Package 'GLMsData'

July 21, 2025

Version 1.4

Date 2022-08-22

Title Generalized Linear Model Data Sets

Description
Data sets from the book Generalized Linear Models with Examples in R by Dunn and Smyth.

Author Peter K. Dunn [cre,aut], Gordon K. Smyth [aut]

Maintainer Peter K. Dunn <pdunn2@usc.edu.au>

License GPL (>= 2)

RoxygenNote 6.0.1

NeedsCompilation no

Repository CRAN

Date/Publication 2022-08-22 06:20:08 UTC

Contents

`AIS` *Australian Institute of Sports (AIS) data*

Description

Physical measurements and blood measurements from high performance athletes at the AIS

Usage

```
data(AIS)
```

Format

A data frame containing 202 observations with the following 13 variables.

`Sex` the sex of the athlete: `F` means female, and `M` means male

`Sport` the sport of the athlete; one of `BBall` (basketball), `Field`, `Gym` (gymnastics), `Netball`, `Rowing`, `Swim` (swimming), `T400m`, (track, further than 400m), `Tennis`, `TPSprnt` (track sprint events), `WPolo` (waterpolo)

`LBM` lean body mass, in kg

`Ht` height, in cm

`Wt` weight, in kg

`BMI` body mass index, in kg per metre-squared

`SSF` sum of skin folds

`PBF` percentage body fat

`RBC` red blood cell count, in 10^{12} per litre

`WBC` white blood cell count, in 10^{12} per litre

`HCT` hematocrit, in percent

`HGB` hemoglobin concentration, in grams per decilitre

`Ferr` plasma ferritins, in ng per decilitre

Details

The data give measurements from high-performance athletes from the Australian Institute of Sport (AIS), for 202 athletes (102 males; 100 females) on 13 variables. Telford and Cunningham (1991) provide more information on how the data were collected.

From the paper: "The main aim of the statistical analysis was to determine whether there were any hematological differences, on average, between athletes from the various sports, between the sexes, and whether there was an effect of mass or height" (p. 789).

Source

OZDASL, available on-line at http://www.statsci.org/data/.

References

Telford, R. D. and Cunningham, R. B. (1991) Sex, sport, and body-size dependency of hematology in highly trained athletes. *Medicine and Science in Sports and Exercise*, **23**(7):788–794.

Examples

```
data(AIS)
summary(AIS)
```

ants *Ants species richness*

Description

The number of ant species in New England (USA)

Usage

```
data(ants)
```

Format

A data frame containing 44 observations with the following 5 variables.

`Site` an abbreviation for the site name

`Srich` species richness (number of ant species); a numeric vector

`Habitat` the habitat type: a factor with levels `Bog` and `Forest`

`Latitude` the latitude (in decimal degrees) for the site; a numeric vector

`Elevation` the elevation, in metres above sea level; a numeric vector

Details

The data give the ant species richness (number of ant species) found in 64 square metre sampling grids, in 22 bogs and 22 forests surrounding the bogs, in Connecticut, Massachusetts and Vermont (USA). The sites span a 3-degrees of latitude in New England.

Source

N. J. Gotelli and A. M. Ellison (2002). Biogeography at a regional scale: determinants of ant species density in bogs and forests of New England. *Ecology*, **83**, 1604–1609.

References

Aaron M. Ellison (2004) Bayesian inference in ecology. *Ecology Letters*, **7**, 509–520.

Examples

```
data(ants)
summary(ants)
```

`apprentice` *Apprentice migration to Edinburgh*

Description

The number of apprentices migrating to Edinburgh

Usage

```
data(apprentice)
```

Format

A data frame with 33 observations on the following 5 variables.

`Dist` the distance from Edinburgh (unit unknown, presumably miles); a numeric vector

`Apps` the number of apprentices moving to Edinburgh from the given county (given in row labels); a numeric vector

`Pop` the population (in thousands) of the given county; a numeric vector

`Urban` the degree of urbanization as measured by the percentage of the population living in urban settlements; a numeric vector

`Locn` the location of the county relative to Edinburgh; a factor with levels `North`, `South` and `West`

Details

The data record the number of apprentices moving to Edinburgh between 1775 and 1799 from other Scottish counties.

Source

Andrew Lovett and Robin Flowerdew (1989) Analysis of count data using Poisson regression. *Professional Geographer*, **4**1(2), 190–198.

Examples

```
data(apprentice)
summary(apprentice)
```

babblers *Feeding rates of babblers*

Description

The daily individual feeding rates of chestnut-crowned babblers

Usage

```
data(babblers)
```

Format

A data frame containing 97 observations with the following 8 variables.

`ObsTime` the length of observation (in decimal hours); a numeric vector

`Sex` the sex of the bird; one of `f` (female) or `m` (male)

`Age` the age of non-breeding group members; one of `adult` or `yearling`

`Relatedness` the pedigree-based relatedness to the brood; one of `0.5` (first-order relatives); `0.25` (second-order relatives) or `0` (more distant relatives)

`ChickAge` the age of the brood, in days; a numeric vector

`BroodSize` the size of the brood: a numeric vector

`UnitSize` the number of individuals in the unit; a numeric vector

`FeedingRate` the daily individual feeding rates, in feeds per hour; a numeric vector

Details

The data relate to a population of colour-ringed population of chestnut-crowned babblers in an area of the University of New South Wales Arid Zone Research Station, (Fowlers Gap, western New South Wales, Australia). The study determined whether, where and how often non-breeding group members contributed to providing for nestlings by monitoring the visit rate of tagged birds during 2007 and 2008. These data are extracted from a larger data set, extracted so that there is one (randomly chosen) observation for each individual bird.

Source

L. E. Browning, S. C. Patrick, L. A. Rollins, S. C. Griffith, and A. F. Russell (2012) Kin selection, not group augmentation, predicts helping in an obligate cooperatively breeding bird. *Proceedings of the Royal Society B*, **279**(1743): 3861–3869. doi: 10.1098/rspb.2012.1080

L. E. Browning, S. C. Patrick, L. A. Rollins, S. C. Griffith, and A. F. Russell (2012) Data from: Kin selection, not group augmentation, predicts helping in an obligate cooperatively breeding bird. *Dryad Digital Repository*. doi: 10.5061/dryad.ff868

References

L. E. Browning, S. C. Patrick, L. A. Rollins, S. C. Griffith, and A. F. Russell (2012) Kin selection, not group augmentation, predicts helping in an obligate cooperatively breeding bird. *Proceedings of the Royal Society B*, **279**(1743): 3861–3869. doi: 10.1098/rspb.2012.1080

Examples

```
data(babblers)
summary(babblers)
```

`belection` *British election candidates*

Description

The number of candidates in the British general election in 1992

Usage

```
data(belection)
```

Format

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

`Region` the region; a factor with levels `EastAnglia`, `EastMidlands`, `GreaterLondon`, `NorthWest`, `Scotland`, `SouthEast`, `SouthWest`, `Wales`, `WestMidlands` and `YorksHumbers`

`Party` the political party; a factor with levels `Cons`, `Green`, `Labour`, `LibDem` and `Other`

`Females` the number of female candidates; a numeric vector

`Males` the number of male candidates; a numeric vector

Details

The data give the number of male and females candidates in the British general election held April 9, 1992.

Source

D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway, and E. Ostrowski (1994) *A Handbook of Small Data Sets*, London: Chapman and Hall. Dataset 374.

References

The data originally came from: *The Independent*, Friday 27th March, 1992.

Examples

```
data(belection)
plot(Females/(Females+Males) ~ Party, data=belection)
```

`blocks` *Blocks stacked by children*

Description

The number of blocks stacked by children, and the time taken

Usage

```
data(blocks)
```

Format

A data frame with 100 observations on the following 6 variables.

`Child` the child; an identifier from `A` to `Y`

`Number` the number of blocks the child could successfully stack; a numeric vector

`Time` the time in seconds taken for the children to make their stack of blocks; a numeric vector

`Trial` the trial number on which the data were gathered (see Details); a factor with levels `1` and `2`

`Shape` the shape of the blocks being stacked; a factor with levels `Cube` and `Cylinder`

`Age` the age of the child in completed years; a numeric vector

Details

Children were seated a a small table, and "told" to build a tower from the blocks as high as they could. This was demonstrated for the child. The time taken and the number of blocks used were recorded. The cubes were always presented first, then cylinders. The second trial was conducted one month later.

The blocks were "half inch cubes and cylinders included in Mrs. Hailmann's Beads No. 470 of Bradley's Kindergarten Material". Throughout the article, the children are referred to using male pronouns, but (in keeping with the custom at the time) it is unclear whether all children were males or not. However, since gender is not recorded the children may all have been boys.

The source (Johnson and Courtney 1931) gives the age in years and months. Here they have been converted to decimal years.

Note

The means given in Table 1 in Johnson and Courtney (1931) do not agree in every case with the data given in that same table.

Source

Buford Johnson and Dorothy Moore Courtney (1931) Tower building, *Child Development*, **2**(2), 161–162

References

Judith D. Singer and John B. Willett (1990) Improving the teaching of applied statistics: Putting the data back into data analysis. *The American Statistician*, **44**(3), 223–230.

Examples

```
data(blocks)
plot( Time ~ Age, data=blocks)
```

`boric` *Dead embryos after exposure to boric acid*

Description

The number of mice embryos dead after exposure to four different doses of boric acid

Usage

```
data(boric)
```

Format

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

`Dose` the dose of boric acid (in percent of boric acid in feed); a numeric vector

`Dead` the number of embryos dead *in utero*; a numeric vector

`Implants` the total number of embryos; a numeric vector

Details

Mice were fed doses of boric acid in their feed during the first 17 days of gestation; the mice were then sacrificed and the embryos examined. Boric acid is widely used in pesticides and household products.

Source

Terra L. Slaton, Walter W. Piegorsch and Stephen D. Durham (2000) Estimation and testing with overdispersed proportions using the beta-logistic regression model of Heckman and Willis. *Biometrics*, **56**(1), 125–133, Table 4.

References

J. H. Hiendel, C. J. Price, E. A. Field, M. C. Marr, C. B. Myers, R. E. Morrissey, and B. A. Schwetz (1992) Developmental toxocity of boric acid in mice and rats. *Fundamental and Applied Toxicology*, **18**, 266–277.

Examples

```
data(boric)
plot( Dead/Implants ~ Dose, data=boric)
```

`breakdown` *Dialetric breakdown data*

Description

The dialetric breakdown strength of electrical insulation

Usage

```
data(breakdown)
```

Format

A data frame containing 128 observations with the following 3 variables.

`Strength` the dialetric breakdown strength, in kiloVolts

`Time` the time exposure in weeks; one of `1`, `2`, `4`, `8`, `16`, `32`, `48`, or `64`

`Temperature` the temperature, in degrees Celsius; one of `180`, `225`, `250` or `275`

Details

The data come from a study of performance degradation of electrical insulation from accelerated tests. The study can be considered as a 8-by-4 factorial experiment, with four measurements for each time–temperature combination.

Source

OZDASL, available on-line at http://www.statsci.org/data/general/dialectr.html.

References

Nelson, W. (1981) Analysis of performance-degradation data. *IEEE Transactions on Reliability*, **2**, R-30, 149–155.

The Statistical Reference Datasets page: http://www.itl.nist.gov/div898/strd/nls/data/nelson.shtml.

Examples

```
data(breakdown)
summary(breakdown)
```

bttstudy *The South African Birth to Ten (BTT) study*

Description

The data record details about the Birth to Ten study (BTT) in South Africa during 1990

Usage

```
data(bttstudy)
```

Format

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

`Counts` the number of subjects in the given classification; a numeric vector

`Group` the group the mother belongs to; a numeric vector with levels `1` (mothers not followed up), `2` (mothers followed up five years later)

`MedicalAid` whether or not the mother had medical aid; a factor with levels `No` and `Yes`

`Race` the mother's race; a factor with levels `Black` and `White`

Details

The data record details about the Birth to Ten study (BTT), performed in the greater Johannesburg/Soweto metropolitan area of South Africa during 1990. In the study, all mothers of singleton births were interviewed during a seven-week period between April and June to women with permanent addresses in a defined area (a total of 4019 births). Five years later, 964 of these mothers were re-interviewed. If the mothers interviewed later and representative of the original populations, the two groups should show similar characteristics. One of those characteristics is documented here: the proportion with and without medical aid.

Source

Christopher H. Morrell (1999) Simpson's Paradox: An example from a longitudinal study in South Africa. *Journal of Statistics Education*, **7**(3).

Examples

```
data(bttstudy)
summary(bttstudy)
```

budworm *Insecticide doses and tobacco budworm*

Description

The number of tobacco budworms dying at various doses of pyrethroid

Usage

```
data(budworm)
```

Format

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

`Killed` the number of budworms killed at each dose; a numeric vector

`Number` the number of budworms exposed at each dose; a numeric vector

`Dose` the dose of pyrethroid *trans*-cypermethrin in micrograms; a numeric vector

`Gender` the gender of the budworms; a factor with levels `F` (female) and `M` (male)

Details

The data concern the tobacco budworm *Heliothis virescens* and the doses of pyrethroid *trans*-cypermethrin (to which the moths were beginning to show resistance). Twenty male and twenty female moths were exposed at each of six doses of the pyrethroid, and the number that were killed recorded.

Source

W. N. Venables and B. D. Ripley (1997). *Modern Applied Statistics with* S-PLUS, second edition. Springer-Verlag: New York (p 230)

D. Collett (1991). *Modelling Binary Data*. Chapman and Hall: London.

Examples

```
data(budworm)
summary(budworm)
```

`butterfat` *Butterfat and dairy cattle*

Description

The average butterfat content for dairy cattle

Usage

```
data(butterfat)
```

Format

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

`Butterfat` the average butterfat percentage; a numeric vector

`Breed` the cattle breed; a factor with levels `Ayrshire`, `Canadian`, `Guernsey`, `Holstein-Fresian` and `Jersey`

`Age` the age of the cow; a factor with levels `2year` and `Mature`

Details

The data give the average butterfat content (percentages) 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.

Source

D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway, and E. Ostrowski (1994) *A Handbook of Small Data Sets*, London: Chapman and Hall. Dataset 23.

R. R. Sokal and F. J. Rohlf (1981) *Biometry*, 2nd edition, San Fransisco: WH Freeman.

Examples

```
data(butterfat)
summary(butterfat)
```

ccancer *Canadian cancers*

Description

The estimated number of deaths from cancer in three regions of Canada by cancer site and gender

Usage

```
data(ccancer)
```

Format

A data frame with 30 observations on the following 5 variables.

`Count` the estimated number of deaths by the given cancer; a numeric vector

`Gender` gender; a factor with levels either \codeF (female) or codeM (male)

`Region` the region; a factor with levels `Ontario`, `Newfoundland` or `Quebec`

`Site` the cancer site; a factor with levels `Lung`, `Colorectal`, `Breast`, `Prostate` or `Pancreas`

`Population` the estimated population of the region in 2000/20001; a numeric vector

Details

The cancer data are *estimated* number of deaths in 2000 from the five leading cancer sites

Source

Cancer estimates from: Canadian Cancer Society. Canadian cancer statistics 2000. Published on the internet: `http://www.cancer.ca/stats2000/tables/tab5e.htm`. Accessed 19 September 2001.

Population estimates from: *The Daily*, Tuesday September 25, 2001. (Accessed on the internet: `http://www.statcan.gc.ca/daily-quotidien/010925/dq010925a-eng.htm` now https://www150.statcan.gc.ca/n1/daily-quotidien/010925/dq010925a-eng.htm)

Examples

```
data(ccancer)
summary(ccancer)
```

ceo *CEO salaries*

Description

The age and salary of CEOs of small companies

Usage

```
data(ceo)
```

Format

A data frame with 60 observations on the following 2 variables.

`Age` the age of the CEO in completed years; a numeric vector

`Salary` the salary of the CEO (including bonuses) in thousands of dollars; a numeric vector

Details

The age and salary of CEOs of small companies (annual sales greater than 5 and less than 350 million dollars); companies were ranked according to 5-year average return on investment. The first 60 firms are listed.

Source

The Data and Story Library (DASL) (formerly `http://lib.stat.cmu.edu/DASL/` now https://dasl.datadescription.com)

References

Originally from *Forbes*, November 8, 1993 "America's Best Small Companies".

Examples

```
data(ceo)
plot(ceo)
```

`cervical` *Deaths from cervical cancer*

Description

The number of deaths from cervical cancer in four countries

Usage

```
data(cervical)
```

Format

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

`Country` the country; a factor with levels `EngWales` (England and Wales), `Belgium`, `France` and `Italy`

`Age` the age group; a factor with levels `25to34`, `35to44`, `45to54`, `55to64`

`Deaths` the number of deaths; a numeric vector

`Wyears` the woman-years of risk; a numeric vector

Details

The data give the number of deaths from cervical cancer, and the woman-years of risk, for various age groups and four countries.

Source

A. S. Whittermore and G. Gong (1991) Poisson regression with misclassified counts: Applications to cervical cancer mortality rates. *Applied Statistics*, **40**(1), 81–93.

Examples

```
data(cervical)
with( cervical, plot(Deaths/Wyears ~ Age) )
```

cheese *Tasting cheese*

Description

The taste of cheddar cheese

Usage

```
data(cheese)
```

Format

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

`Taste` the combined taste scores from several judges (presumably higher scores correspond to better taste); a numeric vector

`Acetic` the concentration of acetic acid in the cheese (units unknown); a numeric vector

`H2S` the concentration of hydrogen sulphide (units unknown); a numeric vector

`Lactic` the concentration of lactic acid (units unknown): a numeric vector

Details

The data give information on taste and concentration of various chemical components of matured 30 cheddar cheeses from the LaTrobe Valley in Victoria, Australia.

The final `Taste` score is a combination of the taste scores from several tasters.

Source

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

The Statlib data base: formerly `http://lib.stat.cmu.edu/DASL/Datafiles/Cheese.html` now https://dasl.datadescription.com.

References

G. P. McCabe, L. McCabe, A. Miller. Analysis of taste and chemical composition of cheddar cheese 1982–83 experiments, CSIRO Division of Mathematics and Statistics Consulting Report VT85/6.

I. Barlow, et al. (1989) Correlations and changes in flavour and chemical parameters of cheddar cheeses during maturation. *Australian Journal of Dairy Technology*, **44**, 7–18.

According to Moore and McCabe (1993), the data are based on the experiments of G. T. Lloyd and E. H. Ramshaw.

Examples

```
data(cheese)
plot(cheese)
```

cins *Canadian car insurance data*

Description

Details of the Canadian car insurance industry

Usage

```
data(cins)
```

Format

A data frame with 20 observations on the following 6 variables.

`Merit` the merit rating; a factor with levels `Merit3` (licensed and accident free 3 or more years), `Merit2` (licensed and accident free 2 or more years), `Merit1` (licensed and accident free 1 or more years), `Merit0` (all others)

`Class` the vehicle class; a factor with levels `Class1` (pleasure, no male operator under 25), `Class2` (pleasure, non-principal male operator under 25), `Class3` (business use), `Class4` (unmarried owner or principal operator under 25), `Class5` (married owner or principal operator under 25)

`Insured` the earned car-years; a numeric vector

`Premium` earned premiums in 1000s of dollars (adjusted to equivalent 2001 rates); a numeric vector

`Claims` the number of claims; a numeric vector

`Cost` total cost of the claim in 1000s of dollars; a numeric vector

Details

The data are for all of Canada except Saskatchewan, and refer to private passenger automobile liability for non-farmers. The data are for policy years 1956 and 1957, as of 30 June 1959.

Source

The data was downloaded from OZDASL http://www.statsci.org/data/general/carinsca.html where it was prepared by Gordon Smyth from Bailey and Simon (1960).

References

Robert A. Bailey and LeRoy J. Simon (1960) Two studies in automobile insurance ratemaking. *ASTIN Bulletin*, **I(IV)**:192-217.

Examples

```
data(cins)
summary(cins)
```

`crawl` *The age at which babies start to crawl*

Description

The age at which babies start to crawl, the birth month and average monthly temperature six months after the birth month

Usage

```
data(crawl)
```

Format

A data frame with 12 observations on the following 5 variables.

`BirthMonth` the baby's birth month; levels such as `January` and `July`

`Age` the mean age (in completed weeks) at which the babies born this month started to crawl; a numeric vector

`SD` the standard deviation (in completed weeks) of the crawling ages for babies born this month; a numeric vector

`SampleSize` the number of babies in the study born in the given month; a numeric vector

`Temp` the monthly average temperature (in degrees F) six months after the birth month; a numeric vector

Details

The data come from a study which hypothesized that babies would take longer to learn to crawl in colder months because the extra clothing restricts their movement. From 1988–1991, recorded were the babies' first crawling age and the average monthly temperature 6 months after birth (when "infants presumably enter the window of locomotor readiness"). The parents reported the birth month, and age when their baby first crept or crawled a distance of four feet in one minute. Data were collected at the University of Denver Infant Study Center on 208 boys and 206 girls, and summarized by the birth month.

Source

Janette Benson (1993) Season of birth and onset of locomotion: Theoretical and methodological implications. *Infant Behavior and Development*, **16**(1), 69–81.

Thanks to Janette Benson for granting permission to use this data set.

Examples

```
data(crawl)
plot(Age ~ Temp, data=crawl, cex=0.05*SampleSize, pch=19)
```

cyclones *Cyclones near Australia*

Description

The data give the number of severe and non-severe tropical cyclones from 1969 to 2005 in the Australian region

Usage

```
data(cyclones)
```

Format

A data frame with 37 observations on the following 8 variables.

`Year` the year

`Severe` the number of severe cyclones recorded; a numeric vector

`NonSevere` the number of non-severe cyclones; a numeric vector

`Total` the total number of cyclones (the sum of `Severe` and `NonSevere`); a numeric vector

`JFM` the Ocean Niño Index, or ONI, averaged over the months January to March; a numeric vector

`AMJ` the Ocean Niño Index, or ONI, averaged over the months April to June; a numeric vector

`JAS` the Ocean Niño Index, or ONI, averaged over the months July to September; a numeric vector

`OND` the Ocean Niño Index, or ONI, averaged over the months October to December; a numeric vector

Details

The data give the number of severe and non-severe cyclones tropical cyclones from 1970 to 2005 in the Australian region (south of equator; 105 to 160 degrees E). Severe cyclones are defined as those with a minimum central pressure less than 970 hPa.

The ONI is based on a three-month running mean of ERSST.v3b Sea Surface Temperature (SST) anomalies in the Niño 3.4 region (5 degrees N to 5 degrees S, 120 degrees to 170 degrees W), based on the 1971 to 2000 base period.

Source

Cyclone information: http://www.bom.gov.au/cyclone/climatology/trends.shtml (accessed 04 April 2011).

Ocean Niño Index: http://www.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ensoyears.shtml (accessed 04 April 2011).

Examples

```
data(cyclones)
plot(Severe~JFM, data=cyclones )
```

danishlc *Danish lung cancer*

Description

The number of cases of lung cancer in four Danish cities

Usage

```
data(danishlc)
```

Format

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

`Cases` the number of lung cancer cases; a numeric vector

`Pop` the population of each age group in each city; a numeric vector

`Age` the age group; a factor with levels `40-54`, `55-59`, `60-64`, `65-69`, `70-74` and `>74`

`City` the city; a factor with levels `Fredericia`, `Horsens`, `Kolding` and `Vejle`

Details

The data gives the number of cases of lung cancer in four Danish cities between 1968 and 1971 inclusive.

Source

James K. Lindsey (1995) *Modelling frequency and count data*. Clarendon Press, page 157.

References

The original source is: E. B. Andersen (1977) Multiplicative Poisson models with unequal cell rates. *Scandinavian Journal of Statistics*, **4**, 153–158.

Examples

```
data(danishlc)
plot(Cases/Pop ~ City, data=danishlc)
```

dental *Decayed, missing and filled teeth*

Description

The data give the estimates of the mean number of decayed, missing and filled teeth (DMFT) at age 12 years, and the mean annual sugar consumption in the previous five years for 90 countries

Usage

```
data(dental)
```

Format

A data frame with 90 observations on the following 4 variables.

`Country` the country; a factor

`Indus` whether the country is considered an industrialized country; a factor with levels `Ind` (industrialized) or `NonInd` (not industrialized)

`Sugar` the mean annual sugar consumption in kilograms per person per year, computed over the five years (or as much as available) prior to the survey; a numeric vector

`DMFT` estimates of the mean number of decayed, missing and filled teeth at age 12; a numeric vector

Details

The data give the estimates of the mean number of decayed, missing and filled teeth (DMFT) at age 12 years, and the mean annual sugar consumption in the previous five years for 90 countries. For some countries, data on sugar consumption was unavailable for the previous five years, and the average was computed for the available data; see Woodward and Walker (1994) for details.

Source

M. Woodward and A. R. P. Walker (1994) Sugar consumption and dental caries: evidence from 90 countries. *British Dental Journal*, **176**, 297–302.

References

M. Woodward (2004) *Epidemiology: Study Design and Data Analysis*, second edition. Chapman and Hall.

Examples

```
data(dental)
plot(DMFT ~ Sugar, data=dental )
```

`deposit` *Insecticides*

Description

The number of insects killed at various doses of insecticide

Usage

```
data(deposit)
```

Format

A data frame with 18 observations on the following 4 variables.

`Killed` the number of insects killed at each poison level; a numeric vector

`Number` the number of insects exposed at each poison level; a numeric vector

`Insecticide` the insecticide used; a factor with levels `A`, `B` and `C`

`Deposit` the amount of deposit (insecticide) used in milligrams; a numeric vector

Details

Fifty insects were exposed to various deposits of insecticides. The proportions of the insects killed after six days exposure were recorded.

Source

P. S. Hewlett and T. J. Plackett (1950) Statistical aspects of the independent joint action of poisons, particularly insecticides. II. Examination of data for agreement with hypothesis. *Annals of Applied Biology*, **37**, 527–552.

References

Wotjek J. Krzanowski (1998) *An Introduction to Statistical Modelling*, Arnold: London.

Examples

```
data(deposit)
summary(deposit)
```

downs *Downs Syndrome cases in British Columbia*

Description

The number of Downs Syndrome cases in British Columbia, Canada

Usage

```
data(downs)
```

Format

A data frame with 30 observations on the following 3 variables.

`Age` the average age of the mother in each group, in completed years; a numeric vector

`Births` the number of live births; a numeric vector

`DS` the number of Downs Syndrome births; a numeric vector

Details

The data give the number of Downs Syndrome cases from 1961–1970 in British Columbia, Canada, in 30 age categories for the mother.

Note

The ages are the means of the ages in defined groups, rounded to one decimal place.

Source

Charles J. Geyer (1991) Constrained maximum likelihood exemplified by isotonic convex logistic regression. *Journal of the American Statistical Association*, **86**(415), 717–724.

References

The data are originally from the British Columbia Health Surveillance Registry.

The data also appear in A. C. Davison and D. V. Hinkley (1997) *Bootstrap Methods and their Applications*, Cambridge University Press, Table 7.12, though there are very slight differences in their data to ours, in the decimal places for age. (The differences are very minor, and will not affect conclusions.)

Examples

```
data(downs)
plot( DS/Births ~ Age, data=downs)
```

dwomen *Depression and children*

Description

The data give the number of women from a sample in Camberwell, South London, who developed depression in a one-year period

Usage

```
data(dwomen)
```

Format

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

`Counts` the counts in each category; a numeric vector

`Depression` whether depression was observed; a factor with levels `Yes` and `No`

`SLE` whether a Severe Life Event was observed; a factor with levels `Yes` and `No`

`Children` whether the woman had three children under 14; a factor with levels `Yes` and `No`

Details

The data give the number of women from a sample in Camberwell, South London, who developed depression in a one-year period.

Source

B. S. Everitt and A. M. R. Smith (1979) Interactions in a contingency tables: a brief discussion of alternative definitions. *Psychological Medicine*, **9**, 581–583.

References

D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway, and E. Ostrowski (1994) *A Handbook of Small Data Sets*, London: Chapman and Hall. Dataset 391 (second table).

Examples

```
data(dwomen)
summary(dwomen)
```

dyouth *Depression in adolescents*

Description

The number of seriously emotionally disturbed and learning disabled adolescents and their reported depression levels.

Usage

```
data(dyouth)
```

Format

A data frame with 24 observations on the following 5 variables.

`Obs` the number of observed adolescents in the given category; a numeric vector

`Age` the age group; a factor with levels `12-14`, `15-16` and `17-18`

`Group` the group; a factor with levels `LD` (learning disabled) and `SED` (serious emotionally disturbed)

`Gender` the gender; a factor with levels `F` (female) and `M` (male)

`Depression` the depression level; a factor with levels `H` (high) and `L` (low)

Details

The data come from a study of seriously emotionally disturbed and learning disabled adolescents and their reported depression levels. The adolescents were classified by age and gender and their depression levels.

Source

J. W. Maag and J. T. Behrens (1989) Epidemiologic data on seriously emotionally disturbed and learning disabled adolescents: reporting extreme depressive symptomatology. *Behavioral Disorders*, **15**(1).

References

D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway, and E. Ostrowski (1994) *A Handbook of Small Data Sets*, London: Chapman and Hall.

Examples

```
data(dyouth)
summary(dyouth)
```

`earinf` *Ear infections in swimmers*

Description

The number of ear infections in swimmers

Usage

```
data(earinf)
```

Format

A data frame with 287 observations on the following 5 variables.

`Swim` how often the swimmer swims in the ocean; a factor with levels `Freq` (frequently) and `Occas` (occasionally)

`Loc` the reported usual swimming location; a factor with levels `Beach` and `NonBeach`

`Age` the age group; a factor with levels `15-19`, `20-24` and `25-29`

`Sex` the sex; a factor with levels `Female` and `Male`

`NumInfec` the number of self-diagnosed ear infections; a numeric vector

`Infec` whether there are self-diagnosed ear infections; a numeric vector where `0` means no self-reported infection, and `1` means at least one self-reported ear infection

Details

The data give the number of self-reported ear infections in the 1990 Pilot Surf/Health Study of NSW Water Board.

Source

This data file was downloaded from OZDASL (http://www.statsci.org/data/oz/earinf.html) where it was prepared by Dr Gordon Smyth from Hand et al (1994) Dataset 328.

References

D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway, and E. Ostrowski (1994) *A Handbook of Small Data Sets*, London: Chapman and Hall. Dataset 328.

Examples

```
data(earinf)
summary(earinf)
```

emeraldaug *August monthly rainfall in Emerald*

Description

The total monthly rainfall in Emerald, Australia, and the average monthly SOI

Usage

```
data(emeraldaug)
```

Format

A data frame with 114 observations on the following 3 variables.

`Year` the year; a numeric vector

`Rain` the total monthly rainfall in August of the given year; a numeric vector

`SOI` the monthly average southern oscillation index (SOI); a numeric vector

`Phase` the SOI phase (see Stone and Auliciems, 1992); a factor with these values: `1` (consistently negative), `2` (consistently positive), `3` (rapidly falling), `4` (rapidly rising), or `5` (consistently near zero)

Details

The data give the total monthly rainfall and monthly in Emerald, Queensland, Australia, from 1889 to 2002, and the average SOI for the corresponding month.

Source

Data obtained from the Australian Bureau of Meteorology (http://www.bom.gov.au) and IRI/LDEO Climate Data Library (http://www.longpaddock.qld.gov.au/seasonalclimateoutlook/southernoscillationindex/soidatafiles/inde on 21 December 2010, then compiled.

References

R. C. Stone and A. Auliciems (1992) SOI phase relationships with rainfall in eastern Australia, *International Journal of Climatology*, **12**, 625–636.

Examples

```
data(emeraldaug)
plot(emeraldaug)
```

`energy` *Energy expenditure*

Description

The energy expenditure for 104 females at rest for a 24 hour period

Usage

```
data(energy)
```

Format

A data frame with 104 observations on the following 3 variables.

`Energy` the energy expenditure (units not given); a numeric vector

`Fat` the mass of fat tissue (units not given); a numeric vector

`NonFat` the mass of fat-free tissue (units not given); a numeric vector

Details

The data give the energy expenditure for 104 females at rest over a 24 hour period; the mass of fat and fat-free tissue was also recorded.

Note that the total mass of each subject is the sum of the fat and fat-free tissue masses.

Source

B. Joergensen (1992) Exponential dispersion models and extensions: A review. *International Statistical Review*, **60**(1), 5–20.

References

L. Garby, J. S. Garrow, B. Joergensen, O. Lammert, K. Madsen, P. Soerensen and J. Webster (1988) Relation between energy expenditure and body composition in man: Specific energy expenditure in *vivo* of fat and fat-free tissue. *European Journal of Clinical Nutrition*, **42**, 301–305.

Examples

```
data(energy)
summary(energy)
```

failures *Failures of electronic equipment*

Description

The number of failures of electronic equipment operating in two modes

Usage

```
data(failures)
```

Format

A data frame with 18 observations on the following 4 variables.

`Period` the time period; a numeric vector

`Time1` the time spent in Mode 1 in the given period (units not given); a numeric vector

`Time2` the time spent in Mode 2 in the given period (units not given); a numeric vector

`Failures` the number of failures in the given period; a numeric vector

Details

The data give the number of failures of a piece of electronic equipment after operating in two modes.

Source

Dale W. Jorgensen (1961) Multiple regression analysis of a Poisson process. *Journal of the American Statistical Association*, **56**(294), 235–245.

Examples

```
data(failures)
summary(failures)
```

feedrates *Feeding rates of birds*

Description

The daily individual feeding rates of chestnut-crowned babblers

Usage

```
data(feedrates)
```

Format

A data frame containing 1293 observations with the following 11 variables.

`SocGroup` the social group for the bird; 27 levels

`NestID` the nest identifier; 61 levels

`ObsTime` the length of observation (in decimal hours); a numeric vector

`Ring` an identifier for individual birds; 97 levels

`Sex` the sex of the bird; one of `f` (female) or `m` (male)

`Age` the age of non-breeding group members; one of `adult` or `yearling`

`Relatedness` the pedigree-based relatedness to the brood; one of `0.5` (first-order relatives); `0.25` (second-order relatives) or `0` (more distant relatives)

`ChickAge` the age of the brood, in days; a numeric vector

`BroodSize` the size of the brood: a numeric vector

`UnitSize` the number of individuals in the unit; a numeric vector

`FeedingRate` the daily individual feeding rates, in feeds per hour; a numeric vector

Details

The data relate to a population of colour-ringed population of chestnut-crowned babblers in an area of the University of New South Wales Arid Zone Research Station, (Fowlers Gap, western New South Wales, Australia). The study determined whether, where and how often non-breeding group members contributed to providing for nestlings by monitoring the visit rate of tagged birds during 2007 and 2008.

Source

L. E. Browning, S. C. Patrick, L. A. Rollins, S. C. Griffith, and A. F. Russell (2012) Kin selection, not group augmentation, predicts helping in an obligate cooperatively breeding bird. *Proceedings of the Royal Society B*, **279**(1743): 3861–3869. doi: 10.1098/rspb.2012.1080

L. E. Browning, S. C. Patrick, L. A. Rollins, S. C. Griffith, and A. F. Russell (2012) Data from: Kin selection, not group augmentation, predicts helping in an obligate cooperatively breeding bird. *Dryad Digital Repository*. doi: 10.5061/dryad.ff868

References

L. E. Browning, S. C. Patrick, L. A. Rollins, S. C. Griffith, and A. F. Russell (2012) Kin selection, not group augmentation, predicts helping in an obligate cooperatively breeding bird. *Proceedings of the Royal Society B*, **279**(1743): 3861–3869. doi: 10.1098/rspb.2012.1080

Examples

```
data(feedrates)
summary(feedrates)
```

`fineroot` *The root length density of apple trees*

Description

The root length density of apple trees

Usage

```
data(fineroot)
```

Format

A data frame with 511 observations on the following 5 variables.

`Plant` the plant number; a numeric vector

`Rstock` the root stock; a factor with levels `Mark`, `MM106` or `M26`

`Spacing` the plant spacing; a factor with levels `5x3` or `4x2` (measured in metres)

`Zone` the zone relative to the plant from which the soil core is taken; a factor with levels `Inner` or `Outer`

`RLD` the root length density in centimetres per cubic centimetre; a numeric vector

Details

The data concern the underground root system of eight apple trees. Three different root stocks and two plant spacings are used; the root length density (the density of the fine roots) is measured in one of the two zones.

The design is not full factorial: plants 1 and 2 are for `Mark` rootstock at `5x3` spacing; plants 3 and 4 are for `Mark` rootstock at `4x2` spacing; plants 5 and 6 are for `MM106` rootstock at `5x3` spacing; plants 7 and 8 are for `M26` rootstock at `4x2` spacing.

Source

Personal communication from Nihal de Silva.

References

H. N. de Silva, A. J. Hall, D. S. Tustin and P. W. Gandar (1999) Analysis of distribution of root length density of apple trees on different dwarfing rootstocks. *Annals of Botany*, **83**, 335–345.

P. K. Dunn and G. K. Smyth (2005) Series evaluation of Tweedie exponential dispersion model densities. *Statistics and Computing*, **15**(4), 267–280.

Examples

```
data(fineroot)
summary(fineroot)
```

`fishfood` *Food consumption for fish*

Description

The food consumption for various fish species

Usage

```
data(fishfood)
```

Format

A data frame with 33 observations on the following 6 variables.

`Species` the fish species; an identifier

`MaxWt` the mean asymptotic (or maximum) weight of the fish in grams; a numeric vector

`Temp` the mean habitat temperature in degrees Celsius; a numeric vector

`AR` the aspect ratio of the fish; a numeric vector

`Food` the food type for the fish; a factor with levels `C` for carnivores, and `H` for herbivores

`FoodCon` the daily food consumption of a fish population as a percentage of its biomass; a numeric vector

Details

The computation of the aspect ratio is detailed in the source.

Source

M. L. Palomares and D. Pauly (1989) A multiple regression model for predicting the food consumption of marine fish population. *Australian Journal of Marine and Freshwater Research*, **40**(3), 259–284.

Examples

```
data(fishfood)
summary(fishfood)
```

`flathead` *Tiger flathead from trawls*

Description

Information about tiger flathead trawls

Usage

```
data(flathead)
```

Format

A data frame containing 169 observations with the following 7 variables.

`Lon` the longitude of the trawl

`Lat` the latitude of the traw

`Depth` the depth (bathymetry) of the trawl, in metres

`Distance` the distance along a 100 metre depth contour for the trawl (northwards of all trawls from an arbitrary origin), in metres

`Area` the area swept, in hectares

`Number` the number of tiger flathead caught

`Biomass` the total biomass of tiger flathead caught, in kg

Details

The data give details of trawls in the South East Fisheries ecosystem off Australia. The data were originally collected by Bax and Williams (2000).

Source

R package `fishMod`: Foster (2016).

References

Nicholas~J. Bax and Alan Williams (2000) Habitat and fisheries production in the South East fishery ecosystem. Final Report 1994/040, Fisheries Research and Development Corporation.

Scott~D. Foster and Mark~V. Bravington (2013) A Poisson–Gamma model for analysis of ecological data. *Environmental and Ecological Statistics*, 20(4):533–552.

Scott D. Foster (2016) fishMod: Fits Poisson-Sum-of-Gammas GLMs, Tweedie GLMs, and Delta Log-Normal Models. R package version 0.29. https://CRAN.R-project.org/package=fishMod

Examples

```
data(flathead)
summary(flathead)
```

`flowers` *The average number of meadowfoam flowers*

Description

The average number of meadowfoam flowers in given light conditions

Usage

```
data(flowers)
```

Format

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

`Flowers` the mean number of flowers per meadowfoam plant, averaged over ten seedlings; a numeric vector

`Light` the light intensity in μmol per square metre per second; a numeric vector

`Timing` when the light treatment was applied; a factor with levels `PFI` (photoperiodic floral induction) or `Before` (24 days before PFI)

Details

The data are collected from an experiment to study how to maximize Mermaid meadowfoam production. (Meadowfoam is a small plant from which a vegetable oil can be extracted.)

These data are consistent with those in Seddigh and Joliff (1994). The data were estimated from their Figure 3, and then adjusted to produce, as closely as possible, the statistics given on those graphs.

Source

M. Seddigh and G.D. Joliff (1994) Light intensity effects on meadowfoam growth and flowering. *Crop Science*, **34**: 497–503.

Examples

```
data(flowers)
summary(flowers)
```

`fluoro` *The time of fluoroscopy and total radiation*

Description

The data give the total procedure time during CT fluoroscopic scanning, and the radiation dose received.

Usage

```
data(fluoro)
```

Format

A data frame with 19 observations on the following 2 variables.

`Time` the total procedure time (in minutes); a numeric vector

`Dose` the total radiation dose received (in rads); a numeric vector

Details

The data are given in the Table as the natural log of `Time` and the natural log of `Dose`. Here the data have been transformed back to the original scale. The source claims the purpose of the data collection was "to assess whether radiation dose could be estimated by simply measuring the total CT fluoroscopic procedure time". The procedure was performed in the abdomen.

Source

Kelly H. Zou, Kemal Tuncali, and Stuart G. Silverman (2003) Correlation and simple linear regression. *Radiology*, **227**, 617–628.

References

The data were originally used, but not given, in S. G. Silverman, K. Tuncali, D. F. Adams, R. D. Nawfel, K. H. Zou, and P. F. Judy (1999) CT fluoroscopy-guided abdominal interventions: techniques, results, and radiation exposure. *Radiology*, **212**, 673–681.

Examples

```
data(fluoro)
plot(fluoro)
```

`galapagos` *Gal\'apagos Island species data*

Description

The number of species on the Gal\'apagos Islands

Usage

```
data(galapagos)
```

Format

A data frame containing 29 observation with the following 11 variables.

`Island` the name of the island

`Plants` the number of plant species; a numeric vector

`PlantEnd` the number of endemic plant species; a numeric vector

`Finches` the number of finch species; a numeric vector

`FinchEnd` the number of endemic finch species; a numeric vector

`FinchGenera` the number of finch genera; a numeric vector

`Area` the area of each island in square kilometres; a numeric vector

`Elevation` the maximum elevation of each island in metres; a numeric vector

`Nearest` the distance to the nearest island; a numeric vector

`StCruz` the distance to Santa Cruz Island in kilometres; a numeric vector

`Adjacent` the area of adjacent island in square kilometres; a numeric vector

Details

The data give the number of plant species and related variables for 29 different islands. Counts are given for both the total number of species and the number of species that occur only in the Gal\'apagos (the endemics).

Note

Elevations for Baltra and Seymour obtained from web searches. Elevations for four other small islands obtained from large-scale maps.

Source

Michael P. Johnson and Peter H. Raven (1973) Species number and endemism: The Gal\'apagos Archipelago revisited. *Science*, **179**(4076), 893–895.

Examples

```
data(galapagos)
summary(galapagos)
```

germ *Germination of seeds*

Description

In an experiment, the number of seeds germination was recorded for two types of seeds and two types of root extracts

Usage

```
data(germ)
```

Format

A data frame with 21 observations on the following 4 variables.

`Germ` the number seeds germinating; a numeric vector

`Total` the number of seeds planted; a numeric vector

`Extract` the extract type; a factor with levels `Bean` and `Cucumber`

`Seeds` the type of seed; a factor with levels `OA75` (*O. aegyptiaca 75*) and `OA73` (*O. aegyptiaca 73*)

Details

The data gives the total number of seeds and the number germinating, for two types of seeds and two types of root stocks; the dilution is 1 in 25 in all cases.

Note

An alternative representation of these data are given in `germBin`.

Source

Martin J. Crowder (1978) Beta-binomial anova for proportions. *Applied Statistics*, **27**(1), 34–37.

References

The following sources also quote the data, but have reversed the two seed types from the original source:

P. J. Smith and D. F. Heitjan (1993). Testing and adjusting for departures from nominal dispersion in generalized linear models. *Applied Statistics*, **42**, 31–41 (Table 1).

D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway, and E. Ostrowski (1994). *A Handbook of Small Data Sets*, London: Chapman and Hall. Dataset 420.

Examples

```
data(germ)
summary(germ)
```

germBin *Germination of seeds*

Description

In an experiment, the number of seeds germination was recorded for two types of seeds and two types of root extracts

Usage

```
data(germ)
```

Format

A data frame with 831 observations on the following 3 variables.

`Extract` the extract type; a factor with levels `Bean` and `Cucumber`

`Seeds` the type of seed; a factor with levels `OA75` (*O. aegyptiaca 75*) and `OA73` (*O. aegyptiaca 73*)

`Result` the result of the experiment: either `Germ` (the seed germinated) or `NotGerm` (the seed did not germinate)

Details

The data gives the total number of seeds and the number germinating, for two types of seeds and two types of root stocks; the dilution is 1 in 25 in all cases.

Note

These data are the same as `germ` but with one row for each seed.

Source

Martin J. Crowder (1978) Beta-binomial anova for proportions. *Applied Statistics*, **27**(1), 34–37.

References

The following sources also quote the data, but have reversed the two seed types from the original source:

P. J. Smith and D. F. Heitjan (1993). Testing and adjusting for departures from nominal dispersion in generalized linear models. *Applied Statistics*, **42**, 31–41 (Table 1).

D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway, and E. Ostrowski (1994). *A Handbook of Small Data Sets*, London: Chapman and Hall. Dataset 420.

Examples

```
data(germBin)
summary(germBin)
```

`gestation` *Gestation time*

Description

The gestation time for 1513 infants

Usage

```
data(gestation)
```

Format

A data frame with 21 observations on the following 4 variables.

`Age` the gestational age in weeks; a numeric vector

`Births` the number of births; a numeric vector

`Weight` the mean birthweight in kilograms; a numeric vector

`SD` the standard deviation of the birthweight in each group in kilograms; a numeric vector

Details

The gestation time for 1513 infants born in St George's Hospital, London, to Caucasian mothers willing to participate between August 1982 and March 1984.

Source

J. M. Bland, J. L. Peacock, H. R. Anderson, and O. G. Brooke (1990) The adjustment of birthweight for very early gestational ages: two related problems in statistical analysis. *Applied Statistics*, **39**(2), 229–239.

Examples

```
data(gestation)
summary(gestation)
```

`gforces` *G-induced loss of consciousness*

Description

Loss of consciousness induced by G-forces)

Usage

```
data(gforces)
```

Format

A data frame containing 8 observations with the following 3 variables.

`Subject` the initials of the subject; a text identifier

`Age` the age of the subject, in years; a numeric vector

`Signs` whether the subject showed syncopal blackout-related signs: a factor with levels `0` (No) and `1` (Yes)

Details

Military pilots sometimes black out when their brains are deprived of oxygen due to G-forces during violent manoeuvres. Glaister and Miller (1990) produced similar symptoms by exposing volunteers' lower bodies to negative air pressure, likewise decreasing oxygen to the brain. The data lists the subjects' ages and whether they showed syncopal blackout related signs (pallor, sweating, slow heartbeat, unconsciousness) during an 18 minute period.

Source

The data were obtained electronically from OZDASL (http://www.statsci.org/data/). The Details above were obtained from this webpage.

References

D. H. Glaister and N. L. Miller (1990) Cerebral tissue oxygen status and psychomotor performance during lower body negative pressure (LBNP). *Aviation, Space and Environmental Medicine*. **61**(2), 99–105.

L. C. Hamilton (1992) *Regression with Graphics: a second course in applied statistics*. Duxbury, page 243.

Examples

```
data(gforces)
summary(gforces)
```

gopher *Clutch sizes of Gopher tortoises*

Description

The clutch sizes from various studies of Gopher tortoises

Usage

```
data(gopher)
```

Format

A data frame with 19 observations on the following 6 variables.

`Site` the site number (an identifier); a numeric vector

`Latitude` the latitude at which the study was conducted; a numeric vector

`Evap` the mean total annual actual evapotranspiration (in mm); a numeric vector

`Temp` the mean annual temperature in degrees Celsius; a numeric vector

`ClutchSize` the mean clutch size; a numeric vector

`SampleSize` the size of the sample upon which the `ClutchSize` was computed; a numeric vector

Details

Nineteen populations of Gopher tortoises were examined across 17 different studies; from each study, the mean clutch size and various other variables were compiled.

Source

K. G. Ashton, R. L. Burke, and J. N. Layne (2007) Geographic variation in body and clutch size of Gopher tortoises. *Copeia*, May 16, Number 2, 355–363.

Examples

```
data(gopher)
summary(gopher)
```

gpsleep *Sleep times for guinea pigs*

Description

Amount of sleep in guinea pigs after receiving ketamine

Usage

```
data(gpsleep)
```

Format

A data frame with 30 observations on the following 2 variables.

`Sleep` the minutes of sleep (zero means the guinea pig did not sleep); a numeric vector

`Dose` the dose of ketamine in mg/kg body weight; a numeric vector

Source

R. C. Bailey, J. P. Summe, L. D. Homer, and L. E. McCraken (1978) A model for analysis of the anesthetic response, *Biometrics*. **34**(2), 223–232.

References

The original source is: L. E. McCracken, R. E. Toby, and R. Bailey (1977) Ketamine and thiopental sleep responses in hyperbaric helium oxygen in guinea pigs. *Undersea Biomedical Research*, **6**(4), 329–338.

Examples

```
data(gpsleep)
plot(Sleep~Dose, data=gpsleep)
```

grazing *Bird abundance in grazing areas*

Description

The density of understorey birds at a series of sites in two areas either side of a stockproof fence

Usage

```
data(grazing)
```

Format

A data frame with 62 observations on the following 3 variables.

`Birds` the number of understorey birds; a numeric vector

`When` when the bird count was conducted; a factor with levels `Before` (before herbivores were removed) and `After` (after herbivores were removed)

`Grazed` which side of the stockproof fence; a factor with levels `Reference` (grazed by native herbivores) and `Feral` (grazed by feral herbivores, mainly horses)

Details

In this experiment, the density of understorey birds at a series of sites in two areas either side of a stockproof fence were compared. Once side had limited grazing (mainly from native herbivores), and the other was heavily grazed by feral herbivores, mostly horses. Bird counts were done at the sites either side of the fence (the `Before` measurements). Then the herbivores were removed, and bird counts done again (the `After` measurements). The measurements are the total number of understorey-foraging birds observed in three 20-minute surveys of two hectare quadrats.

Source

Personal communication from Martine Maron.

References

Alison L. Howes, Martine Maron and Clive A. McAlpine (2010) Bayesian networks and adaptive management of wildlife habitat. *Conservation Biology*. **24**(4), 974–983.

Examples

```
data(grazing)
plot( Birds ~ When, data=grazing)
```

---

hcrabs *Males attached to female horseshoe crabs*

Description

The number of male crabs attached to female horseshoe crabs

Usage

```
data(hcrabs)
```

Format

A data frame with 173 observations on the following 5 variables.

`Col` the color of the female; a factor with levels `LM` (light medium), `M` (medium), `DM` (dark medium) or `D` (dark)

`Spine` the spine condition; a factor with levels `BothOK`, `OneOK` or `NoneOK`

`Width` the carapace width of the female crab in cm; a numeric vector

`Wt` the weight of the female crab in grams; a numeric vector

`Sat` the number of male crabs attached to the female ('satellites'); a numeric vector

Details

The data come from an observational study of nesting horseshoe crabs: "The study was conducted at two beaches on the Delaware shore, Breakwater Harbor at Cape Henlopen Park in Lewes and Fowler's Beach, 32 km north on the same shoreline (Sussex County, Delaware, USA). In 1991 observations were made from 7 to 17 June, in1992 from 28 May to 3 June and from 11 to 14 June, and in 1993 from 18 May to 11 June. At these sites the crabs were most active on the higher of the two daily high tides (which at this time of year are at night between 1700 and 0200 h EST)" (Brockmann, 1996; p. 4).

Source

H. J. Brockmann (1996) Satellite male groups in horseshoe crabs, *Limulus polyphemus*. *Ethology*, **102**(1), 1–21.

Examples

```
data(hcrabs)
plot(Sat ~ Wt, data=hcrabs)
```

---

heatcap *Heat capacity of hydrobromic acid*

Description

The heat capacity of hydrobromic acid measured at various temperatures

Usage

```
data(heatcap)
```

Format

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

`Cp` the heat capacity (in calories per mole per degree Kelvin); a numeric vector

`Temp` the temperature (in Kelvin); a numeric vector

Details

The data give the heat capacity for hydrobromic acid at various temperatures.

Source

M. Shacham and N. Brauner (1997) Minimizing the effects of collinearity in polynomial regression. *Industrial and Engineering Chemical Research*, **36**, 4405–4412.

References

The original source is: W. F. Giauque and R. Wiebe (1929) The heat capacity of hydrogen bromide from $15°$K to its boiling point and its heat of vaporization. The entropy from spectroscopic data. *Journal of the American Chemical Society*, **51**(5), 1441–1449.

Examples

```
data(heatcap)
plot(heatcap)
```

---

humanfat *Human age and fatness*

Description

The age and percent body fat for 18 adults

Usage

```
data(humanfat)
```

Format

A data frame with 18 observations on the following 4 variables.

`Age` the age of the subject in completed years; a numeric vector

`Percent.Fat` the body fat percentage; a numeric vector

`Gender` the gender; a factor with levels `F` (females) or `M` (males)

`BMI` the body mass index in metres per kilogram-squared; a numeric vector

Details

The data come from a study investigating a new method of measuring body composition. The body fat percentage, age and gender is given for 18 adults aged between 23 and 61. "Eighteen normal adult subjects were measured including four young males and 14 females (age 25 to 60 years). None of these subjects had chronic diseases, were taking medications, or had skeletal fractures indicative of osteoporosis" (Mazess et al. (1984), p. 835). The BMI is computed from the weights and heights given in the original source.

Source

R. B. Mazess, W. W. Peppler, and M. Gibbons (1984) Total body composition by dualphoton (^{153}Gd) absorptiometry. *American Journal of Clinical Nutrition*, **40**, 834–839.

References

D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway, and E. Ostrowski (1994) *A Handbook of Small Data Sets*, London: Chapman and Hall. Dataset 17.

Examples

```
data(humanfat)
summary(humanfat)
```

---

janka *Janka hardness*

Description

The Janka hardness of Australian hardwoods

Usage

```
data(janka)
```

Format

A data frame containing 36 observations with the following 2 variables.

`Density` the hardwood density (units unknown); a numeric vector

`Hardness` the Janka hardness (units unknown); a numeric vector

Details

The data give the Janka hardness (which is hard to measure) and the density of Australian hardwoods (which is easier to measure).

Source

W. N. Venables (1998) Exegeses on linear models. In *S-Plus User's Conference*, Washington DC.

References

Williams, E. J. (1959) *Regression Analysis*, Wiley, New York.

Examples

```
data(janka)
plot(janka)
```

kstones *Treating kidney stones*

Description

Treatment of kidney stones

Usage

```
data(kstones)
```

Format

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

`Counts` the number of subjects in the given classification; a numeric vector

`Size` whether the subject has kidney stones with mean diameter less than 2cm (coded as `Small`) or greater than or equal to 2cm (coded as `Large`); a factor with levels `Large` and `Small`

`Method` the treatment method; a factor with levels `A` (open surgery) or `B` (percutaneous nephrolithotomy)

`Outcome` the outcome of the stated treatment; a factor with levels `Failure` and `Success`

Details

The data give the success rates of two methods of treating kidney stones: open surgery methods, and percutaneous nephrolithotomy.

The given data are a subset of that reported by Charig et al. (1986), who also include two other methods of treatment, and also break up the open surgery methods into three sub-groups. The two methods here were chosen because they demonstrate Simpson's paradox.

Source

C. R. Charig, D. R. Webb, S. R. Payne, and J. A. E. Wickham (1986) Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorpeal shockwave lithotripsy. *British Medical Journal*, **292**, 29 March, 879–882.

Steven A. Julious and Mark A. Mullee (1994) Confounding and Simpson's paradox. *British Medical Journal*, **309**(1480):1480–1481.

Examples

```
data(kstones)
summary(kstones)
```

lactation *Lactation of dairy cows*

Description

Lactation of dairy cows over time

Usage

```
data(lactation)
```

Format

A data frame containing 35 observations on the following 2 variables.

`Yield` the average daily far yield from a dairy cow, in kg/day

`Week` the week in which the data were collected

Details

The data give data from a lactating dairy cow, recording the average daily fat yield over 35 weeks.

Source

Harold V. Henderson and Charles E. McCulloch (1990) Transform or link? Technical Report BU-049-MA, Cornell University.

Examples

```
data(lactation)
plot(lactation$Yield ~ lactation$Week)
```

leafblotch *Percentage leaf area of leaf blotch*

Description

The percentage leaf area of barley infected with leafblotch

Usage

```
data(leafblotch)
```

Format

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

`Area` the percentage area infected with leaf blotch; a numeric vector

`Site` the site; a factor with levels `A`, `B` up to `I`

`Variety` the variety of barley; a factor with levels `1`, `2`, up to `9`

Details

The data give the percentage leaf area of barley infected with *Rhynchosporium secalis*, or leaf blotch, for ten different barley varieties grown at nine different sites.

Source

R. W. M. Wedderburn (1974) Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, **61**(3), 439–447.

References

The data also appear in McCullagh and Nelder, p 329, and in Faraway (2006), Exercise 7.5.

Examples

```
data(leafblotch)
plot( Area ~ Site, data=leafblotch)
```

---

leukwbc *Leukaemia survival times*

Description

The times to death and white blood cell counts for two groups of leukaemia patients

Usage

```
data(leukwbc)
```

Format

A data frame with 33 observations on the following 3 variables.

`WBC` the white blood cell count; a numeric vector

`Time` the time to death in weeks; a numeric vector

`AG` the morphological variable, the AG factor; a numeric vector where `1` means AG-positive and `2` means AG-negative

Details

The data gives the times to death (in weeks) and white blood cell counts for two groups of leukaemia patients, AG-positive and AG-negative. The two groups have not been created by random allocation.

Source

P. Feigl and M. Zelen (1965) Estimation of exponential survival probabilities with concomitant information. *Biometrics*, **21**, 826–838.

References

D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway, and E. Ostrowski (1994) *A Handbook of Small Data Sets*, London: Chapman and Hall. Dataset 424.

Examples

```
data(leukwbc)
summary(leukwbc)
```

---

lime *Small-leaved lime trees*

Description

Data from small-leaved lime trees grown in Russia

Usage

```
data(lime)
```

Format

A data frame containing 385 observations with the following 4 variables.

`Foliage` the foliage biomass, in kg (oven dried matter)

`DBH` the tree diameter, at breast height, in cm

`Age` the age of the tree, in years

`Origin` the origin of the tree; one of `Coppice`, `Natural`, `Planted`

Details

The data give measurements from small-leaved lime trees (*Tilia cordata*) growing in Russia.

Source

Schepaschenko, Dmitry; Shvidenko, Anatoly; Usoltsev, Vladimir A; Lakyda, Petro; Luo, Yunjian; Vasylyshyn, Roman; Lakyda, Ivan; Myklush, Yuriy; See, Linda; McCallum, Ian; Fritz, Steffen; Kraxner, Florian; Obersteiner, Michael (2017): Biomass tree data base. doi:10.1594/PANGAEA.871491, In supplement to: Schepaschenko, D et al. (2017): A dataset of forest biomass structure for Eurasia. *Scientific Data*, 4, 170070, doi:10.1038/sdata.2017.70. Extracted from https://doi.pangaea.de/10.1594/PANGAEA.871491

References

The source (Schepaschenko et al.) obtains the data from various sources:

- Dylis N.V., Nosova L.M. (1977) *Biomass of forest biogeocenoses under Moscow region*. Moscow: Nauka Publishing.
- Gabdelkhakov A.K. (2015) *Tilia cordata Mill.* tree biomass in plantations and coppice forests. *Eco-potential*. No. 3 (11). p. 7–16.
- Gabdelkhakov A.K. (2005) *Tilia cordata Mill.* tree biomass in plantations. *Ural forests and their management*. Issue 26. Yekaterinburg: USFEU. p. 43–51.
- Polikarpov N.P. (1962) Scots pine young forest dynamics on clear cut. *Moscow: Academy of Sci.* USSR.
- Prokopovich E.V. (1995) Ecological conditions of soil forming and biological cycle of matters in spruce forests of the Middle Ural. Ph.D. Thesis. Ekaterinburg: Plant and Animals Ecology Institute.
- Remezov N.P., Bykova L.N., Smirnova K.M. (1959) Uptake and cycling of nitrogen and ash elements in forests of European part of USSR. Moscow: State University.
- Smirnov V.V. (1971) Organic mass of certain forest phytocoenoses at European part of USSR. Moscow: Nauka.
- Uvarova S.S. (2005) Biomass dynamics of *Tilia cordata* trees on the example of Achit forest enterprise of Sverdlovsk region. *Ural forests and their management*. Issue 26. Ekaterinburg: State Forest Engineering University, p. 38–40.
- Uvarova S.S. (2006) Growth and biomass of *Tilia cordata* forests of Sverdlovsk region Dissertation. Ekaterinburg: State Forest Engineering University. (USFEU library)

Examples

```
data(lime)
summary(lime)
```

---

lungcap *Lung capacity and smoking in youth*

Description

The health and smoking habits of 654 youth

Usage

```
data(lungcap)
data(lungcapsub)
```

Format

A data frame with 654 observations on the following 5 variables. (The data frame `lungcapsub` contains the data only for smokers, and hence does not contain the variable `Smoke`.)

`Age` the age of the subject in completed years; a numeric vector

`FEV` the forced expiratory volume in litres, a measure of lung capacity; a numeric vector

`Ht` the height in inches; a numeric vector

`Gender` the gender of the subjects: a numeric vector with females coded as `0` and males as `1`

`Smoke` the smoking status of the subject: a numeric vector with non-smokers coded as `0` and smokers as `1`

Details

The data give information on the health and smoking habits of a sample of 654 youths, aged 3 to 19, in the area of East Boston during middle to late 1970s.

Source

Kahn, Michael (2005) An exhalent problem for teaching statistics. *The Journal of Statistical Education*, **13**(2). Available on-line.

References

Kahn, M. (2003) Data Sleuth, *STATS*, **37**, 24.

Ira B. Tager, Scott T. Weiss, Alvaro Munoz, Bernard Rosner, and Frank E. Speizer (1983) Longitudinal study of the effects of maternal smoking on pulmonary function in children. *New England Journal of Medicine*, **309**(12):699–703.

Examples

```
data(lungcap)
summary(lungcap)
```

`mammary` *Adult mammary stem cells*

Description

Assay results from a study of adult mammary stem cells

Usage

```
data(mammary)
```

Format

A data frame containing results from 81 assays, compiled into five rows of data, with the following 3 variables.

`N.Cells` the average number of calls in each assay

`N.Assays` the number of assays at that cell number

`N.Outgrowths` the number of assays giving a positive outcome (i.e. seeing a milk gland outgrowth)

Details

The data give measurements from an assay analysis of adult mammary stem cells.

Source

Mark Shackleton, Francois Valliant, Kaylene J. Simpson, John Sting, Gordon K. Smyth, Marie-Liesse Asselin-Labat, Li Wu, Geoffrey J. Lindeman, and Jane E. Visvader (2006). Generation of a functional mammary gland from a single stem cell. *Nature*, **439**:84–88.

References

Mark Shackleton, Francois Vaillant, Kaylene J. Simpson, John Sting, Gordon K. Smyth, Marie-Liesse Asselin-Labat, Li Wu, Geoffrey J. Lindeman, and Jane E. Visvader (2006) Generation of a functional mammary gland from a single stem cell. *Nature*, **439**:84–88.

Examples

```
data(mammary)
summary(mammary)
```

mandible *Mandible length and gestational age*

Description

The data give the mandible length and gestational age for 167 foetuses from the 12th week of gestation onwards

Usage

```
data(mandible)
```

Format

A data frame with 167 observations on the following 2 variables.

`Age` the gestational age (in weeks); a numeric vector

`Length` the mandible length (in mm); a numeric vector

Details

The data give the mandible length and gestational age for 167 foetuses from the 12th week of gestation onwards, measured using ultrasound.

Source

Patrick Royston and Douglas G. Altman (1994) Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. *Applied Statistics*, **43**(3), 429–467.

Examples

```
data(mandible)
plot(mandible)
```

manuka *Manuka honey and wound healing*

Description

The pH and wound size of wounds before and after treatment with Manuka honey

Usage

```
data(manuka)
```

Format

A data frame containing 20 observations (from 17 patients) with the following 6 variables.

`Aetiology` the aetiology of the wound; one of `V` (venous), `A` (arterial), `M` (mixed) or `P` (pressure ulcer)

`Duration` the duration of the wound; units not given

`Size0` the initial wound size, in square centimetres

`pH0` the initial wound pH

`Size2` the wound size after 2 weeks, in square centimetres

`pH2` the wound pH after 2 weeks

Details

The data give the pH and wound size for 20 lower-leg wounds on 17 patients, giving 20 observations on 6 variables. The `Duration` is never explained or used.

The article Gethin et al. (2008) is subject to a retraction notice.

Source

Gethin, Cowman and Conroy (2008), Table 1.

References

Gethin, G. T., Cowman, S., and Conroy, R. M. (2008) The impact of Manuka honey dressings on the surface pH of chronic wounds. *International Wound Journal*, 5(2):185–194.

International Wound Journal (2014), Retraction. 11: 342. doi:10.1111/iwj.12275

Examples

```
data(manuka)
summary(manuka)
```

---

motorins *Swedish third-party car insurance*

Description

The data give details of third-party motor insurance claims in Sweden for the year 1977.

Usage

```
data(motorins)
data(motorins1)
```

Format

A data frame with 2182 observations on the following 7 variables

`Kilometres` the number of kilometres travelled per year; a numeric vector with levels `1` (less than 10000), `2` (from 10000 to 15000), `3` (15000 to 20000), `4` (20000 to 25000) or `5` (more than 25000)

`Zone` geographical zone (only in `motorins`); a numeric vector with levels `1` to `7` ()see Details below)

`Bonus` no claim bonus; a numeric vector equal to the number of years plus one since the last claim

`Make` the make of vehicle; a numeric vector with levels from `1` to `8` representing eight common cart models, and `9` representing all other models

`Insured` the number of insured in policy-years; a numeric vector

`Claims` the number of claims; a numeric vector

`Payment` the total value of payments in Skoner; a numeric vector

Details

For variable `Zone`, the geographical zones are:

1	Stockholm, Goteborg, Malmo with surroundings
2	Other large cities and surroundings
3	Small cities in northern Sweden
4	Small cities in southern Sweden
5	Rural areas in northern Sweden
6	Rural areas in southern Sweden
7	Gotland

The file `motorins1` only contains the data from Zone 1 (and hence `Zone` is not one of the variables in that data set).

"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" (Andrews and Herzberg (1985), p. 413).

Make 4 is the Volkswagen 1200, which was discontinued shortly after 1977. The other makes could not be identified because of the potential for the data to impact on sales of those cars.

For this data, the number of claims has a Poisson distribution, and the amount of each claim follows a gamma distribution very nicely. The total claim has a Tweedie distribution.

Source

The OZDASL datasets. The data were obtained electronically from the Statlib database by Dr Gordon Smyth for OZDASL (`http://www.statsci.org/data/`).

References

M. Hallin and J.-F. Ingenbleek (1983) The Swedish automobile portfolio in 1977. A statistical study. *Scandinavian Actuarial Journal*, 49–64. The data are not listed in this reference.

D. F. Andrews and A. M. Herzberg (1985) *Data. A collection of problems from many fields for the student and research worker*. Springer, New York, pages 413–421. Only the data from Zone 1 are listed (that is, corresponds to `motorins1`).

Examples

```
data(motorins)
summary(motorins)
```

`mutagen` *Mutagenicity assay*

Description

The number of revertant colonies for various doses of quinoline for TA98 Salmonella

Usage

```
data(mutagen)
```

Format

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

`Dose` the dose of quinoline; a numeric vector

`Colonies` the number of revertant colonies; a numeric vector

Details

The number of revertant colonies (colonies that revert to their former gentype) for various doses of quinoline for TA98 Salmonella.

The given data represent only one replicate of the three given in Margolin, Kim and Risko (1984), but are as given in Breslow (1989).

Three plates were used for each dose, hence the three observations per dose. The data are given in order of increasing numbers of colonies.

Theory suggests one model for the data is Count $= T[1 - \exp(a - bx)]\exp(-cx)$, for b and c greater than or equal to zero, where x is the dose of quinoline. A good approximation to this is the log-linear model $\log(\text{Count}) = A + B\log(x + C) - Dx$.

Source

N. E. Breslow (1984) Extra-Poisson variation in log-linear models. *Applied Statistics*, **33**(1), 38–44.

References

B. H. Margolin, N. Kaplan, E. and Zeiger (1981).\ Statistical analysis of the Ames Salmonella/microsome test. *Proceedings of the National Academy of Science* USA, **76**, 3779–3783.

Examples

```
data(mutagen)
summary(mutagen)
```

`mutantfreq` *Cell mutant frequencies in children*

Description

The "somatic cell mutant frequencies at the *hprt* locus of the X-chromosome" in healthy children

Usage

```
data(mutantfreq)
```

Format

A data frame with 49 observations on the following 5 variables.

`Donor` the donor identifier; a factor

`Sex` the sex of the child; a factor with levels `F` (females) or `M` (males)

`Age` the age of the child in completed years; a numeric vector

`Ceff` the mean unselected cloning efficiency; a numeric vector

`Mfreq` the mutant frequencies $\times 10^{-6}$; a numeric vector

Details

In the original paper, the children are sometimes referred to as belonging to Group II (Ages 0 to 5), Group III (Ages 6 to 11) or Group IV (Ages 12 to 17). (Group I refers to cord data referenced to another article.) `Age` may be treated as categorical with these categories.

Source

B. A. Finette, L. M. Sullivan, J. P. O'Neill, J. A. Nicklas, P. M. Vacek and R. J. Albertini (1994) Determination of *hprt* mutant frequencies in T-lymphocytes from a healthy pediatric population: statistical comparison between newborn, children and adult mutant frequencies, cloning efficiency and age. *Mutation Research*, **308**, 223–231.

Examples

```
data(mutantfreq)
summary(mutantfreq)
```

nambeware *Nambeware products*

Description

Information about the production of various tableware products

Usage

```
data(nambeware)
```

Format

A data frame with 59 observations on the following 4 variables.

`Type` the type of product; a factor with levels `Bowl`, `CassDish`, `Dish`, `Plate` and `Tray`

`Diam` the diameter of the product in inches; a numeric vector

`Time` the total grinding and polishing time in minutes; a numeric vector

`Price` the price in US dollars; a numeric vector

Details

The data come from Nambe Mills (https://www.nambe.com/), manufacturers of tableware made from sand casting a special alloy of several metals. The polishing times for the products are thought to be related to the size of the item, as indicated by the diameter. After casting, the pieces go through a series of shaping, grinding, buffing, and polishing steps. In 1989 the company began a program to rationalize its production schedule of some 100 items in its tableware line. The total grinding and polishing times listed here were a major output of this program.

Source

The data are originally from the Nambe Mills company, as quoted as the DASL website (https://dasl.datadescription.com/datafile/nambe/).

Examples

```
data(nambeware)
summary(nambeware)
```

nhospital *Naval hospital maintenance*

Description

The monthly maintenance hours associated with maintaining the anaesthesiology service for twelve naval hospitals

Usage

```
data(nhospital)
```

Format

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

`MainHours` the monthly maintenance hours associated with maintaining the anaesthesiology service for twelve naval hospitals in the USA; a numeric vector

`Cases` the number of surgical cases; a numeric vector

`Eligible` the eligible population per thousand; a numeric vector

`OpRooms` the number of operating rooms; a numeric vector

Details

The monthly maintenance hours associated with maintaining the anaesthesiology service for twelve naval hospitals in the USA was measured, together with some explanatory variables

Source

D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway, and E. Ostrowski (1994) *A Handbook of Small Data Sets*, London: Chapman and Hall. Dataset 269.

References

Raymond H. Myers (1990) *Classical and Modern Regression with Applications*, second edition, Duxbury: Belmont, CA.

Examples

```
data(nhospital)
summary(nhospital)
```

`nitrogen` *Soil nitrogen*

Description

The soil nitrogen after applying different fertilizer doses

Usage

```
data(nitrogen)
```

Format

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

`Fert` the fertilizer dose, in kilograms of nitrogen per hectare; a numeric vector

`SoilN` the soil nitrogen, in kilograms of nitrogen per hectare; a numeric vector

`Source` the fertilizer source: a factor with levels `0` (inorganic) and `1` (organic; farmyard manure)

Details

The data give the soil inorganic nitrogen content for various fertilizer doses, including a control. One application is from an organic source. Each level of fertilizer has data from three replications.

Source

P. W. Lane (2002) Generalized linear models in soil science. *European Journal of Soil Science*, **53**:241–251.

References

Glendining, M.J., Poulton, P.R. & Powlson, D.S. (1992) The relationship between inorganic N in soil and the rate of fertilizer N applied on the Broadbalk Wheat Experiment. *Aspects of Applied Biology*, **30**, 95–102.

Examples

```
data(nitrogen)
summary(nitrogen)
```

`nminer` *Noisy miner abundance*

Description

The number of noisy miners detected in various 2 hectare transects in buloke woodland patches within the Wimmera Plains of western Victoria, Australia

Usage

```
data(nminer)
```

Format

A data frame with 31 observations on the following 9 variables.

`Miners` the presence or absence of noisy miners; a numeric vector with levels `1` (present) or `0` (absent)

`Eucs` the number of eucalypts in each 2 hectare transect; a numeric vector

`Area` the area in hectares of contiguous remnant patch of vegetation in which the transect was located; a numeric vector

`Grazed` whether the area was grazed or not; a numeric vector with levels `0` (grazed) or `1` (not grazed)

`Shrubs` whether shrubs were present in the transect or not; a numeric vector with levels `1` (shrubs present) or `0` (shrubs not present)

`Bulokes` the number of buloke trees in each 2 ha transect; a numeric vector

`Timber` the number of pieces of fallen timber in the transect; a numeric vector

`Minerab` the number of noisy miners (abundance) observed in three 20 minute surveys; a numeric vector

Details

The data gives the number of noisy miners detected in various two hectare transects in buloke woodland patches within the Wimmera Plains of western Victoria, Australia. The noisy miner is a small but aggressive native Australian bird.

Source

Personal communication from Martine Maron.

References

Martine Maron (2007) Threshold effect of eucalypt density on an aggressive avian competitor. *Biological Conservation*, **136**, 100–107.

Examples

```
data(nminer)
summary(nminer)
```

---

paper *The tensile strength of paper*

Description

The tensile strength of Kraft paper with varying hardwood concentrations

Usage

```
data(paper)
```

Format

A data frame with 19 observations on the following 2 variables.

`Strength` the paper strength (in pounds per square inch (psi)); a numeric vector

`Hardwood` the hardwood concentration in the paper in percent; a numeric vector

Details

The data give the strength of 25 samples of Kraft paper (a strong, coarse, usually brownish type of paper) for varying amounts of hardwood.

Source

G. Joglekar, J. H. Schuenemeyer and V. LaRicca (1989) Lack-of-fit testing when replicates are not available. *American Statistician*, **43**, 135–143.

References

D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway, and E. Ostrowski (1994) *A Handbook of Small Data Sets*, London: Chapman and Hall. Dataset 271. (The response and explanatory variables are reversed from those in the original article.)

D. C. Montgomery and E. A. Peck (1982) *Introduction to linear regression analysis*, New York: John Wiley.

Examples

```
data(paper)
plot(paper)
```

`perm` *Permeability of building materials*

Description

The permeability of building materials

Usage

```
data(perm)
```

Format

A data frame with 81 observations on the following 3 variables.

`Day` the day; a factor with levels `1` up to `9`

`Mach` the machine used for measurement; a factor with levels `A`, `B` or `C`

`Perm` the permeability in seconds: a numeric vector

Details

The data give the average permeability (in seconds) of eight sheets of building materials, using random samples of 81 sheets in three machines over nine days, with three measurements for each machine–day combination.

Source

Bent Joergensen (1992) Exponential dispersion models and extensions: A review. *International Statistical Review*, **60**(1), 5–20.

References

A. Hald (1952) *Statistical theory with engineering applications*. New York: Wiley.

Examples

```
data(perm)
summary(perm)
```

phosphorus *Soil phosphorus*

Description

The amount of phosphorus in soil samples

Usage

```
data(phosphorus)
```

Format

A data frame with 18 observations on the following 4 variables.

`Sample` an identifier, the sample ID; a numeric vector

`Inorg` the amount of inorganic phosphorus chemically determined in ppm (parts per million); a numeric vector

`Org` the amount of organic phosphorus chemically determined in ppm; a numeric vector

`PA` the amount of plant-available phosphorus of corn grown in the soil in ppm; a numeric vector

Details

Chemical determinations of the phosphorus in the soil at 18 locations in Iowa were determined, including the amount of available phosphorus for growing corn at 20 degrees C.

Source

S. M. Snappin and R. D. Small (1986) Tests of significance using regression models for ordered categorical data. *Biometrics*, **42**, 583–592.

References

D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway, and E. Ostrowski (1994) *A Handbook of Small Data Sets*, London: Chapman and Hall. Dataset 237.

Examples

```
data(phosphorus)
summary(phosphorus)
```

pock *Pock counts*

Description

In an experiment, "viral activity was assessed from pock counts at a series of dilutions of the viral medium"

Usage

```
data(pock)
```

Format

A data frame with 48 observations on the following 2 variables.

`Count` the number of membrane pock counts; a numeric vector

`Dilution` the dilution factor; a numeric vector

Details

The data come from a titration bioassay, in which viral activity was assessed from pock counts at different dilutions of the viral medium.

Source

P. J. Smith and D. F. Heitjan (1993) Testing and adjusting for departures from nominal dispersion in generalized linear models. *Applied Statistics*, **42**, 31–41 (Table 1).

Examples

```
data(pock)
with( pock, tapply( Count, list(Dilution), mean) )
with( pock, tapply( Count, list(Dilution), var) )
```

poison *Survival times of animals*

Description

The survival times of animals under various treatments and poisons

Usage

```
data(poison)
```

Format

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

`Psn` the type of poison; a vector with levels `I`, `II` or `III`

`Trmt` the type of treatment; a vector with levels `A`, `B`, `C` or `D`

`Time` the time to death in ten-hour units; a numeric vector

Details

The data give the time to death of animals using one of three different poisons and one of four treatments. For each of the twelve combinations, four times are recorded.

Source

G. E. P. Box and D. R. Cox (1964) An analysis of transformations (with discussion). *Journal of the Royal Statistical Society*, Series A. **143**, 383–430.

References

The data also appear in D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway, and E. Ostrowski (1994) *A Handbook of Small Data Sets*, London: Chapman and Hall. Dataset 403.

Examples

```
data(poison)
summary(poison)
```

---

polyps *The number of polyps and suldinac*

Description

The number of polyps in people with familial adenomatous polyposis, after being given a placebo or a new drug

Usage

```
data(polyps)
```

Format

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

`Number` the number of polyps; a numeric vector

`Treatment` the treatment group; a factor with levels `Drug` (suldinac), `Placebo`

`Age` the age of the person; a numeric vector

Details

The data give the number of polyps in people with famial adenomatous polyposis, after being given a placebo or a new drug (suldinac).

Source

B. S. Everitt and T. Hothorn (2006) *A Handbook of Statistical Analyses Using* R Chapman & Hall/CRC, Table 6.1.

References

F. N. Giardiello, S. R. Hamilton, A. J. Krush, S. Piantadosi, L. M. Hylind, P. Celano, S. V. Booker, C. R. Robinson, and G. J. A. Offerhaus (1993) Treatment of colonic and rectal adenomas with suldindac in famial adenomatous polyposis, *New England Journal of Medicine*, **328**(18), 1313–1316.

S. Piantadosi (1997) *Clinical trials: A methodologic perspective*, New York: John Wiley and Sons.

Examples

```
data(polyps)
coplot( Number ~ Age | Treatment, data=polyps )
```

---

polythene *Cosmetic company use of polythene*

Description

The usage in tonnes of polythene as a packaging materials for 23 UK cosmetic companies (year unknown)

Usage

```
data(polythene)
```

Format

A data frame with 23 observations on the following 3 variables.

`Company` the UK cosmetic company identifier; a numeric vector with levels from `1` to `23`

`Polythene` the amount of polythene used in tonnes for packaging; a numeric vector

`Turnover` the annual company turnover in hundreds of thousands of pounds; a numeric vector

Source

Robert Gilchrist (2000) Regression models for data with a non-zero probability of a zero response. *Communications in Statistics—Theory and Methods*, **29**, 1987–2003.

Examples

```
data(polythene)
summary(polythene)
```

`punting` *Football punting*

Description

The right- and left-leg strengths of 13 American footballers (measured using a weight lifting test), plus the distance they punt a football (with their right leg).

Usage

```
data(punting)
```

Format

A data frame with 13 observations on the following 3 variables.

`Left` left-leg strength in pounds; a numeric vector

`Right` right-leg strength in pounds; a numeric vector

`Punt` punting distance in feet; a numeric vector

Source

Raymond H. Myers (1990) *Classical and modern regression with applications*, second edition. Duxbury; page 75.

References

These appear to come from a larger data set, available from (for example) OZDASL at http://www.statsci.org/data/general/punting.html.

Examples

```
data(punting)
plot(Punt ~ Right, data=punting)
```

quilpie *Total July rainfall at Quilpie*

Description

The total July rainfall at Quilpie, Queensland, Australia from 1921 to 1988

Usage

```
data(quilpie)
```

Format

A data frame with 68 observations on the following 6 variables.

`Year` the year; a numeric vector

`Rain` the total monthly July rainfall in millimetres; a numeric vector

`SOI` the July average southern oscillation index, or SOI; a numeric vector

`Phase` the SOI phase (see Stone and Auliciems, 1992); a factor with these values: `1` (consistently negative), `2` (consistently positive), `3` (rapidly falling), `4` (rapidly rising), or `5` (consistently near zero)

`Exceed` an indicator for whether or not the total monthly July rainfall exceeds 10 millimetres: a factor where `Yes` means the rainfall exceeds 10mm, and `No` means the rainfall is 10mm or less

`y` an indicator for whether or not the total monthly July rainfall exceeds 10 millimetres: a factor where `1` means the rainfall exceeds 10mm, and `0` means the rainfall is 10mm or less

Source

Data obtained from IRI/LDEO Climate Data Library (formerly `http://ingrid.ldgo.columbia.edu` now http://iridl.ldeo.columbia.edu) on 26 May 2009.

References

R. C. Stone and A. Auliciems (1992) SOI phase relationships with rainfall in eastern Australia, *International Journal of Climatology*, **12**, 625–636.

Examples

```
data(quilpie)
plot( Rain ~ SOI, data=quilpie)
plot( Rain ~ factor(Phase), data=quilpie)
```

`ratliver` *Drugs present in rat livers*

Description

The data describe an experiment conducted to investigate the amount of drug present in the liver of a rat.

Usage

```
data(ratliver)
```

Format

A data frame with 19 observations on the following 4 variables.

`BodyWt` the body weight of each rat in grams; a numeric vector

`LiverWt` the weight of each liver in grams; a numeric vector

`Dose` the relative dose of the drug given to each rat as a fraction of the largest dose; a numeric vector

`DoseInLiver` the proportion of the dose in the liver; a numeric vector

Details

The data describe an experiment conducted to investigate the amount of drug present in the liver of a rat. Nineteen rats were randomly selected, weighed, and placed under a light anesthetic and given an oral dose of the drug. Because it was thought that large livers would absorb more of a given dose than a small liver, the actual dose given was approximately determined as 40mg of the drug per kilogram of body weight. After a fixed length of time, each rat was sacrificed and the liver weighed, and the percent dose in the liver was determined.

Source

Sanford Weisberg (1985) *Applied Linear Regression*, second edition, New York: John Wiley and Sons, page 122.

Examples

```
data(ratliver)
summary(ratliver)
```

rootstock *Rootstock data*

Description

The data come from an experiment to investigate the propogation of plum root-stocks

Usage

```
data(rootstock)
```

Format

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

`Count` the number in each category; a numeric vector

`Time` the time of planting; a numeric vector with levels `Now` (straight away) or `Spring` (spring)

`Length` the length of the cutting; a numeric vector with levels `Long` or `Short`

`Condition` the condition of the cutting at the end of the experiment; a numeric vector with levels `Alive` or `Dead`

Source

M. S. Bartlett (1935) Contingency table interactions. *Journal of the Royal Statistical Society Supplement*, **2**, 248–252.

Examples

```
data(rootstock)
summary(rootstock)
```

rrates *Oxidation rate of benzene*

Description

The initial rate of benzene oxidation over a vanadium oxide catalyst using three different reaction temperatures and varying oxygen and benzene concentrations

Usage

```
data(rrates)
```

Format

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

`Run` An identifier; a numeric vector

`Conc.O` the oxygen concentration (by 10000 gmole per litre); a numeric vector

`Temp` the temperature in degrees Kelvin; a numeric vector

`Rate` the reaction rate (by 10^9 gmole per gram) of catlyst per second; a numeric vector

Source

D. J. Pritchard, J. Downie, and D. W. Bacon (1977) Further consideration of heteroscedasticity in fitting kinetic models. *Technometrics*, **19**(3), 227–236.

References

Originally from Jaswal, Mann, Juusola and Downie (1969) The vapour-phase oxidation of benzene over a vandium pentoxide catalyst. *Canadian Journal of Chemical Engineering*, **47**(3), 284–287.

Examples

```
data(rrates)
summary(rrates)
```

---

rtrout *Weights of rainbow trout*

Description

Weights of rainbow trout at various doses of DCA

Usage

```
data(rtrout)
```

Format

A data frame containing 96 observations with the following 2 variables.

`Weight` the weight of the rainbow trout, in grams; a numeric vector

`Dose` the dose of 3, 4-dichloroaniline (DCA), in micrograms per litre; one of `0` (control), `19`, `39`, `39`, `71`, `120`, or `210`

Details

The data give the weight of 95 rainbow trout after exposure to DCA for 28 days (note that one observation is missing at a dose of 39). The aim of the study was to "determine the concentration level which causes 25% inhibition [i.e. weight loss] from the control" (Maul, p. 161).

Source

Crossland, N.O. (1985) A method to evaluate effects of toxic chemicals on fish growth. *Chemosphere*, **14**(11-12), 1855–1870.

References

Maul A. (1992) Application of generalized linear models to the analysis of toxicity test data. *Environmental Monitoring and Assessment*, **23**(1), 153–163.

Examples

```
data(rtrout)
summary(rtrout)
```

---

ruminant *Energy in ruminant's diets*

Description

Energy measurements on various ruminant diets

Usage

```
data(ruminant)
```

Format

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

`DryMatterDigest` the dry matter digestibility in feed (in percent)

`EnergyDigest` the energy digestibility in feed (in percent)

`Energy` the digestible energy content (in calories per gram)

Details

The data give measurements of energy of dry feed fed to Merino wethers aged 2 to 2.5 years.

Source

R. J. Moir (1961) A note on the relationship between the digestible dry matter and the digestible energy content of ruminant diets. *Australian Journal of Experimental Agriculture and Animal Husbandry*, **1**, 24–26.

Examples

```
data(ruminant)
plot(ruminant)
```

satiswt *Satisfaction with weight in youth*

Description

The number of children and youth aged 12–17 who are satisfied with their weight

Usage

```
data(satiswt)
```

Format

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

`Counts` the number of youth in the indicated category; a numeric vector

`Gender` gender; a factor with levels `F` (female) or `M` (male)

`WishWt` the youths' wish for their weight relative to now; a factor with levels `Thinner`, `Same` or `Heavier`

`Matur` when sexual maturity reached; a factor with levels `Late`, `Mid`, and `Early`

Details

The data come from a study of children and youth aged 12–17, sampled from the population of the United States in 1963.

Source

Paula Duke Duncan, Philip L. Ritter, Sanford M. Dornbusch, Ruth T. Gross, and J. Merrill Carlsmith (1985) The effects of pubertal timing on body image, school behavior, and deviance. *Journal of Youth and Adolescence*, **14**(3), 227–235. The data are inferred from Table II.

Examples

```
data(satiswt)
summary(satiswt)
```

sdrink *Soft drink delivery times*

Description

The time taken to deliver soft drinks to vending machines

Usage

```
data(sdrinks)
```

Format

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

`Time` the time taken to service the soft drink vending machine (in minutes); a numeric vector

`Cases` the number of cases of product stocked; a numeric vector

`Distance` the distance walked by the driver to service the vending machines (in feet); a numeric vector

Details

A soft drink bottler is analyzing vending machine service routes in his distribution system. He is interested in predicting the amount of time required by the route driver to service the vending machines in an outlet. The service activity includes the time taken to stock the machine with beverage products, and for minor maintenance and housekeeping.

The industrial engineer responsible for the study has suggested that the two most important variables affecting the delivery time are the number of cases of product stocked and the distance walked by the route driver.

Source

The data were obtained electronically from OZDASL (http://www.statsci.org/data/). The Details above were obtained from this webpage.

References

D. C. Montgomery and E. A. Peck (1992) *Introduction to Regression Analysis*. Wiley, New York. Example 4.1

Examples

```
data(sdrink)
summary(sdrink)
```

`seabirds` *Counts of seabirds*

Description

The number of four species of seabirds

Usage

```
data(seabirds)
```

Format

A data frame with 40 observations on the following 3 variables.

`Quadrat` the quadrat; a numeric factor with levels `0` through `10`

`Species` the species; a factor with levels `M` (murre), `CA` (crested auklet), `LA` (least auklet) and `P` (puffin)

`Count` the number of seabirds of the given species in the given quadrat; a numeric vector

Details

The data are counts of four seabird species in ten 0.25 square-km quadrats in the Anadyr Strait (off the Alaskan coast) during summer, 1998.

Source

Andrew R. Solow and Woollcott Smith (1991) Cluster in a heterogeneous community sampled by quadrats. *Biometrics*, **47**(1), 311–317.

References

D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway, and E. Ostrowski (1994) *A Handbook of Small Data Sets*, London: Chapman and Hall. Dataset 215.

Examples

```
data(seabirds)
summary(seabirds)
```

serum *Mice surviving doses of antipneumococcus serum*

Description

The number of mice surviving a test dose of culture with five different doses of antipneumococcus serum

Usage

```
data(serum)
```

Format

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

`Dose` the dose of antipneumococcus serum in cc; a numeric vector

`Number` the number of surviving mice; a numeric vector

`Survivors` the number of mice in each group; a numeric vector

Details

The number of mice surviving a test dose of culture with five different doses of antipneumococcus serum prior to being infected with pneumocci.

Source

J. O. Irwin and E. A. Cheeseman (1939) On the maximum-likelihood method of determining dosage–response curves and approximations to the median-effective dose, in cases of a quantal response. *Supplement to the Journal of the Royal Statistical Society*, **6**(2), 174–185.

Examples

```
data(serum)
summary(serum)
```

`setting` *Heat evolved by setting cement*

Description

The heat evolved from different formulations of Portland cement

Usage

```
data(setting)
```

Format

A data frame with 13 observations on the following 5 variables.

`A` the percentage by weight of tricalcium aluminate; a numeric vector

`B` the percentage by weight of tricalcium silicate; a numeric vector

`C` the percentage by weight of tetracalcium alumino ferrite; a numeric vector

`D` the percentage by weight of dicalcium silicate; a numeric vector

`Heat` the heat evolved in calories per gram of cement; a numeric vector

Source

D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway, and E. Ostrowski (1994) *A Handbook of Small Data Sets*, London: Chapman and Hall. Dataset 454.

References

The data are originally from H. Woods, H. H. Steinour, and H. P. Starke (1932) Effects of composition of Portland Cement on heat evolved during hardening. *Industrial and Engineering Chemistry*, **24**, 1207–1214.

Examples

```
data(setting)
summary(setting)
```

sharpener *Sharpener data*

Description

The sharpener data

Usage

```
data(sharpener)
```

Format

A data frame with 15 observations on the following 11 variables.

`Y` the measured response; a numeric vector

`X1` a measured predictor; a numeric vector

`X2` a measured predictor; a numeric vector

`X3` a measured predictor; a numeric vector

`X4` a measured predictor; a numeric vector

`X5` a measured predictor; a numeric vector

`X6` a measured predictor; a numeric vector

`X7` a measured predictor; a numeric vector

`X8` a measured predictor; a numeric vector

`X9` a measured predictor; a numeric vector

`X10` a measured predictor; a numeric vector

Details

The data come from a study about making a point.

Examples

```
### The data are actually random numbers, generated in R as follows:
nxvars <- 10   # The number of explanatory variables
nobs <- 15    # The number of observations
set.seed(5000) # To ensure reproducibility

# Ensure the response is normally distributed
y  <- round( rnorm( nobs,0,1), 2) + 10

# The explanatory variables
rd <- runif( nxvars*nobs, 0, 1)
rd <- round( matrix( rd, ncol=nxvars), 2)

# Convert to a dataframe
```

```
rdf <- data.frame( Y=y )

for (i in (1:nxvars)){
code <- paste( "rdf$X",i," <- rd[,",i,"]", sep="")
eval( parse(text=code))
}

head( rdf )
data(sharpener)
head( sharpener )
```

---

sheep *The daily energy requirements for wethers*

Description

The daily energy requirements for wethers at various weights

Usage

```
data(sheep)
```

Format

A data frame with 64 observations on the following 2 variables.

`Weight` the weight of each sheep in kg; a numeric vector

`Energy` the daily energy requirements in Mcal per day; a numeric vector

Details

The data measure the daily energy requirement of castrated male (wethers) grazing Merino sheep at various weights (measured by radioassay of urinary carbon dioxide). The energy requirements are useful for predicting meat production.

Source

D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway, and E. Ostrowski (1994) *A Handbook of Small Data Sets*, London: Chapman and Hall. Dataset 241.

D. Wallach and B. Goffinet (1987) Mean square error of prediction in models for studying ecological systems and agronomic systems. *Biometrics*, **43**(3), 561–573.

References

B. A. Young and J. L. Corbett (1972) Maintenance energy requirement of grazing sheep in relation to herbage availability. *Australian Journal of Agricultural Research*, **23**(1), 57–76.

Examples

```
data(sheep)
plot(Energy ~ Weight, data=sheep, pch=19)
```

`shuttles` *O-rings on the space shuttles*

Description

The number of O-rings damaged for 23 space shuttle launches

Usage

```
data(shuttles)
```

Format

A data frame containing 23 observation with the following 2 variables.

`Temp` the ambient air temperature in degrees Fahrenheit; a numeric vector

`Damaged` the number of primary O-rings damaged for 23 space shuttle launches

Details

The data give the ambient temperature and the number of primary O-rings damaged for 23 of the 24 space shuttle launches before the launch of the space shuttle *Challenger* on January 28, 1986. (Challenger was the 25th shuttle. One engine was lost at sea and could not be examined.) Each space shuttle contains 6 primary O-rings.

Source

Samprit Chatterjee, Mark S. Handcock and Jeffrey S. Simonoff (1995) *A Casebook for a First Course in Statistics and Data Analysis*, Wiley.

Siddhartha R. Dalal, Edward B. Fowlkes and Bruce Hoadley (1989) Risk analysis of the space shuttle: Pre-*Challenger* prediction of failure. *Journal of the American Statistical Association*, **84**(408), 945–957; Table 1.

Examples

```
data(shuttles)
plot(Damaged/6 ~ Temp, data=shuttles)
```

`teenconcerns` *Concerns of teenagers*

Description

Health concerns of teenagers

Usage

```
data(teenconcerns)
```

Format

A data frame with 16 rows, on the following 4 variables.

`Counts` the average number of calls in each assay

`Sex` the sex of the teenagers; one of `M` or `F`

`Age` the age groups of the teenagers; one of `12-15` or `16-17`

`Concern` the type of health concerns; one of `Sex`, `Menstrual`, `Healthy` or `Nothing`

Details

The data give the numbers of teenagers of two age groups with health concerns in specific areas: `Sex`, `Menstrual`, `Healthy` (that is, how healthy they are) or `Nothing` (no concerns at all). More specifically, these are the number of teens who would like to discuss these topics with their doctor. For males `M`, menstrual concerns can be treated as structural zeros.

Source

Brunswick, Ann F. (1971) Adolescent health, sex, and fertility. *American Journal of Public Health*, 61(4): 711–729. The numbers are inferred from the percentages in Table 3.

References

Christen, R. (2013) *Log-Linear Models*, Springer Texts in Statistics, Springer: New York.

Fienberg, S. E. (2007) *The Analysis of Cross-Classified Categorical Data*, Springer: New York.

Examples

```
data(teenconcerns)
summary(teenconcerns)
```

`toothbrush` *Effectiveness of toothbrushes*

Description

The effectiveness of two types of toothbrushes for males and females

Usage

```
data(toothbrush)
```

Format

A data frame with 52 observations on the following 5 variables.

`Subject` an identifier

`Sex` the sex of the subject; a factor with levels `F` (female) or `M` (male)

`Toothbrush` the type of toothbrush; a factor with levels `Hugger` or `Conventional`

`Before` the dental plaque index before brushing; a numeric vector

`After` the dental plaque index after brushing; a numeric vector

Details

The data give the plaque index before and after brushing for two types of toothbrushes for males and females. Each subject uses both toothbrushes. A dental plaque index of zero is the best possible score; brushing cannot make the score worse; `Before - After` is positive continuous with one exact zero.

Source

Reiko Aoki, Jorge A. Achcar, Heleno Bolfarine, and Julio M. Singer (2003) Bayesian analysis of null-intercept errors-in-variables regression for pretest/post-test data. *Journal of Applied Statistics*, **30**(1), 3–12.

References

J. M. Singer and D. F. Andrade (1997) Regression models for the analysis of pretest-posttest data. *Biometrics*, **53**, 729–735.

Examples

```
data(toothbrush)
with(toothbrush, plot(Before-After ~ Sex) )
with(toothbrush, plot(Before-After ~ Toothbrush) )
```

toxo *Toxoplasmosis and rainfall*

Description

The proportion of people sampled in 34 cities in El Salvador who tested positive for toxoplasmosis.

Usage

```
data(toxo)
```

Format

A data frame with 34 observations on the following 5 variables.

`City` the city from which the data comes; a numeric vector

`Rainfall` the recorded rainfall in millimetres at each city, presumably annual; a numeric vector

`Proportion` the proportion of those sampled who tested positive to toxoplasmosis; a numeric vector

`Sampled` the number of people sampled in each city; a numeric vector of integers

`Positive` the number of people who tested positive to toxoplasmosis; a numeric vector of integers

Details

The subjects are not randomly sampled within city.

Source

Bradley Efron (1986) Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association*, **81**(395), 709–721.

Examples

```
data(toxo)
summary(toxo)
```

`triangle` *Artificial data from triangles*

Description

The data are the lengths of three sides of (hypothetical) right-angled triangles

Usage

```
data(triangle)
```

Format

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

`y` the length of the hypotenuse; a numeric vector

`x1` the length of one side of the triangle; a numeric vector

`x2` the length of the third side of the triangle; a numeric vector

Details

The data give the three sides of hypothetical right-angled triangles. The data are randomly generated so that y is the square root of $x_1^2 + x_2^2$, plus a small amount of error. The idea is from Gelman and Nolan (2002).

Source

The data are artificial; generated using R.

References

The idea is from Andrew Gelman and Deborah Nolan (2002) *Teaching Statistics: A bag of tricks*. Oxford University Press.

Examples

```
data(triangle)
plot(triangle)
```

trout *The effect of potassium cyanate on trout eggs*

Description

The survival of trout eggs exposed to potassium cyanate

Usage

```
data(trout)
```

Format

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

`Conc` the concentration of potassium cyanate in mg/litre; a numeric variable

`When` when the toxicant is applied; a factor with levels `Now` or `Later` (after the eggs have water-hardened)

`Number` the number of eggs used; a numeric variable

`Dead` the number of eggs dead; a numeric variable

Details

The data show the number of trout eggs that are dead at Day 19 after exposure to potassium cyanate (KSCN). Half the eggs in each vial were first allowed to water-harden before the toxicant was applied; the other were exposed immediately.

Source

R. J. O'Hara Hines and E. M. Carter (1993) Improved added variable and partial residual plots for detection of influential observations in generalized linear models. *Applied Statistics*, **42**(1), 3–20.

D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway, and E. Ostrowski (1994) *A Handbook of Small Data Sets*, London: Chapman and Hall. Dataset 418.

Examples

```
data(trout)
summary(trout)
```

`turbines` *Fissures in turbine wheels*

Description

In an experiment, turbine wheels were run for a number of hours, and the number of fissures developed was counted

Usage

```
data(turbines)
```

Format

A data frame with 11 observations on the following 3 variables.

`Hours` the number of hours the turbine was run; a numeric vector

`Turbines` the number of turbines run for the given amount of hours; a numeric vector

`Fissures` the number of turbines wheels with fissures; a numeric vector

Details

The data give the midpoints of running times; that is, the first row (where `Hours=400`) actually corresponds to a running time of 0 to 800 hours. The final class is 4400+ hours, taken as `4600` for convenience.

Source

Raymond H. Myers, Douglas C. Montgomery, and G. Geoffrey Vining (2002) *Generalized linear models with applications in engineering and the sciences*, Wiley.

References

The original source is Wayne Nelson (1982) *Applied Life Data*, Wiley, 407–409.

Examples

```
data(turbines)
summary(turbines)
```

`urinationD` *Urination time*

Description

The urination times of animals

Usage

```
data(urinationD)
```

Format

A data frame containing 35 observations with the following 5 variables.

`Animal` the type of animals; some are repeated

`Sex` the sex of the animal; one of `F` or `M`

`Mass` the mass of the animal (or mean mass of the animals, when multiple animals are represented), in kg

`Duration` the urination time of the animal (or the mean, when multiple animals are represented), in seconds

`SampleSize` the size of the sample represented by the data, usually 1

Details

The data give the duration time for urination for animals of different sex and mass. The data were collected using numerous methods (including YouTube videos); see details in Yang et al. (2014). From the paper: "we discover that all mammals above 3 kg in weight empty their bladders over nearly constant duration of 21 ± 13 s." (p. 11932)

Source

Yang et al. (2014) supplementary information Table S1.

References

Patricia J. Yang, Jonathan Pham, Jerome Choo, and David L. Hu (2014) Duration of urination does not change with body size. *Proceedings of the National Academy of Sciences*, **111**(33), 11932–11937.

Examples

```
data(urinationD)
summary(urinationD)
```

`urinationL` *Urethral length*

Description

The urethral length of animals

Usage

```
data(urinationL)
```

Format

A data frame containing 35 observations with the following 5 variables.

`Animal` the type of animals; some are repeated

`Sex` the sex of the animal; one of `F` or `M`

`Mass` the mass of the animal (or mean mass of the animals, when multiple animals are represented), in kg

`Length` the urethral length of the animal (or the mean, when multiple animals are represented), in mm

`SampleSize` the size of the sample represented by the data, usually 1

Details

The data give the urethral length for animals of different sex and mass. The data were collected using numerous methods; see details in Yang et al. (2014).

Source

Yang et al. (2014) supplementary information Table S2.

References

Patricia J. Yang, Jonathan Pham, Jerome Choo, and David L. Hu (2014) Duration of urination does not change with body size. *Proceedings of the National Academy of Sciences*, **111**(33), 11932–11937.

Examples

```
data(urinationL)
summary(urinationL)
```

wacancer *Cancer in Western Australia*

Description

Diagnoses of cancer in Western Australia for males and females in 1996

Usage

```
data(wacancer)
```

Format

A data frame with 14 observations on the following 3 variables.

`Cancer` the type of cancer; a factor with levels `Prostate`, `Breast`, `Colorectal`, `Lung`, `Melanoma`, `Cervix`, and `Other`

`Gender` the gender; a factor with levels `M` (males) and `F` (females)

`Counts` the number of people in the designated category; a numeric vector

Details

The data gives the number of diagnoses of the designated cancers in Western Australia in 1996.

Source

Health Department of Western Australia Annual Report 1997/1998—health of Western Australians—mortality and survival. Published on the internet `http://www.health.wa.gov.au/Publications/annualreport_9798/`, accessed 19~September 2001.

Examples

```
data(wacancer)
summary(wacancer)
```

wheatrain *Annual rainfall in the NSW wheat belt*

Description

The annual rainfall for stations in the wheat-belt in the north and centre of New South Wales (Australia)

Usage

```
data(wheatrain)
```

Format

A data frame with 24 observations on the following 6 variables.

`Station` the station name; a text vector

`Alt` the station altitude (in metres); a numeric vector

`Lat` the station latitude (in degrees south); a numeric vector

`Lon` the station longitude (in degrees east); a numeric vector

`AR` the stations' mean annual rainfall (in mm) between 1916 and 1990; a numeric vector

`Region` the station's region, as computed by Boer et al. (1993) using a principal component analysis based on monthly rainfall; a numeric vector with levels `1`, `2` and `3`

Details

The data gives the mean annual rainfall for 24 stations in the wheat-belt of NSW. The mean rainfall is based on the year 1916 to 1990, apart from Station 1 (1907 to 1983), Station 10 (1916 to 1965) and Station 11 (1935 to 1976).

Source

Rizaldi Boer, David J. Fletcher, and Lindsay C. Campbell (1993) Rainfall patterns in a major wheat-growing region of Australia. *Australian Journal of Agricultural Research*, **44**(2), 606–624.

Examples

```
data(wheatrain)
plot(AR ~ Region, data=wheatrain)
```

---

windmill *Power generation by windmills*

Description

The amount of direct current (DC) output from windmills for varying wind velocities

Usage

```
data(windmill)
```

Format

A data frame with 25 observations on the following 2 variables.

`Wind` the wind velocity in miles per hours; a numeric vector

`DC` the DC output; a numeric vector

Details

The wind velocity and corresponding direct current (DC) output from windmills was recorded.

Source

G. Joglekar, J. H. Schuenemeyer and V. LaRicca (1989) Lack-of-fit testing when replicates are not available. *American Statistician*, **43**, 135–143.

References

D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway, and E. Ostrowski (1994) *A Handbook of Small Data Sets*, London: Chapman and Hall. Dataset 271.

D. C. Montgomery and E. A. Peck (1982) *Introduction to Linear Regression Analysis*. New York: John Wiley.

Examples

```
data(windmill)
summary(windmill)
```

---

wwomen *Smoking and survival*

Description

The smoking habits and survival of women in Whickham

Usage

```
data(wwomen)
```

Format

A data frame with 14 observations on the following 4 variables.

`Age` the age of the women in completed years in the *original* survey; a factor with levels `18-24`, `25-34`, `35-44`, `45-54`, `55-64`, `65-74` and `75+`

`Smoking` the smoking status of the women in the *original* survey; a factor with levels `NonSmoker` and `Smoker`

`Status` the status of the women twenty years after the original survey; a factor with levels `Dead` or `Alive`

`Count` the number of women in each category; a numeric vector

Details

The data gives the smoking and survival data for 1314 women in Whickham (north England). A survey was originally conducted in 1972–1974; a subsequent survey twenty years later followed up the women to determine how many women from the original survey had died. (Of the original women in the survey, 180 have been excluded here: 18 whose smoking habits were not recorded, and 162 who were smokers before the first survey but were non-smokers at the time of the second survey.)

Source

D. R. Appleton, J. M. French, and M. P. J. Vanderpump (1996) Ignoring a covariate: An example of Simpson's paradox. *The American Statistician*, **50**, 340–341.

References

The data also appear in Anthony C. Davison. *Statistical Models* (2003) Number 11 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, UK.

Examples

```
data(wwomen)
summary(wwomen)
```

---

yieldden *Yield of onions at various densities*

Description

The mean yields per plant for three onion varieties

Usage

```
data(yieldden)
```

Format

A data frame with 30 observations on the following 3 variables.

`Yield` the yield per plant in grams; a numeric vector

`Dens` the planting density in plants per square foot; a numeric vector

`Var` the variety; a numeric vector with levels `1`, `2` or `3`

Source

R. Mead (1970) Plant density and crop yield. *Applied Statistics*, **19**(1), 64–81.

Examples

```
data(yieldden)
summary(yieldden)
```

Index