Package 'SMPracticals'

February 5, 2024

Version 1.4-3.1

Date 2024-01-13

Title Practicals for Use with Davison (2003) Statistical Models

Author Anthony Davison <Anthony.Davison@epfl.ch>

Maintainer Alessandra R. Brazzale <brazzale@stat.unipd.it>

Depends R (>= 1.14.0), ellipse

Imports MASS, nlme, survival

Description Contains the datasets and a few functions for use with
the practicals outlined in Appendix A of the book
Statistical Models (Davison, 2003, Cambridge University Press),
which can be found at <doi:10.1017/CBO9780511815850>.

License GPL (>= 2)

URL doi:10.1017/CBO9780511815850

LazyLoad yes

LazyData yes

NeedsCompilation no

Repository CRAN

Date/Publication 2024-02-05 14:30:06 UTC

R topics documented:

add.exp.lines   *Add Exponential Lines in Practical 11.3*

Description

Adds lines to density plot used in Practical 11.3

Usage

```
add.exp.lines(exp.out, i, B = 10)
```

Arguments

`exp.out`	Gibbs sampler output
`i`	Variable index (=1, 2)
`B`	Upper bound for truncated exponential density

Author(s)

Anthony Davison

Examples

```
B <-10; I <- 15; S <- 500

exp.out <- exp.gibbs(B=B,I=I,S=S)

hist(exp.out[1,,I],prob=TRUE,nclass=15,xlab="u1",ylab="PDF",xlim=c(0,B),ylim=c(0,1))

add.exp.lines(exp.out,1)
```

---

alofi   *Daily Rainfall at Alofi*

Description

Three-state data derived from daily rainfall over three years at Alofi in the Niue Island group in the Pacific Ocean. The states are 1 (no rain), 2 (up to 5mm rain), 3 (over 5mm).

Usage

```
data(alofi)
```

Source

Avery, P. J. and Henderson, D. A. (1999) Fitting Markov chain models to discrete state series such as DNA sequences. *Applied Statistics*, **48**, 53–61.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 294.

Examples

```
data(alofi)
fit <- MClik(alofi)
fit$df
fit$AIC
plot(fit$df,fit$AIC)  # best model has minimal AIC?
```

---

aml   *Remission Times for Acute Myelogenous Leukaemia*

Description

A clinical trial to evaluate the efficacy of maintenance chemotherapy for acute myelogenous leukaemia was conducted by Embury et al. (1977) at Stanford University. After reaching a stage of remission through treatment by chemotherapy, patients were randomized into two groups. The first group received maintenance chemotherapy and the second group did not. The aim of the study was to see if maintenance chemotherapy increased the length of the remission. The data here formed a preliminary analysis which was conducted in October 1974.

Usage

```
data(aml)
```

Format

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

time The length of the complete remission (in weeks).

cens An indicator of right censoring. 1 indicates that the patient had a relapse and so 'time' is the length of the remission. 0 indicates that the patient had left the study or was still in remission in October 1974, that is the length of remission is right-censored.

group The group into which the patient was randomized. Group 1 received maintenance chemotherapy, group 2 did not.

Source

Miller, R.G. (1981) *Survival Analysis*. John Wiley: New York. Page 49.

References

Embury, S.H, Elias, L., Heller, P.H., Hood, C.E., Greenberg, P.L. and Schrier, S.L. (1977) Remission maintenance therapy in acute myelogenous leukaemia. *Western Journal of Medicine*, **126**, 267–272.

---

arithmetic   *Teaching Arithmetic Data*

Description

45 school pupils were divided at random into 5 groups of size 9. Groups A and B were taught
arithmetic in separate classes by the usual method. Groups C, D, and E were taught together for
several days. On each day group C were publically praised, group D were publically reproved, and
group E were ignored. The responses are from a standard test taken by all pupils at the end of the
period.

Usage

```
data(arithmetic)
```

Format

A data frame with 45 observations on the following 2 variables.

group a factor with levels A B C D E

y a numeric vector

Source

Unpublished lecture notes, Imperial College, London.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 427.

Examples

```
data(arithmetic)
attach(arithmetic)
plot(y~group)
anova(lm(y~group,data=arithmetic))
summary(lm(y~group,data=arithmetic))  # two different parametrisations
summary(lm(y~group-1,data=arithmetic)) # for ANOVA
```

---

bard	*Shakespeare's Word Type Frequencies*

Description

These are the frequencies with which Shakespeare used word types. There are 846 word types which appear more than 100 times in his total works, giving an overall total of 31534 word types.

Usage

```
data(bard)
```

Format

A data frame with 100 observations on the following 2 variables.

r Number of times a word type is used

n Number of word types used r times

Details

The canon of Shakespeare's accepted works contains 884,647 words, with 31,534 distinct word types. A word type is a distinguishable arrangement of letters, so 'king' is different from 'kings' and 'alehouse' different from both 'ale' and 'house'.

Source

Efron, B. and Thisted, R. (1976) Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, **63**, 435–448.

Thisted, R. and Efron, B. (1987) Did Shakespeare write a newly-discovered poem? *Biometrika*, **74**, 445–455.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 629.

---

barley   *Spring Barley Data*

Description

The spatial layout and plot yield at harvest in a final assessment trial of 75 varietes of spring barley. The varieties are sown in three blocks, each with 75 plots, and each variety is replicated thrice. The yield for variety 27 is missing in block 3.

Usage

```
data(barley)
```

Format

A data frame with 225 observations on the following 4 variables.

Block a factor with three levels

Location a numeric vector with 75 values giving the plot

Variety a factor with 75 levels giving the variety of barley sown in the plot

y yield at harvest, standardised to have unit crude variance

Source

Besag, J. E., Green, P. J., Higdon, D. and Mengersen, K. (1995) Bayesian computation and stochastic systems (with Discussion). *Statistical Science*, **10**, 3–66.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Pages 534–535.

Examples

```
data(barley)
```

---

beaver   *Body Temperatures for a Female Beaver*

Description

Data comprise 100 consecutive elemetric measurements of the body temperature of a female beaver, at 10-minute intervals. The animal remained in its lodge for the first 38 recordings, and then went outside.

Usage

```
data(beaver)
```

Format

A data frame with 100 observations on the following 4 variables.

day Day number

time Time of day (hhmm)

temp Body temperature (degrees Celsius)

activ Indicator of activity outside the lodge

Source

Reynolds, P. S. (1994) Time-series analyses of beaver body temperatures. In *Case Studies in Biometry*, eds N. Lange, L. Ryan, L. Billard, D. R. Brillinger, L. Conquest and J. Greenhouse, pp. 211–228. New York: Wiley.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 266.

Examples

```
data(beaver)
plot(beaver$temp,type="l",xlab="Time",ylab="Temperature")
```

---

beaver.gibbs *Gibbs Sampler for Normal Changepoint Model, Practical 11.7*

Description

This function implements a Gibbs sampler for the normal changepoint model applied to the beaver temperature data used in Example 6.22 and Practical 11.7 of Davison (2003), which should be consulted for details.

Usage

```
beaver.gibbs(init, y, R = 10, a = 1, b = 0.05)
```

Arguments

init	Initial values for parameters
y	A series of normal observations
R	Number of iterations of sampler
a	Value of a hyperparameter
b	Value of a hyperparameter

Details

This is provided simply so that readers spend less time typing. It is not intended to be robust and general code.

Value

A matrix of size R x 6, whose first four columns contain the values of the parameters for the iterations. Columns 5 and 6 contain the log likelihood and log prior for that iteration.

Author(s)

Anthony Davison (anthony.davison@epfl.ch)

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Practical 11.7.

Examples

```
## From Example 11.7:
data(beaver)
system.time( gibbs.out <- beaver.gibbs(c(36, 40, 3, 38), beaver$temp, R=1000))
par(mfrow=c(2,3))
plot.ts(gibbs.out[,1],main="mu1") # time series plot for mu1
plot.ts(gibbs.out[,2],main="mu2") # time series plot for mu2
plot.ts(gibbs.out[,3],main="lambda") # time series plot for lambda
plot.ts(gibbs.out[,4],main="gamma") # time series plot for gamma
plot.ts(gibbs.out[,5],main="log likelihood")  # and of log likelihood
```

---

beetle   *Japanese beetle data*

Description

Numbers of Japanese beetle larvae per square found in the top foot of soil of an 18 x 8 foot area of a field planted with maize. The columns of the matrix correspond to the direction of cultivation of the field; the maize rows were sown 4 feet apart.

Usage

```
data(beetle)
```

Format

The format is: num [1:18, 1:8] 0 2 3 1 5 3 5 3 2 3 ...

Source

Unpublished lecture notes, Imperial College, London

Examples

```
data(beetle)
```

---

bike	*Bicycling Times*

Description

The times taken to cycle up a hill, as function of the bicycle seat height, use of dynamo, and tyre pressure. 16 runs were made using a factorial design.

Usage

```
data(bike)
```

Format

A data frame with 16 observations on the following 11 variables.

day Day of run

run Order of run

seat Seat height: -1 indicates 26 inches, 1 indicates 30 inches

dynamo Use of dynamo: -1 indicates not used

tyre Tyre pressure: -1 indicates 40 psi, 1 indicates 55 psi

dayf factor corresponding to day

runf factor corresponding to run

seatf factor corresponding to seat height

dynamof factor corresponding to use of dynamo

tyref factor corresponding to tyre pressure

time Run time (seconds)

Source

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

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 357.

Examples

```
data(bike)
anova(lm(time~dayf+runf+seat+dynamo+tyre,data=bike))
```

---

births   *Birth Times*

Description

Times spent in delivery suite by 95 women giving birth at the John Radcliffe Hospital, Oxford. The data were kindly provided by Ethel Burns.

Usage

```
data(births)
```

Format

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

day Day on which woman arrived

time Time (hours) spent on delivery suite

Source

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 18.

Examples

```
data(births)
```

---

blalock   *Blalock–Taussig Shunt Data*

Description

The Blalock–Taussig shunt is an operative procedure for infants with congenital cyanotic heart disease. The data are the survival times in months for shunts in 81 infants, divided into two age groups.

Usage

```
data(blalock)
```

Format

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

group 1 indicates infants aged over 1 month at time of the operation. 2 indicates those aged 30 or fewer days at time of operation.

months survival time in months

cens censoring indicator: 1 indicates observed failure time

Source

Oakes, D. (1991) Life-table analysis. In *Statistical Theory and Modelling: In Honour of Sir David Cox, FRS*, eds D. V. Hinkley, N. Reid and E. J. Snell, pp. 107–128. London: Chapman and Hall/CRC Press.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 192.

Examples

```
data(blalock)
library(survival)
plot(survfit(Surv(months,cens)~group,data=blalock),conf.int=TRUE,col=c(2,3))
```

---

bliss	*Bliss data on deaths of flour beetles*

Description

These are the number of adult flour beetles which died following a 5-hour exposure to gaseous carbon disulphide.

Usage

```
data(bliss)
```

Format

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

dose concentration of carbon disulphide(mg. per litre)

m Numbers of beetles exposed

r Numbers of beetles dying

Source

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

Examples

```
data(bliss)
attach(bliss)
plot(log(dose),r/m,ylim=c(0,1),ylab="Proportion dead")
fit <- glm(cbind(r,m-r)~log(dose),binomial)
summary(fit)
```

---

blood   *Blood Group Data*

Description

Data on the incidence of blood groups O, A, B, and AB in 12 studies on people living in Britain or of British origin living elsewhere.

Usage

```
data(blood)
```

Format

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

O Number of persons with blood group O

A Number of persons with blood group A

B Number of persons with blood group B

AB Number of persons with blood group AB

Source

Taylor, G. L. and Prior, A. M. (1938) Blood groups in England. *Annals of Eugenics*, **8**, 343–355.

---

breast   *Breast Cancer Data*

Description

Initial and follow-up status for 37 breast cancer patients treated for spinal metastases. The status is able to walk unaided (1), unable to walk unaided (2), dead (3). The follow-up times are 0, 3, 6, 12, 24, and 60 months after treatment began.

Usage

```
data(breast)
```

Format

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

j Case number

init Initial status

x0 Status immediately after treatment started

x1 Status after 3 months

x2 Status after 6 months

x3 Status after 12 months

x4 Status after 24 months

x5 Status after 60 months

Details

Woman 24 was alive after 6 months but her ability to walk was not recorded (she was in state 1 or 2).

NA indicates that a woman has previously died, or that her status is unknown.

Source

de Stavola, B. L. (1988) Testing departures from time homogeneity in multistate Markov processes. *Applied Statistics*, **37**, 242–250.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 227.

Examples

```
data(breast)
```

---

burt	*IQs of identical twins*

Description

These are said to be measurements of IQ scores for pairs of identical twins, the first raised by foster parents and the second raised by natural parents, published by Sir Cyril Burt. Cases are divided into groups according to parents' social class, A-C, labelled 1-3. The general objective is to assess the impact of social class, and in particular the effect of environment, on IQ.

Usage

```
data(burt)
```

Format

A data frame with 27 observations on the following 3 variables.

y IQ score for twin raised in foster home

x IQ score for twin raised by natural parents

class Social class of twins

Details

Burt used these and similar data to argue that IQ was largely inherited, a view which strongly influenced British education through the creation of the 11+ exam, which was used to decide which children should be given different forms of education. However after his death it was suggested that the data were fake, a view accepted by some and strongly rebutted by others.

Source

Unpublished lecture notes of David Hinkley.

References

For information about Burt, see www.indiana.edu/~intell/burt.shtml

Examples

```
data(burt)
attach(burt)
par(pty="s")
plot(x,y,type="n",xlim=c(60,140),ylim=c(60,140))
text(x,y,class,cex=0.8)
abline(0,1,lty=2)
```

---

cake	*Breaking of Chocolate Cakes*

Description

Data on breaking angles of chocolate cakes made using different recipes, mixes, and cooking temperatures.

Usage

```
data(cake)
```

Format

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

recipe Recipe used

mix mix, a factor with 15 levels

temp temperature (degrees Fahrenheit) at which cake baked

y breaking angle (degrees)

Details

These are data from an experiment in which six different temperatures for cooking three recipes for chocolate cake were compared. Each time a mix was made using one of the recipes, enough batter was prepared for six cakes, which were then randomly allocated to be cooked at the different temperatures. The response is the breaking angle, found by fixing one half of a slab of cake, then pivoting the other half about the middle until breakage occurs.

Source

Cochran, W. G. and Cox, G. M. (1959) *Experimental Designs*. Second edition. New York: Wiley.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 454.

Examples

```
data(cake)
```

---

calcium	*Calcium Uptake Data*

Description

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

Usage

```
data(calcium)
```

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.

References

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

Examples

```
data(calcium)
summary(nls(cal~beta0*(1-exp(-time/beta1)),data=calcium,start=list(beta0=5,beta1=5)))
```

---

cardiac   *Mortality Rates for Cardiac Surgery on Babies at 12 Hospitals*

Description

The title should be self-explanatory.

Usage

```
data(cardiac)
```

Format

A data frame with 12 observations on the following 2 variables.

r Number of deaths

m Number of operations

Source

Spiegelhalter, D. J., Thomas, A., Best, N. G. and W. R. Gilks (1996) *BUGS 0.5 Examples Volume 1 (Version ii)*. Cambridge: MRC Biostatistics Unit. Page 15.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 579.

---

cat.heart   *Cat Heart Data*

Description

Data from a Latin square experiment on the potencies of cardiac drugs given to anesthetized cats.

Usage

```
data(cat.heart)
```

Format

A data frame with 64 observations on the following 6 variables.

Day on which experiment performed

Time morning or afternoon

Observer four observers took part

Drug cardiac drug given to cat

y 100 times log dose in micrograms at which cat died

x 100 times log cat heart weight in grams

Details

These are results from an experiment to determine the relative potencies of eight similar cardiac drugs, labelled A–H, where A is a standard. The method used was to infuse slowly a suitable dilution of the drug into an anaesthetized cat. The dose at which death occurred and the weight of the cat's heart were recorded. Four observers each made two determinations on each of eight days, with a Latin square design used to eliminate observer and time differences. The heart weight cannot be known at the start of the experiment, but might be expected to affect comparisons among the treatments; it is assumed that heart weight is unaffected by the treatments.

Source

Unpublished lecture notes, Imperial College, London.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 447.

Examples

```
data(cat.heart)
anova(lm(y~Observer+Time+Day+Drug+Observer:Time,data=cat.heart))
```

---

cement   *Hald Cement Data*

Description

Heat evolved in setting of cement, as a function of its chemical composition.

Usage

```
data(cement)
```

Format

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

x1 percentage weight in clinkers of 3CaO.Al2O3

x2 percentage weight in clinkers of 3CaO.SiO2

x3 percentage weight in clinkers of 4CaO.Al2O3.Fe2O3

x4 percentage weight in clinkers of 2CaO.SiO2

y heat evolved (calories/gram)

Source

Woods, H., Steinour, H. H. and Starke, H. R. (1932) Effect of composition of Portland cement on heat evolved during hardening. *Industrial Engineering and Chemistry*, **24**, 1207–1214.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 355.

Examples

```
data(cement)
lm(y~x1+x2+x3+x4,data=cement)
```

---

chicks   *Chick Bone Data*

Description

Balanced incomplete block design on the effect of amino acids on growth of chick bones.

Usage

```
data(chicks)
```

Format

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

Treat Treatment with levels `All` (all amino acids present), `Arg-`(all acids present except Arg), etc.

Pair bones were taken in pairs from 15 chicks

y Log10 dry weight of bones at end of experiment

Details

Bones from 7-day-old chick embryos were cultivated over a nutrient chemical medium. Two bones were available from each chick, and the experiment was set out in a balanced incomplete block design with two units per block. The treatments were growth in the complete medium, with about 30 nutrients in carefully controlled quantities, and growth in five other media, each with a single amino acid omitted. Thus His-, Arg-, and so forth denote media without particular amino acids.

Source

Cox, D. R. and Snell, E. J. (1981) *Applied Statistics: Principles and Examples*. London: Chapman and Hall/CRC Press. Page 95.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 432.

---

chimps   *Chimpanzee Learning Data*

Description

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

Usage

```
data(chimps)
```

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.

References

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

Examples

```
data(chimps)
anova(glm(y~chimp+word,Gamma(log),data=chimps),test="F")
anova(glm(y~word+chimp,Gamma(log),data=chimps),test="F")
```

---

cloth   *Numbers of Flaws in Lengths of Cloth*

Description

The data comprise lengths of cloth samples and the numbers of flaws found in them.

Usage

```
data(cloth)
```

Format

A data frame with 32 observations on the following 2 variables.

x The length of the roll of cloth.

y The number of flaws found in the roll.

Source

Bissell, A. F. (1972) A negative binomial model with varying element size. *Biometrika*, **59**, 435–441.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 515.

Examples

```
data(cloth)
attach(cloth)
plot(x,y)
# Comparison of Poisson and quasilikelihood fits
summary(glm(y~x-1,family=poisson(identity)))
summary(glm(y~x-1,family=quasipoisson(identity)))
```

---

coal   *Data on UK coal mining disasters*

Description

The 'coal' data frame has 191 rows and 1 column.

This data frame gives the dates of 191 explosions in UK coal mines which resulted in 10 or more fatalities. The time span of the data is from March 15, 1851 until March 22 1962.

Usage

```
data(coal)
```

Format

This data frame contains the following column:

date The date of the disaster. The integer part of 'date' gives the year. The day is represented as the fraction of the year that had elapsed on that day.

Source

The data were obtained from

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

References

Jarrett, R.G. (1979) A note on the intervals between coal-mining disasters. *Biometrika*, **66**, 191–193.

Examples

```
data(coal)
plot(density(coal$date))
rug(coal$date)
```

---

coin.spin   *Function for Coin Spinning, Practical 11.1*

Description

This function computes the posterior distribution of the success probability theta when a coin is spun on its edge (or tossed), when the prior density for that probability is a mixture of beta densities.

Usage

```
coin.spin(para, r = 0, n = 0, n.points = 199)
```

Arguments

para	A matrix with 3 columns and k rows, where k is the number of components of the mixture. The first column contains the probabilities, and the next two the shape parameters a and b for the components.
r	Number of successes
n	Number of trials
n.points	The number of values of theta, equally-spaced between 0 and 1.

Details

This is provided simply so that readers spend less time typing. It is not intended to be robust and general code.

Value

x	Values of theta
y	Values of posterior density for theta

Author(s)

Anthony Davison (anthony.davison@epfl.ch)

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Practical 11.1.

Examples

```
## From Practical 11.1:
para <- matrix( c(0.5, 10, 20, 0.5, 20, 10), nrow=2, ncol=3, byrow=TRUE)
prior <- coin.spin(para)
plot(prior, xlab="theta",ylab="PDF", type="l",ylim=c(0,6))
post <- coin.spin(para, r=4, n=10)
```

---

danish   *Danish Fire Insurance Claims*

Description

Data on major insurance claims due to fires in Denmark, 1980–1990. The values of the claims have been rescaled for commercial reasons.

Usage

```
data(beaver)
```

Format

An irregular time series.

Source

Embrechts, P., Kluppelberg, C. and Mikosch, T. (1997) *Modelling Extremal Events for Insurance and Finance*. Berlin: Springer.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 278.

Examples

```
data(danish)
plot(danish,type="h")
```

darwin   *Darwin's Maize Data*

Description

The heights in eighths of inches of young maize plants put by Charles Darwin in four pots. He planted 15 pairs of plants together, one of each pair being cross-fertilised, and the other being self-fertilised.

Usage

```
data(darwin)
```

Format

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

pot a factor giving the pot

pair a factor giving the pair

type a factor giving the type of fertilisation

height height of plant in eighths of inches

Source

Fisher, R. A. (1935) *Design of Experiments*. Edinburgh: Oliver and Boyd. Page 30.

References

The original book is reprinted as part of Fisher, R. A. (1990) *Statistical Methods, Experimental Design, and Scientific Inference*. Oxford University Press.

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 2.

Examples

```
data(darwin)
attach(darwin)
plot(height~type)
anova(lm(height~pot+pair+type,data=darwin))
```

---

exp.gibbs	*Gibbs Sampling for Two Truncated Exponential Variables, Practical 11.3*

Description

Performs Gibbs sampling for problem with two truncated exponential variables. See Practical 11.3 of Davison (2003) for details.

Usage

```
exp.gibbs(u1 = NULL, u2 = NULL, B, I = 100, S = 100)
```

Arguments

u1	Initial values for variable 1
u2	Initial values for variable 2
B	Value at which exponential distribution is truncated
I	Number of iterations of sampler
S	Number of replicates of sampler

Details

This is provided simply so that readers spend less time typing. It is not intended to be robust and general code.

Value

A 2 x S x I array containing the values of the variables for the successive iterations

Author(s)

Anthony Davison (anthony.davison@epfl.ch)

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press.Practical 11.3.

Examples

```
add.exp.lines <- function( exp.out, i, B=10)
{
  dexp.trunc <- function( u, lambda, B )
    dexp(u, rate=lambda)/(1-exp(-lambda*B))
  S <- dim(exp.out)[2]
  I <- dim(exp.out)[3]
  u <- seq(0.0001,B,length=1000)
  fu <- rep(0,1000)
```

```
    for (s in 1:S) fu <- fu + dexp.trunc(u,exp.out[3-i,s,I],B)/S
    lines(u,fu,col="red")
    invisible()
}
par(mfrow=c(3,2))
B <-10; I <- 15; S <- 500
exp.out <- exp.gibbs(B=B,I=I,S=S)
hist(exp.out[1,,I],prob=TRUE,nclass=15,xlab="u1",ylab="PDF",xlim=c(0,B),ylim=c(0,1))
add.exp.lines(exp.out,1)
hist(exp.out[2,,I],prob=TRUE,nclass=15,xlab="u2",ylab="PDF",xlim=c(0,B),ylim=c(0,1))
add.exp.lines(exp.out,2)
```

---

eyes   *Visual Impairment Data*

Description

Joint distribution of visual impairment on both eyes, by race and age.

Usage

```
data(eyes)
```

Format

A data frame with 32 observations on the following 6 variables.

L Impairment (+) or not (-) for left eye.

R Impairment (+) or not (-) for right eye.

age a factor with levels 40-50 51-60 61-70 70+

colour White (W) or black (B)

a mid-point for age groups, as numeric vector

y Number of individuals in each class

Source

K.-Y. Liang, S. L. Zeger and B. Qaqish (1992) Multivariate regression analyses for categorical data (with Discussion). *Journal of the Royal Statistical Society, series B*, **54**, 3–40.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 505.

Examples

```
data(eyes)
eyes.glm <- glm(y~age*colour+L*R+(L+R):poly(a,2)+colour:(L+R),poisson,data=eyes)
anova(eyes.glm,test="Chi")   # analysis of deviance for loglinear model
```

---

field.concrete   *Field Concrete Mixture Data*

Description

Data from a 4x4 Latin square experiment on the efficiency of a field concrete mixer.

Usage

```
data(field.concrete)
```

Format

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

efficiency a numeric vector

speed a factor with levels 4, 8, 12, 16 mph.

run order in which runs were performed each day

day day on which runs were performed

Source

Unpublished lecture notes, Imperial College, London.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 435.

Examples

```
data(field.concrete)
fit <- lm(efficiency~run+day+speed,data=field.concrete)
anova(fit)
summary(fit)
fit <- lm(efficiency~run+day+poly(as.numeric(speed),3),data=field.concrete)
summary(fit)
```

---

fir   *Counts of Balsam-fir Seedlings*

Description

The number of balsam-fir seedlings in each quadrant of a grid of 50 five foot square quadrants were counted. The grid consisted of 5 rows of 10 quadrants in each row.

Usage

```
data(fir)
```

Format

A data frame with 50 observations on the following 3 variables.

count The number of seedlings in the quadrant

row The row number of the quadrant

col The quadrant number within the row

Source

Davison, A.C. and Hinkley, D.V. (1997) *Bootstrap Methods and Their Application*. Cambridge University Press.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 641.

---

forbes   *Atmospheric Pressure and Boiling Point in the Alps*

Description

James Forbes measured the atmospheric pressure and boiling point of water at 17 locations in the Alps.

Usage

```
data(forbes)
```

Format

A data frame with 17 observations on the following 2 variables.

bp Boiling point (Fahrenheit)

pres Pressure (inches of mercury)

Source

Atkinson, A. C. (1985) *Plots, Transformations, and Regression*. Oxford University Press.

Examples

```
data(forbes)
plot(forbes)
fit <- lm(bp~pres,data=forbes)
fit
plot(forbes$pres,resid(fit))  # model OK?
# try refitting with transformation
fit <- lm(log(bp)~log(pres),data=forbes)
```

---

frets	*Head Dimensions in Brothers*

Description

The 'frets' data frame has 25 rows and 4 columns.

The data consist of measurements of the length and breadth of the heads of pairs of adult brothers in 25 randomly sampled families. All measurements are expressed in millimetres.

Usage

```
data(frets)
```

Format

This data frame contains the following columns:

l1 The head length of the eldest son.

b1 The head breadth of the eldest son.

l2 The head length of the second son.

b2 The head breadth of the second son.

Source

Frets, G.P. (1921) Heredity of head form in man. *Genetica*, **3**, 193.

References

Whittaker, J. (1990) *Graphical Models in Applied Multivariate Statistics*. John Wiley.

Examples

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

---

ftse   *FTSE Daily Returns*

Description

Daily returns (index, 1991–1998.

Usage

```
data(ftse)
```

Format

A time series.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 266.

Examples

```
data(ftse)
plot(ftse,type="l",xlab="Time",ylab="Percent return")
plot(exp(cumsum(ftse/100)),type="l",xlab="Time",ylab="Relative closing value")
```

---

galaxy   *Galaxy Velocity Data*

Description

Velocities (km/second) of 82 galaxies in a survey of the Corona Borealis region.

Usage

```
data(galaxy)
```

Format

The format is: num [1:82] 9.17 9.35 9.48 9.56 9.78 ...

Source

Roeder, K. (1990) Density estimation with confidence sets exemplified by superclusters and voids in galaxies. *Journal of the American Statistical Association*, **85**, 617–624.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 214.

Examples

```
data(galaxy)
plot(density(galaxy))
rug(galaxy)
```

---

get.alpha   *Estimate Alpha from Data*

Description

Estimate prior value of a parameter from the data.

Usage

```
get.alpha(d)
```

Arguments

d   A data frame with vector components y and x of the same length.

Value

Prior value of a parameter, estimated from the data.

See Also

poi.beta.laplace

---

glm.diag   *Generalized Linear Model Diagnostics*

Description

Calculates jackknife deviance residuals, standardized deviance residuals, standardized Pearson residuals, approximate Cook statistic, leverage and estimated dispersion.

Usage

```
## S3 method for class 'diag'
glm(glmfit)
```

Arguments

glmfit   glmfit is a glm.object or a lm.object - the result of a call to glm() or lm()

Value

A list containing the following items:

res   The vector of jackknife deviance residuals.

rd   The vector of standardized deviance residuals.

rp   The vector of standardized Pearson residuals.

cook   The vector of approximate Cook statistics.

h   The vector of leverages of the observations.

sd   The value used to standardize the residuals. This is the the estimate of residual
                standard deviation in the Gaussian family and is the square root of the estimated
                shape parameter in the Gamma family. In all other cases it is 1.

Note

See the helpfile for glm.diag.plots for an example of the use of glm.diag.

Author(s)

Anthony Davison <anthony.davison@epfl.ch>

References

Davison, A.C. and Snell, E.J. (1991) Residuals and diagnostics. In *Statistical Theory and Modelling: In Honour of Sir David Cox.* D.V. Hinkley, N. Reid and E.J. Snell (editors), 83-106. Chapman and Hall.

See Also

glm,lm,plot.glm.diag,summary.glm

---

hus   *Haemolytic Uraemic Syndrome*

Description

Annual numbers of cases of 'diarrhoea-associated haemolytic uraemic syndrome' treated in clinics in Birmingham and Newcastle from 1970–1989.

Usage

data(hus)

Format

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

year a numeric vector

birmingham Number of cases treated in Birmingham

newcastle Number of cases treated in Newcastle

Source

Henderson, R. and Matthews, J. N. S. (1993) An investigation of changepoints in the annual number of cases of haemolytic uraemic syndrome. *Applied Statistics*, **42**, 461–471.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 142.

Examples

```
data(hus)
plot(hus$year,hus$birmingham,ylab="Annual number of cases",type="s")
```

---

hus.gibbs	*Gibbs Sampler for Poisson Changepoint Model, Practical 11.6*

Description

This function implements a Gibbs sampler for the Poisson changepoint model applied to the HUS data used in Example 4.40 and Practical 11.6 of Davison (2003), which should be consulted for details.

Usage

```
hus.gibbs(init, y, R = 10, a1 = 1, a2 = 1, c = 0.01, d = 0.01)
```

Arguments

init	Initial values for parameters
y	A series of Poisson counts
R	Number of iterations of sampler
a1	Value of a hyperparameter
a2	Value of a hyperparameter
c	Value of a hyperparameter
d	Value of a hyperparameter

Details

This is provided simply so that readers spend less time typing. It is not intended to be robust and general code.

Value

A matrix of size R x 7, whose first five columns contain the values of the parameters for the iterations. Columns 6 and 7 contain the log likelihood and log prior for that iteration.

Author(s)

Anthony Davison (anthony.davison@epfl.ch)

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Practical 11.6.

Examples

```
## From Example 11.6:
hus <- c(1,5,3,2,2,1,0,0,2,1,1,7,11,4,7,10,16,16,9,15)
system.time( gibbs.out <- hus.gibbs(c(5, 5, 1, 1, 2), hus, R=1000))
plot.ts(gibbs.out[,1], main="lambda1") # time series plot for lam1
plot.ts(gibbs.out[,2], main="lambda1") # time series plot for lam2
plot.ts(gibbs.out[,6], main="log lik") # and of log likelihood
table(gibbs.out[,5])  # tabulate observed values of tau
rm(hus)
```

---

ihess   *Inverse Hessian*

Description

Inverse Hessian matrix, useful for obtaining standard errors

Usage

```
ihess(f, x, ep = 1e-04, ...)
```

Arguments

f	Usually a negative log likelihood
x	Usually maximum likelihood estimates for f
ep	Step length used to compute numerical second derivatives
...	Extra arguments for f, if any

Value

Matrix of dimension dim(x) times dim(x), containing inverse Hessian matrix of f at x.

Note

This is not needed in R, where hessian matrices are obtained by setting hessian=T in calls to optimisation functions.

Author(s)

Anthony Davison

References

Based on code written by Stuart Coles of Padova University

Examples

```
# ML fit of t distribution
nlogL <- function(x, data) # negative log likelihood
{ mu <- x[1]
  sig <- x[2]
  df <- x[3]
  -sum(log( dt((data-mu)/sig, df=df)/sig )) }
y <- rt(n=100, df=10) # generate t data
# this is Splus code.....so remove the #'s for it to work in R
# fit <- nlminb(c(1,1,4), nlogL, upper=c(Inf,Inf,Inf), lower=c(-Inf,0,0),
#               data=y)
# fit$parameters # maximum likelihood estimates
# J <- ihess(nlogL, fit$parameters, data=y)
#  sqrt(diag(J)) # standard errors based on observed information
#
# In this example the standard error can be a bad measure of
# uncertainty for the df.
```

---

intron   *Intron Gene Sequence*

Description

Sequence of 1572 bases from first human preproglucagon gene

Usage

```
data(intron)
```

Format

This is a factor with 4 levels, "A","C","G","T"

Source

Avery, P. J. and Henderson, D. A. (1999) Fitting Markov chain models to discrete state series such as DNA sequences. *Applied Statistics*, **48**, 53–61.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 226.

Examples

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

---

jacamar   *Jacamar Learning Ability Data*

Description

Response of a rufous-tailed jacamar to butterflies, by not attacking them, by attacking but not eating them, and by attacking and eating them.

Usage

```
data(jacamar)
```

Format

A data frame with 48 observations on the following 5 variables.

species Butterfly species: Aphrissa boisduvalli (Ab), Phoebis argante (Pa), Dryas iulia (Di), Pierella luna (Pl), Consul fabius (Cf), Siproeta stelenes (Ss)

colour colour butterfly wings were painted: Unpainted, Brown, Yellow, Blue, Green, Red, Orange, Black

N Number not attacked

S Number attacked but rejected

E Number eaten

Details

As part of a study of the learning ability of tropical birds, Peng Chai of the University of Texas at Austin collected data on the response of a rufous-tailed jacamar to butterflies. He used marker pens to paint the underside of the wings of eight species of butterflies, and then released each butterfly in the cage where the bird was confined. The bird responded in three ways: by not attacking the butterfly (N); by attacking the butterfly, then sampling but rejecting it (S); or by attacking and eating the butterfly, usually after removing some or all of the wings (E).

Source

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 470.

Examples

```
data(jacamar)
```

---

jelinski	*Jelinski and Moranda Data on Software Failures*

Description

These are times in days between sucessive failures of a piece of software developed as part of a large data system. The software was released after the first 31 failures. The last three failures occurred after release.

Usage

```
data(jelinski)
```

Format

The format is: num [1:34] 9 12 11 4 7 2 5 8 5 7 ...

Source

Jelinski, Z. and Moranda, P. B. (1972) Software reliability research. In W. Freiberger (ed), *Statistical Computer Performance Evaluation*. London: Academic Press. Pages 465–484.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 299.

---

leuk	*Survival Times and White Blood Counts for Leukaemia Patients*

Description

A data frame of data from 33 leukaemia patients.

Usage

```
data(leuk)
```

Format

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

wbc white blood cell count

ag a test result, '"present"' or '"absent"'

time survival time in weeks

Details

Survival times are given for 33 patients who died from acute myelogenous leukaemia. Also measured was the patient's white blood cell count at the time of diagnosis. The patients were also factored into 2 groups according to the presence or absence of a morphologic characteristic of white blood cells. Patients termed AG positive were identified by the presence of Auer rods and/or significant granulation of the leukaemic cells in the bone marrow at the time of diagnosis.

Source

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

References

Cox, D. R. and Oakes, D. (1984) *Analysis of Survival Data*. Chapman & Hall, p. 9.

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 542.

Examples

```
data(leuk)
library(survival)
plot(survfit(Surv(time) ~ ag, data = leuk), lty = 2:3, col = 2:3)
# fit of exponential model
summary(glm(time~ag+log10(wbc),data=leuk,family=Gamma(log)),dispersion=1)
# now Cox models
leuk.cox <- coxph(Surv(time) ~ ag + log(wbc), leuk)
summary(leuk.cox)
```

---

lik.ci   *Likelihood Confidence Intervals for Scalar Parameter*

Description

A simple function for computing confidence intervals from the values of a likelihood function for a scalar parameter. It prints the maximum likelihood estimate (MLE) and its standard error, and confidence intervals based on normal approximation to the distribution of the MLE and on the chi-squared approximation to the distribution of the likelihood ratio statistic.

Usage

```
lik.ci(psi, logL, conf = c(0.975, 0.025))
```

Arguments

psi	Vector containing parameter values, the range of which contains the MLE
logL	Vector containing corresponding log likelihood values
conf	Vector containing levels for which confidence interval limits needed

Value

See above

Note

This uses the spline functions in library(modreg).

Author(s)

Anthony Davison (Anthony.Davison@epfl.ch)

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Sections 4.4.2, 4.5.1.

Examples

```
# likelihood analysis for mean of truncated Poisson data
y <- c(1:6)
n <- c(1486,694,195,37,10,1)
logL <- function(x, y, n.obs)      # x is theta
{  f <- dpois(y,x)/(1-dpois(0,x))  # dpois is Poisson PDF
   sum(n*log(f))  }                # log likelihood
theta <- seq(from=0.8, to=1, length=200)
L <- rep(NA, 200)
for (i in 1:200) L[i] <- logL(theta[i], y, n)
plot(theta, L, type="l", ylab="Log likelihood")
lik.ci(theta, L)
```

---

limits	*Swedish Speed Limit Data*

Description

The data are numbers of traffic accidents with personal injuries, reported to the police, on Swedish roads on 92 days in 1961 and 92 matching days in 1962. On some of these days a general speed limit of 90 or 100 km/hour was imposed.

Usage

```
data(limits)
```

Format

A data frame with 92 observations on the following 5 variables.

day A factor indicating the day, coded 1-92.

lim1 1 indicates a limit imposed in 1961, 0 not.

lim2 1 indicates a limit imposed in 1962, 0 not.

y1 Number of accidents on this day in 1961.

y2 Number of accidents on this day in 1962.

Source

Svensson, A. (1981) On a goodness-of-fit test for multiplicative Poisson models. *Annals of Statistics*, **9**, 697–704.

Examples

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

---

    lizards   *Lizard Count Data*

Description

These are data on the structural habitat of two species of lizards in Whitehouse, Jamaica. They comprise observed counts for perch height, perch diameter, insolation, and time of day, for both species. The data can be represented as a 2 x 2 x 2 x 3 x 2 contingency table.

Usage

```
data(lizards)
```

Format

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

height high indicates perch at height 5 or more feet, low indicates perch below 5 feet.

diameter large indicates perch diameter 2 inches or more, small indicates perch diameter less than 2 inches.

sun Is the perch in a shady or a sunny location?

time Time of day when lizard observed: early, late or midday.

species Species of lizard: grahami or opalinus.

y Number of lizards seen.

Source

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975) *Discrete Multivariate Analysis*. Cambridge, Mass.: MIT Press. Page 164.

Examples

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

---

lung.cancer   *Lung Cancer Deaths among UK Physicians*

Description

The data give the number of deaths due to lung cancer in British male physicians, as a function of years of smoking and cigarette consumption.

Usage

```
data(lung.cancer)
```

Format

A data frame with 63 observations on the following 4 variables.

years.smok a factor giving the number of years smoking

cigarettes a factor giving cigarette consumption

Time man-years at risk

y number of deaths

Source

Frome, E. L. (1983) The analysis of rates using Poisson regression models. *Biometrics*, **39**, 665–674.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 8.

Examples

```
data(lung.cancer)
```

---

magnesium   *Magnesium Treatment for Heart Attack Patients*

Description

Data from 11 clinical trials to compare magnesium treatment for heart attacks with control.

Usage

```
data(magnesium)
```

Format

A data frame with 22 observations on the following 4 variables.

trial a factor with levels 1–11

group Treatment indicator (factor)

m Total patients in group

r Number of deaths in group

Source

Copas, J. B. (1999) What works?: Selectivity models and meta-analysis. *Journal of the Royal Statistical Society series A*, **162**, 96–109.

References

Davison, A. C. (2003) Statistical Models. Cambridge University Press. Page 208.

Examples

```
data(magnesium)
fit <- glm(cbind(r,m-r)~trial+group,binomial,data=magnesium[1:20,])
anova(fit,test="Chi")
summary(fit)
```

| manaus | *Average Heights of the Rio Negro river at Manaus* |

Description

The 'manaus' time series is of class '"ts"' and has 1080 observations on one variable, which is the monthly average of the daily stages (heights) of the Rio Negro at Manaus, from January 1903 until December 1992. The units are metres.

Usage

```
data(manaus)
```

Format

A time series

Details

The data values are monthly averages of the daily stages (heights) of the Rio Negro at Manaus. Manaus is 18km upstream from the confluence of the Rio Negro with the Amazon but because of the tiny slope of the water surface and the lower courses of its flatland affluents, they may be regarded as a good approximation of the water level in the Amazon at the confluence. The data here cover 90 years from January 1903 until December 1992.

The Manaus gauge is tied in with an arbitrary bench mark of 100m set in the steps of the Municipal Prefecture; gauge readings are usually referred to sea level, on the basis of a mark on the steps leading to the Parish Church (Matriz), which is assumed to lie at an altitude of 35.874 m according to observations made many years ago under the direction of Samuel Pereira, an engineer in charge of the Manaus Sanitation Committee Whereas such an altitude cannot, by any means, be considered to be a precise datum point, observations have been provisionally referred to it. The measurements are in metres.

Source

The data were kindly made available by Professors H. O'Reilly Sternberg and D. R. Brillinger of the University of California at Berkeley.

References

Davison, A.C. and Hinkley, D.V. (1997) *Bootstrap Methods and Their Application*. Cambridge University Press.

Sternberg, H. O'R. (1987) Aggravation of floods in the Amazon river as a consequence of deforestation? *Geografiska Annaler*, 69A, 201-219.

Sternberg, H. O'R. (1995) Waters and wetlands of Brazilian Amazonia: An uncertain future. In *The Fragile Tropics of Latin America: Sustainable Management of Changing Environments*, Nishizawa, T. and Uitto, J.I. (editors), United Nations University Press, 113-179.

Examples

```
data(manaus)
plot(manaus)
acf(manaus)
pacf(manaus)
```

---

marking   *Examination Marking Data*

Description

The data are from an experiment to compare how different markers assess examination scripts, some of which were original and others of which were photocopies.

Usage

```
data(marking)
```

Format

A data frame with 32 observations on the following 5 variables.

Exam Two exams were marked

Script Scripts from 8 persons were marked, coded 1-8.

Marker Coded 1-4

Original Is the script an original (1) or a photocopy (0)?

y The mark out of 80 attributed by the marker.

Details

Normally each marker had a different batch of scripts, but for the experiment one script was taken at random from each batch and replaced after three copies of it had been made. The three copies were sent to the other three markers who assessed them, while the original was replaced and assessed in the usual way. Each of the four copies was therefore assessed by a single marker, but the three markers who had a copy knew that the script was part of the experiment, while the person marking the original did not know it to be part of the experiment. The experiment was repeated at another examination, with the same examiners, but different scripts.

Source

Lindley, D. V. (1961) An experiment in the marking of an examination (with Discussion). *Journal of the Royal Statistical Society, series A*, **124**, 285–313.

Examples

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

mathmarks	*Math Marks Data*

Description

Marks out of 100 for 88 students taking examinations in mechanics (C), vectors (C), algebra (O), analysis (O), statistics (O), where C indicates closed and O indicates open book examination.

Usage

```
data(mathmarks)
```

Format

A data frame with 88 observations on the following 5 variables.

mechanics mark out of 100 for mechanics

vectors mark out of 100 for vectors

algebra mark out of 100 for algebra

analysis mark out of 100 for analysis

statistics mark out of 100 for statistics

Source

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979) *Multivariate Analysis*. London: Academic Press. Pages 3–4.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 256.

Examples

```
data(mathmarks)
pairs(mathmarks)
var(mathmarks)
```

MClik   *Likelihood Estimation for Markov Chains*

Description

Computes maximum likelihood estimates of transition probabilities for stationary Markov chain models, of order 0 (independence) to 3.

This is intended for use with Practical 6.1 of Davison (2003), not as production code.

Usage

```
MClik(d)
```

Arguments

d   A sequence containing successive states of the chain

Value

order   order of fitted chain

df   degrees of freedom using in fitting

L   maximum log likelihood for each order

AIC   Akaike information criterion for each order

one   one-way marginal table of counts

two   two-way margin table of transitions

three   three-way marginal table of transitions

four   four-way marginal table of transitions

Author(s)

A. C. Davison (Anthony.Davison@epfl.ch)

References

Avery, P. J. and Henderson, D. A. (1999) Fitting Markov chain models to discrete state series such as DNA sequences. *Applied Statistics*, **48**, 53–61.

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Section 6.1.

Examples

```
data(intron)

fit <- MClik(intron)
```

---

mice *Mice Deaths from Radiation*

Description

RFM male mice were exposed to 300 rads of x-radiation at 5–6 weeks of age. The causes of death were thymic lymphoma, reticulum cell sarcoma, and other. Some of the mice were kept in a conventional environment, and the others in a germ-free environment.

Usage

```
data(mice)
```

Format

A data frame with 177 observations on the following 4 variables.

type Environment type (factor)

cause Cause of death

status Censoring indicator, with 1 indicating death

y Age at death (weeks)

Source

Hoel, D. G. and Walburg, H. E. (1972) Statistical analysis of survival experiments. *Journal of the National Cancer Institute*, **49**, 361–372.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 200.

Examples

```
data(mice)
library(survival)
fit <- survfit(Surv(y,status)~cause,data=mice[1:95,]) # first group
plot(fit,lty=c(3,2,1))
```

```
millet                    Millet Data
```

Description

Data from an experiment conducted to determine the optimal planting distance between plants in rows of millet. The rows were 1 foot apart. The design was a 5 x 5 Latin square.

Usage

```
data(millet)
```

Format

A data frame with 25 observations on the following 4 variables.

row Row label, coded 1-5.

col Column label, coded 1-5.

dist distances between plants:2, 4, 6, 8, or 10 inches.

y Average yield (grams) of three central rows, 15 feet long after allowing for discards, from each plot.

Source

Unpublished lecture notes, Imperial College, London.

Examples

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

```
motorette                 Motorette Failure Data
```

Description

Times to failure of motorettes tested at different temperatures.

Usage

```
data(motorette)
```

Format

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

x Temperature in degrees Fahrenheit

cens Censoring indicator

y Failure time in hours

Source

Nelson, W. D. and Hahn, G. J. (1972) Linear estimation of a regression relationship from censored data. Part 1 — simple methods and their application (with Discussion). *Technometrics*, **14**, 247–276.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 615.

Examples

```
data(motorette)
library(survival)
motor.fit <- survreg(Surv(y,cens)~log(x),dist="weibull",data=motorette)
summary(motor.fit)
```

---

nematode   *Nematode Data*

Description

Numbers of nematodes invading individual fly larvae for various numbers of initial challengers.

Usage

```
data(nematode)
```

Format

A data frame with 29 observations on the following 3 variables.

m Number of challengers

r Number of invading nematodes

y Number of fly larvae

Source

Faddy, M. J. and Fenlon, J. S. (1999) Stochastic modelling of the invasion process of nematodes in fly larvae. *Applied Statistics*, **48**, 31–37.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 295.

Examples

```
data(nematode)
```

---

nodal   *Nodal Involvement in Prostate Cancer*

Description

The 'nodal' data frame has 53 rows and 7 columns.

The treatment strategy for a patient diagnosed with cancer of the prostate depend highly on whether the cancer has spread to the surrounding lymph nodes. It is common to operate on the patient to get samples from the nodes which can then be analysed under a microscope but clearly it would be preferable if an accurate assessment of nodal involvement could be made without surgery.

For a sample of 53 prostate cancer patients, a number of possible predictor variables were measured before surgery. The patients then had surgery to determine nodal involvement. It was required to see if nodal involvement could be accurately predicted from the predictor variables and which ones were most important.

Usage

```
data(nodal)
```

Format

A data frame with 53 observations on the following 7 variables.

m A column of ones.

r An indicator of nodal involvement.

aged The patients age dichotomized into less than 60 ('0') and 60 or over '1'.

stage A measurement of the size and position of the tumour observed by palpatation with the fingers via the rectum. A value of '1' indicates a more serious case of the cancer.

grade Another indicator of the seriousness of the cancer, this one is determined by a pathology reading of a biopsy taken by needle before surgery. A value of '1' indicates a more serious case of the cancer.

xray A third measure of the seriousness of the cancer taken from an X-ray reading. A value of '1' indicates a more serious case of the cancer.

acid The level of acid phosphatase in the blood serum.

Source

Brown, B.W. (1980) Prediction analysis for binary data. In *Biostatistics Casebook*. R.G. Miller, B. Efron, B.W. Brown and L.E. Moses (editors), 3-18. John Wiley.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 491.

Examples

```
data(nodal)
nodal.glm <- glm(r~aged+stage+grade+xray+acid,binomial,data=nodal)
summary(nodal.glm,correlation=FALSE)
```

---

nuclear	*Nuclear Power Station Construction Data*

Description

The data relate to the construction of 32 light water reactor (LWR) plants constructed in the U.S.A in the late 1960's and early 1970's. The data was collected with the aim of predicting the cost of construction of further LWR plants. 6 of the power plants had partial turnkey guarantees and it is possible that, for these plants, some manufacturers' subsidies may be hidden in the quoted capital costs.

Usage

```
data(nuclear)
```

Format

A data frame with 32 observations on the following 11 variables.

cost The capital cost of construction in millions of dollars adjusted to 1976 base.

date The date on which the construction permit was issued. The data are measured in years since January 1 1990 to the nearest month.

t1 The time between application for and issue of the construction permit.

t2 The time between issue of operating license and construction permit.

cap The net capacity of the power plant (MWe).

pr A binary variable where '1' indicates the prior existence of a LWR plant at the same site.

ne A binary variable where '1' indicates that the plant was constructed in the north-east region of the U.S.A.

ct A binary variable where '1' indicates the use of a cooling tower in the plant.

bw A binary variable where '1' indicates that the nuclear steam supply system was manufactured by Babcock-Wilcox.

cum.n The cumulative number of power plants constructed by each architect-engineer.

pt A binary variable where '1' indicates those plants with partial turnkey guarantees.

Source

Cox, D.R. and Snell, E.J. (1981) *Applied Statistics: Principles and Examples*. Chapman and Hall.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 401.

Examples

```
data(nuclear)
pairs(nuclear)
fit <- lm(log(cost)~date+t1+t2+cap+pr+ne+ct+bw+cum.n+pr,data=nuclear)
step(fit)  # stepwise model selection
```

---

old.age   *Estimates of Hazard Function for Old Age*

Description

Historical estimates of the force of mortality (hazard function) averaged for 5-year age groups. The data are taken from various historical sources.

Usage

```
data(old.age)
```

Format

A data frame with 14 observations on the following 8 variables.

age Age group (5-year intervals)

hungary Data estimated from Hungarian graveyards, 900–1100

eng.1640 Data estimated from England, 1640–1689

breslau Data estimated from Breslau, 1687–1691

engm.1841 Data from England and Wales, males, 1841

engf.1841 Data from England and Wales, females, 1841

engm.1980 Data from England and Wales, males, 1980–1982

engf.1980 Data from England and Wales, females, 1980–1982

Details

The estimated numbers of people on which the data in the columns are based are 2300, 3133, 2675, 71,000, 74,000, 834,0000, and 828,000.

Source

Thatcher, A. R. (1999) The long-term pattern of adult mortality and the highest attained age (with Discussion). *Journal of the Royal Statistical Society series A*, **16**, 5–43.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 194.

Examples

```
data(old.age)
```

---

pairs.mod   *Modified Scatterplot Matrix*

Description

Plots a scatterplot matrix in which panels below the diagonal show ordinary scatterplots of pairs of variables, and those above the diagonal show scatterplots of residuals after regression on all the other variables.

Usage

```
## S3 method for class 'mod'
pairs(x, format = "MC", labelnames = names(x), highlight = NULL, level = 0.9, ...)
```

Arguments

x	A matrix whose rows correspond to units and whose columns correspond to variables measured on those units.
format	'MM' for marginal (that is, standard) scatterplots above and below the diagonal, 'MC' for marginal below and conditional (= partial) above, etc. 'MC' by default.
labelnames	Names of the variables.
highlight	Indexes of observations (rows) to be highlighted.
level	Scalar giving the level for the contour, 0.9 by default.
...	The plotting symbol and other arguments for the points can be controlled by 'pch=', etc.

Details

The diagonal shows histograms of the original data, and (in black) histograms of the partial residuals after adjustment on all the other variables, shifted to have the same mean as the original data. Also given are the original

The below-diagonal panels contain the numerical value of the correlation, and those above the diagonal contain the partial correlation, that is, the correlation of the residuals after linear regression on the remaining variables. The panels show ellipses which would contain 90 percent of the observations in a large normal sample with the same mean and covariance matrix as the data.

Value

Produces the scatterplot matrix, and prints the marginal and partial standard deviations of the variables.

Note

pairs.mod calls library(ellipse) and will give an error if this is unavailable.

Author(s)

Sylvain Sardy (Sylvain.Sardy@epfl.ch)

References

Davison, A. C. and Sardy, S. (2000) The partial scatterplot matrix. *Journal of Computational and Graphical Statistics*, **9**, 750–758.

Examples

```
library(ellipse)
data(mathmarks)
pairs.mod(mathmarks)
```

---

paulsen   *Neurotransmission in Guinea Pig Brains*

Description

The 'paulsen' data frame has 346 rows and 1 columns.

Sections were prepared from the brain of adult guinea pigs. Spontaneous currents that flowed into individual brain cells were then recorded and the peak amplitude of each current measured. The aim of the experiment was to see if the current flow was quantal in nature (i.e. that it is not a single burst but instead is built up of many smaller bursts of current). If the current was indeed quantal then it would be expected that the distribution of the current amplitude would be multimodal with modes at regular intervals. The modes would be expected to decrease in magnitude for higher current amplitudes.

Usage

```
data(paulsen)
```

Format

This data frame contains the following column:

y The current flowing into individual brain cells. The currents are measured in pico-amperes.

Source

The data were kindly made available by Dr. O. Paulsen of the Department of Pharmacology, University of Oxford.

Paulsen, O. and Heggelund, P. (1994) The quantal size at retinogeniculate synapses determined from spontaneous and evoked EPSCs in guinea-pig thalamic slices. *Journal of Physiology*, **480**, 505-511.

Examples

```
data(paulsen)
hist(paulsen$y,prob=TRUE)
```

---

pbc	*Mayo Clinic Primary Biliary Cirrhosis Data*

Description

Followup of 312 randomised and 108 unrandomised patients with primary biliary cirrhosis, a rare autoimmune liver disease, at Mayo Clinic.

Usage

```
data(pbc)
```

Format

A data frame with 418 observations on the following 20 variables.

age in years

alb serum albumin

alkphos alkaline phosphotase

ascites presence of ascites

bili serum bilirubin

chol serum cholesterol

edema presence of edema

edtrt 0 no edema, 0.5 untreated or successfully treated 1 unsuccessfully treated edema

hepmeg enlarged liver

time survival time

platelet platelet count

protime standardised blood clotting time

sex 1=male

sgot liver enzyme (now called AST)

spiders blood vessel malformations in the skin

stage histologic stage of disease (needs biopsy)

status censoring status

trt 1/2/-9 for control, treatment, not randomised

trig triglycerides

copper urine copper

Source

Fleming, T. R. and Harrington, D. P. (1991) *Counting Processes and Survival Analysis*. Wiley: New York.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 549.

Examples

```
data(pbc)
# to make version of dataset used in book
pbcm <- pbc[(pbc$trt!=-9),]
pbcm$copper[(pbcm$copper==-9)] <- median(pbcm$copper[(pbcm$copper!=-9)])
pbcm$platelet[(pbcm$platelet==-9)] <- median(pbcm$platelet[(pbcm$platelet!=-9)])
attach(pbcm)

library(survival)
par(mfrow=c(1,2),pty="s")
plot(survfit(Surv(time,status)~trt),ylim=c(0,1),lty=c(1,2),
   ylab="Survival probability",xlab="Time (days)")
plot(survfit(coxph(Surv(time,status)~trt+strata(sex))),ylim=c(0,1),lty=c(1,2),
   ylab="Survival probability",xlab="Time (days)")
lines(survfit(coxph(Surv(time,status)~trt)),lwd=2)
# proportional hazards model fit
fit <- coxph(formula = Surv(time, status) ~ age + alb + alkphos + ascites +
      bili + edtrt + hepmeg + platelet + protime + sex + spiders, data=pbcm)
summary(fit)
step.fit <- step(fit,direction="backward")
```

---

pigeon   *Homing Pigeon Data*

Description

Bearings (degrees) of 29 homing pigeons 30, 60, 90 after their release, and on vanishing from sight.

Usage

```
data(pigeon)
```

Format

A data frame with 29 observations on the following 4 variables.

s30 Bearing after 30 seconds

s60 Bearing after 60 seconds

s90 Bearing after 90 seconds

van Bearing on vanishing from sight

Source

Artes, R. (1997) Extensoes da Teoria das Equacoes de Estimacao Generalizadas a Dados Circulares e Modelos de Dispersao. Ph.D. thesis, University of Sao Paulo.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 173.

Examples

```
data(pigeon)
plt <- function( ang, r=c(1,2,3,4), lty=1,... )
{
  si <- sin(2*pi*ang/360)
  co <- cos(2*pi*ang/360)
  points( r*si,r*co )
  lines( c(0,r*si),c(0,r*co),...)
}
par(pty="s")
plot(c(0,0),c(0,0),xlim=c(-4,4),ylim=c(-4,4),
    xlab="Easting",ylab="Northing")
for (i in 1:nrow(pigeon)) plt( pigeon[i,],col=i )
```

---

pigs   *Pig Diet Data*

Description

Data on weight gains in 32 pigs, divided into eight groups of four, and with 4 different diets allocated to the group members.

Usage

```
data(pigs)
```

Format

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

group a factor with 8 levels

diet a factor with levels `I–IV`

gain weight gain (units unknown)

Source

Unpublished lecture notes, Imperial College, London

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 431.

Examples

```
data(pigs)
anova(lm(gain~group+diet,data=pigs))
```

---

pipette   *Red Blood Cell Data*

Description

Numbers of red blood cells counted by five doctors using ten sets of apparatus.

Usage

```
data(pipette)
```

Format

A data frame with 50 observations on the following 3 variables.

apparatus Factor with ten levels

doctor Factor with five levels

y Number of red blood cells

Source

Unpublished lecture notes, Imperial College, London.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 462.

| plot.glm.diag | *Diagnostic plots for generalized linear models* |

Description

Makes plot of jackknife deviance residuals against linear predictor, normal scores plots of standardized deviance residuals, plot of approximate Cook statistics against leverage/(1-leverage), and case plot of Cook statistic.

Usage

```
## S3 method for class 'glm.diag'
plot(x, glmdiag = glm.diag(x), subset = NULL, iden = FALSE,
                    labels = NULL, ret = FALSE, ...)
```

Arguments

x	glm.object or lm.object : the result of a call to glm() or lm()
glmdiag	Diagnostics of x obtained from a call to glm.diag. If it is not supplied then it is calculated.
subset	Subset of data for which glm fitting performed: should be the same as the subset option used in the call to glm() or lm() which generated x. Needed only if the subset= option was used in the call to glm or lm.
iden	A logical argument. If TRUE then, after the plots are drawn, the user will be prompted for an integer between 0 and 4. A positive integer will select a plot and invoke identify() on that plot. After exiting identify(), the user is again prompted, this loop continuing until the user responds to the prompt with 0. If iden is FALSE (default) the user cannot interact with the plots.
labels	A vector of labels for use with identify() if iden is TRUE. If it is not supplied then the labels are derived from x.
ret	A logical argument indicating if glmdiag should be returned. The default is FALSE.
...	Other arguments, which are ignored. This is included only for compatibility with S3 methods.

Details

The plot on the top left is a plot of the jackknife deviance residuals against the fitted values.

The plot on the top right is a normal QQ plot of the standardized deviance residuals. The dotted line is the expected line if the standardized residuals are normally distributed, i.e. it is the line with intercept 0 and slope 1.

The bottom two panels are plots of the Cook statistics. On the left is a plot of the Cook statistics against the standardized leverages. In general there will be two dotted lines on this plot. The horizontal line is at $8/(n-2p)$ where n is the number of observations and p is the number of parameters estimated. Points above this line may be points with high influence on the model. The vertical line

is at 2p/(n-2p) and points to the right of this line have high leverage compared to the variance of the raw residual at that point. If all points are below the horizontal line or to the left of the vertical line then the line is not shown.

The final plot again shows the Cook statistic this time plotted against case number enabling us to find which observations are influential.

Use of iden=T is encouraged for proper exploration of these four plots as a guide to how well the model fits the data and whether certain observations have an unduly large effect on parameter estimates.

Value

If ret is TRUE then the value of glmdiag is returned otherwise there is no returned value.

Author(s)

Angelo Canty

References

Davison, A.C. and Snell, E.J. (1991) Residuals and diagnostics. In *Statistical Theory and Modelling: In Honour of Sir David Cox*. D.V. Hinkley, N. Reid, and E.J. Snell (editors), 83-106. Chapman and Hall.

See Also

[glm](#),[glm.diag](#),[identify](#)

Examples

```
   # leukaemia data
require(MASS)
data(leuk, package="MASS")
leuk.mod <- glm(time~ag-1+log10(wbc),family=Gamma(log),data=leuk)
leuk.diag <- glm.diag(leuk.mod)
plot.glm.diag(leuk.mod,leuk.diag)
```

---

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

```
data(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.

References

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

Examples

```
data(pneu)
summary(glm(cbind(Present+Severe,Normal)~log(Years),data=pneu,binomial))
summary(glm(cbind(Severe,Normal+Present)~log(Years),data=pneu,binomial))
```

---

poi.beta.laplace	*Laplace Approximation for Posterior Density, Practical 11.2*

Description

This function computes the Laplace approximation to the posterior density of the parameter beta in a Poisson regression model. For more details see Practical 11.2 of Davison (2003).

Usage

```
poi.beta.laplace(data, alpha = get.alpha(data), phi = 1, nu = 0.1,
                 beta = seq(from = 0, to = 7, length = 1000))
```

Arguments

data	A data frame with vector components y and x of the same length. y contains the numbers of counts, and x the corresponding time intervals.
alpha	Prior value of a parameter, estimated from the data by default.
phi	Prior value of a parameter.
nu	Prior value of a parameter.
beta	Values for which posterior density of beta should be provided.

Details

This is provided simply so that readers spend less time typing. It is not intended to be robust and general code.

Value

int	Estimated integral of posterior density.
conv	Did the routine for the Laplace optimization converge?
x	Values of beta
y	Values of posterior density

Author(s)

Anthony Davison (anthony.davison@epfl.ch)

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Practical 11.2.

Examples

```
## From Practical 11.2:
get.alpha <- function(d)
{  # estimate alpha from data
  rho <- d$y/d$x
  n <- length(d$y)
  mean(rho)^2/( (n-1)*var(rho)/n - mean(rho)*mean(1/d$x) )
}
data(cloth)
attach(cloth)
plot(x,y)
beta <- seq(from=0,to=10,length=1000)
beta.post <- poi.beta.laplace(cloth,beta=beta,nu=1)
plot(beta.post,type="l",xlab="beta",ylab="Posterior density")
beta.post <- poi.beta.laplace(cloth,beta=beta,nu=5)
lines(beta.post,lty=2)
```

---

poi.gibbs   *Gibbs Sampler for Hierarchical Poisson Model, Practical 11.5*

Description

This function implements Gibbs sampling for the hierarchical Poisson model described in Example 11.19 and Practical 11.5 of Davison (2003), which should be consulted for more details.

Usage

```
poi.gibbs(d, alpha, gamma, delta, I, S)
```

Arguments

d	A data frame with vector components y containing the numbers of counts and x the period for which the n Poisson processes are observed.
alpha	A hyperparameter of the prior density
gamma	A hyperparameter of the prior density
delta	A hyperparameter of the prior density
I	Number of iterations for which sampler is run
S	Number of independent replicates of sampler

Details

This is provided simply so that readers spend less time typing. It is not intended to be robust and general code.

Value

An I x S x (n+1) array containing the successive iterations of the samplers, for the I iterations, S independent replicates, and n rate parameters plus the parameter beta of the prior distribution.

Author(s)

Anthony Davison (anthony.davison@epfl.ch)

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Practical 11.5.

Examples

```
 ## From Practical 11.5:
data(pumps)
system.time( pumps.sim <- poi.gibbs(pumps, alpha=1.8, delta=0.1, gamma=1,
             I=1000, S=5) )
par(mfrow=c(2,3))
plot.ts(pumps.sim[,1,1])
acf(pumps.sim[,1,1])
pacf(pumps.sim[,1,1])
plot.ts(pumps.sim[,1,11])
acf(pumps.sim[,1,11])
pacf(pumps.sim[,1,11])
```

---

poisons   *Survival Times for Poisoned Animals*

Description

In an experiment to assess the usefulness of treatments for poisons, 48 animals were split randomly into 12 groups of 4. Each group was administered one of three poisons, and one of four treatments, giving a 3x4 factorial design with 4 replicates.

Usage

```
data(poisons)
```

Format

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

time Survival time (units of 10 hours)

poison Factor giving poison

treat Factor giving treatment

Source

Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations (with Discussion). *Journal of the Royal Statistical Society series B*, **26**, 211–246.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 391.

Examples

```
data(poisons)
fit <- lm(time~poison+treat,data=poisons)
library(MASS)
boxcox(time~poison+treat,data=poisons)
```

---

pollution   *Air Pollution and Mortality*

Description

Data on the relation between weather, socioeconomic, and air pollution variables and mortality rates in 60 Standard Metropolitan Statistical Areas (SMSAs) of the USA, for the years 1959-1961. Some of the variables are highly collinear.

Usage

```
data(pollution)
```

Format

A data frame with 60 observations on the following variables.

prec Average annual precipitation in inches

jant Average January temperature in degrees F

jult Average July temperature in degrees F

ovr95 Percentage of 1960 SMSA population aged 65 or older

popn Average household size

educ Median school years completed by those over 22

hous percentage of housing units which are sound and with all facilities

dens Population per square mile in urbanized areas, 1960

nonw Percentage non-white population in urbanized areas, 1960

wwdrk Percentage employed in white collar occupations

poor Percentage of families with income < 3000 dollars

hc Relative hydrocarbon pollution potential

nox Same for nitric oxides

so Same for sulphur dioxide

humid Annual average percentage relative humidity at 1pm

mort Total age-adjusted mortality rate per 100,000

Source

McDonald, G. C. and Schwing, R. C. (1973) Instabilities of regression estimates relating air pollution to mortality, *Technometrics*, **15**, 463-482.

Examples

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

---

pumps   *Pump Failure Data*

Description

The data give numbers of failures of ten pumps from several systems in the nuclear plant Farley 1.
Pumps 1, 3, 4, and 6 operate continuously, while the rest operate only intermittantly or on standby.

Usage

```
data(pumps)
```

Format

A data frame with 10 observations on the following 2 variables.

x Operating time (in thousands of operatin hours)

y Number of failures

Source

Gaver, D. P. and O'Muircheartaigh, I. G. (1987) Robust empirical Bayes analysis of event rates.
Technometrics, **29**, 1–15.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 600.

---

qqexp   *Exponential Quantile-Quantile Plots*

Description

Exponential probability plot of data.

Usage

```
qqexp(y, line = FALSE, ...)
```

Arguments

y	Vector for which plot is required
line	Add line to plot (no line by default)
...	Other options for plot command

Value

A exponential probablity plot of the data in y; that is, a plot of the ordered values of y against the quantiles of the standard exponential distribution.

See Also

[qqnorm](#)

Examples

```
qqexp(rexp(50))
qqexp(rgamma(50,shape=2),line=TRUE)
```

---

quake   *Japanese Earthquake Data*

Description

Times and magnitudes (Richter scale) of 483 shallow earthquakes in an offshore region east of Honshu and south of Hokkaido, for the period 1885–1980.

Usage

```
data(quake)
```

Format

An irregular time series with earthquake

time in days since start of 1885

mag magnitude (Richter scale)

Source

Ogata, Y. (1988) Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, **83**, 9–27.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 289.

---

rat.growth   *Rat Growth Data*

Description

Data on the weights of 30 rats each week for 5 weeks.

Usage

```
data(rat.growth)
```

Format

A data frame with 150 observations on the following 3 variables.

rat a factor with levels 1-30

week takes values 0-4

y rat weight (units unspecified)

Source

Gelfand, A. E., Hills, S. E., Racine-Poon, A. and Smith, A. F. M. (1990) Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association*, **85**, 972–985.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 460.

Examples

```
data(rat.growth)
library(nlme)
rat.fit <- groupedData( y~poly(week,2) | rat,
                  data = rat.growth,
                  labels = list( x = "Week",
                    y = "Weight" ),
                  units = list( x = "", y = "(?)") )

summary(lme(rat.fit))
```

salinity	*Water Salinity and River Discharge*

Description

The 'salinity' data frame has 28 rows and 4 columns.

Biweekly averages of the water salinity and river discharge in Pamlico Sound, North Carolina were recorded between the years 1972 and 1977. The data in this set consists only of those measurements in March, April and May.

Usage

```
data(salinity)
```

Format

This data frame contains the following columns:

sal The average salinity of the water over two weeks.

lag The average salinity of the water lagged two weeks. Since only spring is used, the value of 'lag' is not always equal to the previous value of 'sal'.

trend A factor indicating in which of the 6 biweekly periods between March and May, the observations were taken. The levels of the factor are from 0 to 5 with 0 being the first two weeks in March.

dis The amount of river discharge during the two weeks for which 'sal' is the average salinity.

Source

The data were obtained from

Ruppert, D. and Carroll, R.J. (1980) Trimmed least squares estimation in the linear model. *Journal of the American Statistical Association*, **75**, 828–838.

Examples

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

---

seeds   *Germination of seeds*

Description

These are the number of seeds germinating when subjected to extracts of certain roots.

Usage

```
data(seeds)
```

Format

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

r Number of seeds germinating

m Total number of seeds

seed Seed type: O. aegyptiaco 75 or O. aegyptiaco 73

root Root extract

Source

Crowder, M. J. (1978) Beta-binomial ANOVA for proportions. *Applied Statistics*, **27**, 34–37.

References

Cox, D. R. and Snell, E. J. (1989) *Analysis of Binary Data*, second edition. London: Chapman and Hall. Section 3.2.

Examples

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

---

shoe   *Shoe Wear Data*

Description

Amount of wear in a paired comparison of two materials used for soling the shoes of 10 boys. The materials were allocated randomly to the left and right feet.

Usage

```
data(shoe)
```

Format

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

material factor giving the shoe sole material

boy factor with 10 levels

foot factor giving left or right foot

y amount of shoe wear

Source

Box, G. E. P., Hunter, W. G. and Hunter, J. S. (1978) *Statistics for Experimenters*. New York: Wiley. Page 100.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 422.

Examples

```
data(shoe)
attach(shoe)
d <- y[material=="B"]-y[material=="A"]  # difference
t.test(d)  # t test of hypothesis that B wears quicker
```

---

shuttle   *O-ring Thermal Distress Data for Space Shuttle*

Description

Data on the number of rubber O-rings showing thermal distress for 23 flights of the space shuttle, with the ambient temperature and pressure at which tests on the putty next to the rings were performed.

Usage

```
data(shuttle)
```

Format

A data frame with 23 observations on the following 4 variables.

m Number of rings

r Number of rings showing thermal distress

temperature ambient temperature (degrees Fahrenheit)

pressure pressure (pounds per square inch)

Source

Dalal, S. R., Fowlkes, E. B. and Hoadley, B. (1989) Risk analysis of the space shuttle: Pre-Challenger prediction of failure. *Journal of the American Statistical Association*, **84**, 945–957.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 7.

Examples

```
data(shuttle)
attach(shuttle)
plot(temperature, r/m,ylab="Proportion of failures")
```

---

smoking   *Survival and Smoking*

Description

Twenty-year survival and smoking status for 1314 women from Whickham, near Newcastle-upon-Tyne.

Usage

```
data(smoking)
```

Format

A data frame with 14 observations on the following 4 variables.

age Age group (factor)

smoker Smoking status (1=smoker, 0=non-smoker)

alive Number alive after 20 years

dead Number dead after 20 years

Source

Appleton, D. R., French, J. M. and Vanderpump, M. P. J. (1996) Ignoring a covariate: An example of Simpson's paradox. *The American Statistician*, **50**, 340–341.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 258.

Examples

```
data(smoking)
summary(glm(cbind(dead,alive)~smoker,data=smoking,binomial))
# note sign change for smoker covariate, due to Simpson's paradox
summary(glm(cbind(dead,alive)~age+smoker,data=smoking,binomial))
```

---

soccer *Soccer Scores from English Premier League, 2000-2001 Season*

Description

These are scores for the 380 fixtures in the English Premier League, 2000–2001.

Usage

```
data(soccer)
```

Format

A data frame with 380 observations on the following 7 variables.

month Month of match

day Day of match

year Year of match

team1 Home team

team2 Away team

score1 Goals scored by home team

score2 Goals scored by away team

Source

http://www.soccerbase.com/footballlive/

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 499.

---

springs   *Spring Failure Data*

Description

Failure times of 60 springs divided into 6 groups of 10, with each group subject to a different level of stress. Some of the times are right-censored.

Usage

```
data(springs)
```

Format

A data frame with 60 observations on the following 3 variables.

cycles failure times (in units of 10^3 cycles of loading)

cens censoring indicator, with 0 indicating right-censoring

stress a factor giving the stress (N/mm^3)

Source

Cox, D. R. and Oakes, D. (1984) *Analysis of Survival Data*. London: Chapman and Hall/CRC Press.

Examples

```
data(springs)
attach(springs)
plot(cycles~stress)
plot(cycles~stress,log="y")
```

---

sticky   *Stickiness of blood data*

Description

Data on stickiness of blood for six subjects

Usage

```
data(sticky)
```

Format

A data frame with 42 observations on the following 2 variables.

subject factor with levels 1–6

y measurement of a property related to stickiness of blood

Source

Unpublished lecture notes, Imperial College London.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 450.

Examples

```
data(sticky)
anova(lm(y~subject,data=sticky))
```

---

survival *Survival of Rats After Radiation Doses*

Description

The 'survival' data frame has 14 rows and 2 columns.

The data measured the survival percentages of batches of rats who were given varying doses of radiation. At each of 6 doses there were two or three replications of the experiment.

Usage

```
data(survival)
```

Format

A data frame with 14 observations on the following 2 variables.

dose The dose of radiation administered (rads).

surv The survival rate of the batches expressed as a percentage.

Source

Efron, B. (1988) Computer-intensive methods in statistical regression. *SIAM Review*, **30**, 421-449.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 376.

Examples

```
data(survival)
plot(survival$dose,survival$surv,log="y")  # note the obvious outlier
lm(log(surv)~dose,data=survival)
```

---

teak   *Teak Plant Data*

Description

These are data from an experiment on the growth of teak plants after one season, using two planting methods and three root lengths. Plants were laid out in four randomised blocks, each consisting of 6 plots with 50 plants in each plot.

Usage

```
data(teak)
```

Format

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

Block Block labels.

Plant A indicates planting using pits, B using crowbar.

Root length, 4, 6 or 8 inches.

y mean height (inches) of the 50 plants grown on each plot.

Source

Unpublished lecture notes, Imperial College, London.

Examples

```
data(teak)
anova(lm(y~Block*Plant*Root,data=teak),test="F")
```

---

tide *Annual Maximum Sea Levels*

Description

Annual maximum sea levels (m) at seven locations near to or in south-east England, between 1819–1986. There are many missing values.

Usage

```
data(tide)
```

Format

A data frame with 168 observations on the following 8 variables.

year Year

Yarmouth Annual maximum high tide at Yarmouth

Lowestoft Annual maximum high tide at Lowestoft

Harwich Annual maximum high tide at Harwich

Walton Annual maximum high tide at Walton

Holland Annual maximum high tide at a site in Holland

Southend Annual maximum high tide at Southend

Sheerness Annual maximum high tide at Sheerness

Source

The data were kindly provided by Professor Jonathan Tawn of Lancaster University.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 281.

Examples

```
data(tide)
plot(tide$year,tide$Yarmouth,type="l")
```

---

toxo   *Toxoplasmosis Data*

Description

Data on the relation between rainfall and the numbers of people testing positive for toxoplasmosis in 34 cities in El Salvador.

Usage

```
data(toxo)
```

Format

A data frame with 34 observations on the following 3 variables.

rain Annual rainfall (mm)

m Number of persons tested

r Number of persons testing positive for toxoplasmosis

Source

Efron, B. (1986) Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association*, **82**, 171–200.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 516.

Examples

```
data(toxo)
anova(glm(cbind(r,m-r)~poly(rain,3),data=toxo,family=binomial),test="Chi")
fit <- glm(cbind(r,m-r)~poly(rain,3),data=toxo,family=quasibinomial)
anova(fit,test="F")
summary(fit)
```

---

ulcer   *Recurrent Bleeding from Ulcers*

Description

Data from 40 experiments to compare a new surgery for stomach ulcer with an older surgery.

Usage

```
data(ulcer)
```

Format

A data frame with 80 observations on the following 9 variables.

author Author of study from which data taken

year Year of publication

quality Assessment of quality of trial on which data based

age Mean age of patients

r Number of patients without recurrent bleeding

m Total number of patients

bleed a numeric vector

treat Factor giving control (C) or variants of new treatment

table Factor giving 2x2 table corresponding to each trial

Source

Efron, B. (1996) Empirical Bayes methods for combining likelihoods (with Discussion). *Journal of the American Statistical Association*, **91**, 538–565.

Errors in the data given in the paper have been corrected here.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 496.

Examples

```
data(ulcer)
glm(cbind(r,m-r)~table+treat,data=ulcer,family=binomial)
```

urine	*Urine Analysis Data*

Description

The 'urine' data frame has 79 rows and 7 columns.

79 urine specimens were analyzed in an effort to determine if certain physical characteristics of the urine might be related to the formation of calcium oxalate crystals. Cases 1 and 55 have missing covariates.

Usage

```
data(urine)
```

Format

This data frame contains the following columns:

r Indicator of the presence of calcium oxalate crystals.

gravity The specific gravity of the urine.

ph The pH reading of the urine.

osmo The osmolarity of the urine. Osmolarity is proportional to the concentration of molecules in solution.

cond The conductivity of the urine. Conductivity is proportional to the concentration of charged ions in solution.

urea The urea concentration in millimoles per litre.

calc The calcium concentration in millimoles per litre.

Source

Andrews, D.F. and Herzberg, A.M. (1985) *Data: A Collection of Problems from Many Fields for the Student and Research Worker*. Springer-Verlag. Pages 249–251.

Examples

```
data(urine)
glm(r~gravity+ph+osmo+cond+urea+calc,binomial,data=urine,subset=-c(1,55))
```

---

venice   *Extreme Sea Levels at Venice*

Description

The ten highest annual sea levels (cm) at Venice, from 1887–1981.

Usage

```
data(venice)
```

Format

A data frame with 95 observations on the following 11 variables.

year 1887–1981
y1 Annual maximum sea level (cm)
y2 Second largest sea level (cm)
y3 Third largest sea level (cm)
y4 Fourth largest sea level (cm)
y5 Fifth largest sea level (cm)
y6 Sixth largest sea level (cm)
y7 Seventh largest sea level (cm)
y8 Eighth largest sea level (cm)
y9 Ninth largest sea level (cm)
y10 Tenth largest sea level (cm)

Details

There are missing values in 1922 and 1935.

Source

Pirazzoli, P. A. (1982) Maree estreme a Venezia (periodo 1872–1981). *Acqua Aria*, **10**, 1023–1039.

References

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. Page 162.

Examples

```
data(venice)
attach(venice)
y <- y1[year>1930]  # for analysis in Section 5 of Davison (2003)
x <- year[year>1930]-1956
plot(x+1956,y,ylab="Sea level (cm)",xlab="Year")
lm(y~x)
```

---

yahoo   *Yahoo Closing Prices*

Description

Daily closing prices (US dollars) of Yahoo.com shares from 12 April 1996 to 26 April 2000.

Usage

```
data(yahoo)
```

Format

An irregular time series with 1017 values.

Examples

```
data(yahoo)
plot(yahoo,type="l",ylab="Yahoo closing prices")
plot(diff(100*log(yahoo)),type="l",ylab="Yahoo log returns (percent)")
```

Index