Package ‘gamlss.data’

May 15, 2019

Description  Data used as examples to demonstrate GAMLSS models.
Title  GAMLSS Data
LazyData  yes
Version  5.1-4
Date  2019-05-14
Depends  R (>= 2.10)
Author  Mikis Stasinopoulos <d.stasinopoulos@londonmet.ac.uk>, Bob
       Rigby, Fernanda De Bastiani
Maintainer  Mikis Stasinopoulos <d.stasinopoulos@londonmet.ac.uk>
License  GPL-2 | GPL-3
URL  http://www.gamlss.org/
NeedsCompilation  no
Repository  CRAN
Date/Publication  2019-05-15 09:40:07 UTC

R topics documented:

abdom ................................................................. 2
acidity ............................................................... 3
aep ................................................................. 4
aids ................................................................. 5
aircond ............................................................ 6
alveolar ......................................................... 7
brownfat ....................................................... 8
bush2000 ......................................................... 9
cable ......................................................... 10
CD4 ............................................................ 11
computer ....................................................... 12
cysts .......................................................... 12
db ............................................................. 13
dbbmi ......................................................... 14
Description

The `abdom` data frame has 610 rows and 2 columns. The data are measurements of abdominal circumference (response variable) taken from fetuses during ultrasound scans at Kings College Hospital, London, at gestational ages (explanatory variable) ranging between 12 and 42 weeks.
Usage

data(abdom)

Format

This data frame contains the following columns:

- **y**: abdominal circumference: a numeric vector
- **x**: gestational age: a numeric vector

Details

The data were used to derived reference intervals by Chitty et al. (1994) and also for comparing different reference centile methods by Wright and Royston (1997), who also commented that the distribution of Z-scores obtained from the different fitted models 'has somewhat longer tails than the normal distribution'.

Source

Dr. Eileen M. Wright, Department of Medical Statistics and Evaluation, Royal Postgraduate Medical School, Du Cane Road, London, W12 0NN.

References


Examples

data(abdom)
attach(abdom)
plot(x,y)
detach(abdom)

---

**acidity**

*The Acidity Data files for GAMLSS*

Description

The data shows the acidity index for 155 lakes in the Northeastern United States (previously analysed as a mixture of gaussian distributions on the log scale by Crawford et al.(1992, 1994)). These 155 observations are the log acidity indices for the lakes.

Usage

data(acidity)
Format

A data frame with 155 observations on the following variable.

\textit{y} a numeric vector showing the acidity index for 155 lakes in the Northeastern United States

References


Examples

\begin{verbatim}
data(acidity)
with( acidity, hist(y))
\end{verbatim}

\begin{verbatim}
aep  The Hospital Stay Data
\end{verbatim}

Description

The data, 1383 observations, are from a study at the Hospital del Mar, Barcelona during the years 1988 and 1990, Gange et al. (1996).

Usage

\begin{verbatim}
data(aep)
\end{verbatim}

Format

A data frame with 1383 observations on the following 8 variables.

\textbf{los} the total number of days patients spent in hospital: a discrete vector

\textbf{noinap} the number of inappropriate days spent in hospital: a discrete vector

\textbf{loglos} the log(los/10): a numeric vector

\textbf{sex} the gender of patient: a factor with levels 1=male, 2=female

\textbf{ward} the type of ward in the hospital: a factor with levels 1=medical 2=surgical, 3=others

\textbf{year} the specific year 1988 or 1990: a factor with levels 88 and 90

\textbf{age} the age of the patient subtracted from 55: a numeric vector

\textbf{y} the response variable a matrix with 2 columns, the first is noinap the second is equal to (los-noinap)
Details

Gange et al. (1996) used a logistic regression model for the number of inappropriate days (noinap) out of the total number of days spent in hospital (los), with binomial and beta binomial errors and found that the later provided a better fit to the data. They modelled both the mean and the dispersion of the beta binomial distribution (BB) as functions of explanatory variables.

Source


References


Examples

data(aep)
attach(aep)
pro<-noinap/los
plot(ward,pro)
rm(pro)
detach(aep)

---

**aids**  
*Aids Cases in England and Wales*

Description

The quarterly reported AIDS cases in the U.K. from January 1983 to March 1994 obtained from the Public Health Laboratory Service, Communicable Disease Surveillance Centre, London.

Usage

data(aids)

Format

A data frame with 45 observations on the following 3 variables.

- **y** the number of quarterly aids cases in England and Wales: a numeric vector
- **x** time in months from January 1983, 1:45: a numeric vector
- **qrt** the quarterly seasonal effect a factor with 4 levels, [1=Q1 (Jan-March), 2=Q2 (Apr-June), 3=Q3 (July-Sept), 4=Q4 (Oct-Dec)]
Details

The counts \( y \) can be modelled using a (smooth) Poisson regression model in time \( x \) with the quarterly effects i.e. \( \text{cs}(x, df=7) + \text{qrt} \). Overdispersion persists, so use a Negative Binomial distribution of type I or II. The data also can be used to find a break point in time, see Rigby and Stasinopoulos (1992).

Source

Public Health Laboratory Service, Communicable Disease Surveillance Centre, London.

References


Examples

data(aids)
attach(aids)
plot(x,y,pch=21,bg=c("red","green3","blue","yellow")[unclass(qrt)])
detach(aids)

---

### Description

These data, reported by Proschan (1963, *Technometrics* 5, 375-383), refer to the intervals, in service-hours, between failures of the air-conditioning equipment in a Boeing 720 aircraft. (Proschan reports data on 10 different aircraft. The data from only one of the aircraft is used here. Cox and Snell (1981, *Applied Statistics: principles and examples*, Chapman and Hall, London) discuss the analysis of the data on all 10 aircraft.) The dataset consists of a single vector of data. They are used in the book ‘Distributions for location, scale and shape: Using GAMLSS in R’ to demonstrate the likelihood function and maximum likelihood estimation.

Usage

data(“aircond”)

Format

A data frame with 24 observations on the following variable.

<table>
<thead>
<tr>
<th>aircond</th>
<th>Air-conditioning data</th>
</tr>
</thead>
</table>

Source

The data were taken from the R package rpanel where they refer to as aircon.
alveolar

References


Examples

data(alveolar)

-----------------------------------------------------------------------------------------------------------------
alveolar The Alveolar Data files for GAMLSS

Description

alveolar : alveolar-bronchiolar adenomas data used by Tamura and Young (1987) and also reproduce in Hand et al. (1994), data set 256. The data are the number of mice out of certain number of mice (the binomial denominator) in 23 independent groups, having alveolar-bronchiolar adenomas.

Usage

data(alveolar)

Format

Data frames each with the following variable.

r  a numeric vector showing the number of mice out of n number of mice (the binomial denominator below) in 23 independent groups, having alveolar-bronchiolar adenomas.

n  a numeric vector showing the total number of mice

Details

Data sets usefull for the GAMLSS booklet

References


Examples

data(alveolar)
with(alveolar, hist(r/n))
The brown fat data set

Description

Brown fat (or brown adipose tissue) is found in hibernating mammals, its function being to increase tolerance to the cold. It is also present in newborn humans. In adult humans it is more rare and is known to vary considerably with ambient temperature. RouthierLabadie2011 analysed data on 4,842 subjects over the period 2007-2008, of whom 328 (6.8%) had brown fat. Brown fat mass and other demographic and clinical variables were recorded. The purpose of the study was to investigate the factors associated with brown fat occurrence and mass in humans.

Usage

data("brownfat")

Format

A data frame with 4842 observations on the following 14 variables.

- **sex** 1=female, 2=male
- **diabetes** 0=no, 1=yes
- **age** age in years
- **day** day of observation (1=1 January, ..., 365=31 December)
- **exttemp** external temperature (degrees Centigrade)
- **season** Spring=1, Summer=2, Autumn=3, Winter=4
- **weight** weight in kg
- **height** height in cm
- **BMI** body mass index
- **glycemy** glycemia (mmol/L)
- **LBW** lean body weight
- **cancerstatus** 0=no, 1=yes, 99=missing
- **brownfat** presence of brown fat (0=no, 1=yes)
- **bfmass** brown fat mass (g) (zero if brownfat=0)

Source

References


Examples

data(brownfat)

bush2000  The Bush 2000 election data

Description


Usage

data("bush2000")

Format

A data frame with 51 observations on the following 10 variables.

- state: name of state a factor with levels 51 levels.
- bush: proportion of state’s vote for George Bush
- male: percentage of population male
- pop: population
- rural: percentage of population living in rural areas
- bpoval: percentage of population with income below the poverty level
- c1fu: unemployment rate (%)  
- mgt18: percentage of male population older than 18 years
- pgt65: percentage of population older than 65 years
- numgt75: percentage of population with income greater than 75K

Details

The US election data, at the state level, in the 2000 Presidential Election. The response variable is the proportion of the state that voted for George Bush; and the predictors are state demographic indicators.
Source
Kieschnick and McCullough (2003)

References

Examples
data(bush2000)
plot(bush~bpo1, data=bush2000)

---

cable  The cable data set

description
The penetration of cable television in 283 market areas in the USA.

Usage
data("cable")

Format
A data frame with 283 observations on the following 6 variables.

pen5  proportion of households having cable TV in market area
ln1n  log median income
child  percentage of households with children
ltv   number of local TV stations
dis   consumer satisfaction index with values 0 and 1
agehe age of cable TV headend

Details
The cable data set concerns the penetration of cable television in 283 market areas in the USA. The data were collected in a mailed survey questionnaire in 1992 Kieschnick and McCullough (2003). The aim of the study was to explain cable television uptake (the proportion pen5) as a function of area demographics.

Source
Kieschnick and McCullough (2003)
References

Kieschnick, R. and McCullough, B. D. (2003) Regression analysis of variates observed on (0, 1): 
percentages, proportions and fractions, Statistical Modelling, 3, Vol 3, pp 193-213, Sage Publica-

Examples

data(cable)

CD4

The CD4 Count Data files for GAMLSS

Description

CD4: The data were given by Wade and Ades (1994) and refer to cd4 counts from uninfected 
children born to HIV-1 mothers and the age of the child.

Usage

data(CD4)

Format

Data frames each with the following variable.

    cd4  a numeric vector showing the CD4 counts
    age  the age of the child

Details

Data sets useful for the GAMLSS booklet

References

and confidence intervals for centiles. Statistics in Medicine, 13, pages 2359-2367.

Examples

data(CD4)
with(CD4,plot(cd4~age))
The Computer Failure Data files for GAMLSS

Description

computing: The data relate to DEC-20 computers which operated at the Open University in the 1980. They give the number of computers that broke down in each of the 128 consecutive weeks of operation, starting in late 1983, see Hand et al. (1994) page 109 data set 141.

Usage

data(computer)

Format

Data frames each with the following variable.

failure a numeric vector showing the number of times computers failed

Details

Data sets usefull for the GAMLSS booklet

References


Examples

data(computer)
with(computer, plot(table(failure)))

cysts Data for count data

Description

The cysts data set is a univariate sample of 110 counts of kidney cysts in mice fetuses, Para and Jan (2016).

Usage

data("cysts")
Format

The cysts data frame has 12 observations on the following 2 variables.

y the counts
f the frequency

Source

For systs Para and Jan (2016)

References


Examples

data(cysts)
barplot(cysts$f, names.arg=cysts$y)

---

*db*  
*Head Circumference of Dutch Boys*

Description

The data are coming from the Fourth Dutch Growth Study, Fredriks et al. (2000a, 2000b), which is a cross-sectional study that measures growth and development of the Dutch population between the ages 0 and 21 years. The study measured, among other variables, height, weight, head circumference and age for 7482 males and 7018 females. Here we have the only the head circumference of Dutch boys.

Usage

data(db)

Format

A data frame with 7040 observations on the following 2 variables.

head head circumference
age age in years

Source

The data were kindly given by professor Stef. van Buuren.
References


Examples

data(db)
attach(db)
plot(age,head)
detach(db)

---

<table>
<thead>
<tr>
<th>dbbmi</th>
<th>BMI of Dutch Boys</th>
</tr>
</thead>
</table>

Description

The data are comming from the Fourth Dutch Growth Study, Fredriks et al. (2000a, 2000b), which is a cross-sectional study that measures growth and development of the Dutch population between the ages 0 and 21 years. The study measured, among other variables, height, weight, head circumference and age for 7482 males and 7018 females. Here we have the only the BMI of Dutch boys.

Usage

data(dbbmi)

Format

A data frame with 7294 observations on the following 2 variables.

- age a numeric vector
- bmi a numeric vector

Source

The data were kindly given by professor Stef. van Buuren.
References

Examples
```r
data(dbbmi)
plot(bmi~age, data=dbbmi)
```

---

**dbhh**

*Head circumference and height of Dutch Boys*

Description
The data are coming from the Fourth Dutch Growth Study, Fredriks et al. (2000a, 2000b), which is a cross-sectional study that measures growth and development of the Dutch population between the ages 0 and 21 years. The study measured, among other variables, height, weight, head circumference and age for 7482 males and 7018 females. Here we have the only the head circumference and height of Dutch boys.

Usage
```r
data("dbhh")
```

Format
A data frame with 6885 observations on the following 3 variables.

- head  head circumference
- age  age in years
- ht  height

Source
The data were kindly given by professor Stef. van Buuren.
References


Examples

data(dbhh)
plot(dbhh$age, dbhh$head)
plot(dbhh$age, dbhh$ht)

---

eu15 GDP of 15 EU countries from 1960 to 2009

Description

The purpose of this data is to estimate the importance of labor, capital and useful energy in explaining economic growth (quantified by the GDP) of the EU 15 from 1960 to 2009. The response variable is the GDP while the independent variables are the labor, capital and useful energy. The EU 15 includes Austria, Belgium, Benmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, Sweden and UK. The data was analysed by Voudouris et al.[2015].

Usage

data("eu15")

Format

A data frame with 50 observations on the following 5 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td>the year from 1960 to 2009</td>
</tr>
<tr>
<td>UsefullEnergy</td>
<td>the total amount of useful energy (energy that performs some short of work) for the EU 15 countries</td>
</tr>
<tr>
<td>GDP</td>
<td>the sum of the GDP of the EU 15 countries</td>
</tr>
<tr>
<td>Labor</td>
<td>the sum of total hours worked of the EU 15 countries.</td>
</tr>
<tr>
<td>Capital</td>
<td>the sum of the net capital stock of the EU 15 countries.</td>
</tr>
</tbody>
</table>

Source

The Fabric Data

Examples

```r
data(fabric)

fabric
```

Description

The data are 32 observations on faults in rolls of fabric

Usage

```r
data(fabric)
```

Format

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

- **leng** the length of the roll: a numeric vector
- **y** the number of faults in the roll of fabric: a discrete vector
- **x** the log of the length of the roll: a numeric vector

Details

The data are 32 observations on faults in rolls of fabric taken from Hinde (1982) who used the EM algorithm to fit a Poisson-normal model. The response variable is the number of faults in the roll of fabric and the explanatory variable is the log of the length of the roll.

Source

John Hinde

References


Examples

```r
data(fabric)
attach(fabric)
plot(x, y)
detach(fabric)
```
**film30**  
*Film revenue data for the 1930’s*

---

**Description**

Data from film revenues from the 1930s’.

**Usage**

```r
data(film30)
```

**Format**

A data frame with 969 observations on the following 3 variables.

- **film** a factor with the name of the film
- **total** a numeric vector
- **opening** a numeric vector

**Source**

The data were collected by Prof. John Sedgwick

**References**


**Examples**

```r
data(film30)
## maybe str(film30); plot(film30) ...
```
Description
Data from film revenues from the 1990s’.

Usage
data(film90)

Format
A data frame with 4031 observations on the following 14 variables.

lnosc  the log of the number of screens
lboopen the log of box office opening revenues
lborev1 the log of box office revenues after the first week
dist  a factor indicating whether Independent or Major distributor

Details
Those data are data analysed in Voudouris et. al. (2011) suitably anonymised.

Source
Data collected by Prof. John Sedgwick

References


Examples
data(film90)
Description

glass: show the strength of glass fibres, measured at the National Physical Laboratory, England, see Smith and Naylor (1987), (the unit of measurement were not given in the paper).

Usage

data(glass)

Format

Data frames each with the following variable.

strength a numeric vector showing the strength of glass fibres

Details

Data sets usefull for the GAMLSS booklet

References


Examples

data(glass)
with(glass, hist(strength))

Description

The Blue Mountains Eye Study.

Usage

data("glasses")
**grip**

**Format**

A data frame with 1016 observations on the following 3 variables.

- **age** The age of the participants in the Blue Mountains Eye Study
- **sex** the gender of the participants, a factor with levels 1='male' 2='female'.
- **ageread** the age in which reading glasses were required.

**References**


**Examples**

```r
data(glasses)
plot(ageread~sex, data=glasses)
```

---

**grip**

*The hand grip strength data*

**Description**

The data is a subset (only boys) from the data analysed by Cohen et al. (2010).

**Usage**

```r
data("grip")
```

**Format**

A data frame with 3766 observations on the following 2 variables.

- **age** the age of the participant
- **grip** the handgrip strength

**Details**

Cohen et al. (2010) analysed the of hand grip (HG) strength in relation to gender and age in English schoolchildren. Here there are 3766 observations of the boys.

**References**


**Examples**

```r
data(grip)
```
**Description**

There two data sets contain data used in Hodges (1998). In addition to the data used in that manuscript, it contains other data items.

The original data consists of two matrices of dimensions of 341x6 and a 45x4 respectively.

The first matrix `hodges` describes plans. The information for each plan is: the state, a two-character code that identifies plans within state, the total premium for an individual, the total premium for a family, the total enrollment of federal employees as individuals, and the total enrollment of federal employees as families.

The second matrix, `hodges`, describes states. The information for each state is: its two-letter abbreviation, the state average expenses per admission (from American Medical Association 1991 Annual Survey of Hospitals), population (1990 Census), and the region (from the Marion Merrill Dow Managed Care Digest 1991).

The Hodges manuscript used these variables: Plan level: individual premium, individual enrollment. State level: expenses per admission, region.

**Usage**

```r
data(hodges)
```

**Format**

Two data frames the first with 341 observations on the following 6 variables.

- **state** a factor with 45 levels AL AZ CA CO CT DC DE FL GA GU HI IA ID IL IN KS KY LA MA MD ME MI MN MO NC ND NE NH NJ NM NV NY OH OK OR PA PR RI SC TN TX UT VA WA WI
- **plan** a two-character code that identifies plans within state declared here as factor with 325 levels.
- **prind** a numeric vector showing the total premium for an individual
- **prfam** a numeric vector showing the total premium for a family
- **enind** a numeric vector showing the total enrollment of federal employees as individuals
- **enfam** a numeric vector showing the total enrollment of federal employees as families.

and the second with 45 observations on the following 4 variables

- **state** a factor with levels same as state above
- **expe** a numeric vector showing the state average expenses per admission (from American Medical Association 1991 Annual Survey of Hospitals)
- **pop** a numeric vector showing the population (1990 Census)
- **region** the region (from the Marion Merrill Dow Managed Care Digest 1991), a factor with levels MA MT NC NE PA SA SC
InfMort

Source

http://www.biostat.umn.edu/~hodges/

References


Examples

data(hodges)
attach(hodges)
plot(prind~state, cex=1, cex.lab=1.5, cex.axis=1, cex.main=1.2)
str(hodges)
data(hodges1)
str(hodges1)

InfMort  

*Infant Mortality Data*

Description

The following data set is not a real data set but is created for the purpose of demonstrating a binomial type response variable. The data set is based on some real data obtained from the Parana State in Brazil in 2010.

Usage

data("InfMort")

Format

A data frame with 399 observations on the following 11 variables.

- **x**: the x-coordinate
- **y**: the y-coordinate
- **dead**: the number of dead infants
- **bornalive**: the number of infants born alive
- **ifdm**: FIRJAN index of city development
- **illit**: the illiteracy index
- **lgdp**: the logarithm of the gross national product
- **cli**: the proportion of children living in a household with half the basic salary
- **lpop**: the logarithm of the number of people living in each city
- **psf**: the proportion covered by the family health program
- **poor**: the proportion of individuals low household income per capita
Details

There is geographical information given by the x and y coordinates and also several social-economics variables.

References


Examples

data(Hinfmort)

Description

The data set, kindly provided to us by Dr Maria Durban, is based on a study conducted at Harvard University with girls affected by Acute lymphoblastic leukaemia. The obesity and short stature are common effects on teens who have or have had the disease, and the treatments applied trying to minimize this type of side effects without compromising its effectiveness. In one of the clinical trials conducted, 618 children were studied between the years 1987 and 1995 and three different treatments were applied: intracranial therapy without radiation, conventional intracranial radiation therapy and intracranial radiation therapy twice a day. Approximately every 6 months the children height was measured. For children the height increases smoothly along the years. In this example, (the data have been changed for confidentiality) 197 girls diagnosed with Acute lymphoblastic leukaemia between 2 and 9 years old are measured. The height of the children was measured at different times and in total 1988 observations were collected. The number of observations per child varies between 1 and 21.

Usage

data("Leukemia")

Format

A data frame with 1988 observations on the following 4 variables.

case  a factor with levels 1 to 197 indicating the participant
treatment a factor with levels 1 2 3
height the height of the participants
age    the age of the participants

Source

Dr Maria Durban
References

Examples
data(Leukemia)

---

The LGA Claims Data files for GAMLSS

Description
These are several small data files useful for gamlss fits.

LGAclaims: the data were given by Gillian Heller and can be found in de Jong and Heller (2007). This data set records the number of third party claims, Claims, in a twelve month period between 1984-1986 in each of 176 geographical areas (local government areas) in New South Wales, Australia. Areas are grouped into thirteen statistical divisions (SD). Other recorded variables are the number of accidents, Accidents, the number of people killed or injured and population with all variables classified according to area.

Usage
data(LGAclaims)

Format
Data frames each with the following variable.

Claims the number of third party claims
LGA Local government areas in New South Wales
SD statistical divisions
Pop\_density population density
KI the number of people killed or injured
Accidents the number of accidents
Population population size
L\_KI log of KI
L\_Accidents the log of the number of accidents
L\_Population log Population
Details

Data sets useful for the GAMLSS booklet

References


Examples

data(LGAclaims)
with(LGAclaims, plot(data.frame(Claims, Pop_density, KI, Accidents, Population)))

Description

lice: The data come from Williams (1944) (also used by Stein and Juritz (1988).) and they are lice per head of Hindu male prisoners in Cannamore, South India, 1937-1939.

Usage

data(lice)

Format

Data frames each with the following variable.

head: a numeric vector showing the number lice per head of Hindu male prisoners in Cannamore, South India, 1937-1939.

freq: a numeric vector showing the frequency of lice per head

Details

Data sets useful for the GAMLSS booklet

References


Examples

data(lice)
lungFunction

The lung function data

Description

3164 male observations of lung function data previously analysed by Stanojevic et al. 2008 and Hossain et al. 2016.

Usage

data("lungFunction")

Format

A data frame with 3164 observations on the following 3 variables.

- slf  the spirometric lung function, FEV_1 / FVC, which is an established index for diagnosing airway obstruction (males only)
- height the height in centimetres
- age the age

Details

The response variable is slf=FEV_1/FVC and the explanatory variable is height. The response variable slf is a ratio of forced expiratory volume in 1 second, FEV_1, to forced vital capacity, FVC. Spirometric lung function slf is an established index for diagnosing airway obstruction, e.g. Quanjer et al. 2010. The purpose here is to create centile curves of slf against height. More details about the analysis using GAMLSS of the FEV_1/FVC data can be found in Hossain et al. 2016.

Source

The data were kindly provided by Dr Sanja Stanojevic.

References


margolin

Examples

```r
data(lungFunction)
plot(lungFunction)
```

---

**The Margolin Data files for GAMLSS**

**Description**

margolin: Margolin et al. (1981) present data from an Ames Salmonella assay, where \( y \) is the number of revertant colonies observed on a plate given a dose \( x \) of quinoline. The data were subsequently analysed by Breslow (1984), Lawless (1987) and Saha and Paul (2005).

**Usage**

```r
data(margolin)
```

**Format**

Data frames each with the following variable.

- \( y \) a numeric vector showing the number of revertant colonies observed on a plate given a dose \( x \) of quinoline.
- \( x \) a numeric vector showing a dose \( x \) of quinoline.

**Details**

Data sets usefull for the GAMLSS booklet

**References**


**Examples**

```r
data(margolin)
with(margolin, plot(y~x))
```
A Meta Analysis on Smoking Cessation

Description
The data here are coming from a statistical meta analysis problem. In meta analysis we combine the evidence from different studies to obtain an overall treatment effect. The data from Silagy et al. (2003) consist of different clinical trials of nicotine replacement therapy for smoking cessation. In each trial the patient was randomized into a treatment or control group. The treatment group were given a nicotine gum. In the majority of studies the control group receive the same appearance gum but without the ingredients but in some they were given no gum. The outcome, whether the participant is smoking or not, was observed after six months. The data were previously analysed by Aitkin (1999) and by Skrondal and Rabe-Hesketh (2004).

Usage
```r
data("meta")
```

Format
A data frame with 54 observations on the following 6 variables.
- **studyname**: a factor the name of the place of the different studies (note that the values of `studyname` is the same for studies at the same place in different years)
- **year**: the year of the study
- **d**: the number of quitters (non-smokers) after six months
- **n**: the total number of participants in the study
- **fac**: a factor with two levels indicating whether control, 1 or treatment 2
- **study**: a factor with levels from 1 to 27 indicating the different studies (that is, the interaction of `studyname` and `year`)

References

Examples
```r
data(meta)
## maybe str(meta) ; plot(meta) ...
```
Mothers encouragement data

Description

Mothers encouragement for participation in Higher Education. The response variable is mums a three level factor which can be used in a multinomial Logistic model or mumsB a two level factor suitable for binary logistic model.

Usage

data(Mums)

Format

A data frame with 871 observations on the following 7 variables.

mums mothers encouragement: factor with levels 1 is for strong encouragement, 2 is for some encouragement and 3 for no encouragement/discouragement
class social class: a factor with levels 1 is C1, 2 is C2, 3 is D and 4 is E
age age of the participants: a factor with levels 1 is 16-18, 2 is 19-20 and 3 is 20-30
gender a factor with levels 1 is male and 2 is female
ethn ethnicity of the participants: a factor with levels 1 is white, 2 is black, 3 is asian and 4 is other
qual qualifications of the participants: a factor with levels, 1 is greater or equal to 2 A levels, 2 is HND or more than 5 GCSE’s, 3 is less than 5 GSCSE’s ar none above and 4 no formal qualification
mumsb mothers encouragement: a factor with levels, 0 is no encouragement or some encouragement 1 is for strong encouragement

Details

The data were collected as part of the Social Class and widening Participation in Higher Education Project based at the University of North London (now London Metropolitan University) and supported by the University’s Development and Diversity Fund over the period 1998-2000.

Source

Professor Robert Gilchrist director of STORM at London Metropolitan

References

Examples

```r
data(Mums)
MM <- xtabs(~mums+qual, data=Mums)
mosaicplot(MM, color=TRUE)
MM <- xtabs(~mums+ethn+gender, data=Mums)
mosaicplot(MM, color=TRUE)
```

Description

The motor vehicle insurance data are motor vehicle insurance policies. `mvi` is a sample of 2000 observations from `mviBig` which has 67143 observations.

Usage

```r
data(mvi)
data(mviBig)
```

Format

Two data frames with 2000 or 67143 observations on the following 14 variables.

- `retval` a numeric vector showing the value of the vehicle
- `whetherclm` a numeric vector showing whether a claim is made, 0 no claim, 1 at least one claim
- `numclaims` a numeric vector showing the number of claims
- `claimcst` a numeric vector showing the total amount of claim, i.e. for `numclaims=0` is zero.
- `vehmake` a factor showing the make of the car with levels `BMW DAEWOO FORD Mitsubisi` `H`
- `vehbody` a factor showing the type of the car, with levels `BUS CONT COUPE HACK HDTOP HRSE MCARA MIBUS PANVN RDSTR SEDAN STNWG TRUCK UTE`
- `vehage` a numeric vector showing the age of the car
- `gender` a factor showing the gender of the policy holder with levels `F M`
- `area` a factor showing the Area of residence of the policy holder with levels `A B C D E F`
- `agecat` a factor showing the age band of the policy holder with levels `1 2 3 4 5 6` one is youngest
- `exposure` a numeric vector showing the time of exposure with values from zero to one

Details

The motor vehicle insurance data are motor vehicle insurance policies from an insurance company over a twelve-month period in 2004-05. The original data are 67143 observation but here we also include a random sample of 2000.
References


Examples

data(mvi)
## a histogram of claims with fitted gamma distribution
## library(gamlss)
## with(mvi, histDist(claimst[whetherclm==1&claimst<15000], family=GA, main="Claims"))

o!l

The oil price data

Description

The Oil data: Using model selection to discover what affects the price of oil. The data s contains the daily prices of front month WTI (West Texas Intermediate) oil price traded by NYMEX (New York Mercantile Exchange). The front month WTI oil price is a futures contract with the shortest duration that could be purchased in the NYMEX market. The idea is to use other financially traded products (e.g., gold price) to discover what might affect the daily dynamics of the price of oil.

Usage

data("oil")

Format

A data frame with 1000 observations on the following 25 variables.

OILPRICE the log price of front month WTI oil contract traded by NYMEX - in financial terms, this is the CL1. This is the response variable.

CL2_log, CL3_log, CL4_log, CL5_log, CL6_log, CL7_log, CL8_log, CL9_log, CL10_log, CL11_log, CL12_log, CL13_log, numeric vectors which are the log prices of the 2 to 15 months ahead WTI oil contracts traded by NYMEX. For example, for the trading day of 2nd June 2016, the CL2 is the WTI oil contract for delivery in August 2016.

BDI_Y_log the Baltic Dry Index, which is an assessment of the price of moving the major raw materials by sea.

SPX_log the S&P 500 index

DX1_log the US Dollar Index.

GC1_log the log price of front month gold price contract traded by NYMEX

HO1_log the log price of front month heating oil contract traded by NYMEX
USCI_log  the United States Commodity Index
GNR_log  the S&P Global Natural Resources Index
SHCOMP_log  the Shanghai Stock Exchange Composite Index.
FTSE_log  the FTSE 100 Index
resplag  the lag 1 of OILPRICE - lagged version of the response variable.

Source
The dataset was downloaded from https://www.quandl.com/.

Examples

```r
data(oil)
plot(OILPRICE~SPX_log, data=oil)
```

parzen  The Parzen Data File for GAMLSS

Description
Parzen: Parzen (1979) and also contained in Hand et al. (1994), data set 278. The data give the annual snowfall in Buffalo, NY (inches) for the 63 years, from 1910 to 1972 inclusive.

Usage
```
data(parzen)
```

Format
Data frames each with the following variable.

`snowfall`  the annual snowfall in Buffalo, NY (inches) for the 63 years, from 1910 to 1972 inclusive, 63 observations

Details
Data sets useful for the GAMLSS booklet

References

Examples
```
data(parzen)
with(parzen, hist(snowfall))
```
plasma

The plasma data set

Description
A cross-sectional study to investigate the relationship between personal characteristics and dietary factors, and plasma concentrations.

Usage
data("plasma")

Format
A data frame with 315 observations on the following 14 variables.

age  age (years)
sex  sex, 1=male, 2=female
smokstat smoking status 1=never, 2=former, 3=current Smoker
bmi  body mass index weight/(height^2)
vituse vitamin use 1=yes, fairly often, 2=yes, not often, 3=no
calories number of calories consumed per day
fat  grams of fat consumed per day
fiber grams of fiber consumed per day
alcohol number of alcoholic drinks consumed per week
cholesterol cholesterol consumed (mg per day)
betadiet dietary beta-carotene consumed (mcg per day)
retdiet dietary retinol consumed (mcg per day)
betaplasma plasma beta-carotene (ng/ml)
retplasma plasma retinol (ng/ml)

Details
“Observational studies have suggested that low dietary intake or low plasma concentrations of retinol, beta-carotene, or other carotenoids might be associated with increased risk of developing certain types of cancer \... We designed a cross-sectional study to investigate the relationship between personal characteristics and dietary factors, and plasma concentrations of retinol, beta-carotene and other carotenoids.” Harrell (2002)

Source
Harrell (2002)
References

Examples
data(plasma)

dataset

polio Poliomyelitis cases in US

Description
Poliomyelitis cases reported to the U.S. Centers for Disease Control for the years 1970 to 1983, that is, 168 observations.

Usage
data(polio)

Format
The format is: Time-Series [1:168] from 1970 to 1984: 0 1 0 0 1 3 9 2 3 5 ...

Details
The data were originally modelled by Zeger (1988) who used a parameter driven approach, in which a first order autoregressive model was used for the latent process, to conclude that there is evidence of a decrease in the polio infection rate. The data were analysed also by Li (1994), Zeger and Qaqish (1988), Davis et al. (1999), and by Benjamin et al (2003).

Source
Zeger (1988) w

References
Examples

```r
data(polio)
plot(polio)
```

---

**Description**

A survey was conducted in April 1993 by Infratest Sozialforschung. A random sample of accommodation with new tenancy agreements or increases of rents within the last four years in Munich was selected including: i) single rooms, ii) small apartments, iii) flats, iv) two-family houses. Accommodation subject to price control rents, one family houses and special houses, such as penthouses, were excluded because they are rather different from the rest and are considered a separate market. For the purpose of this study, 1967 observations of the variables listed below were used, i.e. the rent response variable R followed by the explanatory variables found to be appropriate for a regression analysis approach by Fahrmeir et al. (1994, 1995):

**Usage**

```r
data(rent)
```

**Format**

A data frame with 1969 observations on the following 9 variables.

- **R**: rent response variable, the monthly net rent in DM, i.e. the monthly rent minus calculated or estimated utility cost
- **Fl**: floor space in square meters
- **A**: year of construction
- **Sp**: a variable indicating whether the location is above average, 1, (550 observations) or not, 0, (1419 observations)
- **Sm**: a variable indicating whether the location is below, 1, average (172 obs.) or not, 0, (1797 obs.)
- **B**: a factor with levels indicating whether there is a bathroom, 1, (1925 obs.) or not, 0, (44 obs.)
- **H**: a factor with levels indicating whether there is central heating, 1, (1580 obs.) or not, 0, (389 obs.)
- **L**: a factor with levels indicating whether the kitchen equipment is above average, 1, (161 obs.) or not, 0, (1808 obs.)
- **loc**: a factor (combination of Sp and Sm) indicating whether the location is below, 1, average, 2, or above average 3

**Details**

This set of data were used by Stasinopoulos et al. (2000) to fit a model where both the mean and the dispersion parameter of a Gamma distribution were modelled using the explanatory variables.
Rent99

Source

Provide by Prof. L. Fahrmeir

References

Fahrmeir L., Gieger C., Mathes H. and Schneeweiss H. (1994) Gutachten zur Erstellung des Miet-
Munchen, Sozialreferat-Amt fur Wohnungswesen.


mean and dispersion additive models, Statistician, 49, 479-493.

org/v23/i07.

Examples

data(rent)
attach(rent)
plot(F1,R)

----

Rent99

Munich rent data of 1999

Description

The Munich rent data and boundaries files of of 1999 survey.

Usage

data(rent99)

Format

A data frame with 3082 observations on the following 9 variables.

rent the monthly net rent per month (in Euro).
rentsqm the net rent per month per square meter (in Euro).
area Living area in square meters.
yearc year of construction.
location quality of location: a factor indicating whether the location is average location, 1, good
location, 2, and top location, 3.
bath quality of bathroom: a a factor indicating whether the bath facilities are standard, 0, or pre-
mium, 1.
kitchen  Quality of kitchen: 0 standard 1 premium.
cheating central heating: a factor 0 without central heating, 1 with central heating.
district  District in Munich.

Details
See Fahrmeir et. al., (2013) page 5, for more details about the data.

Source
Thanks to Thomas Kneib who provide us with the data.

References

Examples
data(rent99)
plot(rent-area, data=rent99)

rent99.polys  The boundaries file for Munich rent data from the 1999 survey.

Description
The boundaries files of 1999 Munich survey.

Usage
data(rent99.polys)

Format
This data frame contains the boundaries of the Munich data.

Details
See Fahrmeir et. al., (2013) page 5, for more details about the data.

Source
Thanks to Thomas Kneib who provide us with the data.

References
respInf

Examples

```r
data(rent99.polys)
## library(gamlss.spatial); draw.polys(rent99.polys)
```

---

respInf  
Respiratory Infection in Indonesian Children.

Description

This is cohort study of 275 Indonesian preschool children, ($J=1,2,\ldots,275$), examined on up to six, consecutive quarters for the presence of respiratory infection. Sommer et al. (1983) describe the study, while Zeger and Karim (1991) and Diggle et al (2002) among others analyzed it. The data were also analyzed by Skrondal and Rabe-Hesketh (2004).

Usage

```r
data("respInf")
```

Format

A data frame with 1200 observations on the following 14 variables.

- **id**: a factor with 275 levels identifying the individual children
- **time**: the binary response variable identifying the presence of respiratory infection
- **resp**: a vector of ones (not used further)
- **age**: the age in months (centered around 36)
- **xero**: a factor variable for the present of xerophthalmia with levels 0 1
- **cosine**: a cosine term of the annual cycle
- **sine**: a sin term of the annual cycle
- **female**: a gender factor with levels 0 is male 1 is female
- **height**: height for age as percent of the National Center for health Statistics standard centered at 90%
- **stunted**: a factor whether below 85% in height for age 0 1
- **time1**: the time that the children has been examine, 1 to 6
- **age1**: the age of the child at the fist time of examination
- **season**: a variable taking the values 1,2,3,4 indicating the season
- **time2**: the time in months
References


Examples

```r
data(respInf)
## maybe str(respInf); plot(respInf) ...
```

---

**sleep**  
*Data on sleep*

---

**Description**

Data from a study conducted on 133 patients thought to have the condition Obstructive Sleep Apnea (OSA). These patients have undergone a sleep study at a Canadian sleep clinic Ahmadi at al. (2008). While the focus on the study was the relationship between the Berlin Questionnaire for sleep apnea to polysomnographic measurements of respiratory disturbance, in particular the arousal index, we will analyse the proportion of sleep time that is REM sleep (REM). This variable is in the interval [0,1), so necessitates the use of zero-inflated models. We have removed patients with missing values, giving n=106 observations.

**Usage**

```r
data("sleep")
```

**Format**

A data frame with 106 observations on the following 9 variables.

- **age** age in years
- **gender** 1=female, 0=male
- **bmi** body mass index
- **necksize** neck circumference (cm)
- **sbp** systolic blood pressure (mmHg)
- **alcohol** alcohol usage (1=yes, 0=no)
- **caffeine** caffeine usage (1=yes, 0=no)
- **REM** proportion of rapid eye movement (REM) sleep time
- **AI** arousal index (number of arousals from sleep per hour of sleep)
Source

see references

References


Examples

data(sleep)

```
  species       The Fish Species Data files for GAMLSS
```

Description

species: The number of different fish species ($y=$fish) was recorded for 70 lakes of the world together with explanatory variable $x=$log(lake) area. The data are given and analyzed by Stein and Juritz (1988).

Usage

data(species)

Format

Data frames each with the following variable.

fish   a numeric vector showing the number of different species in 70 lakes in the world
lake   a numeric vector showing the lake area

Details

Data sets useful for the GAMLSS booklet

References

Examples

```r
data(species)
with(species, plot(fish~log(lake)))
```

stylo

The Stylometric Data files for GAMLSS

Description

stylo: the data were given by Dr Mario Corina-Borja, see Chappas and Corina-Borja (2006), and has the number of a word appearing in a text.

Usage

```r
data(stylo)
```

Format

Data frames each with the following variable.

- **word**: a numeric vector showing the number a word appearing in a text
- **freq**: a numeric vector showing the frequency of the number a word appearing in a text

Details

Data sets usefull for the GAMLSS booklet

References


Examples

```r
data(stylo)
plot(freq~word, type="h", data=stylo)
```
tensile

The Tensile Data files for GAMLSS

Description

tensile: These data come from Quesenberry and Hales (1980) and were also reproduced in Hand et al. (1994), data set 180, page 140. They contain measurements of tensile strength of polyester fibres and the authors were trying to check if they were consistent with the lognormal distribution. According to Hand et al. (1994) "these data follow from a preliminary transformation. If the lognormal hypothesis is correct, these data should have been uniformly distributed".

Usage

data(tensile)

Format

Data frames each with the following variable.

str a numeric vector showing the tensile strength

Details

Data sets usefull for the GAMLSS booklet

References


Examples

data(tensile)
with(tensile,hist(str))
The tidal data set

Description

The dataset tidal, McArdle and Anderson (2004), gives counts of the organism "intertidal bivalve A. Stutchburyi" in three tidal areas in the Bay of Plenty, New Zealand.

Usage

data("tidal")

Format

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

number  count of A. Stutchburyi organisms
vertlt  vertical tidal height (m)
ht      tidal area, a factor with three level

Details

The dataset gives counts of the organism "intertidal bivalve A. Stutchburyi" in three tidal areas in the Bay of Plenty, New Zealand. Each observation is the count of the number of these organisms in a 0.25 m quadrat, as well as the vertical tidal height of the quadrat. The vertical heights have been classified into three tidal areas: upper (vertical height > 0.66 m), middle (0.33 - 0.66 m) and lower (<0.33 m). Ecologists are interested in the effect of tidal height (either raw or classified) on the number of organisms.

Source

McArdle and Anderson (2004)

References


Examples

str(tidal)
plot(number~vertlt, data=tidal)
plot(number~ht, data=tidal)
Description

The Tokyo rainfall data from Kitagawa (1987), analysed also by Rue and Held (2005) and Fahrmeir and Tutz (2013).

Usage

data("trd")

Format

The format is: num [1:366] 0 0 1 0 1 1 0 0 0 ...

Details

The data taken from Kitagawa (1987) contain observations from two years 1983-1984. They record whether there is more than 1 mm rainfall in Tokyo. The data consists of 366 observations of one (response) variable, \( y \), which takes values 0, 1, 2 on whether there was rain at the specific day of the year (during the two year period). The observation number 60 corresponds to the 29th of February therefore only on day is observed during the two years. The data can be analysed using a binomial distribution with a binomial denominator equal to 2 (apart from the 29th of February which has 1). The data were analysed by Rue and Held (2005) and Fahrmeir and Tutz (2013).

Source


References


Examples

data(trd)
plot(trd)
tse

*The Turkish stock exchange index*

**Description**

The Turkish stock exchange index, was recorded daily from 1/1/1988 to 31/12/1998. The daily returns, \( \text{ret} = \log(I_{i+1}/I_i) \), were obtained for \( i = 1, 2, \ldots, 2868 \).

**Usage**

```r
data(tse)
```

**Format**

A data frame with 2868 observations on the following 4 variables.

- **year** the year
- **month** the month
- **day** the day
- **ret** day returns \( \text{ret}[t] = \log(c_{currency[t]}) - \log(c_{currency[t-1]}) \)
- **currency** the currency exchange rate
- **tl** day return \( \text{ret}[t] = \log(10(c_{currency[t]}) - \log(10(c_{currency[t-1]}) \)

**References**


**Examples**

```r
data(tse)
plot(ts(tse$ret))
```

usair

*US air pollution data set*

**Description**


**Usage**

```r
data(usair)
```
Format

A data frame with 41 observations on the following 7 variables.

- **y**: a numeric vector: sulphur dioxide concentration in air mgs. per cubic metre in 41 cities in the USA
- **x1**: a numeric vector: average annual temperature in degrees F
- **x2**: a numeric vector: number of manufacturers employing >20 workers
- **x3**: a numeric vector: population size in thousands
- **x4**: a numeric vector: average annual wind speed in miles per hour
- **x5**: a numeric vector: average annual rainfall in inches
- **x6**: a numeric vector: average number of days rainfall per year

Source

Hand et al. (1994) data set 26, USAIR.DAT, originally from Sokal and Rohlf (1981)

References


Examples

data(usair)
str(usair)
plot(usair)

# a possible gamlss model
# gamlsslibrary)
#ap<-gamlss(y~cs(x1,2)+x2+x3+cs(x4,2)*x5+cs(x6,3)*x4:x5,
#    data=usair, family=GA(mu.link="inverse"))
#

---

**Visual analog scale (VAS) data**

Description

In the original data 368 patients, measured at 18 times after treatment with one of 7 drug treatments (including placebo), plus a baseline measure (time=0) and one or more pre-baseline measures (time=-1). Here for illustration we will ignore the repeated measure nature of the data and we shall use data from time 5 only (364 observations). The VAS scale response variable, Y, is assumed to be distributed as BEINF(mu, sigma, nu, tau) where any of the distributional parameters mu, sigma, nu and tau are modelled as a constant or as a function of the treatment.

Usage

data(vas5)
**Format**

A data frame with 364 observations on the following 3 variables.

- **patient**: a factor indicating the patient
- **treat**: the treatment factor with levels 1 2 3 4 5 6 7
- **vas**: the response variable

**Details**

The Visual analog scale is used to measure pain and quality of life. For example patients are required to indicate in a scale from 0 to 100 the amount of discomfort they have. This can be easily translated to a value from 0 to 1 and consequently analyzed using the beta distribution. Unfortunately if 0’s or 100’s are recorded the beta distribution is not appropriate since the values 0 and 1 are not allowed in the definition of the beta distribution. Note that the inflated beta distribution allows values at 0 and 1. This is a mixed distribution (continuous and discrete) having four parameters, nu for modelling the probability at zero p(Y=0) relative to p(0<Y<1), tau for modelling the probability at one p(Y=1) relative to p(0<Y<1), and mu and sigma for modelling the between values, 0<Y<1, using a beta distributed variable BE(mu, sigma) with mean mu and variance sigma*mu*(1-mu).

**Source**

The data were provided by Dr. Peter Lane

**Examples**

```r
data(vas5)
```

---

**VictimsOfCrime**  
*Reported victims of crime data*

**Description**

The data shows whether victims of crime were reported in the local media.

**Usage**

```r
data(VictimsOfCrime)
```

**Format**

A data frame with 10590 observations on the following 2 variables.

- **reported**: Whether the crime was reported in local media.
- **age**: the age of the victim

**Details**

Whether the crime was reported in local media.
Source
The data were given by Prof Brian Francis of Lancaster University. They can be used to demonstrate the usefulness of smoothing techniques with a binary response variable.

References

Examples
data(VictimsOfCrime)
Index

*Topic datasets
   abdom, 2
   acidity, 3
   aep, 4
   aids, 5
   aircond, 6
   alveolar, 7
   brownfat, 8
   bush2000, 9
   cable, 10
   CD4, 11
   computer, 12
   cysts, 12
   db, 13
   dbbmi, 14
   dbhh, 15
   eu15, 16
   fabric, 17
   film30, 18
   film90, 19
   glass, 20
   glasses, 20
   grip, 21
   Hodges, 22
   InfMort, 23
   Leukemia, 24
   LGaclaims, 25
   lice, 26
   lungFunction, 27
   margolin, 28
   meta, 29
   Mums, 30
   mvi, 31
   oil, 32
   parzen, 33
   plasma, 34
   polio, 35
   rent, 36
   rent99, 37
   rent99.polys, 38
   respInf, 39
   sleep, 40
   species, 41
   stylo, 42
   tensile, 43
   tidal, 44
   trd, 45
   tse, 46
   usair, 46
   vas5, 47
   VictimsOfCrime, 48

abdom, 2
acidity, 3
aep, 4
aids, 5
aircond, 6
alveolar, 7
brownfat, 8
bush2000, 9
cable, 10
CD4, 11
computer, 12
cysts, 12
db, 13
dbbmi, 14
dbhh, 15
eu15, 16
fabric, 17
film30, 18
film90, 19
glass, 20
glasses, 20
grip, 21
Hodges, 22
InfMort, 23
Leukemia, 24
LGaclaims, 25
lice, 26
lungFunction, 27
margolin, 28
meta, 29
Mums, 30
mvi, 31
oil, 32
parzen, 33
plasma, 34
polio, 35
rent, 36
rent99, 37
rent99.polys, 38
respInf, 39
sleep, 40
species, 41
stylo, 42
tensile, 43
tidal, 44
trd, 45
tse, 46
usair, 46
vas5, 47
VictimsOfCrime, 48
hodges, 22
hodges1 (hodges), 22

InfMort, 23

Leukemia, 24
LGAclaims, 25
lice, 26
lungFunction, 27

margolin, 28
meta, 29
Mums, 30
mvi, 31
mviBig (mvi), 31

oil, 32

parzen, 33
plasma, 34
polio, 35

rent, 36
rent99, 37
rent99.polys, 38
respInf, 39

sleep, 40
species, 41
stylo, 42

tensile, 43
tidal, 44
trd, 45
tse, 46

usair, 46

vas5, 47
VictimsOfCrime, 48