

# Package ‘SMIR’

January 2, 2012

**Type** Package

**Title** Companion to Statistical Modelling in R

**Version** 0.02

**Date** 2009-5-05

**Author** Murray Aitkin, Brian Francis, John Hinde, Ross Darnell <ross.darnell@csiro.au>

**Maintainer** Ross Darnell <ross.darnell@csiro.au>

**Depends** R (>= 2.6.0),

**Suggests** lattice, foreign, gdata, car, dglm, gnm, MASS, npmlreg,survival

**LazyLoad** no

**LazyData** no

**Description** This package accompanies Aitkin et al, Statistical Modelling in R, OUP, 2009. The package contains some functions and datasets used in the text.

**License** GPL (>= 2)

**Repository** CRAN

**Date/Publication** 2009-05-11 13:16:53

## R topics documented:

betablok . . . . .	2
bronchitis . . . . .	3
byssinosis . . . . .	4
cars . . . . .	4
chd . . . . .	5
claims . . . . .	6
coxph.disparity . . . . .	7
disparity . . . . .	8
disparity.glm . . . . .	9

disparity.lm . . . . .	9
faults . . . . .	10
feigl . . . . .	10
gehan . . . . .	11
ghq . . . . .	12
hostility . . . . .	12
insult . . . . .	13
lsat . . . . .	14
miners . . . . .	15
NPL.bands . . . . .	16
poison . . . . .	17
prentice . . . . .	18
R2 . . . . .	18
R2CV . . . . .	19
Rsq . . . . .	20
solv . . . . .	20
stackloss . . . . .	21
stan . . . . .	22
statlab . . . . .	23
summary.treg . . . . .	24
toxaemia . . . . .	25
toxoplas . . . . .	26
trees . . . . .	26
treg . . . . .	27
trypanos . . . . .	28
vaso . . . . .	29
vietnam . . . . .	30
woolson . . . . .	31
<b>Index</b>	<b>32</b>

---

betablok	<i>Beta-blockers for myocardial infarction</i>
----------	--

---

## Description

A 22-centre clinical trial of beta-blockers for reducing mortality after myocardial infarction, described by Yusuf *et. al.* (1984). The important issue is the generalizability of the treatment effect across different patient populations.

## Usage

data(betablok)

**Format**

A dataframe with 44 obs. and 4 variables:

```
[,1] r      integer
[,2] n      integer
[,3] centre int
[,4] treat  factor w /2 levels "C","T"
```

**Note**

See p.524 in SMIR

**Source**

Yusuf, S., Peto, R., Lewis, J., Collins, R. and Sleight, P. (1984). "Beta blockade during and after myocardial infarction: an overview of the randomized trials", *Progress in Cardiovascular Diseases*, **27**, 335–371.

---

bronchitis

*Chronic bronchitis in a sample of mean in Cardiff*

---

**Description**

The data consist of observations on three variables for each of 212 men in a sample of Cardiff enumeration districts.

**Usage**

```
data(bronchit)
```

**Format**

A data.frame of 212 obs of 3 variables:

```
[,1] r      integer, 1= respondent suffered from chronic bronchitis
[,2] cig    numeric, the number of cigarettes per day
[,3] poll   numeric, the smoke level in the locality
```

**Note**

See p.224 in SMIR

**Source**

Jones, K. (1975), *A geographical contribution to the aetiology of chronic bronchitis*, Unpublished BSc dissertation, University of Southampton. Published in Wrigley, N. (1976). *Introduction to*

*the use of logit models in geography*, Geo.ABSTRACTS Ltd, CATMOG 10, University of East Anglia, Norwich.

---

byssinosis                      *Byssinosis in the cotton industry*

---

### Description

The dataset contains the number of workers in a survey of the US cotton industry suffering and not suffering from the lung disease byssinosis, together with the values of five cross-classifying categorical explanatory variables: the race, sex and smoking habit of the worker, the length of employment in three categories, and the dustiness of the workplace in three categories.

### Usage

data(byssinosis)

### Format

A data.frame of 72 obs. of 7 variables:

[,1]	dust	Factor w/ 3 levels "most", "less"
[,2]	race	Factor w/ 2 levels "white", "non-white"
[,3]	sex	Factor w/ 2 levels "male", "female"
[,4]	smok	Factor w/ 2 levels "smoker", "non"
[,5]	emp	Factor w/ 3 levels "<10", "10-20", ...
[,6]	yes	int, Number of workers who suffered byssinosis
[,7]	no	int, Number of workers who did not suffer from byssinosis

### Note

See p.255 in SMIR

### Source

Higgins, J.E. and Koch, G.G. (1977), "Variable selection and generalized chi-square analysis of categorical data to a large cross-sectional occupational health survey", *Int. Statist. Rev.*, **45**, 51–62.

---

cars                                      *Performance data for cars from Motor Trend magazine*

---

### Description

The data are quarter-mile acceleration time in seconds and fuel consumption in miles per (US) gallon for 32 cars tested by the US *Motor Trend* magazine in 1974. Nine explanatory variables are

given: shape of engine, number of cylinders, transmission type, number of gears, engine displacement in cubic inches, horsepower, number of carburettor barrels, final drive ratio, and weight of the car in thousands of pounds.

### Usage

```
data(cars)
```

### Format

A data.frame of 32 obs. of 11 variables:

[,1]	s	integer, shape of engine (straight = 1, vee = 0)
[,2]	c	integer, number of cylinders
[,3]	t	integer, transmission type, (automatic = 0, manual = 1)
[,4]	g	integer, number of gears
[,5]	disp	numeric, engine displacement in cubic inches
[,6]	hp	integer, horsepower
[,7]	cb	integer, number of carburettors
[,8]	drat	numeric, final drive ratio
[,9]	wt	numeric, weight of car in thousands of pounds
[,10]	qmt	numeric, quarter-mile acceleration time in seconds
[,11]	mpg	numeric, fuel consumption in miles per gallon
[,12]	model	Factor w/ 32 levels "AMC Javelin" ...:18 19 ...

### Note

See p.144 in SMIR

### Source

Henderson, H.V. and Velleman, P.F. (1981), "Building multiple regression models interactively", *Biometrics*, **37**, 391–411.

---

chd

*Coronary heart disease*

---

### Description

The file gives the number of men diagnosed as having coronary heart disease (CHD) in an American study of 1329 men (the data are presented and analysed in Ku and Kullback, 1974). The serum cholesterol level and blood pressure in mm mercury were recorded for each man, and are reported in one of four categories, giving a 4X4 cross-classified in each cell of which the number of men with CHD and the total number of men examined are given.

### Usage

```
data(chd)
```

**Format**

A data.frame of 16 obs. of 4 variables:

[,1]	chol	Factor w/ 4 levels "<200","200-219",...
[,2]	bp	Factor w/ 4 levels "<127","127-146",...
[,3]	r	integer, number of men with CHD
[,4]	t	integer, total number of men in study

**Note**

See p.248 in SMIR

**Source**

Ku, H.H. and Kullback, S. (1974), "Loglinear models in contingency table analysis", *The American Statistician*, **28**, 115–22.

---

claims

*Insurance claims data*

---

**Description**

The file gives the number of policyholders of an insurance company who were “exposed to risk”, and the number of car insurance claims made in the third quarter of 1973 by these policyholders, arranged as a contingency table cross-classified by three four-level factors: `dist`, the district in which the policyholder lived (1: rural, 2: small towns, 3: large towns, 4: major cities), `car`, the engine capacity of the car (1:  $\leq$  1 litre, 2: 1 – 1.5 litres, 3: 1.5 – 2 litres, 4:  $\geq$  2 litres), and `age`, the age of the policyholder (1:  $\leq$  25, 2: 25 – 29, 3: 30 – 35, 4:  $\geq$  35)

**Usage**

`data(claims)`

**Format**

[,1]	n	integer, number of policy holders
[,2]	c	integer, number of claims
[,3]	age	Factor w/ 4 levels "<25","25-29","30-35", ">35"
[,4]	dist	Factor w/ 4 levels "rural", "small towns", "large towns", "major cities"
[,5]	car	Factor w/ 4 levels "<1", "1-1.5", "1.5-2", ">2"

**Note**

See p.271 in SMIR

**Source**

Baxter, L.A., Coutts, S.M. and Ross, G.A.F., 1980, "Application of linear models in motor insurance", *Proceedings of the 21st International Congress of Actuaries*, Zurich, 11–29.

---

coxph.disparity

*Check disparity in a Cox Proportional Hazard Model*

---

**Description**

The `coxph.disparity()` function returns the disparity from the piecewise exponential model, including all the terms in the likelihood, and is directly comparable to the disparity for the fit of other models used in this chapter.

**Usage**

```
coxph.disparity(fit)
```

**Arguments**

`fit` name of an object of class “coxph”

**Details**

This form of the likelihood, allows the Cox proportional hazards model to be compared directly to fully parametric models. (Note that log-likelihood value stored in `coxph.object` is not comparable as it is based on the proportional hazards function and does not include the baseline hazard, this cancels out in the conditional probabilities that form the partial likelihood.)

**Value**

a num vector

**Author(s)**

<john.hinde@nuigalway.ie>

**References**

Aitkin, M., Francis, B., Hinde, J. and Darnell, R. (2009). *Statistical modelling in R*, OUP.

**Examples**

```
require(survival)
data(feigl)
feigl <- within(feigl, {lwbc <- log(wbc)})
feigl.cph <- coxph(Surv(time) ~ ag * lwbc, data = feigl,
                  method = "breslow")
coxph.disparity(feigl.cph)
```

---

disparity

*Model Disparities*

---

**Description**

disparity is a generic function used to produce the disparities of the results of various models.

**Usage**

```
disparity(model)
```

**Arguments**

model            a valid model lm or glm object

**Author(s)**

<ross.darnell@csiro.au>

**References**

Aitkin, M., Francis, B., Hinde, J. and Darnell, R. (2009). *Statistical Modelling in R*, OUP.

**Examples**

```
## The function is currently defined as
function(model, ...)
  UseMethod("disparity")
```

---

disparity.glm	<i>Disparities for Generalized Linear Model Fits</i>
---------------	--

---

**Description**

This function is a methods for class glm objects.

**Usage**

```
## S3 method for class 'glm'  
disparity(model)
```

**Arguments**

model            an object of class "glm".

**Details**

disparity prints  $-2 \times$  log-likelihood.

**Author(s)**

<ross.darnell@csiro.au>

---

disparity.lm	<i>Disparities for Linear Model Fits</i>
--------------	--

---

**Description**

This function is a method for class lm objects.

**Usage**

```
## S3 method for class 'lm':  
## S3 method for class 'lm'  
disparity(model)
```

**Arguments**

model            an object of class "lm".

**Details**

disparity prints  $-2 \times$  log-likelihood.

**Author(s)**

<ross.darnell@csiro.au>

---

 faults

*Faults in rolls of material*


---

**Description**

Bissell gives the numbers of yarn breaks observed in a roll of fabric whilst a textile process was running, as well as the length of the roll of fabric.

**Usage**

```
data(faults)
```

**Format**

A data.frame of 32 obs. of 2 variables:

```
[,1]  l  integer, roll length (m)
[,2]  n  integer, number of faults
```

**Note**

See pages 269 and 474 of SMIR

**Source**

Bissell, A. F. (1972), "A negative binomial model with varying elemental sizes", *Biometrika*, **59**, 435–441.

---

 feigl

*Leukaemia survival times — Feigl & Zelen*


---

**Description**

The file contains the survival times in weeks of 33 patients suffering from acute myelogenous leukaemia, and the values of two explanatory variables, white blood cell count in thousands and a positive or negative factor, positive values being defined by the presence of Auer rods and/or significant granulation of the leukaemic cells in the bone marrow at diagnosis, and negative values if both Auer rods and granulation are absent.

**Usage**

```
data(feigl)
```

**Format**

A data.frame of 33 obs. of 3 variables:

[,1]	wbc	numeric, white blood cell count in thousands
[,2]	time	integer, survival time in weeks
[,3]	ag	Factor w/ 2 levels "+", "-"

**Note**

See Ch.6 of SMIR

**Source**

Feigl, P. and Zelen, M. (1965). "Estimation of exponential probabilities with concomitant information", *Biometrics*, **21**, 826–38.

---

gehan

*Remission times of acute leukemia patients — Gehan et al*

---

**Description**

Data from a clinical trial which compared 6-mercaptopurine (6-MP) to a placebo in the maintenance of remissions in acute leukemia. The remission times in weeks one year after the start of the study were recorded. Participants were paired according to remission status, an aspect not described in Gehan (1965).

**Usage**

`data(gehan)`

**Format**

A dataframe containing 42 obs. of 5 variables:

[1,]	pair	numeric defining pair according to remission status
[2,]	time	numeric time to remission available at the time the trial was stopped
[3,]	cens	numeric "0" indicating censored, "1" uncensored
[4,]	treat	factor w/ 2 levels "6-MP", "control"

**Note**

See Ch. 6 of SMIR

**Source**

Gehan, E. A. (1965), "A generalized Wilcoxon test for comparing arbitrarily singly-censored samples", *Biometrika*, **52**, 203–233.

ghq

*Psychiatric diagnosis based on GHQ***Description**

These data were published by Silvapulle, and come from a psychiatric study of the relation between psychiatric diagnosis (as case or non-case) and the value of the score on a 12-item General Health Questionnaire (GHQ), for 120 patients attending a general practitioner's surgery. Each patient was administered the GHQ, resulting in a score between 0 and 12, (however there were no cases or non-cases with GHQ scores of 11 or 12) and was subsequently given a full psychiatric examination by a psychiatrist who did not know the patient's GHQ score. The patient was classified by the psychiatrist as either a "case", requiring psychiatric treatment, or a "non-case".

**Usage**

```
data(ghq)
```

**Format**

[,1]	sex	Factor w/ 2 levels "men","women"
[,2]	ghq	integer, score from 0,...,12
[,3]	c	integer, number of patients considered a "case"
[,4]	nc	integer, number of patients considered a "non-case"

**Note**

See p.235 in SMIR

**Source**

Silvapulle, M. J. (1981), "On the existence of maximum likelihood estimators for the binomial response model", *J. Roy. Statist. Soc. B.*, **43**, 310–13.

hostility

*Bennett's hostility data***Description**

A measure of hostility based on word use exhibiting hostility by husbands of wives who had been admitted to hospital after suicide attempts by taking drug overdoses compared to a "control" group of husbands.

**Usage**

```
data(hostility)
```

**Format**

A data frame with 67 observations on the following 10 variables.

group a numeric vector

nationality a numeric vector

po a factor with levels none previous

in.host a numeric vector

amb.host a numeric vector

out.host a numeric vector

covert.host a numeric vector

positivity a numeric vector

g a factor with levels overdoses F controls T controls

nation a factor with levels Australian British

**Note**

See p.168 of SMIR

**Source**

Bennett, M. D. (1974). *The Emotional Response of Husbands to Suicide Attempts by Their Wives*, Sydney University, Unpublished MD thesis.

**Examples**

```
data(hostility)
## maybe str(hostility)
plot(hostility)
```

---

insult

*Effects of unprovoked verbal attack*

---

**Description**

Ten male and nine female subjects were asked to fill out a questionnaire which mixed innocuous questions with questions attempting to assess the subject's self-reported hostility. A hostility score for each individual was calculated from these responses. After completing the questionnaire, the subjects were then left waiting for a long time, and were subjected to insults and verbal abuse by the experimenter when the questionnaire was eventually collected. All subjects were told that they had filled out the questionnaire incorrectly, and were instructed to fill it out again. A second hostility score was then calculated from these later responses.

**Usage**

```
data(insult)
```

**Format**

A data.frame of 19 obs. of 3 variables:

[,1]	hbefore	integer, hostility score before verbal attack
[,2]	hafter	integer, hostility score after verbal attack
[,3]	sex	Factor w/ 2 levels "female","male"

**Note**

See pages 5 and 17 of SMIR

**Source**

Erickson, B. H. and Nosanchuk, T. A., (1979). *Understanding Data*, Milton Keynes, UK, Open University Press, UK.

---

 lsat

*LSAT*


---

**Description**

The original dataset consists of responses from 1,000 subjects to five dichotomous items from section 6 of the LSAT exam. The version here is presented as frequencies of unique patterns of responses. The data is from Bock and Lieberman 1970.

**Usage**

```
data(lsat)
```

**Format**

A data frame with 32 observations on the following 7 variables. The variable wt7 represents the number with each pattern.

y1 a numeric vector  
 y2 a numeric vector  
 y3 a numeric vector  
 y4 a numeric vector  
 y5 a numeric vector  
 wt6 a numeric vector  
 wt7 a numeric vector

**Note**

See p.547 in SMIR

**Source**

Bock, R. and Leiberhan, M. (1970), "Fitting a response model for a dichotomously scored items." *Psychometrika*, **35**, 179–197.

**References**

Bock, R. D. and Aitkin, M. (1981). "Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm." *Psychometrika*, **46**, 443–459.

**Examples**

```
data(lsat)
```

---

miners

*Pneumoconiosis in coal miners*

---

**Description**

The file gives the numbers of coalminers classified by radiological examination into one of three categories of pneumoconiosis, normal, mild pneumoconiosis and severe pneumoconiosis, and by years spent working at the coalface (interval midpoint).

**Usage**

```
data(miners)
```

**Format**

A data.frame of 8 obs. of 4 variables:

[,1]	years	numeric, years (midpoint) of years spent at coalface
[,2]	n	integer, number of miners classified as normal
[,3]	m	integer, number of miners with mild pneumoconiosis
[,4]	s	integer, number of miners with severe pneumoconiosis

**Note**

See p.279 in SMIR

**Source**

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

---

NPL.bands

*Nonparametric likelihood confidence bands*


---

**Description**

Computes the confidence bands for the empirical distribution function as described by Owen, A. (1997) *JASA* **90**:516–521.

**Usage**

```
NPL.bands(x, conf.level)
```

**Arguments**

x	a numeric vector
conf.level	Either 0.95 (default) or 0.99

**Value**

x	The unique values of x
lower	The lower bound
upper	The upper bound

**Author(s)**

```
<ross.darnell@csiro.au>
```

**Examples**

```
### Empirical distribution of a gamma variable
### and comparing to a normal
library(lattice)
y <- round(rgamma(100, shape=1.4, scale=20))
meany <- mean(y)
sdy <- sd(y)
print(xyplot(qnorm(lower)+qnorm(upper)~x, data=NPL.bands(y),
  panel=function(x,y,...){
    panel.xyplot(x,y,...)
    panel.curve(qnorm(pnorm(x, mean=meany, sd=sdy))))))
### and for a larger sample
yy <- round(rgamma(1000, shape=1.4, scale=20))
meanyy <- mean(yy)
sdy <- sd(yy)
print(xyplot(qnorm(lower)+qnorm(upper)~x, data=NPL.bands(yy),
  panel=function(x,y,...){
    panel.xyplot(x,y,...)
    panel.curve(qnorm(pnorm(x, mean=meanyy, sd=sdy))))))
### and for a t-distributed variable with df=10
```

```

yyy <- round(rt(1000,df=10),1)
meanyyy <- mean(yyy)
sdyyy <- sd(yyy)
print(xyplot(qnorm(lower)+qnorm(upper)~x,data=NPL.bands(yyy),
panel=function(x,y,...){
panel.xyplot(x,y,...)
panel.curve(qnorm(pnorm(x,mean=meanyyy,sd=sdyyy))))))
### and for a mixture of t-distributed variables with df=5
yyyy <- round(c(rt(100,df=5)*5+20,rt(100,df=5)*5+40))
meanyyyy <- mean(yyyy)
sdyyyy <- sd(yyyy)
print(xyplot(qnorm(lower)+qnorm(upper)~x,data=NPL.bands(yyyy),
panel=function(x,y,...){
panel.xyplot(x,y,...)
panel.curve(qnorm(pnorm(x,mean=meanyyyy,sd=sdyyyy))))))
#

```

---

poison

*Box Cox Poison Dataset*


---

### Description

Survival times (units 10 hrs) of animals in a 3 X 4 factorial experiment, the factors being (a) three poisons and (b) four treatments given in Box and Cox (1964). Each combination of the two factors is used for four animals, the allocation to animals being completely randomized.

### Usage

```
data(poison)
```

### Format

A dataframe containing 48 observations for 2 factors type and treat and the vector time.

```

[1,] type   Factor w/ 3 levels "I","II","III"
[2,] treat  Factor w/ 4 levels "A","B","C","D"
[3,] time   numeric (units 10 hrs)

```

### Note

See pp.161, 180 and 184 in SMIR

### Source

Box, G. E. P. and Cox, D. R. (1964). "An analysis of transformations", *Journal of the Royal Statistical Society, B*, **26**, 211–252.

---

```
prentice
```

---

*Veteran's Administration Lung Cancer Trial*

---

### Description

The file consists of survival times in days of 137 lung cancer patients from a Veteran's Administration Lung Cancer trial, together with explanatory variables: performance status, a measure of general medical status on a continuous scale 1–9.9, with 1–3 completely hospitalized, 4–6 partial confinement to hospital, 7–9.9 able to care for self; age in years; time in months from diagnosis to starting on the study; a factor prior therapy (1 no, 2 yes); a factor treatment (1 standard, 2 test) and a factor tumour type (1 squamous, 2 small, 3 adeno, 4 large). There are three censored observations.

### Usage

```
data(prentice)
```

### Format

A data.frame of 137 obs. of 8 variables:

[,1]	treat	integer, 1= standard, 2= test
[,2]	type	integer, 1= squamous, 2 =small, 3= adeno, 4= large
[,3]	time	integer, survival time in days
[,4]	sensor	integer, censoring indicator
[,5]	status	integer, general medical status on a scale 1–9.9
[,6]	mfd	integer, time in months from diagnosis
[,7]	age	integer, age in years
[,8]	prior	integer, prior therapy 1=no, 2=yes

### Note

See p.414 in SMIR

### Source

Prentice, R. L. (1973), "Exponential survivals with censoring and explanatory variables", *Biometrika*, **60**, 279–88.

---

```
R2
```

---

*Coefficient of determination of linear models*

---

### Description

This function provides the coefficient of determination for `lm` objects that may not have an intercept

**Usage**

```
R2(model)
```

**Arguments**

model            an object as returned by 'lm'

**Author(s)**

<ross.darnell@csiro.au>

**References**

Aitkin, M., Francis, B., Hinde, J. and Darnell, R. (2009). *Statistical Modelling in R*, UOP.

**Examples**

```
data(trees)
R2(lm(v ~ d + h - 1, data=trees))
```

---

R2CV

*Cross-validated coefficient of determination*

---

**Description**

This function provides the leave-one-out crossvalidation version of the coefficient of determination for regression models

**Usage**

```
R2CV(model)
```

**Arguments**

model            an object as returned by 'lm'

**References**

Aitkin, M., Francis, B., Hinde, J. and Darnell, R. (2009). *Statistical Modelling in R*, UOP.

**Examples**

```
data(trees)
R2CV(lm(v ~ d + h, data=trees))
```

---

Rsq	<i>Coefficient of determination, R-squared</i>
-----	--

---

**Description**

Calculates the coefficient of determination for any model of class 'lm'.

**Usage**

```
Rsq(model)
```

**Arguments**

model            a model object list with argument `y` and method `fitted`

**See Also**

cor

---

solv	<i>Children's block design test</i>
------	-------------------------------------

---

**Description**

A sample of twenty-four children was randomly drawn from the population of fifth-grade children attending a state primary school in a Sydney suburb. Each child was assigned to one of two experimental groups, and given instructions by the experimenter on how to construct, from nine differently coloured blocks, one of the 3X3 square designs in the Block Design subtest of the Wechsler Intelligence Scale for Children (WISC). Children in the first group were told to construct the design by starting with a row of three blocks (row group), and those in the second group were told to start with a corner of three blocks (corner group). The total time in seconds to construct four different designs was then measured for each child.

Before the experiment began, the extent of each child's "field dependence" was tested by the Embedded Figures Test (EFT), which measures the extent to which subjects can abstract the essential logical structure of a problem from its context (high scores corresponding to high field dependence and low ability).

**Usage**

```
data(solv)
```

**Format**

A data.frame of 24 obs. of 4 variables:

[,1]	child	integer, child id
[,2]	group	Factor w/ 2 levels "corner","row"
[,3]	time	integer, time in seconds
[,4]	eft	integer, EFT score

**Note**

See p.97 of SMIR

**Source**

Aitkin, M. Anderson, D., Francis, B. and Hinde, J. (1981), *Statistical Modelling in GLIM*, Oxford University Press

---

stackloss	<i>The Brownlee stackloss data</i>
-----------	------------------------------------

---

**Description**

The data consist of 21 observations on stack-loss (the loss of acid through the stack) in a chemical plant for the conversion of ammonia to nitric acid, with three explanatory variables: air flow(x1), cooling water inlet temperature(x2) and acid concentration(x3).

**Usage**

data(stackloss)

**Format**

A data.frame of 21 obs. of 4 variables:

[,1]	y	integer, loss of acid through the stack
[,2]	x1	integer, air flow
[,3]	x2	integer, cooling water inlet temperature
[,4]	x3	integer, acid concentration

**Note**

See p.469 of SMIR

**Source**

Lange, K L and Little, R J A and Taylor, J M G, (1989). "Robust statistical modeling using the \$t\$ distribution", *J Amer Statist Assoc*, **84**, 881–896.

---

stan

*Stanford Heart Transplantation Programme*

---

### Description

The file contains the data on 65 transplanted patients, consisting of the patient's age at transplantation, prior open-heart surgery (1 = yes, 0 = no), a censoring indicator (1 = yes, 0 = no), the survival time in days after transplant, a score representing the mismatch between the patient's and the donor's tissue type (values range from 0.00 to 3.05), and an indicator for death by rejection (1 = yes, 0 = no). One zero survival time is recoded to 0.5. There are 41 deaths and 24 censored survivals, with 39 distinct death times.

### Usage

```
data(stan)
```

### Format

A data frame with 65 observations on the following 12 variables.

```
id  a numeric vector
za  a numeric vector
zb  a numeric vector
age a numeric vector
surg a numeric vector
acc a numeric vector
died a numeric vector
surv a numeric vector
nmm a numeric vector
hla a numeric vector
mm  a numeric vector
rej a numeric vector
```

### Note

See p.422 in SMIR

### Source

Crowley, J. and Hu, M. (1977), Covariance analysis of heart transplant survival data. *Journal of the American Statistical Association*, **72**, 27–36.

### Examples

```
data(stan)
## maybe str(stan)
```

statlab

*STATLAB census data***Description**

The STATLAB Census covers 1296 member families of the Kaiser Foundation Health Plan (a pre-paid medical care program) living in the San Francisco Bay Area during the years 1961 - 1972. These families were participating members of the Child Health and Development Study conceived and directed by Jacob Yerushalmy, for many years Professor of Biostatistics in the School of Public Health, University of California, Berkeley.

On her first visit to the Oakland hospital of the Health Plan after pregnancy was diagnosed, each woman was interviewed intensively on a wide range of medical and socioeconomic matters relating both to herself and to her husband. In addition, various physical and physiological measures were made. When her child was born, further data about her and her newborn baby were recorded. Approximately 10 years later the child and mother were called in for follow-up testing, interviewing, and measurement. In some instances, the husband was also interviewed and measured.

The 1296 families of the STATLAB Census are divided into two equal subpopulations: 648 families consisting of a mother, father, and female child; and 648 families of a mother, father, and male child. The children were all born in the Kaiser Foundation Hospital, Oakland, California, between 1 April 1961 and 15 April 1963. The Census does not cover any other children who may also have existed in these families.

**Usage**

```
data(statlab)
```

**Format**

A data.frame of 1296 obs. of 34 variables:

[,]	id	integer,
[,]	c.b.blood	Factor w/ 9 levels
[,]	c.b.lgth	numeric,
[,]	c.b.wgt	numeric,
[,]	c.b.mo	integer,
[,]	c.b.day	integer,
[,]	c.b.hour	integer,
[,]	c.t.hght	numeric,
[,]	c.t.wgt	integer,
[,]	c.t.l	Factor w/ 8 levels
[,]	c.t.pea	integer,
[,]	c.t.ra	integer,
[,]	m.b.blood	Factor w/ 9 levels
[,]	m.b.ag	integer,
[,]	m.b.wgt	integer,
[,]	m.b.o	Factor w/ 8 levels
[,]	m.b.sm	Factor w/ 31 levels

[,]	m.t.hght	numeric,
[,]	m.t.wgt	integer,
[,]	m.t.e	Factor w/ 5 levels
[,]	m.t.o	Factor w/ 8 levels
[,]	m.t.sm	Factor w/ 26 levels
[,]	f.b.ag	integer,
[,]	f.b.o	Factor w/ 9 levels
[,]	f.b.sm	Factor w/ 32 levels
[,]	f.t.hght	numeric,
[,]	f.t.wgt	integer,
[,]	f.t.e	Factor w/ 5 levels
[,]	f.t.o	Factor w/ 9 levels
[,]	f.t.sm	Factor w/ 32 levels
[,]	family.i.b	integer,
[,]	family.i.t	integer,
[,]	family.c	Factor w/ 6 levels
[,]	sex	Factor w/ 2 levels "boy","girl"

**Note**

See pp.31 and 443 in SMIR

**Source**

Hodges, J.L., Krech, D. and Crutchfield, R.S. (1975). *StatLab: An Empirical Introduction to Statistics*, McGraw-Hill Ryerson, Toronto

---

summary.treg

*Summarizing Robust Regression Models*

---

**Description**

This function is a method for class treg.

**Usage**

```
## S3 method for class 'treg'
summary(object, ...)
```

**Arguments**

object            an object of class 'treg'.  
 ...                further arguments passed to or from other methods.

**Value**

The function summary.treg computes and prints statistics of "lm" class objects as well as the robust estimates of coefficients, the disparity and 'r', the degrees of freedom.

**Author(s)**

<ross.darnell@csiro.au>

**References**

Aitkin, M., Francis, B., Hinde, J. and Darnell, R. (2009). *Statistical Modelling in R*, OUP.

**See Also**

[treg](#)

---

toxaemia

*Bradford toxaemia data*

---

**Description**

The number of women giving birth to their first child who showed toxaemic signs (hypertension and/or proteinurea, classified as Yes or No) during pregnancy.

**Usage**

```
data(toxaemia)
```

**Format**

A data frame with 60 observations on the following 4 variables.

response a factor with levels HN HU NN NU

smoke a factor with levels 0 1-19 20+

class a factor with levels I II III IV V

count a numeric vector

**Note**

See p.330 in SMIR

**Source**

Brown, P.J., Stone, J., and Ord-Smith, C. (1983). Toxaemic signs during pregnancy. *Applied Statistics*, **32**, 69–72.

**Examples**

```
data(toxaemia)
tox.prop.table1 <- with(toxaemia, prop.table(tapply(count,
  list(class = class, response = response, smoke = smoke),
  sum), c(1, 3))[, c(2, 1, 4, 3), 1:2])
tox.prop.table2 <- with(toxaemia, prop.table(tapply(count,
  list(class = class, response = response, smoke = smoke),
  sum), c(1, 3))[, c(2, 1, 4, 3), 3, drop = FALSE])
```

---

 toxoplas

*Toxoplasmosis in El Salvador*


---

**Description**

The file shows the number of men tested and the number with a positive test for toxoplasmosis in 34 cities in El Slavador, together with the annual rainfall in metres.

**Usage**

```
data(toxoplas)
```

**Format**

A data.frame of 34 obs. of 3 variables:

```
[,1] y integer, number of men with a positive test
[,2] n integer, number of men tested
[,3] x numeric, annual rainfall in metres,
```

**Note**

See p.484 in SMIR

**Source**

Efron, B., (1986), Double exponential families and their use in generalized linear regression, *J Amer Statist Assoc*, **81**, 709–721.

---

 trees

*The Minitab cherry tree data*


---

**Description**

The volume of usable wood  $v$  in cubic feet (1 foot  $\backslash$  = 30.48 cm) is given for each of a sample of 31 black cherry trees, and the height  $h$  in feet and the diameter  $d$  in inches (1 inch = 2.54 cm) at a height 4.5 feet above the ground.

**Usage**

```
data(trees)
```

**Format**

A data.frame of 31 obs. of 3 variables:

```
[,] d  numeric, diameter in inches
[,] h  integer, height in feet
[,] v  numeric volume of usable wood,
```

**Note**

See pp.126 and 191 in SMIR

**Source**

Ryan, T. and Joiner, B. and Ryan, B., (1976). *Minitab Students Handbook*, Duxbury Press, North Scituate, Mass

---

treg	<i>t-regression model fit</i>
------	-------------------------------

---

**Description**

Robust regression by modelling errors as  $t$ -distributed with known degrees of freedom rather than normal

**Usage**

```
treg(lm.object, r, verbose=TRUE)
```

**Arguments**

lm.object	An object of class "lm"
r	a vector of degrees of freedom
verbose	TRUE prints estimates for $-2 \times$ log likelihood, sigma, and r at each iteration.

**Details**

Fits the  $t$  distribution for known degrees of freedom,  $r$ , and computes the profile likelihood and obtains the joint MLEs of the regression coefficients, sigma and disparity of a *robust* regression.

**Value**

an object of class “treg”

weights	working weights
disparity	disparity, i.e. full likelihood
tcoef	robust regression parameter estimates
r	degrees of freedom
sigma	estimate of residual standard deviation

**Author(s)**

<ross.darnell@csiro.au>

**References**

Aitkin, M., Francis, B., Hinde, J. and Darnell, R. (2008). *Statistical modelling in R*, OUP.

**See Also**

SMIR::summary.treg

**Examples**

```
library(SMIR)
data(stackloss)
stackloss.lm <- lm(y ~ x1 + x2 + x3, data = stackloss)
(stackloss.treg1.1 <- treg(stackloss.lm , r=1.1, verbose = FALSE) )
```

---

trypanos

*The trypanosome data*

---

**Description**

Follman and Lambert (1989) gave an example of a logistic regression with a varying intercept term. The data consist of numbers  $y_i$  of trypanosomes killed out of  $n_i$  treated at a treatment dose  $x_i$ .

**Usage**

```
data(trypanos)
```

**Format**

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

x a numeric vector

n a numeric vector

y a numeric vector

**Note**

See p.500 in SMIR

**Source**

Follman, D.A. and Lambert, D. (1989). Generalizing logistic regression by nonparametric mixing. *Journal of the American Statistical Association*, **84**, 295–300.

**Examples**

```
data(trypanos)
library(npmlreg)
(trypanos.np1 <- alldist(cbind(y, (n - y)) ~ log(x),
  random = ~1, data = trypanos, family = binomial,
  plot.opt = 0, verbose = FALSE,k=1))
(trypanos.np2 <- update(trypanos.np1,k=2))
```

---

 vaso

*Gilliatt's vaso-constriction data*


---

**Description**

These data were obtained in a carefully controlled study of the effect of the rate and volume of air inspired by human subjects on the occurrence or non-occurrence of a transient vasoconstriction response in the skin of the fingers.

**Usage**

```
data(vaso)
```

**Format**

A data.frame of 39 obs. of 4 variables:

[,1]	Subject	Factor w/ 3 levels "1","2","3", ...
[,2]	Volume	numeric, volume of air expired
[,3]	Rate	numeric, rate of air expired
[,4]	Y	integer, transient vasoconstriction response, 1=yes, 0=no

**Note**

See p.209 in SMIR

**Source**

Finney, D.J. (1947)., "The estimation from individual records of the relationship between dose and quantal response", *Biometrika*, **34**, 320–34.

## References

Gilliatt, R.W. (1948). "Vaso-constriction in the finger after deep inspiration", *J. Physiol.*, **107**, 76–88.

---

vietnam

*University of North Carolina Vietnam War Student Survey*

---

## Description

A survey of student opinion on the Vietnam War was taken at the University of North Carolina in Chapel Hill in May 1967 and published in the student newspaper. Students were asked to fill in “ballot papers”, available in the Student Council building, stating which policy out of A, B, C or D they supported. Responses were cross-classified by sex and by undergraduate year or graduate status. The policies were:

A: The US should defeat the power of North Vietnam by widespread bombing of its industries, ports and harbours and by land invasion.

B: The US should follow the present policy in Vietnam.

C: The US should de-escalate its military activity, stop bombing North Vietnam, and intensify its efforts to begin negotiation.

D: The US should withdraw its military forces from Vietnam immediately.

## Usage

`data(vietnam)`

## Format

A data.frame of 40 obs. of 4 variables:

[,1]	policy	Factor w/ 4 levels "A","B","C","D"
[,2]	year	Factor w/ 5 levels "1","2","3","4","5"
[,3]	sex	Factor w/ 2 levels "female","male"
[,4]	count	integer , the number of students in each cell

## Note

See p.310 in SMIR

## Source

Aitkin, M. (1996). "A short history of a Vietnam War attitude survey.", *Stats*, **17**, 1–9.

---

woolson

*Woolson and Clarke's Obesity Study*

---

**Description**

This is a subset of the Obesity dataset. Binary indicators of obesity on 1014 children who were 7-9 years old in 1977, and were followed up in 1979 and 1981. Children were classified as obese if their weights were more than 210% of the population median weight for their gender and height.

**Usage**

```
data(woolson)
```

**Format**

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

x a numeric vector

y a numeric vector

sex a numeric vector

age a numeric vector

**Note**

See p.539 in SMIR

**Source**

Woolson, R.F. and Clark, W.R.(1984). Analysis of categorical incomplete longitudinal data. *JRSS A*. **147**, 87–99.

**Examples**

```
data(woolson)
```

# Index

## \*Topic **datasets**

- betablok, 2
- bronchitis, 3
- byssinosis, 4
- cars, 4
- chd, 5
- claims, 6
- faults, 10
- feigl, 10
- ghq, 12
- hostility, 12
- insult, 13
- lsat, 14
- miners, 15
- poison, 17
- prentice, 18
- solv, 20
- stackloss, 21
- stan, 22
- statlab, 23
- toxaemia, 25
- toxoplas, 26
- trees, 26
- trypanos, 28
- vaso, 29
- vietnam, 30
- woolson, 31

## \*Topic **distribution**

- NPL.bands, 16

## \*Topic **models**

- disparity, 8

## \*Topic **regression**

- R2, 18
- R2CV, 19
- summary.treg, 24

## \*Topic **robust**

- treg, 27

## \*Topic **survival**

- coxph.disparity, 7

- gehan, 11

- betablok, 2
- bronchit (bronchitis), 3
- bronchitis, 3
- byssinosis, 4

- cars, 4
- chd, 5
- claims, 6
- coxph.disparity, 7

- disparity, 8
- disparity.glm, 9
- disparity.lm, 9

- faults, 10
- feigl, 10

- gehan, 11
- ghq, 12

- hostility, 12

- insult, 13

- lsat, 14

- miners, 15

- NPL.bands, 16

- poison, 17
- prentice, 18

- R2, 18
- R2CV, 19
- Rsq, 20

- solv, 20
- stackloss, 21
- stan, 22

statlab, 23  
summary.treg, 24

toxaemia, 25  
toxoplas, 26  
trees, 26  
treg, 25, 27  
trypanos, 28

vaso, 29  
vietnam, 30

woolson, 31