Package 'MLGdata'

July 21, 2025

Type Package

Title Datasets for Use with Salvan, Sartori and Pace (2020)

Version 0.1.0

Author Nicola Sartori, Alessandra Salvan, Luigi Pace

Maintainer Nicola Sartori <nicola.sartori@unipd.it>

Description Contains the datasets for use with the book Salvan, Sartori and Pace (2020, ISBN:978-88-470-4002-1) ``Modelli Lineari Generalizzati".

License GPL (>= 2)

Encoding UTF-8

LazyData true

RoxygenNote 7.1.1

NeedsCompilation no

Depends R (>= 3.5.0)

Repository CRAN

Date/Publication 2020-09-30 08:50:12 UTC

Contents

`Abrasion` *Abrasion loss*

Description

Data on the weight loss due to abrasion, hardness and tensile strength for 30 samples of rubber.

Usage

```
Abrasion
```

Format

A data frame with 30 observations on the following 3 variables

`perdita` weight loss (in grams per hour)

`D` hardness (in degrees Shore)

`Re` tensile strength (in kg/cm^2)

Source

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

`Aids` *Aids mortality*

Description

Number of AIDS deaths in a sequence of three-months periods between 1983 and 1986.

Usage

```
Aids
```

Format

Data frame with 14 observations on the following 2 variables

`cases` number of deaths

`periodo` number of period

Source

Dobson, A.J. (1990). *An Introduction to Generalized Linear Models*. London: CRC Press.

Alligators *Alligator food choice data*

Description

Alligator food choice data

Usage

```
Alligators
```

Format

A data frame with 40 rows and 4 variables:

`foodchoice` primary food type, in volume, found in an alligator's stomach, with levels `fish`, `invertebrate`, `reptile`, `bird`, `other`

`lake` lake of capture with levels `Hancock`, `Oklawaha`, `Trafford`, `George`

`size` size of the alligator with levels `<=2.3` meters long and `>2.3` meters long

`Freq` number of alligators for each foodchoice, lake, gender and size combination

Source

The alligators data set is analysed in Agresti (2002, Subsection 7.1.2).

This is an edited version of the original data set, which is available at http://www.stat.ufl.edu/~aa/glm/data/

References

Agresti, A. (2002). *Categorical Data Analysis*. New York: Wiley.

Ants *Ants and sandwiches*

Description

The dataset refers to an experiment carried out by some students of an Australian university.

Usage

```
Ants
```

Format

Data frame with 48 observations on the following 5 variables

`Bread` integer indicator for the kind of bread (1=rye, 2=wheatmeal, 3=multigrain, 4=white)

`Filling` integer indicator for the kind of filling (1=vegemite, 2=peanut butter, 3=ham and pickles)

`Butter` indicator for butter (1=butter, -1=no butter)

`Ant_count` number of captured ants

`Order` order of the experiment

Source

Mackisack, M. (2017). What is the use of experiments conducted by Statistics students? *Journal of Statistics Education*, **2**, 12-15.

`Aziende` *Number of closed businesses*

Description

The data refers to the number of business that have closed their activity in the first trimester of 2005 in 16 Italian regions.

Usage

```
Aziende
```

Format

Data frame with 16 observations on the following 4 variables

`regione` integer indicator for the region

`numero` number of closed businesses

`dimensione` average dimension of the businesses

`salario` average individual salary

Source

Salvan, A., Sartori, N., Pace, L. (2020). *Modelli lineari generalizzati*. Milano: Springer-Verlag.

`Bartlett` *Bartlett data on plum root cuttings*

Description

In an experiment to investigate the effect of cutting length (two levels) and planting time (two levels) on the survival of plum root cuttings, 240 cuttings were planted for each of the 2 x 2 combinations of these factors, and their survival was later recorded.

Usage

```
Bartlett
```

Format

A 3-dimensional array resulting from cross-tabulating 3 variables for 960 observations. The variable names and their levels are:

No	Name	Levels
1	`Alive`	`"Alive", "Dead"`
2	`Time`	`"Now", "Spring"`
3	`Length`	`"Long", "Short"`

Source

Hand, D. and Daly, F. and Lunn, A. D.and McConway, K. J. and Ostrowski, E. (1994). *A Handbook of Small Data Sets*. London: Chapman & Hall, p. 15, # 19.

Package `vcdExtra`

References

Bartlett, M. S. (1935). Contingency Table Interactions *Journal of the Royal Statistical Society*, Supplement, 1935, 2, 248-252.

See Also

`Bartlett2` for the same data in data frame format

`Bartlett2` *Bartlett data on plum root cuttings*

Description

In an experiment to investigate the effect of cutting length (two levels) and planting time (two levels) on the survival of plum root cuttings, 240 cuttings were planted for each of the 2 x 2 combinations of these factors, and their survival was later recorded.

Usage

```
Bartlett2
```

Format

A data frame with 4 rows and 4 columns related to the cross-classification of 960 observations. The variables are:

`Alive` number of plum root cuttings survived

`Dead` number of plum root cuttings dead

`Time` factor w/ 2 levels (`Now`, `Spring`)

`Length` factor w/ 2 levels (`Long`, `Short`)

Source

Hand, D. and Daly, F. and Lunn, A. D.and McConway, K. J. and Ostrowski, E. (1994). *A Handbook of Small Data Sets*. London: Chapman & Hall, p. 15, # 19.

References

Bartlett, M. S. (1935). Contingency Table Interactions *Journal of the Royal Statistical Society*, Supplement, 1935, 2, 248-252.

See Also

`Bartlett` for the same data in table format

Beetles *Deaths of flour beetles*

Description

Number of adult flour beetles which died following a 5-hour exposure to gaseous carbon disulphide.

Usage

```
Beetles
```

Format

A data frame with 8 observations on the following 3 variables

`num` numbers of beetles exposed

`uccisi` numbers of beetles dying

`logdose` concentration of carbon disulphide (mg. per litre) in logarithmic scale

Source

Bliss, C. I. (1935).The calculation of the dosage-mortality curve. *Annals of Applied BIology*, **22**, 134-167.

See Also

`Beetles10` for an ungrouped version of this data

Beetles10 *Deaths of flour beetles*

Description

Survival adult flour beetles which died following a 5-hour exposure to gaseous carbon disulphide.

Usage

```
Beetles10
```

Format

A data frame with 481 observations on the following 2 variables

`log.dose10` concentration of carbon disulphide (mg. per litre) in logarithmic scale

`ucciso` indicator variable of death (0: survived, 1: dead)

Source

Bliss, C. I. (1935).The calculation of the dosage-mortality curve. *Annals of Applied BIology*, **22**, 134-167.

See Also

`Beetles` for a grouped version of these data

`Bioassay` *Biological experiment*

Description

Number of events observed in a biological experiment with different dose exposure.

Usage

```
Bioassay
```

Format

A data frame with 10 observations on the following 3 variables

`z` dose level

`den` number of exposed

`y` number of observed events

Source

Finney, D.J. (1947). *Probit Analysis*. Cambridge: Cambridge University Press.

`Biochemists` *article production by graduate students in biochemistry Ph.D. programs*

Description

A sample of 915 biochemistry graduate students.

Usage

```
Biochemists
```

Format

Data frame with 915 observations on the following 6 variables

`art` count of articles produced during last 3 years of Ph.D.

`fem` factor indicating gender of student, with levels Men and Women

`mar` factor indicating marital status of student, with levels Single and Married

`kid5` number of children aged 5 or younger

`phd` prestige of Ph.D. department

`ment` count of articles produced by Ph.D. mentor during last 3 years

Source

Package `pscl`

References

Long, J. Scott. 1990. The origins of sex differences in science. *Social Forces*. 68(3):1297-1316.

Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, California: Sage.

`Britishdoc` *British doctors study*

Description

Study on coronary deaths involving British doctors.

Usage

```
Britishdoc
```

Format

A data frame with 10 observations on the following 4 variables

`age` factor with 5 levels (`35-44, 45-54, 55-64, 65-74, 75-84`)

`smoke` factor with 2 levels (`n, y`)

`person.years` total number of observed person-years

`deaths` number of observed deaths by coronary disease

Source

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. Hoboken: Wiley.

Calcium *Calcium Uptake Data*

Description

Data on the uptake of calcium by cells suspended in a radioactive solution, as a function of time.

Format

A data frame with 27 observations on the following 2 variables

`time` The time (in minutes) that the cells were suspended in the solution

`cal` The amount of calcium uptake (nmoles/mg)

Details

Howard Grimes from the Botany Department, North Carolina State University, conducted an experiment for biochemical analysis of intracellular storage and transport of calcium across plasma membrane. Cells were suspended in a solution of radioactive calcium for a certain length of time and then the amount of radioactive calcium that was absorbed by the cells was measured. The experiment was repeated independently with 9 different times of suspension each replicated 3 times.

Source

Rawlings, J.O. (1988) *Applied Regression Analysis*. Wadsworth and Brooks/Cole Statistics/Probability Series.

Package `SMPracticals`

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 469.

Cement *Tensile strength of cement*

Description

Experiment where different batches of cement were tested for tensile strength after different curing times.

Usage

```
Cement
```

Format

An object of class `data.frame` with 21 rows and 2 columns.

Details

`tempo` curing times (in days)

`resistenza` tensile strength (kg/cm^2)

Source

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

`Chimps` *Chimpanzee Learning Data*

Description

These are the times in minutes taken for four chimpanzees to learn each of four words.

Format

A data frame with 40 observations on the following 3 variables

`chimp` a factor with levels 1-4

`word` a factor with 1-10

`y` learning time (minutes)

Source

Brown, B. W. and Hollander, M. (1977) *Statistics: A Biomedical Introduction*. New York: Wiley.

Package `SMPracticals`

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 485.

Chlorsulfuron *Chlorsulfuron Data*

Description

Bioassay on the action of the herbicide chlorsulfuron on the callus area of colonies of Brassica napus L. The experiment consists of 51 measurements for 10 different dose levels. The design is unbalanced: the number of replicates per dose varies from a minimum of 5 to a maximum of 8.

Usage

```
Chlorsulfuron
```

Format

A data frame with 51 observations on the following 3 variables

`gruppo` indicator variable for each tested dose

`dose` the tested dose (nmol/l)

`area` the callus area (mm^2)

Source

Package nlreg

Seiden, P., Kappel, D. e Streibig, J.C. (1998). Response of Brassica napus L. tissue culture to metsulfuron methyl and chlorsulfuron. *Weed Research*, **38**, 221-228.

Clotting *Blood clotting times*

Description

Mean blood clotting times in seconds for nine percentage concentrations of normal plasma and two lots of clotting agent.

Usage

```
Clotting
```

Format

Data frame with 18 observations on the following 3 variables

`u` plasma concentration (in precentage)

`tempo` clotting time (in seconds)

`lotto` lot (factor with two levels: uno, due)

Source

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models* (2nd Edition). London: Chapman and Hall.

`Credit` *Credit Score Data From a South German Bank*

Description

Data for 1000 clients of a south german bank, 700 good payers and 300 bad payers. They are used to construct a credit scoring method.

Format

Data frame with 1000 observations on the following 8 variables

`Y` a factor with levels `buen` `mal`, the response variable. buen is the good payers.

`Cuenta` a factor with levels `no good running` `bad running`, quality of the credit clients bank account.

`Mes` a numeric vector, duration of loan in months.

`Ppag` a factor with levels `pre buen pagador` `pre mal pagador`, if the client previosly have been a good or bad payer.

`Uso` a factor with levels `privado` `profesional`, the use to which the loan is made.

`DM` a numeric vector, the size of loan in german marks.

`Sexo` a factor with levels `mujer` `hombre`, sex of the client.

`Estc` a factor with levels `no vive solo` `vive solo`, civil state of the client.

Source

Fahrmeier, L. and Tutz, G. (2001) *Multivariate Generalized Linear Models*. New York: Springer Verlag.

Package `Fahrmeir`

`Customer` *Bus customer satisfaction*

Description

Survey on the customer satisfaction among passengers of a certain bus line.

Usage

```
Customer
```

Format

A data frame with 12231 observations on the following 2 variables

`y` level of satisfaction, factor with 5 levels (`Neutral`, `Satisfied`, `Unsatisfied`, `Very satisfied`, `Very unsatisfied`)

`delay` bus delay (in minutes)

Source

Madsen, H. e Thyregod, P. (2010). *Introduction to General and Generalized Linear Models*. Boca Raton, CRC Press.

See Also

`Customer3` for the same data in table format

`Customer3` *Bus customer satisfaction*

Description

Survey on the customer satisfaction among passengers of a certain bus line.

Usage

```
Customer3
```

Format

The data are stored as a frequency table. Data frame with 4 observations on the following 6 variables

`delay` bus delay (in minutes)

`Verydissatisfied` frequency of "Very dissatisfied" replies to the survey

`Dissatisfied` frequency of "Dissatisfied" replies to the survey

`Neutral` frequency of "Neutral" replies to the survey

`Satisfied` frequency of "Satisfied" replies to the survey

`Verysatisfied` frequency of "Very satisfied" replies to the survey

Source

Madsen, H. e Thyregod, P. (2010). *Introduction to General and Generalized Linear Models*. Boca Raton, CRC Press.

See Also

`Customer` for the individual level data

`Dogs` *Dogs data*

Description

Measurements of left ventricular volume and parallel conductance volume on five dogs under eight different load conditions

Usage

```
Dogs
```

Format

Data frame with 40 observations on the following 4 variables

`dog` dog number

`condition` load condition

`y` left ventricular volume

`x` parallel conductance volume

Source

Package `dobson`

Dobson, A. J. and Barnett A. (2008). *An Introduction to Generalized Linear Models*, 3rd ed. Boca Raton: CRC Press.

References

Boltwood, C. M., R. Appleyard, and S. A. Glantz (1989). Left ventricular volume measurement by conductance catheter in intact dogs: the parallel conductance volume increases with end-systolic volume. *Circulation* 80, 1360–1377.

`Drugs` *Student Substance Use*

Description

Survey on alcohol, cigarettes, or marijuana use collected on 2276 students in their final year of high school in a rural area near Dayton, Ohio.

Usage

```
Drugs
```

Format

A data frame with 8 observations on the following 4 variables

`alc` alcohol use, factor with 2 levels (`no`, `yes`)

`sig` sigarettes use, factor with 2 levels (`no`, `yes`)

`mar` marijuana use, factor with 2 levels (`no`, `yes`)

`count` frequency of students in the cross classification of the previous three variables

Source

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. Hoboken: Wiley.

See Also

`Drugs2` for a different format of the same data and `Drugs3` for an extended version of the data with additional variables.

Drugs2 *Student Substance Use*

Description

Survey on alcohol, cigarettes, or marijuana use made on 2276 students in their final year of high school in a rural area near Dayton, Ohio.

Usage

```
Drugs2
```

Format

A data frame with 4 observations on the following 5 variables

`alc` alcohol use, factor with 2 levels (`no`, `yes`)

`sig` sigarettes use, factor with 2 levels (`no`, `yes`)

`M_yes` frequency of students that have tried marijuana

`M_no` frequency of students that have never tried marijuana

`n` frequency of students in the cross classification of variables `alc` and `sig`

Source

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. Hoboken: Wiley.

See Also

`Drugs` for a different format of the same data and `Drugs3` for an extended version of the data with additional variables.

Drugs3 *Student Substance Use*

Description

Survey on alcohol, cigarettes, or marijuana use made on 2276 students in their final year of high school in a rural area near Dayton, Ohio.

Usage

```
Drugs3
```

Format

A data frame with 32 observations on the following 6 variables

`alcohol` alcohol use, factor with 2 levels (`no`, `yes`)

`cigarette` cigarettes use, factor with 2 levels (`no`, `yes`)

`marijuana` marijuana use, factor with 2 levels (`no`, `yes`)

`gender` factor with 2 levels (`Female`, `Male`)

`race` factor with 2 levels (`Other`, `White`)

`Freq` frequency of students in the cross classification of the previous five variables

Source

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. Hoboken: Wiley.

See Also

`Drugs` and `Drugs2` for a reduced version of this data, with fewer variables, in two different formats.

`Esito` *Recreational activities and university performance*

Description

Survey on the effect of recreational activities on university performance collected on 485 students.

Usage

```
Esito
```

Format

A data frame with 18 observations on the following 4 variables

`freq` frequency of students in the in the cross classification of the following three variables

`sex` factor with 2 levels (`f`, `m`)

`ore` weekly hours of recreational activities, factor with 3 levels (`m10`, less than 10 hours; `m15`, between 10 and 15 hours; `m20`, more than 15 hours)

`voto` university performance in a given exam, factor with 3 levels (`ins`, not sufficient; `suff`, sufficient; buono, good)

Source

Salvan, A., Sartori, N., Pace, L. (2020). *Modelli lineari generalizzati*. Milano: Springer-Verlag.

Germination *Seed Germination*

Description

Factorial experiment on the germination of two different kind of seeds (Orobanche aegyptiaca 75 and Orobanche aegyptiaca 73) in two different experimental conditions (bean or cucumber root).

Usage

```
Germination
```

Format

Data frame with 21 observations in the following 4 variables

`s` number of germinated seeds

`m` total number of seeds

`seed` seed indicator, factor with 2 levels (`073`, `075`)

`root` root indicator, factor with 2 levels (`C`, `F`)

Source

Cox, D.R. e Snell, E.J. (1989). *Analysis of Binary Data*, 2nd ed. London: Chapman & Hall/CRC.

Heart *Creatinine kinase and heart attacks*

Description

Data on diagnosed heart attacks in a sample of 360 patients hospitalized with suspected heart attack.

Usage

```
Heart
```

Format

Data frame with 13 observations and the following 4 variables

`mck` central value of the class of Creatinine kinase level in variable `ck`

`ck` class of Creatinine kinase level (in IU per litre), factor with 13 levels (`Below 40`, `40-80`, ..., `480 and over`)

`ha` number of patients with diagnosed heart attack

`nha` number of patients without heart attack

Source

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

`Homicide` *Homicide data*

Description

Survey on number of victims of murder known in the past year by race.

Usage

```
Homicide
```

Format

A data frame with 1308 observations on the following 2 variables

`race` indicator of self-identified race (`0`, white; `1`, black)

`count` number of known victims of murder in the last year

Source

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. Hoboken: Wiley.

http://www.stat.ufl.edu/~aa/glm/data

`Infant` *Infant survival*

Description

Study that relates the survival of infants to length of gestation, age and smoking habit of mothers.

Usage

```
Infant
```

Format

A data frame with 16 observations on the following 5 variables

`survival` survival of the infant, factor with 2 levels (`No, Yes`)

`gestation` length of gestation (in days), factor with 2 levels (`<=260, >260`)

`smoking` number of cigarettes per day smoked by the mother, factor with 2 levels (`<5, >5`)

`age` age of the mother (in years), factor with 2 levels (`<30, >30`)

`Freq` frequency of infant in the cross classification of the previous 4 variables

Source

Agresti, A. (2013). *Categorical Data Analysis*, 3rd ed. New York: Wiley.

`Kyphosis` *Data on Children who have had Corrective Spinal Surgery*

Description

Data on children who have had corrective spinal surgery.

Usage

```
Kyphosis
```

Format

Data frame with 81 observations on the following 4 variables

`Kyphosis` a factor with levels `absent` `present` indicating if a kyphosis (a type of deformation) was present after the operation.

`Age` in months

`Number` the number of vertebrae involved

`Start` the number of the first (topmost) vertebra operated on.

Source

Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models*. London: Chapman & Hall/CRC.

`Malaria` *Malaria Transmission in the Western Kenyan Highlands*

Description

The dataset contains information on 8204 individuals enrolled in concurrent school and community cross-sectional surveys, conducted in 46 school clusters in the western Kenyan highlands. Malaria was assessed by rapid diagnostic test (RDT).

Usage

```
Malaria
```

Format

The data frame has 8204 observations on the following variables

`Cluster` unique ID for each of the 46 school clusters

`Long` longitude coordinate of the household location

`Lat` latitude coordinate of the household location

`RDT` binary variable indicating the outcome of the RDT (`1`, positive; `0`, negative)

`Gender` factor variable indicating the gender of the sampled individual (`Female`, `Male`)

`Age` age of the sampled individual (in years)

`NetUse` binary variable indicating whether the sampled individual slept under a bed net the previous night (`1`, yes; `0`, no)

`MosqCntl` binary variable indicating whether the household has used some kind of mosquito control, such as sprays and coils (`1`, yes; `0`, no)

`IRS` binary variables in indicating whether there has been indoor residual spraying (IRS) in the house in the last 12 months (`1`, yes; `0`, no)

`Travel` binary variable indicating whether the sampled individual has travelled outside the village in the last three months (`1`, yes; `0`, no)

`SES` ordinal variable indicating the socio-economic status (SES) of the household. The variable is an integer score from `1` (poor) to `5` (rich)

`District` factor variable indicating the village of the sampled individual (`Kisii Central`, `Rachuonyo`)

`Survey` factor variables indicating the survey in which the participant was enrolled (`community`, `school`)

Source

https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxtYmdnbG9iYWxoZWFsdGh8Z3g6NjZh

References

Stevenson, J.C., Stresman, G.H., Gitonga, C.W., Gillig, J., Owaga, C., Marube, E., Odongo, W., Okoth, A., China, P., Oriango, R. e Brooker, S.J. (2013). Reliability of school surveys in estimating geographic variation in malaria transmission in the western Kenyan highlands. *PLoS One*, **8**, e77641.

`Mental` *Mental impairment*

Description

Study of mental health for a random sample of adult residents of Alachua County, Florida.

Usage

```
Mental
```

Format

Data frame with 40 observations in the following 3 variables

`menom` mental health status on an ordinal scale (`1`, well; `2`, mild symptom formation; `3`, moderate symptom formation; `4`, impaired)

`sse` Socioeconomic status (`1`, high; `0`, low)

`eventi` life events index, a composite measure of the number and severity of important life events that occurred to the subject within the past 3 years, such as the birth of a child, a new job, a divorce, or a death in the family

Source

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. Hoboken: Wiley.

`Neonati` *Weight at birth*

Description

Data on the weight at birth, the duration of the gestation, and the smoke habit of the mother for 32 newborns.

Usage

```
Neonati
```

Format

Data frame with 32 observations on the following 3 variables

`peso` weigth at birth (in grams)

`durata` duration of gestation (in weeks)

`fumo` a factor with levels `F` (smoker), `NF` (non smoker)

Source

Daniel, W.W. (1999). *Biostatistics: A Foundation for Analysis in the Health Sciences*. New York: Wiley.

Ohio *Ohio Children Wheeze Status*

Description

The dataset is a subset of the six-city study, a longitudinal study of the health effects of air pollution.

Usage

```
Ohio
```

Format

Data frame with 2148 observations on the following 4 variables

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

`id` a numeric vector for subject id

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

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

Source

Package geepack

References

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

Halekoh, U., Højsgaard, S. e Yan, J. (2005). The R package geepack for generalized estimating equations. *Journal of Statistical Software*, **15**, 1-11.

Orthodont *Growth curve data on an orthdontic measurement*

Description

Study of the change in an orthdontic measurement over time for 27 young subjects.

Usage

```
Orthodont
```

Format

Data frame with 27 observations in the following 5 variables

`genere` gender of the subject, factor with 2 levels (`F`, `M`)

`dist8a` measurement of the orthodontic distance (in mm) at age 8

`dist10a` measurement of the orthodontic distance (in mm) at age 10

`dist12a` measurement of the orthodontic distance (in mm) at age 12

`dist14a` measurement of the orthodontic distance (in mm) at age 14

Source

Pinheiro, J.C. and Bates, D.M. (2000). *Mixed Effects Models in S and S-PLUS*. New York: Springer.

Package `nlme`

References

Potthoff, R.F. and Roy, S.N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, **51**, 313-326.

See Also

`Orthodont1` for the same data in an different format

Orthodont1 *Growth curve data on an orthdontic measurement*

Description

Study of the change in an orthdontic measurement over time for 27 young subjects.

Usage

```
Orthodont1
```

Format

Data frame with 108 observations in the following 4 variables

`caso` subject index

`genere` gender of the subject, factor with 2 levels (`F`, `M`)

`eta` age of the subject

`y` measurement of the orthodontic distance (in mm)

Source

Pinheiro, J.C. and Bates, D.M. (2000). *Mixed Effects Models in S and S-PLUS*. New York: Springer.

Package `nlme`

References

Potthoff, R.F. and Roy, S.N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, **51**, 313-326.

See Also

`Orthodont` for the same data in a different version

`Pneu` *Pneumoconiosis amongst Coalminers*

Description

This gives the degree of pneumoconiosis (normal, present, or severe) in a group of coalminers as a function of the number of years worked at the coalface. The degree of the disease was assessed radiologically and is qualitative.

Usage

```
Pneu
```

Format

A data frame with 8 observations on the following 4 variables

`Years` Period of exposure (years worked at the coalface)

`Normal` Number of miners with normal lungs

`Present` Number of miners with disease present

`Severe` Number of miners with severe disease

Source

Ashford, J. R. (1959) An approach to the analysis of data for semi-quantal responses in biological assay. *Biometrics*, **15**, 573–581.

Package `SMPracticals`

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 509.

Rats *Teratology study*

Description

Teratology experiment investigating effects of dietary regimens or chemical agents on the fetal development of rats in a laboratory setting. The experiment, as describred in Agresti (2015, Section 8.2.4), regards female rats on iron-deficient diets, assigned to four groups. Rats in group 1 were given placebo injections, and rats in other groups were given injections of an iron supplement. This was done on days 7 and 10 in group 2, on days 0 and 7 in group 3, and weekly in group 4. The 58 rats were made pregnant, sacrificed after 3 weeks, and then the total number of dead fetuses was counted in each litter, as was the mother's hemoglobin level.

Usage

```
Rats
```

Format

A data frame with 58 observations on the following 5 variables

`litter` litter index

`group` group index (1, . . . , 4)

`h` hemoglobin level of the mother

`n` number of fetuses in the litter

`s` number of dead fetuses in the litter

Source

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. Hoboken: Wiley.

Package `catdata`

References

Moore, D.F. and Tsiatis, A. (1991). Robust estimation of the variance in moment methods for extra-binomial and extra-Poisson variation. *Biometrics*, **47**, 383-401.

Seed *Seed germination*

Description

This is an artificial dataset representing an experiment relating probability of germination of seeds to the level of fertilizer used.

Usage

```
Seed
```

Format

A data frame with 20 observations on the following 2 variables

`fert` level of fertilizer used

`x` indicator of germination of the seed(`1`, yes; `0`, no)

Source

Salvan, A., Sartori, N., Pace, L. (2020). *Modelli lineari generalizzati*. Milano: Springer-Verlag.

Snore *Snoring and heart disease*

Description

Data from a report of a survey which investigated whether snoring was related to heart disease. Those surveyed were classified according to the amount they snored, on the basis of reports from their spouses.

Usage

```
Snore
```

Format

Data frame with 8 observations on the following 3 variables

`pat` presence of heart disease, factor with 2 levels (`no, si`)

`russ` level of snoring, factor with 4 levels (`mai`, no snoring; `a volte`, occasional snoring; `spesso`, snoring nearly every night; `sempre`, alwayssnoring;)

`freq` frequency observed in the cross classification of the previous 2 variables

Source

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

`Spending` *Opinions about government spending*

Description

Subjects in a 1989 General Social Survey from the National Opinion Research Center in the U.S. were asked their opinions about government spending on the environment (e), health (h), assistance to big cities (c), and law enforcement (l).

Usage

```
Spending
```

Format

A data frame with 81 observations on the following 5 variables

`e` opinion on spending on the environment (`1`, too little; `2`, about right; `3`, too much)

`h` opinion on spending on the health (`1`, too little; `2`, about right; `3`, too much)

`c` opinion on spending on assistance to big cities (`1`, too little; `2`, about right; `3`, too much)

`l` opinion on spending on law enforcement (`1`, too little; `2`, about right; `3`, too much)

`count` frequency of subjects in the cross classification of the previous 4 variables

Source

Agresti, A. (2013). *Categorical Data Analysis*, 3rd ed. New York: Wiley.

`http://users.stat.ufl.edu/~aa/cda/data.html`

`Stroke` *Stroke data*

Description

Longitudinal data from an experiment to promote the recovery of stroke patients in wide format. The response variable is the Bartel index with higher scores meaning better outcomes and a maximum score of 100.

Usage

```
Stroke
```

Format

A `tibble` with 24 observations and the following 10 variables

`Subject` subject number

`Group` group; A=new occupational therapy intervention, B = existing stroke rehabilitation program in the same hospital as A, C = usual care in a different hospital

`week1` Bartel index in week 1

`week2` Bartel index in week 2

`week3` Bartel index in week 3

`week4` Bartel index in week 4

`week5` Bartel index in week 5

`week6` Bartel index in week 6

`week7` Bartel index in week 7

`week8` Bartel index in week 8

Source

Dobson, A. J. and Barnett A. (2008). *An Introduction to Generalized Linear Models*, 3-rd ed. Boca Raton: CRC Press.

Package `dobson`

See Also

`Stroke1` for the same data in an extended format.

Stroke1 *Stroke data*

Description

Longitudinal data from an experiment to promote the recovery of stroke patients in wide format. The response variable is the Bartel index with higher scores meaning better outcomes and a maximum score of 100.

Usage

```
Stroke1
```

Format

A data frame with 192 observations on the following 4 variables

`Subject` subject indicator

`Group` group indicator, factor with 3 levels (`A`, `B`, `C`)

`Week` week indicator

`y` Bartel index

Source

Dobson, A. J. and Barnett A. (2008). *An Introduction to Generalized Linear Models*, 3-rd ed. Boca Raton: CRC Press.

See Also

`Stroke` for the same data in a different format

`Testingresso` *University admission test*

Description

Admission test for Statistical Sciences bachelor course at University of Padova in 2014/15. The data refers to the answers of 63 candidates to 10 questions on text comprehension.

Usage

```
Testingresso
```

Format

A data frame with 630 observations on the following 3 variables

`y` indicator variable of correct answer (`1`, correct; `0`, wrong)

`subject` candidate indicator (`1`, ..., `63`)

`item` question indicator (`1`, ..., `10`)

Source

Salvan, A., Sartori, N., Pace, L. (2020). *Modelli lineari generalizzati*. Milano: Springer-Verlag.

`Vehicle` *Preferred vehicle*

Description

Data from an insurance company, which record for each contract the kind of vehicle, together with some additional variables.

Usage

```
Vehicle
```

Format

A data frame with 2067 observations on the following 4 variables

`age` age of the owner

`men` gender (`1`, man; `0`, female)

`urban` residential area (`1`, urban; `0`, rural)

`veh` kind of vehicle, factor with 3 levels (`C`, car; `F`, fourwheel; `M`, motorcycle)

Source

http://www.ub.edu/rfa/R/regression_with_categorical_dependent_variables.html

Guillén, M. (2014). Regression with categorical dependent variables. In *Predictive Modeling Applications in Actuarial Science - Volume I: Predictive Modeling Techniques*, E.W. Frees, R.A. Derrig and G. Meyers (Eds.) pp. 65-86. Cambridge: Cambridge University Press.

`Wool` *Wool data*

Description

The data show the number of cycles to failure of samples of worsted yarn under cycles of repeated loading. There are three experimental conditions arranged in a 3 x 3 x 3 factorial design.

Usage

```
Wool
```

Format

Data frame with 27 observations on the following 4 variables

`x1` length of test specimen (`-1`, 250 mm; `0`, 300 mm; `1`, 350 mm)

`x2` amplitude of loading cycle (`-1`, 8 mm; `0`, 9 mm; `1`, 10 mm)

`x3` load (`-1`, 40 g; `0`, 45 g; `1`, 50 g)

`y` cycles to failure

Source

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

Index