita income

perypc percentage growth in per capita income

regtest Regulatory environment index (high values = more regulations)

rcdum Rent control - a factor with levels `0`=no `1`=yes

ajwtr Adjacent to a coastline - a factor with levels `0`=no `1`=yes

msa indicator for the MSA

time Year 1=1986 to 9=1994

Source

Longitudinal and Panel Data: Analysis and Applications in the Social Sciences, by Edward W. Frees, Cambridge University Press, August 2004.

hsb *Career choice of high school students*

Description

Data was collected as a subset of the "High School and Beyond" study conducted by the National Education Longitudinal Studies (NELS) program of the National Center for Education Statistics (NCES).

Format

A data frame with 200 observations on the following 11 variables.

id ID of student

gender a factor with levels `female` `male`

race a factor with levels `african-amer` `asian` `hispanic` `white`

ses socioeconomic class - a factor with levels `high` `low` `middle`

schtyp school type - a factor with levels `private` `public`

prog choice of high school program - a factor with levels `academic` `general` `vocation`

read reading score

write writing score

math math score

science science score

socst social science score

Details

One purpose of the study was to determine which factors are related to the choice of the type of program, academic, vocational or general, that the students pursue in high school.

Source

National Education Longitudinal Studies (NELS) program of the National Center for Education Statistics (NCES).

ilogit *Inverse Logit Transformation*

Description

Computes the inverse logit transformation

Usage

```
ilogit(x)
```

Arguments

x	a numeric vector

Value

exp(x)/(1+exp(x))

Author(s)

Julian Faraway

See Also

logit

Examples

```
ilogit(1:3)
#[1] 0.7310586 0.8807971 0.9525741
```

infmort *Infant mortality according to income and region*

Description

The `infmort` data frame has 105 rows and 4 columns. The infant mortality in regions of the world may be related to per capita income and whether oil is exported. The dataset is not recent.

Format

This data frame contains the following columns:

region Region of the world, Africa, Europe, Asia or the Americas

income Per capita annual income in dollars

mortality Infant mortality in deaths per 1000 births

oil Does the country export oil or not?

Source

Unknown

`insulgas` *Effects of insulation on gas consumption*

Description

Data on natural gas usage in a house. The weekly gas consumption (in 1000 cubic feet) and the average outside temperature (in degrees Celsius) was recorded for 26 weeks before and 30 weeks after cavity-wall insulation had been installed. The house thermostat was set at 20C throughout.

Format

A data frame with 44 observations on the following 3 variables.

Insulate a factor with levels `After` `Before`

Temp Outside temperature

Gas Weekly consumption in 1000 cubic feet

Source

MASS package as whiteside

Examples

```
data(insulgas)
## maybe str(insulgas) ; plot(insulgas) ...
```

irrigation *Irrigation methods in an agricultural field trial*

Description

In an agricultural field trial, the objective was to determine the effects of two crop varieties and four different irrigation methods. Eight fields were available, but only one type of irrigation may be applied to each field. The fields may be divided into two parts with a different variety planted in each half. The whole plot factor is the method of irrigation, which should be randomly assigned to the fields. Within each field, the variety is randomly assigned.

Format

A data frame with 16 observations on the following 4 variables.

field a factor with levels `f1` `f2` `f3` `f4` `f5` `f6` `f7` `f8`

irrigation a factor with levels `i1` `i2` `i3` `i4`

variety a factor with levels `v1` `v2`

yield a numeric vector

Source

Found online but source not recorded.

Examples

```
data(irrigation)
## maybe str(irrigation) ; plot(irrigation) ...
```

jsp *Junior School Project*

Description

Junior School Project collected from primary (U.S. term is elementary) schools in inner London.

Format

A data frame with 3236 observations on the following 9 variables.

school 50 schools code 1-50

class a factor with levels `1` `2` `3` `4`

gender a factor with levels `boy` `girl`

social class of the father I=1; II=2; III nonmanual=3; III manual=4; IV=5; V=6; Long-term unemployed=7; Not currently employed=8; Father absent=9

raven test score

id student id coded 1-1402

english score on English

math score on Maths

year year of school

Source

Mortimore, P., P. Sammons, L. Stoll, D. Lewis, and R. Ecob (1988). School Matters. Wells, UK: Open Books.

References

Goldstein, H. (1995). Multilevel Statistical Models (2 ed.). London: Arnold.

Examples

```
data(jsp)
## maybe str(jsp) ; plot(jsp) ...
```

---

kanga *Kangaroo skull measurements*

Description

Sex and species of an specimens of kangaroo.

Format

A data frame with 148 observations on the following 20 variables.

species a factor with levels `fuliginosus` `giganteus` `melanops`

sex a factor with levels `Female` `Male`

basilar.length a numeric vector

occipitonasal.length a numeric vector

palate.length a numeric vector

palate.width a numeric vector

nasal.length a numeric vector

nasal.width a numeric vector

squamosal.depth a numeric vector

lacrymal.width a numeric vector

zygomatic.width a numeric vector
orbital.width a numeric vector
rostral.width a numeric vector
occipital.depth a numeric vector
crest.width a numeric vector
foramina.length a numeric vector
mandible.length a numeric vector
mandible.width a numeric vector
mandible.depth a numeric vector
ramus.height a numeric vector

Source

Andrews and Herzberg (1985) Chapter 53.

References

Andrews, D. F. and Herzberg, A. M. (1985). Data. Springer-Verlag, New York.

Examples

```
data(kanga)
## maybe str(kanga) ; plot(kanga) ...
```

---

lawn *Cut-off times of lawnmowers*

Description

Data on the cut-off times of lawnmowers was collected. 3 machines were randomly selected from those produced by manufacturers A and B. Each machine was tested twice at low speed and high speed.

Format

A data frame with 24 observations on the following 4 variables.

manufact Manufacturer - a factor with levels `A` `B`
machine Lawn mower - a factor with levels `m1` `m2` `m3` `m4` `m5` `m6`
speed Speed of testing - a factor with levels `H` `L`
time cut-off time

Source

Unknown.

leafblotch *Leaf blotch on barley*

Description

The data gives the proportion of leaf area affected by leaf blotch on 10 varieties of barley at 9 different sites.

Format

A data frame with 90 observations on the following 3 variables.

blotch proportion of the barley leaf affected by blotch

site the physical location - a factor with levels `1 2 3 4 5 6 7 8 9`

variety variety of barley - a factor with levels `1 2 3 4 5 6 7 8 9 10`

Source

R. W. M. Wedderburn (1974) "Quasilikelihood functions, generalized linear models and the Gauss-Newton method" Biometrika, 61, 439-447.

References

P. McCullagh and J. Nelder (1989) "Generalized Linear Models" Chapman and Hall, 2nd ed.

leafburn *Data on the burning time of samples of tobacco leaves*

Description

Data on the burning time of samples of tobacco leaves

Format

A data frame with 30 observations on the following 4 variables.

nitrogen nitrogen content by percentage weight

chlorine chlorine content by percentage weight

potassium potassium content by percentage weight

burntime burn time in seconds

Source

Steel, R. G. D. and Torrie, J. H. (1980), Principles and Procedures of Statistics, Second Edition, New York: McGraw-Hill

`logit` *Logit transformation*

Description

Computes the logit transformation

Usage

```
logit(x)
```

Arguments

x	a numeric vector

Details

x <=0 or >=1 will return NA

Value

log(x/(1-x))

Author(s)

Julian Faraway

See Also

ilogit

Examples

```
logit(c(0.1,0.5,1.0,1.1))
#[1] -2.197225  0.000000        NA        NA
```

mammalsleep *Sleep in Mammals: Ecological and Constitutional Correlates*

Description

The mammalsleep data frame has 62 rows and 10 columns. Sleep in Mammals: Ecological and Constitutional Correlates

Format

This data frame contains the following columns:

body body weight in kg

brain brain weight in g

nondream slow wave ("nondreaming") sleep (hrs/day)

dream paradoxical ("dreaming") sleep (hrs/day)

sleep total sleep (hrs/day) (sum of slow wave and paradoxical sleep)

lifespan maximum life span (years)

gestation gestation time (days)

predation predation index (1-5) 1 = minimum (least likely to be preyed upon) to 5 = maximum (most likely to be preyed upon)

exposure sleep exposure index (1-5) 1 = least exposed (e.g. animal sleeps in a well-protected den) 5 = most exposed

danger overall danger index (1-5) (based on the above two indices and other information) 1 = least danger (from other animals) 5 = most danger (from other animals)

Source

"Sleep in Mammals: Ecological and Constitutional Correlates" by Allison, T. and Cicchetti, D. (1976), Science, November 12, vol. 194, pp. 732-734.

manilius *Mayer's 1750 data on the Manilius crater on the moon*

Description

In 1750, Tobias Mayer collected data on various landmarks on the moon in order to determine its orbit. The data involving the position of the Manilius crater resulted in a least squares like problem. The example is discussed in Steven Stigler's History of Statistics.

Format

A data frame with 27 observations on the following 4 variables.

arc an angle known as h in Stigler's notation

sinang the sin(g-k) where g and k are two angles in Stigler

cosang the cos(g-k) where g and k are two angles in Stigler

group one of three groups determined by Mayer

Details

See Stigler for a detailed description.

Source

Stigler, S. (1986) History of Statistics. Belknap Press, Harvard.

References

Mayer, T. (1750) Abhandlung uber die Umwaltzung des Monds um seine Axe und die scheinbare Bewegung der Mondsflecken published in the Kosmographische Nachrichten und Sammlungen auf das Jahr 1748. 52-183

Examples

```
data(manilius)
```

---

maxadjr *Maximum Adjusted R-squared*

Description

Displays the best models from a leaps object

Usage

```
maxadjr(l, best = 3)
```

Arguments

`l`	A leaps object returned from leaps()
`best`	An optional argument specify the number of models to be returned taking the default value of 3

Details

Requires leaps package

Value

A list of the best models

Author(s)

Julian Faraway

See Also

leaps()

---

meatspec *Meat spectrometry to determine fat content*

Description

A Tecator Infratec Food and Feed Analyzer working in the wavelength range 850 - 1050 nm by the Near Infrared Transmission (NIT) principle was used to collect data on samples of finely chopped pure meat. 215 samples were measured. For each sample, the fat content was measured along with a 100 channel spectrum of absorbances. Since determining the fat content via analytical chemistry is time consuming we would like to build a model to predict the fat content of new samples using the 100 absorbances which can be measured more easily.

Format

Dataset contains the following variables

V1-V100 absorbances across a range of 100 wavelengths

fat fat content

Source

H. H. Thodberg (1993) "Ace of Bayes: Application of Neural Networks With Pruning", report no. 1132E, Maglegaardvej 2, DK-4000 Roskilde, Danmark

melanoma *Melanoma by type and location*

Description

Data comes from a study of Malignant Melanoma involving 400 subjects.

Format

A data frame with 12 observations on the following 3 variables.

count number of cases

tumor type of tumor - a factor with levels `freckle` `indeterminate` `nodular` `superficial`

site location of tumor on the body - a factor with levels `extremity` `head` `trunk`

Source

Dobson A. (2002) An introduction to generalized linear models, Chapman Hall.

motorins *Third party motor insurance claims in Sweden in 1977*

Description

In Sweden all motor insurance companies apply identical risk arguments to classify customers, and thus their portfolios and their claims statistics can be combined. The data were compiled by a Swedish Committee on the Analysis of Risk Premium in Motor Insurance. The Committee was asked to look into the problem of analyzing the real influence on claims of the risk arguments and to compare this structure with the actual tariff.

Format

A data frame with 1797 observations on the following 8 variables.

Kilometres an ordered factor representing kilomoters per year with levels 1: < 1000, 2: 1000-15000, 3: 15000-20000, 4: 20000-25000, 5: > 25000

Zone a factor representing geographical area with levels 1: Stockholm, Goteborg, Malmo with surroundings 2: Other large cities with surroundings 3: Smaller cities with surroundings in southern Sweden 4: Rural areas in southern Sweden 5: Smaller cities with surroundings in northern Sweden 6: Rural areas in northern Sweden 7: Gotland

Bonus No claims bonus. Equal to the number of years, plus one, since last claim

Make A factor representing eight different common car models. All other models are combined in class 9

Insured Number of insured in policy-years

Claims Number of claims

Payment Total value of payments in Skr

perd payment per claim

Source

http://www.statsci.org/data/general/motorins.html

References

Hallin, M., and Ingenbleek, J.-F. (1983). The Swedish automobile portfolio in 1977. A statistical study. Scandinavian Actuarial Journal, 49-64.

---

neighbor *Questionnaire study of neighborly help*

Description

Subjects were asked questions in a study of neighborly help. Questions below are a subset of the full study.

Format

A data frame with 181 observations on the following 8 variables.

longlive About how long have you lived where you do now? Ans is a factor with levels `<6mos` `6-12mos` `1-3yrs` `3-10yrs` `10yrs`

wherebfr Where were you living before you moved to your present house? Ans is a factor with levels `same` `Exeter` `Devon` `Britain` `Abroad`

hownbly How neighborly do you think the area where you now live is? Ans is a factor with levels `Vunfriendly` `NVfriendly` `Average` `FFriendly` `VFriendly`

knowname Roughly how many people in your street, or in the streets just near you, do you know the names of? Ans is a factor with levels `none` `1-5` `6-20` `20+`

callname How many of those people (not counting children) would you call by their first names? Ans is a factor with levels `none` `1-5` `6-20` `20+`

age a factor with levels `-18` `18-30` `31-50` `51-65` `65+`

district a factor with levels `1` `2` `3` `4`

sex a factor with levels `female` `male`

Details

Exeter is a city in the county of Devon which is in Britain. The four districts can be briefly described as follows. District 1 was a long-established residential area near the city centre, with housing dating from the late nineteenth century. Originally working class, it now has a considerable middle class population with some student and other temporary accommodation. District 2 was a working-class housing estate dating from the 1930s, with mainly rented accommodation but some owner occupation. District 3 was the oldest part of a more recently developed, mainly middle-class, almost exclusively owner-occupied estate, dating from the 1960s. District 4 was the most recently developed part of a more sought-after middle-class residential area, with smaller but almost entirely owner-occupied properties dating from the 1970s and 1980s.

Source

P. Webley & S. Lea 1993, Human Relations 46, 65-76.

`nels88` *National Education Longitudinal Study of 1988*

Description

A subset of the National Education Longitudinal Study of 1988

Format

A data frame with 260 observations on the following 5 variables.

sex a factor with levels `Female` `Male`

race a factor with levels `White` `Asian` `Black` `Hispanic`

ses a numeric vector

paredu a factor with levels `ba` `college` `hs` `lesshs` `ma` `phd`

math a numeric vector

Source

http://www.icpsr.umich.edu/icpsrweb/ICPSR/series/107

Examples

```
data(nels88)
## maybe str(nels88) ; plot(nels88) ...
```

`nepali` *Nepali child heath study*

Description

The data are a subset from public health study on Nepalese children.

Format

A data frame with 1000 observations on the following 9 variables.

id There is a six digit code for the child's ID: 2 digits for the panchayat number; 2 digits for the ward within panchayat; 1 digits for the household; 1 digit for child within household.

sex 1 = male; 2 = female

wt Child's weight measured in kilograms

ht Child's height measured in centimeters

mage Mother's age in years

lit Indicator of mother's literacy: 0 = no; 1 = yes

died The number of children the mother has had that died.

alive The number of children the mother has ever had born alive

age age of child

Source

West KP, Jr., LeClerq SC, Shrestha SR, Wu LS, Pradhan EK, Khatry SK, Katz J, Adhikari R, Sommer A. Effects of vitamin A on growth of vitamin A deficient children: field studies in Nepal. J Nutr 1997;10:1957-1965.

---

nes96 *US 1996 national election study*

Description

10 variable subset of the 1996 American National Election Study. Missing values and "don't know" responses have been se deleted. Respondents expressing a voting preference other than Clinton or Dole have been removed.

Format

A data frame with 944 observations on the following 10 variables.

popul population of respondent's location in 1000s of people

TVnews days in the past week spent watching news on TV

selfLR Left-Right self-placement of respondent: an ordered factor with levels extremely liberal, `extLib` < liberal, `Lib` < slightly liberal, `sliLib` < moderate, `Mod` < slightly conservative, `sliCon` < conservative, `Con` < extremely conservative, `extCon`

ClinLR Left-Right placement of Bill Clinton (same scale as selfLR): an ordered factor with levels `extLib` < `Lib` < `sliLib` < `Mod` < `sliCon` < `Con` < `extCon`

DoleLR Left-Right placement of Bob Dole (same scale as selfLR): an ordered factor with levels `extLib` < `Lib` < `sliLib` < `Mod` < `sliCon` < `Con` < `extCon`

PID Party identification: an ordered factor with levels strong Democrat, `strDem` < weak Democrat, `weakDem` < independent Democrat, `indDem` < independent independent`indind` < indepedent Republican, `indRep` < waek Republican, `weakRep` < strong Republican, `strRep`

age Respondent's age in years

educ Respondent's education: an ordered factor with levels 8 years or less, `MS` < high school dropout, `HSdrop` < high school diploma or GED, `HS` < some College, `Coll` < Community or junior College degree, `CCdeg` < BA degree, `BAdeg` < postgraduate degree, `MAdeg`

income Respondent's family income: an ordered factor with levels `$3Kminus` < `$3K-$5K` < `$5K-$7K` < `$7K-$9K` < `$9K-$10K` < `$10K-$11K` < `$11K-$12K` < `$12K-$13K` < `$13K-$14K` < `$14K-$15K` < `$15K-$17K` < `$17K-$20K` < `$20K-$22K` < `$22K-$25K` < `$25K-$30K` < `$30K-$35K` < `$35K-$40K` < `$40K-$45K` < `$45K-$50K` < `$50K-$60K` < `$60K-$75K` < `$75K-$90K` < `$90K-$105K` < `$105Kplus`

vote Expected vote in 1996 presidential election: a factor with levels `Clinton` and `Dole`

Source

Sapiro, Virginia, Steven J. Rosenstone, Donald R. Kinder, Warren E. Miller, and the National Election Studies. AMERICAN NATIONAL ELECTION STUDIES, 1992-1997: COMBINED FILE [Computer file]. 2nd ICPSR version. Ann Arbor, MI: University of Michigan, Center for Political Studies [producer], 1999. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 1999.

References

Found at http://www.stat.washington.edu/

---

newhamp *New Hampshire Democratic Party Primary 2008*

Description

Votes and other demographic information from 276 wards in the 2008 Democratic Party presidential primary.

Format

A data frame with 276 observations on the following 12 variables.

votesys The voting system used where H is counted by hand and D is counted by machine.

Obama The number of votes for Barack Obama.

Clinton The number of votes for Hillary Clinton.

dem The total number of votes cast in the Democratic primary (there were other candidates besides Clinton and Obama).

povrate The poverty rate as a proportion as determined by the 2000 census.

pci Per capita annual income in USD in 1999.

Dean The proportion of voters for Howard Dean in the 2004 Democratic primary.

Kerry The proportion of voters for John Kerry in the 2004 Democratic primary.

white The proportion of non-Hispanic whites according to the 2000 census.

absentee The proportion voting by absentee ballot.

population An estimate of the population from 2002.

pObama Proportion voting for Obama

Details

On the 8th January 2008, primaries to select US presidential candidates were held in New Hampshire. In the Democratic party primary, Hillary Clinton defeated Barack Obama contrary to the expectations pre-election opinion polls. Essentially two different voting technologies were used in New Hampshire. Some wards used paper ballots, counted by hand while others used optically scanned ballots, counted by machine. Among the paper ballots, Obama had more votes than Clinton while Clinton defeated Obama on just the machine counted ballots. Since the method of voting should make no causal difference to the outcome, suspicions have been raised regarding the integrity of the election.

Source

Herron, M., W. M. Jr, and J. Wand (2008). Voting Technology and the 2008 New Hampshire Primary. Wm. & Mary Bill Rts. J. 17, 351-374.

`oatvar` *Yields of oat varieties planted in blocks*

Description

Data from an experiment to compare 8 varieties of oats. The growing area was heterogeneous and so was grouped into 5 blocks. Each variety was sown once within each block and the yield in grams per 16ft row was recorded.

Format

The dataset contains the following variables

yield Yield in grams per 16ft row

block Blocks I to V

variety Variety 1 to 8

Source

"Statistical Theory in Research" by R. Anderson and T. Bancroft, McGraw Hill,1952

odor *Odor of chemical by production settings*

Description

Data from an experiment to determine the effects of column temperature, gas/liquid ratio and packing height in reducing unpleasant odor of chemical product that was being sold for household use

Format

odor Odor score

temp Temperature coded as -1, 0 and 1

gas Gas/Liquid ratio coded as -1, 0 and 1

pack Packing height coded as -1, 0 and 1

Source

"Statistical Design and Analysis of Experiments" by P. John, Macmillan, 1971

ohio *Ohio Children Wheeze Status*

Description

The ohio data frame has 2148 rows and 4 columns. The dataset is a subset of the six-city study, a longitudinal study of the health effects of air pollution.

Format

This data frame contains the following columns:

resp an indicator of wheeze status (1=yes, 0=no)

id a numeric vector for subject id

age a numeric vector of age, 0 is 9 years old

smoke an indicator of maternal smoking at the first year of the study

References

Fitzmaurice, G.M. and Laird, N.M. (1993) A likelihood-based method for analyzing longitudinal binary responses, *Biometrika* **80**: 141–151.

orings *Spache Shuttle Challenger O-rings*

Description

The 1986 crash of the space shuttle Challenger was linked to failure of O-ring seals in the rocket engines. Data was collected on the 23 previous shuttle missions. The launch temperature on the day of the crash was 31F.

Format

A data frame with 23 observations on the following 2 variables.

temp temperature at launch in degrees F

damage number of damage incidents out of 6 possible

Source

Presidential Commission on the Space Shuttle Challenger Accident, Vol. 1, 1986: 129-131.

References

S. Dalal, E. Fowlkes and B. Hoadley (1989) "Risk Analysis of the Space Shuttle: Pre-Challenger Prediction of Failure." Journal of the American Statistical Association. 84: 945-957.

ozone *Ozone in LA in 1976*

Description

A study the relationship between atmospheric ozone concentration and meteorology in the Los Angeles Basin in 1976. A number of cases with missing variables have been removed for simplicity.

Format

A data frame with 330 observations on the following 10 variables.

O3 Ozone conc., ppm, at Sandbug AFB.

vh a numeric vector

wind wind speed

humidity a numeric vector

temp temperature

ibh inversion base height

dpg Daggett pressure gradient

ibt a numeric vector

vis visibility

doy day of the year

Source

Breiman, L. and J. H. Friedman (1985). Estimating optimal transformations for multiple regression and correlation. Journal of the American Statistical Association 80, 580-598.

Examples

```
data(ozone)
## maybe str(ozone) ; plot(ozone) ...
```

`parstum` *Marijuana and parent alcohol and drug use*

Description

445 college students were classified according to both frequency of marijuana use and parental use of alcohol and psychoactive drugs.

Format

A data frame with 9 observations on the following 3 variables.

parent Number of parents using drugs or alcohol - a factor with levels `Both` `Neither` `One`

student Student usage of marijuana - a factor with levels `Never` `Occasional` `Regular`

count the number of cases

Source

Ellis, Godfrey J. and Stone, Lorene H. (1979) Marijuana Use in College: "An Evaluation of a Modeling Explanation" Youth and Society 10, 323-34

`peanut` *Carbon dioxide effects on peanut oil extraction*

Description

The `peanut` data frame has 16 rows and 6 columns. Carbon dioxide effects on peanut oil extraction

Format

This data frame contains the following columns:

press CO2 pressure - two levels low=0, high=1

temp CO2 temperature - two levels low=0, high=1

moist peanut moisture - two levels low=0, high=1

flow CO2 flow rate - two levels low=0, high=1

size peanut particle size - two levels low=0, high=1

solubility the amount of oil that could dissolve in the CO2

Source

Kilgo, M (1989) "An Application of Fractional Factorial Experimental Designs" Quality Engineering, 1, 45-54

`penicillin` *Penicillin yield by block and treatment*

Description

The production of penicillin uses a raw material, corn steep liquor, is quite variable and can only be made in blends sufficient for four runs. There are four processes, A, B, C and D, for the production.

Format

A data frame with 20 observations on the following 3 variables.

treat a factor with levels `A` `B` `C` `D`

blend a factor with levels `Blend1` `Blend2` `Blend3` `Blend4` `Blend5`

yield a numeric vector

Source

Box, G., W. Hunter, and J. Hunter (1978). Statistics for Experimenters. New York: Wiley.

Examples

```
data(penicillin)
## maybe str(penicillin) ; plot(penicillin) ...
```

`phbirths` *Birth weights in Philadelphia*

Description

Data based on a 5

Format

A data frame with 1115 observations on the following 5 variables.

black is the mother Black?

educ mother's years of education

smoke does the mother smoke during pregnancy?

gestate gestational age in weeks

grams birth weight in grams

Source

I. T. Elo, G. Rodriguez and H. Lee (2001). Racial and Neighborhood Disparities in Birthweight in Philadelphia. Paper presented at the Annual Meeting of the Population Association of America, Washington, DC 2001.

Examples

```
data(phbirths)
## maybe str(phbirths) ; plot(phbirths) ...
```

---

pima *Diabetes survey on Pima Indians*

Description

The National Institute of Diabetes and Digestive and Kidney Diseases conducted a study on 768 adult female Pima Indians living near Phoenix.

Format

The dataset contains the following variables

pregnant Number of times pregnant

glucose Plasma glucose concentration at 2 hours in an oral glucose tolerance test

diastolic Diastolic blood pressure (mm Hg)

triceps Triceps skin fold thickness (mm)

insulin 2-Hour serum insulin (mu U/ml)

bmi Body mass index (weight in kg/(height in metres squared))

diabetes Diabetes pedigree function

age Age (years)

test test whether the patient shows signs of diabetes (coded 0 if negative, 1 if positive)

Source

The data may be obtained from UCI Repository of machine learning databases at http://archive.ics.uci.edu/ml/

pipeline — *NIST data on ultrasonic measurements of defects in the Alaska pipeline*

Description

Researchers at National Institutes of Standards and Technology (NIST) collected data on ultrasonic measurements of the depths of defects in the Alaska pipeline in the field. The depth of the defects were then remeasured in the laboratory. These measurements were performed in six different batches. The laboratory measurements are more accurate than the in-field measurements, but more time consuming and expensive.

Format

A data frame with 107 observations on the following 3 variables.

Field measurement of depth of defect on site

Lab measurement of depth of defect in the lab

Batch the batch of measurements

Source

Office of the Director of the Institute of Materials Research (now the Materials Science and Engineering Laboratory) of NIST

pneumo — *Pneumonoconiosis in coal miners*

Description

The data for this example contains the number of coal miners classified by radiological examination into one of three categories of pneumonoultramicroscopicosilicovolcanoconiosis (known as pneumonoconiosis for short) and by number of years spent working at the coal face divided into eight categories.

Format

A data frame with 24 observations on the following 3 variables.

Freq number of miners

status pneumoconiosis status - a factor with levels `mild` `normal` `severe`

year number of years service (midpoint of interval)

Source

M. Aitkin and D. Anderson and B. Francis and J. Hinde (1989) "Statistical Modelling in GLIM" Oxford University Press.

potuse *Marijuana usage by youth*

Description

The National Youth Survey collected a sample of 11 to 17 year olds - 117 boys and 120 girls - asking questions about marijuana usage.

Format

A data frame with 486 observations on the following 7 variables.

sex 1=Male, 2=Female

year.76 1=never used, 2=used no more than once a month, 3=used more than once a month in 1976

year.77 1=never used, 2=used no more than once a month, 3=used more than once a month in 1977

year.78 1=never used, 2=used no more than once a month, 3=used more than once a month in 1978

year.79 1=never used, 2=used no more than once a month, 3=used more than once a month in 1979

year.80 1=never used, 2=used no more than once a month, 3=used more than once a month in 1980

count Number of cases in this category

Source

ICPSR, University of Michigan

References

Lang J., McDonald, J and Smith P. (1999) "Association-Marginal Modeling of Mutlivariate Categorical Responses: A Maximum Likelihood Approach" JASA 94, 1161-

prostate *Prostate cancer surgery*

Description

The `prostate` data frame has 97 rows and 9 columns. A study on 97 men with prostate cancer who were due to receive a radical prostatectomy.

Format

This data frame contains the following columns:

lcavol log(cancer volume)
lweight log(prostate weight)
age age
lbph log(benign prostatic hyperplasia amount)
svi seminal vesicle invasion
lcp log(capsular penetration)
gleason Gleason score
pgg45 percentage Gleason scores 4 or 5
lpsa log(prostate specific antigen)

Source

Andrews DF and Herzberg AM (1985): Data. New York: Springer-Verlag

---

prplot *Partial Residual Plot*

Description

Makes a Partial Residual plot

Usage

```
prplot(g, i)
```

Arguments

g	An object returned from lm()
i	index of predictor

Value

none

Author(s)

Julian Faraway

Examples

```
data(stackloss)
g <- lm(stack.loss ~ .,stackloss)
prplot(g,1)
```

psid *Panel Study of Income Dynamics subset*

Description

The Panel Study of Income Dynamics (PSID), begun in 1968, is a longitudinal study of a representative sample of U.S. individuals. The study is conducted at the Survey Research Center, Institute for Social Research, University of Michigan and is still continuing. The data represents a small subset of the total data.

Format

A data frame with 1661 observations on the following 6 variables.

age age in 1968

educ years of education

sex sex of individual, `F` or `M`

income annual income in dollars

year calendar year

person ID number for individual

Source

Martha S. Hill, The Panel Study of Income Dynamics: A User's Guide, Sage Publications, 1992,Newbury Park, CA.

pulp *Brightness of paper pulp depending on shift operator*

Description

The `pulp` data frame has 20 rows and 2 columns. Data comes from an experiment to test the paper brightness depending on a shift operator.

Format

This data frame contains the following columns:

bright Brightness of the pulp as measured by a reflectance meter

operator Shift operator a-d

Source

"Statistical techniques applied to production situations" F. Sheldon (1960) Industrial and Engineering Chemistry, 52, 507-509

punting *Leg strength and punting*

Description

Investigators studied physical characteristics and ability in 13 (American) football punters. Each volunteer punted a football ten times. The investigators recorded the average distance for the ten punts, in feet.

Format

A data frame with 13 observations on the following 7 variables.

Distance average distance over 10 punts

Hang hang time

RStr right leg strength in pounds

LStr left leg strength in pounds

RFlex right hamstring muscle flexibility in degrees

LFlex left hamstring muscle flexibility in degrees

OStr overall leg strength in foot pounds

Source

Unknown

Examples

```
data(punting)
## maybe str(punting) ; plot(punting) ...
```

pvc *Production of PVC by operator and resin railcar*

Description

Data from an experiment to study factors affecting the production of the plastic PVC, 3 operators used 8 different devices called resin railcars to produce PVC. For each of the 24 combinations, two samples were produced.

Format

Dataset contains the following variables

psize Particle size

operator Operator number 1, 2 or 3

resin Resin railcar 1-8

Source

R. Morris and E. Watson (1998) "A comparison of the techniques used to evaluate the measurement process" Quality Engineering, 11, 213-219

`pyrimidines` *Activity in pyrimidines*

Description

Structural information on 74 2,4-diamino- 5-(substituted benzyl) pyrimidines used as inhibitors of DHFR in E. coli. There are 3 positions where chemical activity occurs and 9 attributes per position leading to 27 total predictors. One predictor had no variability and was removed from the data set. 26 chemical properties of 74 compounds and an activity level

Format

A data frame with 74 observations on the following 27 variables.

p1.polar measured on a [0,1] scale

p1.size measured on a [0,1] scale

p1.flex measured on a [0,1] scale

p1.h.doner measured on a [0,1] scale

p1.h.acceptor measured on a [0,1] scale

p1.pi.doner measured on a [0,1] scale

p1.pi.acceptor measured on a [0,1] scale

p1.polarisable measured on a [0,1] scale

p1.sigma measured on a [0,1] scale

p2.polar measured on a [0,1] scale

p2.size measured on a [0,1] scale

p2.flex measured on a [0,1] scale

p2.h.doner measured on a [0,1] scale

p2.h.acceptor measured on a [0,1] scale

p2.pi.doner measured on a [0,1] scale

p2.pi.acceptor measured on a [0,1] scale

p2.polarisable measured on a [0,1] scale

p2.sigma measured on a [0,1] scale

p3.polar measured on a [0,1] scale

p3.size measured on a [0,1] scale

p3.flex measured on a [0,1] scale

p3.h.doner measured on a [0,1] scale

p3.h.acceptor measured on a [0,1] scale

p3.pi.doner measured on a [0,1] scale

p3.polarisable measured on a [0,1] scale

p3.sigma measured on a [0,1] scale

activity log 1/Ki, where Ki is the inhibition constant as experimentally assayed, scaled to [0,1]

Source

Jonathan D. Hirst, Ross D. King, Michael J. E. Sternberg (1994) Quantitative structure-activity relationships by neural networks and inductive logic programming. I. The inhibition of dihydrofolate reductase by pyrimidines doi:10.1007/BF00125375

Examples

```
data(pyrimidines)
## maybe str(pyrimidines) ; plot(pyrimidines) ...
```

`qqnorml` *Labeled QQ plot*

Description

Makes a labeled QQ plot

Usage

```
qqnorml(
  y,
  main = "Normal Q-Q Plot",
  xlab = "Theoretical Quantiles",
  ylab = "Sample Quantiles",
  ...
)
```

Arguments

`y`	A numeric vector
`main`	main label
`xlab`	x-axis label
`ylab`	y-axis label
`...`	arguments passed to plot()

Value

none

Author(s)

Julian Faraway

See Also

qqnorm

Examples

```
qqnorml(rnorm(16))
```

---

rabbit *Rabbit weight gain by diet and litter*

Description

A nutritionist studied the effects of six diets, on weight gain of domestic rabbits. From past experience with sizes of litters, it was felt that only 3 uniform rabbits could be selected from each available litter. There were ten litters available forming blocks of size three.

Format

The variables in the dataset were

treat Diet a through f

gain Weight gain

block Block (10 litters)

Source

"Experimental Design and Analysis" by M. Lentner and T. Bishop, Valley Book Company, 1986

---

ratdrink *Rat growth weights affected by additives*

Description

The data consist of 5 weekly measurements of body weight for 27 rats. The first 10 rats are on a control treatment while 7 rats have thyroxine added to their drinking water. 10 Rats have thiouracil added to their water.

Format

A data frame with 135 observations on the following 4 variables.

wt Weight of the rat

weeks Week of the study from 0 to 4

subject the rat code number

treat treatment applied to the rat drinking water - a factor with levels `control` `thiouracil` `thyroxine`

Source

Unknown

---

rats *Effect of toxic agents on rats*

Description

An experiment was conducted as part of an investigation to combat the effects of certain toxic agents.

Format

A data frame with 48 observations on the following 3 variables.

time survival time in tens of hours

poison the poison type - a factor with levels `I` `II` `III`

treat the treatment - a factor with levels `A` `B` `C` `D`

Source

Box G and Cox D (1964) "An analysis of transformations" J. Roy. Stat. Soc. Series B. **26** 211.

---

resceram *Shape and plate effects on current noise in resistors*

Description

The `resceram` data frame has 12 rows and 3 columns. Shape and plate effects on current noise in resistors

Format

This data frame contains the following columns:

noise current noise

shape the geometrical shape of the resistor, A, B, C or D

plate the ceramic plate on which the resistor was mounted. Only three resistors will fit on one plate.

Source

Natrella, M (1963) "Experimental Statistics" National Bureau of Standards Handbook 91, Gaithersburg MD.

`salmonella` *Salmonella reverse mutagenicity assay*

Description

The data was collected in a salmonella reverse mutagenicity assay where the numbers of revertant colonies of TA98 Salmonella observed on each of three replicate plates for different doses of quinoline

Format

A data frame with 18 observations on the following 2 variables.

colonies numbers of revertant colonies of TA98 Salmonella

dose dose level of quinoline

Source

Breslow N.E. (1984), Extra-Poisson Variation in Log-linear Models, ApplStat, pp. 38-44.

`sat` *School expenditure and test scores from USA in 1994-95*

Description

The `sat` data frame has 50 rows and 7 columns. Data were collected to study the relationship between expenditures on public education and test results.

Format

This data frame contains the following columns:

expend Current expenditure per pupil in average daily attendance in public elementary and secondary schools, 1994-95 (in thousands of dollars)

ratio Average pupil/teacher ratio in public elementary and secondary schools, Fall 1994

salary Estimated average annual salary of teachers in public elementary and secondary schools, 1994-95 (in thousands of dollars)

takers Percentage of all eligible students taking the SAT, 1994-95

verbal Average verbal SAT score, 1994-95

math Average math SAT score, 1994-95

total Average total score on the SAT, 1994-95

Source

"Getting What You Pay For: The Debate Over Equity in Public School Expenditures" D. Guber, Journal of Statistics Education, 1999

---

savings *Savings rates in 50 countries*

Description

The `savings` data frame has 50 rows and 5 columns. The data is averaged over the period 1960-1970.

Format

This data frame contains the following columns:

sr savings rate - personal saving divided by disposable income

pop15 percent population under age of 15

pop75 percent population over age of 75

dpi per-capita disposable income in dollars

ddpi percent growth rate of dpi

Details

Now also appears as `LifeCycleSavings` in the `datasets` package

Source

Belsley, D., Kuh. E. and Welsch, R. (1980) "Regression Diagnostics" Wiley.

See Also

LifeCycleSavings

seatpos *Car seat position depending driver size*

Description

Car drivers like to adjust the seat position for their own comfort. Car designers would find it helpful to know where different drivers will position the seat depending on their size and age. Researchers at the HuMoSim laboratory at the University of Michigan collected data on 38 drivers.

Format

The dataset contains the following variables

Age Age in years

Weight Weight in lbs

HtShoes Height in shoes in cm

Ht Height bare foot in cm

Seated Seated height in cm

Arm lower arm length in cm

Thigh Thigh length in cm

Leg Lower leg length in cm

hipcenter horizontal distance of the midpoint of the hips from a fixed location in the car in mm

Source

"Linear Models in R" by Julian Faraway, CRC Press, 2004

seeds *Germination of seeds depending on moisture and covering*

Description

A Biologist analyzed an experiment to determine the effect of moisture content on seed germination. Eight boxes of 100 seeds each were treated with the same moisture level. 4 boxes were covered and 4 left uncovered. The process was repeated at 6 different moisture levels (nonlinear scale).

Format

A data frame with 48 observations on the following 3 variables.

germ percentage germinated

moisture moisture level

covered a factor with levels `no` `yes`

Source

Chatfield C. (1995) Problem Solving: A Statistician's Guide, 2ed Chapman Hall.

Examples

```
data(seeds)
## maybe str(seeds) ; plot(seeds) ...
```

`semicond` *Semiconductor split-plot experiment*

Description

The `semicond` data frame has 48 rows and 5 columns.

Format

This data frame contains the following columns:

resistance a numeric vector

ET a factor with levels `1` to `4` representing etch time.

Wafer a factor with levels `1` to `3`

position a factor with levels `1` to `4`

Grp an ordered factor with levels `1/1` < `1/2` < `1/3` < `2/1` < `2/2` < `2/3` < `3/1` < `3/2` < `3/3` < `4/1` < `4/2` < `4/3`

Details

Also found in the `SASmixed` package

Source

Littel, R. C., Milliken, G. A., Stroup, W. W., and Wolfinger, R. D. (1996), *SAS System for Mixed Models*, SAS Institute (Data Set 2.2(b)).

sexab *Post traumatic stress disorder in abused adult females*

Description

The data for this example come from a study of the effects of childhood sexual abuse on adult females. 45 women being treated at a clinic, who reported childhood sexual abuse, were measured for post traumatic stress disorder and childhood physical abuse both on standardized scales. 31 women also being treated at the same clinic, who did not report childhood sexual abuse were also measured. The full study was more complex than reported here and so readers interested in the subject matter should refer to the original article.

Format

The variables in the dataset are

cpa Childhood physical abuse on standard scale

ptsd Post-traumatic stress disorder on standard scale

csa Childhood sexual abuse - abused or not abused

Source

N. Rodriguez and S. Ryan and H. Vande Kemp and D. Foy (1997) "Postraumatic stress disorder in adult female survivors of childhood sexual abuse: A comparison study", Journal of Consulting and Clinical Pyschology, 65, 53-59

sexfun *Marital sex ratings*

Description

Data from a questionaire from 91 couples in the Tucson, Arizona area. Subjects answered the question "Sex is fun for me and my partner". The possible answers were "never or occasionally","fairly often","very often" and "almost always"

Format

A data frame with 16 observations on the following 3 variables.

y the count

husband a factor with levels `never` `fairly` `very` `always`

wife a factor with levels `never` `fairly` `very` `always`

Source

Hout, M., Duncan, O. and Sobel M. (1987) Association and heterogeneity: Structural models of similarities and differences. Sociological Methods. 17, 145-184.

snail *Snail production*

Description

A study was conducted to optimize snail production for consumption. The percentage water content of the tissues of snails grown under three different levels of relative humidity and two different temperatures was recorded. For each combination, 4 snails were observed.

Format

A data frame with 24 observations on the following 3 variables.

water percentage water content

temp temperature in C

humid relative humidity

Source

Unknown

Examples

```
data(snail)
## maybe str(snail) ; plot(snail) ...
```

solder *Solder skips in printing circuit boards*

Description

ATT ran an experiment varying five factors relevant to a wave-soldering procedure for mounting components on printed circuit boards. The response variable, skips, is a count of how many solder skips appeared to a visual inspection.

Format

A data frame with 900 observations on the following 6 variables.

Opening a factor with levels `L` `M` `S`

Solder a factor with levels `Thick` `Thin`

Mask a factor with levels `A1.5` `A3` `A6` `B3` `B6`

PadType a factor with levels `D4` `D6` `D7` `L4` `L6` `L7` `L8` `L9` `W4` `W9`

Panel a numeric vector

skips count of how many solder skips appeared to a visual inspection

Source

Comizzoli, R. B., J. M. Landwehr, and J. D. Sinclair (1990). Robust materials and processes: Key to reliability. AT&T Technical Journal 69(6), 113-128.

Examples

```
data(solder)
## maybe str(solder) ; plot(solder) ...
```

---

solv *Block design task testing child ability*

Description

Behavioural scientists at Macquarie University conducted an experiment to test the time taken to perform a block design task with 24 fifth grade children (12 boys and 12 girls).

Format

A data frame with 24 observations and 3 variables.

group Solution attempted first by row(r) or corner(c)

time Time taken to complete the task in seconds

eft Score on the embedded figures test which is a measure of difficulty in abstracting logical structure of a problem from its context.

Source

Statistical Modelling in GLIM (1989) M. Aitkin and D. Anderson and B. Francis and J. Hinde Oxford University Press

---

sono *Sonoluminescence*

Description

The sono data frame has 16 rows and 8 columns. Sonoluminescence is the process of turning sound energy into light. An experiment was conducted to study factors affecting this process.

Format

This data frame contains the following columns:

Intensity Sonoluminescent light intensity

Molarity Amount of Solute. The coding is "low" for 0.10 mol and "high" for 0.33 mol.

Solute Solute type. The coding is "low" for sugar and "high" for glycerol.

pH The coding is "low" for 3 and "high" for 11.

Gas Gas type in water. The coding is "low" for helium and "high" for air.

Water Water depth. The coding is "low" for half and "high" for full.

Horn Horn depth. The coding is "low" for 5 mm and "high" for 10 mm.

Flask Flask clamping. The coding is "low" for unclamped and "high" for clamped.

Source

Eva Wilcox and Ken Inn of the NIST Physics Laboratory conducted this experiment during 1999 and published in NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/

---

soybean *Germination failures for soybean seeds*

Description

An experiment was conducted to compare the germination rates of the five varieties of soybean. Five plots were available.

Format

A data frame with 25 observations on the following 3 variables.

variety the variety - a factor with levels `arasan` `check` `fermate` `semesan` `spergon`

replicate the plot - a factor with levels `1` `2` `3` `4` `5`

failure the number of failures out of 100 planted seeds

Source

Snedecor G. and Cochran W. (1967) Statistical Methods (6th Ed) Iowa State University Press

spector *Teaching methods in Economics*

Description

A study to determine the effectiveness of a new teaching method in Economics

Format

A data frame with 32 observations on the following 4 variables.

grade 1 = exam grades improved, 0 = not improved

psi 1 = student exposed to PSI (a new teach method), 0 = not exposed

tuce a measure of ability when entering the class

gpa grade point average

Source

Spector, L. and Mazzeo, M. (1980), "Probit Analysis and Economic Education", Journal of Economic Education, 11, 37 - 44.

speedo *Speedometer cable shrinkage*

Description

Speedometer cables can be noisy because of shrinkage in the plastic casing material. An experiment was conducted to find out what caused shrinkage by screening a large number of factors. The engineers started with 15 different factors.

Format

The dataset contains the following variables: (variables a-o are 2 level factors, coded "+" and "-" where "+" indicates a higher value where appropriate)

a liner outer diameter

b liner die

c liner material

d liner line speed

e wire braid type

f braiding tension

g wire diameter

h liner tension

i liner temperature

j coating material

k coating die type

l melt temperature

m screen pack

n cooling method

o line speed

y percentage shrinkage per specimen

Source

G. P. Box and S. Bisgaard and C. Fung (1988) "An explanation and critque of Taguchi's contributions to quality engineering", Quality and reliability engineering international, 4, 123-131

`star` *Star temperatures and light intensites*

Description

Data on the log of the surface temperature and the log of the light intensity of 47 stars in the star cluster CYG OB1, which is in the direction of Cygnus,

Format

A data frame with 47 observations on the following 3 variables.

index a numeric vector

temp temperature

light light intensity

Source

Rousseeuw, P. and A. Leroy (1987). Robust Regression and Outlier Detection. New York: Wiley.

Examples

```
data(star)
## maybe str(star) ; plot(star) ...
```

stat500 *Marks in a statistics class*

Description

Marks from Statistics 500 one year at the University of Michigan

Format

A data frame with 55 observations on the following 4 variables.

midterm a numeric vector

final a numeric vector

hw a numeric vector

total a numeric vector

Source

Julian Faraway

Examples

```
data(stat500)
## maybe str(stat500) ; plot(stat500) ...
```

stepping *Stepping and effect on heart rate*

Description

An experiment was conducted to explore the nature of the relationship between a person's heart rate and the frequency at which that person stepped up and down on steps of various heights.

Format

A data frame with 30 observations on the following 6 variables.

Order running order within the experiment

Block Experimenter used

Height 0 if step at the low (5.75in) height, 1 if at the high (11.5in) height

Frequency the rate of stepping. 0 if slow (14 steps/min), 1 if medium (21 steps/min), 2 if high (28 steps/min)

RestHR the resting heart rate of the subject before a trial, in beats per minute

HR the final heart rate of the subject after a trial, in beats per minute

Source

Unknown

Examples

```
data(stepping)
## maybe str(stepping) ; plot(stepping) ...
```

`strongx` *Strong interaction experiment data*

Description

Example Dataset from "Practical Regression and Anova"

Format

Dataframe with 10 cases

momentum **energy** inverse total energy

crossx Scattering cross-section/sec

sd standard deviation

Source

Weisberg, H., Beier, H., Brody, H., Patton, R., Raychaudhari, K., Takeda, H., Thern, R. and Van Berg, R. (1978). s-dependence of proton fragmentation by hadrons. II. Incident laboratory momenta, 30–250 GeV/c. Physics Review D, 17, 2875–2887.

References

Weisberg, S. (2014). Applied Linear Regression, 4th edition. Hoboken NJ: Wiley.

`suicide` *Suicide method data from the UK*

Description

One year of suicide data from the United Kingdom crossclassified by sex, age and method.

Format

A data frame with 36 observations on the following 4 variables.

y number of people

cause method used - a factor with levels drug (suicide by solid or liquid matter), gas, gun (guns, knives or explosives) hang (hanging, strangling, suffocating or drowning, jump other

age a factor with levels m (middle-aged) o (old) y (young)

sex a factor with levels f m

Source

Everitt B. & Dunn G. (1991) "Applied Multivariate Data Analysis" Edward Arnold

---

sumary *Abbreviated Regression Summary*

Description

Generic summaries for lm, glm and mer objects

Usage

```
sumary(object, ...)
```

Arguments

object	An lm, glm or mer object returned from lm(), glm() or lmer() respectively
...	further arguments passed to or from other methods.

Details

This generic function provides an abbreviated regression output containing the more useful information. Users wanting to see more are advised to use summary()

Value

returns the same as summary()

Author(s)

Julian Faraway

References

This function is adapted from the display() function in the arm package

See Also

summary, lm, glm, lmer

Examples

```
data(stackloss)
object <- lm(stack.loss ~ .,stackloss)
sumary(object)
```

---

teengamb *Study of teenage gambling in Britain*

Description

The teengamb data frame has 47 rows and 5 columns. A survey was conducted to study teenage gambling in Britain.

Format

This data frame contains the following columns:

sex 0=male, 1=female

status Socioeconomic status score based on parents' occupation

income in pounds per week

verbal verbal score in words out of 12 correctly defined

gamble expenditure on gambling in pounds per year

Source

Ide-Smith & Lea, 1988, Journal of Gambling Behavior, 4, 110-118

---

toenail *Toenail infection treatment study*

Description

The data come from a Multicenter study comparing two oral treatments for toenail infection. Patients were evaluated for the degree of separation of the nail. Patients were randomized into two treatments and were followed over seven visits - four in the first year and yearly thereafter. The patients have not been treated prior to the first visit so this should be regarded as the baseline.

Format

A data frame with 1908 observations on the following 5 variables.

ID ID of patient

outcome 0=none or mild seperation, 1=moderate or severe

treatment the treatment A=0 or B=1

month time of the visit (not exactly monthly intervals hence not round numbers)

visit the number of the visit

Source

De Backer, M., De Vroey, C., Lesaffre, E., Scheys, I., and De Keyser, P. (1998). Twelve weeks of continuous oral therapy for toenail onychomycosis caused by dermatophytes: A double-blind comparative trial of terbinafine 250 mg/day versus itraconazole 200 mg/day. Journal of the American Academy of Dermatology, 38, 57-63.

References

Lesaffre, E. and Spiessens, B. (2001). On the effect of the number of quadrature points in a logistic random-effects model: An example. Journal of the Royal Statistical Society, Series C, 50, 325-335.

G. Fitzmaurice, N. Laird and J. Ware (2004) Applied Longitudinal Analysis, Wiley

---

`troutegg` *Survival of trout eggs depending on time and location*

Description

Boxes of trout eggs were buried at five different stream locations and retrieved at 4 different times. The number of surviving eggs was recorded. The box was not returned to the stream.

Format

A data frame with 20 observations on the following 4 variables.

survive the number of surviving eggs

total the number of eggs in the box

location the location in the stream with levels `1 2 3 4 5`

period the number of weeks after placement that the box was withdrawn levels `4 7 8 11`

Source

Manly B. (1978) Regression models for proportions with extraneous variance. Biometrie-Praximetrie, 18, 1-18.

References

Hinde J. and Demetrio C. (1988) Overdispersion: Models and estimation. Computational Statistics and Data Analysis. 27, 151-170.

truck *Truck leaf spring experiment*

Description

Data on an experiment concerning the production of leaf springs for trucks. A 2^{5-1} fractional factorial experiment with 3 replicates was carried out with objective of recommending production settings to achieve a free height as close as possible to 8 inches.

Format

A data frame with 48 observations on the following 6 variables.

B furnace temperature - a factor with levels + -

C heating time - a factor with levels + -

D transfer time - a factor with levels + -

E hold-down time - a factor with levels + -

O quench oil temperature - a factor with levels + -

height leaf spring free height in inches

Source

J. J. Pignatiello and J. S. Ramberg (1985) Contribution to discussion of offline quality control, parameter design and the Taguchi method, Journal of Quality Technology, **17** 198-206.

References

P. McCullagh and J. Nelder (1989) "Generalized Linear Models" Chapman and Hall, 2nd ed.

turtle *Incubation temperature and the sex of turtles*

Description

Incubation temperature can affect the sex of turtles. There are 3 independent replicates for each temperature.

Format

A data frame with 15 observations on the following 3 variables.

temp temperature in degrees centigrade

male number of male turtles hatched

female number of female turtles hatched

Source

Beyond Traditional Statistical Methods Copyright 2000 D. Cook, P. Dixon, W. M. Duckworth, M. S. Kaiser, K. Koehler, W. Q. Meeker and W. R. Stephenson. Developed as part of NSF/ILI grant DUE9751644.

Examples

```
data(turtle)
```

---

tvdoctor *Life, TVs and Doctors*

Description

Life expectancy, doctors and televisions collected on 38 countries in 1993

Format

A data frame with 38 observations on the following 3 variables.

life Life expectancy in years

tv Number of people per television set

doctor Number of people per doctor

Source

Unknown, data for illustration purposes only

Examples

```
data(tvdoctor)
## maybe str(tvdoctor) ; plot(tvdoctor) ...
```

twins *Twin IQs from Burt*

Description

Study of IQ in twins reared apart

Format

A dataframe with the following variables:

Foster IQ of the fostered child

Biological IQ of the biological child

Social social class of natural parents

Source

Burt, C. (1966). The genetic estimation of differences in intelligence: A study of monozygotic twins reared together and apart. Br. J. Psych., 57, 147-153.

References

Weisberg, S. (2014). Applied Linear Regression, 4th edition. Hoboken NJ: Wiley.

uncviet *UNC student opinions about the Vietnam War*

Description

A student newspaper conducted a survey of student opinions about the Vietnam War in May 1967. Responses were classified by sex, year in the program and one of four opinions. The survey was voluntary.

Format

A data frame with 40 observations on the following 4 variables.

y the count

policy a factor with levels `A` (defeat power of North Vietnam by widespread bombing and land invasion) `B` (follow the present policy) `C` (withdraw troops to strong points and open negotiations on elections involving the Viet Cong) `D` (immediate withdrawal of all U.S. troops)

sex a factor with levels `Female` `Male`

year a factor with levels `Fresh` `Grad` `Junior` `Senior` `Soph`

Source

M. Aitkin and D. Anderson and B. Francis and J. Hinde (1989) "Statistical Modelling in GLIM" Oxford University Press.

uswages *Weekly wages of US male workers in 1988*

Description

The uswages data frame has 2000 rows and 10 columns. Weekly Wages for US male workers sampled from the Current Population Survey in 1988.

Format

This data frame contains the following columns:

wage Real weekly wages in dollars (deflated by personal consumption expenditures - 1992 base year)

educ Years of education

exper Years of experience

race 1 if Black, 0 if White (other races not in sample)

smsa 1 if living in Standard Metropolitan Statistical Area, 0 if not

ne 1 if living in the North East

mw 1 if living in the Midwest

we 1 if living in the West

so 1 if living in the South

pt 1 if working part time, 0 if not

Source

Bierens, H.J., and D. Ginther (2001): "Integrated Conditional Moment Testing of Quantile Regression Models", Empirical Economics 26, 307-324

vif *vif*

Description

vif

Usage

```
vif(object)

## Default S3 method:
vif(object)

## S3 method for class 'lm'
vif(object)
```

Arguments

`object`	a data matrix (design matrix without intercept) or a model object

Details

Computes the variance inflation factors

Value

variance inflation factors

Author(s)

Julian Faraway

Examples

```
data(stackloss)
vif(stackloss[,-4])
#  Air.Flow Water.Temp Acid.Conc.
#    2.9065     2.5726     1.3336
```

---

vision *Acuity of vision in response to light flash*

Description

The acuity of vision for seven subjects was tested. The response is the lag in milliseconds between a light flash and a response in the cortex of the eye. Each eye is tested at four different powers of lens. An object at the distance of the second number appears to be at distance of the first number.

Format

A data frame with 56 observations on the following 4 variables.

acuity a numeric vector

power a factor with levels `6/6` `6/18` `6/36` `6/60`

eye a factor with levels `left` `right`

subject a factor with levels `1` `2` `3` `4` `5` `6` `7`

Source

Crowder, M. J. and D. J. Hand (1990). Analysis of Repeated Measures. London: Chapman & Hall.

Examples

```
data(vision)
## maybe str(vision) ; plot(vision) ...
```

wafer *resitivity of wafer in semiconductor experiment*

Description

A full factorial experiment with four two-level predictors.

Format

A data frame with 16 observations on the following 5 variables.

x1 a factor with levels - +

x2 a factor with levels - +

x3 a factor with levels - +

x4 a factor with levels - +

resist Resistivity of the wafer

Source

Myers, R. and Montgomery D. (1997) A tutorial on generalized linear models, Journal of Quality Technology, 29, 274-291.

wavesolder *Defects in a wave soldering process*

Description

Components are attached to an electronic circuit card assembly by a wave-soldering process. The soldering process involves baking and preheating the circuit card and then passing it through a solder wave by conveyor. Defect arise during the process. Design is 2^{7-3} with 3 replicates.

Format

A data frame with 16 observations on the following 10 variables.

y1 Number of defects in the first replicate

y2 Number of defects in the second replicate

y3 Number of defects in the third replicate

prebake prebake condition - a factor with levels 1 2

flux flux density - a factor with levels 1 2

speed conveyor speed - a factor with levels 1 2

preheat preheat condition - a factor with levels 1 2

cooling cooling time - a factor with levels 1 2

agitator ultrasonic solder agitator - a factor with levels 1 2

temp solder temperature - facctor with levels 1 2

Source

L. Condra (1993) Reliability improvement with design of experiments. Marcel Dekker, NY.

References

M. Hamada and J. Nelder (1997) Generalized linear models for quality improvement experiments, Journal of Quality Technology, 29, 292-304

---

wbca *Wisconsin breast cancer database*

Description

Data come from a study of breast cancer in Wisconsin. There are 681 cases of potentially cancerous tumors of which 238 are actually malignant. Determining whether a tumor is really malignant is traditionally determined by an invasive surgical procedure. The purpose of this study was to determine whether a new procedure called fine needle aspiration which draws only a small sample of tissue could be effective in determining tumor status.

Format

A data frame with 681 observations on the following 10 variables.

Class 0 if malignant, 1 if benign

Adhes marginal adhesion

BNucl bare nuclei

Chrom bland chromatin

Epith epithelial cell size

Mitos mitoses

NNucl normal nucleoli

Thick clump thickness

UShap cell shape uniformity

USize cell size uniformity

Details

The predictor values are determined by a doctor observing the cells and rating them on a scale from 1 (normal) to 10 (most abnormal) with respect to the particular characteristic.

Source

Bennett, K.,P., and Mangasarian, O.L., Neural network training via linear programming. In P. M. Pardalos, editor, Advances in Optimization and Parallel Computing, pages 56-57. Elsevier Science, 1992

wcgs *Western Collaborative Group Study*

Description

3154 healthy young men aged 39-59 from the San Francisco area were assessed for their personality type. All were free from coronary heart disease at the start of the research. Eight and a half years later change in this situation was recorded.

Format

A data frame with 3154 observations on the following 13 variables.

age age in years

height height in inches

weight weight in pounds

sdp systolic blood pressure in mm Hg

dbp diastolic blood pressure in mm Hg

chol Fasting serum cholesterol in mm %

behave behavior type which is a factor with levels `A1` `A2` `B3` `B4`

cigs number of cigarettes smoked per day

dibep behavior type a factor with levels `A` (Agressive) `B` (Passive)

chd coronary heat disease developed is a factor with levels `no` `yes`

typechd type of coronary heart disease is a factor with levels `angina` `infdeath` `none` `silent`

timechd Time of CHD event or end of follow-up

arcus arcus senilis is a factor with levels `absent` `present`

Details

The WCGS began in 1960 with 3,524 male volunteers who were employed by 11 California companies. Subjects were 39 to 59 years old and free of heart disease as determined by electrocardiogram. After the initial screening, the study population dropped to 3,154 and the number of companies to 10 because of various exclusions. The cohort comprised both blue- and white-collar employees. At baseline the following information was collected: socio-demographic including age, education, marital status, income, occupation; physical and physiological including height, weight, blood pressure, electrocardiogram, and corneal arcus; biochemical including cholesterol and lipoprotein fractions; medical and family history and use of medications; behavioral data including Type A interview, smoking, exercise, and alcohol use. Later surveys added data on anthropometry, triglycerides, Jenkins Activity Survey, and caffeine use. Average follow-up continued for 8.5 years with repeat examinations

Source

Statistics for Epidemiology by N. Jewell (2004)

References

Coronary Heart Disease in the Western Collaborative Group Study Final Follow-up Experience of 8 1/2 Years Ray H. Rosenman, MD; Richard J. Brand, PhD; C. David Jenkins, PhD; Meyer Friedman, MD; Reuben Straus, MD; Moses Wurm, MD JAMA. 1975;233(8):872-877. doi:10.1001/jama.1975.03260080034016.

Examples

```
data(wcgs)
## maybe str(wcgs) ; plot(wcgs) ...
```

`weldstrength` *welding strength DOE*

Description

An experiment to investigate factors affecting welding strength.

Format

A data frame with 16 observations on the following 10 variables.

Rods a 0-1 predictor

Drying a 0-1 predictor

Material a 0-1 predictor

Thickness a 0-1 predictor

Angle a 0-1 predictor

Opening a 0-1 predictor

Current a 0-1 predictor

Method a 0-1 predictor

Preheating a 0-1 predictor

Strength The welding strength

Source

G. Box and R. Meyer (1986) Dispersion effects from fractional designs, Technometrics, 28, 19-27

`wfat` *Percentage of Body Fat and Body Measurements in Women*

Description

Age, weight, height, and 10 body circumference measurements are recorded for 184 women. Each woman's percentage of body fat was accurately estimated by an underwater weighing technique.

Format

A data frame with 184 observations on the following 19 variables.

siri Percent body fat using Siri's equation

weight Weight (lbs)

height Height (inches)

bmi Body Mass Index

age Age (yrs)

neck Neck circumference (cm)

chest Chest circumference (cm)

calf Calf circumference (cm)

biceps Extended biceps circumference (cm)

hip Hip circumference (cm)

abdom Horizontal minimal measurement, at the end of a normal expiration (cm)

forearm Forearm circumference (cm)

thigh (Proximal Thigh) Horizontal measurement immediately distal to the gluteal furrow (cm)

mthigh (Middle Thigh) Measurement midway between the midpoint of the inguinal crease and the proximal border of the patella (cm)

dthigh (Distal Thigh) Measurement proximal to the femoral epicondyles (cm)

wrist Wrist circumference (cm) distal to the styloid processes

knee Knee circumference (cm)

elbow A minimal circumference measurement with the elbow extended (cm)

ankle Ankle circumference (cm)

Source

Roger W. Johnson (2021): Fitting Percentage of Body Fat to Simple Body Measurements: College Women, Journal of Statistics and Data Science Education, DOI: 10.1080/26939169.2021.1971585 (Note that I have changed some of the variable names to correspond with the older fat data for men)

See Also

`fat`

wheat *Insect damage to wheat by variety*

Description

Insect damage to wheat by variety

Format

A data frame with 13 observations on the following 2 variables.

damage a numeric vector

variety a factor with levels `A` `B` `C` `D`

Source

Unknown

Examples

```
data(wheat)
## maybe str(wheat) ; plot(wheat) ...
```

worldcup *Data on players from the 2010 World Cup*

Description

Data on players from the 2010 World Cup

Format

A data frame with 595 observations on the following 7 variables.

Team Country

Position a factor with levels `Defender` `Forward` `Goalkeeper` `Midfielder`

Time Time played in minutes

Shots Number of shots attempted

Passes Number of passes made

Tackles Number of tackles made

Saves Number of saves made

Details

None

Source

Lost

Examples

```
data(worldcup)
## maybe str(worldcup) ; plot(worldcup) ...
```

Index