Package 'VGAMdata'

January 27, 2015

Version 0.9-5

Date 2014-11-04

Title Data supporting the VGAM package

Author Thomas W. Yee <t.yee@auckland.ac.nz>

Maintainer Thomas Yee <t.yee@auckland.ac.nz>

Depends R (>= 3.0.0)

Suggests VGAM

Description Data sets to accompany the VGAM package.
These are used to illustrate vector generalized
linear and additive models (VGLMs/VGAMs), and associated models
(Reduced-Rank VGLMs, Quadratic RR-VGLMs, Row-Column
Interaction Models, and constrained and unconstrained ordination
models in ecology.

License GPL-2

URL http://www.stat.auckland.ac.nz/~yee/VGAMdata

Repository CRAN

LazyLoad yes

LazyData yes

NeedsCompilation no

Date/Publication 2014-11-04 19:53:13

R topics documented:

`VGAMdata-package` *Data for the VGAM package*

Description

VGAMdata is an assortment of larger data sets which are a useful resource for the **VGAM** package.

Details

This package contains some larger data sets originally distributed with the **VGAM** package. Ideally both packages can be installed and loaded to be fully functional. The main intent was to limit the size of **VGAM** to a bare essential. Many data sets in my monograph will refer to data sets in either package.

Author(s)

Thomas W. Yee, <t.yee@auckland.ac.nz>.

Maintainer: Thomas Yee <t.yee@auckland.ac.nz>.

References

Yee, T. W. Vector Generalized Linear and Additive Models. *Monograph in preparation*.

See Also

`VGAM-package`.

Examples

```
# Example 1; xs.nz
head(xs.nz)
summary(xs.nz)

# Example 2; ugss
head(ugss)
summary(ugss)
```

---

bb.de *Battle of Britain Data (a Luftwaffe subset)*

Description

Luftwaffe losses during a subset of the Battle of Britain.

Usage

```
data(bb.de)
```

Format

The format is a 3-dimensional array. The first dimension is the event (in order: shot down or failed to return, written off, seriously damaged), the second dimension is the day, the third is the aircraft type.

Details

This is a Battle of Britain data set of Luftwaffe losses during operations 26 August–31 August 1940 continued on to 1–7 September 1940. The aircraft types are prefixed `Bf` for Messerschmitt (Bayerische Flugzeugwerke), `Do` for Dornier, `He` for Heinkel, `Ju` for Junkers.

Note that p.151 and p.165 of Bowyer (1990) contain tables (during the first week of September) and almost the same data; however, the former is labelled "shot down" whereas the latter is "shot down or failed to return". The latter is used here. Also, there are some other small discrepancies.

Yet to do: add the data available at other dates, and include the RAF data.

Source

Bowyer, M. J. F. (1990) *The Battle of Britain: 50 years On*. Patrick Stephens Limited, Northamptonshire, U.K.

Examples

```
data(bb.de)
bb.de[,, "Bf109"]

## Not run:
plot(bb.de["sdown",, "Bf109"] ~ as.Date(dimnames(bb.de)[[2]]),
     type = "h", col = "blue", las = 1, lwd = 3,
     ylab = "Frequency", xlab = "1940",
     main = "Numbers shot down (Bf 109)")
abline(h = c(5, 10, 15, 20), lty = "dashed", col = "grey")
points(bb.de["sdown",,"Bf109"] ~ as.Date(dimnames(bb.de)[[2]]), col = "blue")

## End(Not run)
```

`belcap` *BELCAP Dental Data*

Description

A prospective data set containing the DMFT index of childeren in Belo Horizonte at the beginning and end of the BELCAP study.

Usage

```
data(belcap)
```

Format

A data frame with 797 observations on the following 5 variables.

`dmftb` a numeric vector, DMFT-Index at the beginning of the study.

`dmfte` a numeric vector, DMFT-Index at the end of the study.

`gender` a factor with levels `0` = female, `1` = male.

`ethnic` a factor with levels `1` = dark, `2` = white, `3` = black.

`school` the kind of prevention measure. A factor with levels `1` = oral health education, `2` = all four methods, `3` = control group, `4` = enrichment of the school diet with ricebran, `5` = mouthrinse with 0.2% NaF-solution, `6` = oral hygiene.

Details

This data set is from the Belo Horizonte Caries Prevention (BELCAP) study. The data is collected from children in Belo Horizonte (Brazil) aged seven years at the start of the study. In in order to determine which method(s) were best for preventing tooth decay, six treatments were randomized to six separate schools. The measure used is the decayed, missing and filled teeth (DMFT) index—a well known and important measure of a person's dental health. Only the eight deciduous molars are considered, so the lowest value is 0, and the highest is 8.

Source

http://onlinelibrary.wiley.com contains the data file (a supplement of the JRSSA article). Downloaded in January 2014 and formatted into R by J. T. Gray, jamsgry@gmail.com.

References

Bohning, D., D. Ekkehart, P. Schlattmann, L. Mendonca, and U. Kircher (1999). The Zero-Inflated Poisson Model and the Decayed, Missing and Filled Teeth Index in Dental Epidemiology, *Journal of the Royal Statistical Society, A* **162**(2), 195–209.

Examples

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

---

crashf.au *Fatal Crashes on Australian Roads 2010–2012*

Description

The number of fatal road crashes on Australian roads during 2010–2012. They are cross-classified by time of day (in 6 hour blocks) and day of the week.

Usage

```
data(crashf.au)
```

Format

A data frame with 4 observations on the following 7 variables.

Mon, Tue, Wed, Thu, Fri, Sat, Sun Day of the week.

Details

Each cell is the aggregate number of crashes reported in Australia during each six hour time block throughout the years 2010–2012. The rownames are the time period the crashes took place in. Morning is from 3:00am to 8:59am, midday is from 9:00am to 2:59pm, evening is from 3:00pm to 8:59pm and night is from 9:00pm to 2:59am.

Source

http://www.bitre.gov.au/publications/ongoing/files/RDA_Summary_2012_June.pdf

References

Road Deaths Australia; 2012 Statistical Summary. Department of Infrastructure and Transport, Australian Government; ISSN: 1323–3688

Downloaded by J. T. Gray, April 2014.

Examples

```
crashf.au
```

`crime.us` *Estimated Crime in 2009 in USA*

Description

Crime totals and rates, cross-classified by US state, during 2009.

Usage

```
data(crime.us)
```

Format

A data frame with 50 observations on the following 22 variables.

`State` a character vector. White spaces have been replaced by underscores.

`Population` a numeric vector

`ViolentCrimeTotal` a numeric vector

`Murder` a numeric vector

`Rape` a numeric vector

`Robbery` a numeric vector

`Assault` a numeric vector

`PropertyCrimeTotal` a numeric vector

`Burglary` a numeric vector

`LarcenyTheft` a numeric vector

`MotorVehicleTheft` a numeric vector

`ViolentCrimeRate` a numeric vector

`MurderRate` a numeric vector

`RapeRate` a numeric vector

`RobberyRate` a numeric vector

`AssaultRate` a numeric vector

`PropertyCrimeRate` a numeric vector

`BurglaryRate` a numeric vector

`LarcenyTheftRate` a numeric vector

`MotorVehicleTheftRate` a numeric vector

`stateNumber` a numeric vector, running from 1 to 50.

`abbrev` State name as a character vector

Details

Each row is a state of the United States of America. The first half of the columns tend to be totals, and the second half are crime rates per 100,000 population.

The data frame was downloaded as a `.csv` file and edited. The full column names are: State, Population, Violent crime total, Murder and nonnegligent Manslaughter, Forcible rape, Robbery, Aggravated assault, Property crime total, Burglary, Larceny-theft, Motor vehicle theft, Violent Crime rate, Murder and nonnegligent manslaughter rate, Forcible rape rate, Robbery rate, Aggravated assault rate, Property crime rate, Burglary rate, Larceny-theft rate, Motor vehicle theft rate, state Number, abbreviation. Technical details governing the data set are given in the URL.

Source

http://www.ucrdatatool.gov, http://www.ucrdatatool.gov/Search/Crime/State/RunCrimeOneYearofData.cfm

Examples

```
## Not run:   # Louisiana is the one outlier
plot(MurderRate ~ stateNumber, crime.us,
     axes = FALSE, type = "h", col = 1:6,
     main = "USA murder rates in 2009 (per 100,000 population)")
axis(1, with(crime.us, abbrev), at = with(crime.us, stateNumber),
     col = 1:6, col.tick = 1:6, cex.lab = 0.5)
axis(2)
## End(Not run)
tail(crime.us[ sort.list(with(crime.us, MurderRate)), ])
```

`DeLury` *DeLury's Method for Population Size Estimation*

Description

Computes DeLury's method or Leslie's method for estimating a biological population size.

Usage

```
DeLury(catch, effort, type = c("DeLury","Leslie"), ricker = FALSE)
```

Arguments

`catch, effort`	Catch and effort. These should be numeric vectors of equal length.
`type`	Character specifying which of the DeLury or Leslie models is to be fitted. The default is the first value.
`ricker`	Logical. If `TRUE` then the Ricker (1975) modification is computed.

Details

This simple function implements the methods of DeLury (1947). These are called the DeLury and Leslie models. Note that there are many assumptions. These include: (i) Catch and effort records are available for a series of consecutive time intervals. The catch for a given time interval, specified by t, is $c(t)$, and the corresponding effort by $e(t)$. The *catch per unit effort* (CPUE) for the time interval t is $C(t) = c(t)/e(t)$. Let $d(t)$ represent the proportion of the population captured during the time interval t. Then $d(t) = k(t)e(t)$ so that $k(t)$ is the proportion of the population captured during interval t by one unit of effort. Then $k(t)$ is called the *catchability*, and the *intensity* of effort is $e(t)$. Let $E(t)$ and $K(t)$ be the total effort and total catch up to interval t, and $N(t)$ be the number of individuals in the population at time t. It is good idea to plot $\log(C(t))$ against $E(t)$ for `type = "DeLury"` and $C(t)$ versus $K(t)$ for `type = "Leslie"`.

The other assumptions are as follows.

(ii) The population is closed—the population must be closed to sources of animals such as recruitment and immigration and losses of animals due to natural mortality and emigration.

(iii) Catchability is constant over the period of removals.

(iv) The units of effort are independent, i.e., the individual units of the method of capture (i.e., nets, traps, etc) do not compete with each other.

(v) All fish are equally vulnerable to the method of capture—source of error may include gear saturation and trap-happy or trap-shy individuals.

(vi) Enough fish must be removed to substantially reduce the CPUE.

(vii) The catches may remove less than 2% of the population.

Also, the usual assumptions of simple regression such as

(viii) random sampling,

(ix) the independent variable(s) are measured without error—both catches and effort should be known, not estimated,

(x) a line describes the data,

(xi) the errors are independent and normally distributed.

Value

A list with the following components.

`catch, effort`	Catch and effort. Same as the original vectors. These correspond to $c(t)$ and $e(t)$ respectively.
`type, ricker`	Same as input.
`N0`	an estimate of the population size at time 0. Only valid if the assumptions are satisfied.
`CPUE`	Catch Per Unit Effort $= C(t)$.
`K, E`	$K(t)$ and $E(t)$. Only one is computed depending on `type`.
`lmfit`	the `lm` object from the fit of `log(CPUE)` on `K` (when `type = "Leslie"`). Note that the `x` component of the object is the model matrix.

Note

The data in the example below comes from DeLury (1947), and some plots of his are reproduced. Note that he used log to base 10 whereas natural logs are used here. His plots had some observations obscured by the y-axis!

The DeLury method is not applicable to the data frame wffc.nc since the 2008 World Fly Fishing Competition was strictly catch-and-release.

Author(s)

T. W. Yee.

References

DeLury, D. B. (1947) On the estimation of biological populations. *Biometrics*, **3**, 145–167.

Ricker, W. E. (1975) Computation and interpretation of biological statistics of fish populations. *Bull. Fish. Res. Bd. Can.*, **191**, 382–

Yee, T. W. (2010) VGLMs and VGAMs: an overview for applications in fisheries research. *Fisheries Research*, **101**, 116–126.

See Also

wffc.nc.

Examples

```
pounds <- c(  147, 2796, 6888, 7723, 5330, 8839, 6324, 3569, 8120, 8084,
            8252, 8411, 6757, 1152, 1500, 11945, 6995, 5851, 3221, 6345,
            3035, 6271, 5567, 3017, 4559, 4721, 3613,  473,  928, 2784,
            2375, 2640, 3569)
traps  <- c(  200, 3780, 7174, 8850, 5793, 9504, 6655, 3685, 8202, 8585,
            9105, 9069, 7920, 1215, 1471, 11597, 8470, 7770, 3430, 7970,
            4740, 8144, 7965, 5198, 7115, 8585, 6935, 1060, 2070, 5725,
            5235, 5480, 8300)
table1 <- DeLury(pounds/1000, traps/1000)

## Not run:
with(table1, plot(1+log(CPUE) ~ E, las = 1, pch = 19, main = "DeLury method",
     xlab = "E(t)", ylab = "1 + log(C(t))", col = "blue"))

## End(Not run)
omitIndices <- -(1:16)
table1b <- DeLury(pounds[omitIndices]/1000, traps[omitIndices]/1000)
## Not run:
with(table1b, plot(1+log(CPUE) ~ E, las = 1, pch = 19, main = "DeLury method",
     xlab = "E(t)", ylab = "1 + log(C(t))", col = "blue"))
mylmfit <- with(table1b, lmfit)
lines(mylmfit$x[, 2], 1 + predict.lm(mylmfit), col = "red", lty = "dashed")

## End(Not run)
```

```
omitIndices <- -(1:16)
table2 <- DeLury(pounds[omitIndices]/1000, traps[omitIndices]/1000, type = "L")
## Not run:
with(table2, plot(CPUE ~ K, las = 1, pch = 19,
     main = "Leslie method; Fig. III",
     xlab = "K(t)", ylab = "C(t)", col = "blue"))
mylmfit <- with(table2, lmfit)
abline(a = coef(mylmfit)[1], b = coef(mylmfit)[2],
       col = "orange", lty = "dashed")

## End(Not run)
```

---

exam1 *Examination data*

Description

Exam results of 35 students on 18 questions.

Usage

```
data(exam1)
```

Format

A data frame with 35 observations on the following 18 variables.

q01, q02, q03, q04, q05, q06 binary response

q07, q08, q09, q10, q11, q12 binary response

q13, q14, q15, q16, q17, q18 binary response

Details

For each question, a 1 means correct, a 0 means incorrect. A simple Rasch model may be fitted to this dataframe using rcim and binomialff.

Source

Taken from William Revelle's *Short Guide to R*, http://www.unt.edu/rss/rasch_models.htm, http://www.personality-project.org/r/. Downloaded in October 2013.

Examples

```
summary(exam1)  # The names of the students are the row names

# Fit a simple Rasch model.
# First, remove all questions and people who were totally correct or wrong
exam1.1 <- exam1  [, colMeans(exam1  ) > 0]
exam1.1 <- exam1.1[, colMeans(exam1.1) < 1]
exam1.1 <- exam1.1[rowMeans(exam1.1) > 0, ]
exam1.1 <- exam1.1[rowMeans(exam1.1) < 1, ]
Y.matrix <- rdata <- exam1.1

## Not run:  # The following needs: library(VGAM)
rfit <- rcim(Y.matrix, family = binomialff(mv = TRUE), trace = TRUE)

coef(rfit)  # Row and column effects
constraints(rfit, matrix = TRUE)  # Constraint matrices side-by-side
dim(model.matrix(rfit, type = "vlm"))  # 'Big' VLM matrix

## End(Not run)

## Not run:  # This plot shows the (main) row and column effects
par(mfrow = c(1, 2), las = 1, mar = c(4.5, 4.4, 2, 0.9) + 0.1)
saved <- plot(rfit, rcol = "blue", ccol = "orange",
              cylab = "Item effects", rylab = "Person effects",
              rxlab = "", cxlab = "")

names(saved@post)  # Some useful output put here
cbind(saved@post$row.effects)
cbind(saved@post$raw.row.effects)
round(cbind(-saved@post$col.effects), dig = 3)
round(cbind(-saved@post$raw.col.effects), dig = 3)
round(matrix(-saved@post$raw.col.effects, ncol = 1,  # Rename for humans
             dimnames = list(colnames(Y.matrix), NULL)), dig = 3)

## End(Not run)
```

---

hued *Harvard University Degrees Conferred by Student Ethnicity*

Description

A two-way table of counts; there are 7 ethnic groups by 12 degrees.

Usage

```
data(hued)
```

Format

The format is: chr "hued"

Details

The rownames and colnames have been edited. The full names are: Asian/Pacific Islander, Black/Non-Hispanic, Hispanic, International Students, Native American, White/Non-Hispanic, Unknown/Other. The academic year was 2009–2010. GSAS stands for Graduate School of Arts and Sciences. The Other group includes students reported as Two or More Races. See the URL below for more technical details supporting the data.

Source

http://www.provost.harvard.edu/institutional_research/factbook.php

See Also

huie, huse.

Examples

```
print(hued)
```

---

huie *Harvard University International Enrollments*

Description

A two-way table of counts; there are 12 degrees and 8 areas of the world.

Usage

```
data(huie)
```

Format

The format is: chr "huie"

Details

The rownames and colnames have been edited. The full colnames are: Africa, Asia, Europe, Caribbean and Central and and South America, Middle East, North America, Oceania, Stateless.

The data was for the autumn (Fall) of 2010. GSAS stands for Graduate School of Arts and Sciences. See the URL below for more technical details supporting the data.

Source

http://www.provost.harvard.edu/institutional_research/factbook.php

See Also

hued, huse.

Examples

```
print(huie)
## maybe str(huie) ; plot(huie) ...
```

---

huse *Harvard University Numbers of Ladder Faculty by School and Ethnicity*

Description

A two-way table of counts; there are 14 schools and 5 race/ethnicities.

Usage

```
data(huse)
```

Format

The format is: chr "huse"

Details

Ladder faculty members of Harvard University are cross-classified by their school and their race/ethnicity. This was for the period 2010–1. Ladder Faculty are defined as Assistant Professors or Convertible Instructors, Associate Professors, and Professors that have been appointed in certain Schools.

Abbreviations: FAS = Faculty of Arts and Sciences = Humanities + Social Sciences + Natural Sciences + SEAS, Natural Sciences = Life Sciences + Physical Sciences, SEAS = School of Engineering and Applied Sciences, HBS = Harvard Business School, HMS = Harvard Medical School, HSPH = Harvard School of Public Health, HLS = Harvard Law School, HKS = Harvard Kennedy School, HGSE = Harvard Graduate School of Education, GSD = Graduate School of Design , HDS = Harvard Divinity School, HSDM = Harvard School of Dental Medicine.

See the URL below for many technical details supporting the data. The table was constructed from pp.31–2 from the source.

Source

http://www.provost.harvard.edu/institutional_research/factbook.php

References

Harvard University Office of the Senior Vice Provost Faculty Development \& Diversity: 2010 Annual Report.

See Also

hued, huie.

Examples

```
print(huse)
## maybe str(huse) ; plot(huse) ...
```

`oly12` *2012 Summer Olympics: Individuals Data*

Description

Individual data for the Summer 2012 Olympic Games.

Usage

```
data(oly12)
```

Format

A data frame with 10384 observations on the following 14 variables.

`Name` The individual competitor's name.

`Country` Country.

`Age` A numeric vector, age in years.

`Height` A numeric vector, height in m.

`Weight` A numeric vector, weight in kg.

`Sex` A factor with levels `F` and `M`.

`DOB` A Date, date of birth.

`PlaceOB` Place of birth.

`Gold` Numeric vector, number of such medals won.

`Silver` Similar to `Gold`.

`Bronze` Similar to `Gold`.

`Total` A numeric vector, total number of medals.

`Sport` A factor with levels `Archery`, `Athletics`, `Athletics`, `Triathlon`, `Badminton`, etc.

`Event` The sporting event.

Details

This data set represents a very small modification of a `.csv` spreadsheet from the source below. Height has been converted to meters, and date of birth is of a `"Date"` class (see `as.Date`). A few non-ASCII characters have been replaced by some ASCII sequence (yet to be fixed up properly).

Some competitors share the same name. Some errors in the data are likely to exist.

Source

Downloaded from http://www.guardian.co.uk/sport/series/london-2012-olympics-data in 2013-03.

Examples

```
data(oly12)
mtab <- with(oly12, table(Country, Gold))
(mtab <- head(sort(mtab[, "1"] + 2 * mtab[, "2"], decreasing = TRUE), 10))

## Not run:
barplot(mtab, col = "gold", cex.names = 0.8, names = abbreviate(names(mtab)),
        beside = TRUE, main = "2012 Summer Olympic Final Gold Medal Count",
        ylab = "Gold medal count", las = 1, sub = "Top 10 countries")

## End(Not run)
```

---

prison.us *US Prison Data*

Description

Number of prisoners in each North American state, and the populations of those states from years 1977 to 2010

Usage

```
data(prison.us)
```

Format

A data frame with 34 observations on the following 103 variables.

Year a numeric vector, the year

AL.num, AL.pop numeric vectors

AK.num, AK.pop, AZ.num numeric vectors

AZ.pop, AR.num, AR.pop numeric vectors

CA.num, CA.pop, CO.num numeric vectors

CO.pop, CT.num, CT.pop numeric vectors

DE.num, DE.pop, FL.num numeric vectors

FL.pop, GA.num, GA.pop numeric vectors

HI.num, HI.pop, ID.num numeric vectors

ID.pop, IL.num, IL.pop numeric vectors

IN.num, IN.pop, IA.num numeric vectors

IA.pop, KS.num, KS.pop numeric vectors

KY.num, KY.pop, LA.num numeric vectors

LA.pop, ME.num, ME.pop numeric vectors

MD.num, MD.pop, MA.num numeric vectors

MA.pop, MI.num, MI.pop numeric vectors

MN.num, MN.pop, MS.num numeric vectors

MS.pop, MO.num, MO.pop numeric vectors

MT.num, MT.pop, NE.num numeric vectors

NE.pop, NV.num, NV.pop numeric vectors

NH.num, NH.pop, NJ.num numeric vectors

NJ.pop, NM.num, NM.pop numeric vectors

NY.num, NY.pop, NC.num numeric vectors

NC.pop, ND.num, ND.pop numeric vectors

OH.num, OH.pop, OK.num numeric vectors

OK.pop, OR.num, OR.pop numeric vectors

PA.num, PA.pop, RI.num numeric vectors

RI.pop, SC.num, SC.pop numeric vectors

SD.num, SD.pop, TN.num numeric vectors

TN.pop, TX.num, TX.pop numeric vectors

UT.num, UT.pop, VT.num numeric vectors

VT.pop, VA.num, VA.pop numeric vectors

WA.num, WA.pop, WV.num numeric vectors

WV.pop, WI.num, WI.pop numeric vectors

WY.num, WY.pop numeric vectors

US.pop, US.num numeric vectors, overall counts for the whole country

Details

This is a data set of the number of prisoners in each American state and the populations of those states, from 1977 to 2010. The number of prisoners are taken from December 31st, while the populations are estimates taken from July 1st based on the previous Census, except for pop.1980, which uses exact census data from 1980/04/01.

Warning: a scatterplot of `US.pop` shows a discontinuity around 2000.

Source

The prisoner data was compiled from: Bureau of Justice Statistics, http://www.bjs.gov/index.cfm. Downloaded in September 2013 and formatted into R by J. T. Gray, jamsgry@gmail.com.

The population data was compiled from: United States Census Bureau, http://www.census.gov/popest/data. Downloaded in September 2013 by J. T. Gray.

Examples

```
summary(prison.us)
## Not run:  # This plot shows a discontinuity around 2000.
plot(US.pop / 1e6 ~ Year, prison.us, main = "US population (millions)",
     las = 1, type = "b", col = "blue")
## End(Not run)
```

---

profs.nz *Professors of Statistics in New Zealand*

Description

This data set contains information on about 22 past or present professors of statistics in New Zealand universities.

Usage

```
data(profs.nz)
```

Format

A data frame with 22 observations on the following 7 variables.

`pubtotal` a numeric vector, the total number of publications.

`cites` a numeric vector, the number of citations.

`initials` character, first and middle and surname initials.

`Surname` character, the surname.

`firstyear` a numeric vector, the earliest indexed publication.

`ID` a numeric vector, the unique MR Author ID for each professor.

`ARPtotal` a numeric vector, the total number of author/related publications.

`institution` character, with values `"MU"`, `"UA"`, `"UC"`, `"UO"`, `"UW"`, `"VU"`, the university affiliation. The abbreviations are for: Massey University, University of Auckland, University of Canterbury, University of Otago, University of Waikato and Victoria University Wellington.

Details

This data set contains information taken from the MathSciNet database on professors of statistics (and some related fields) affiliated with New Zealand universities.

In the future the following names may be added: C. F. Ansley, P. C. B. Phillips, B. S. Weir.

Source

The data is compiled from http://www.ams.org/mathscinet by J. T. Gray in April 2014.

Examples

```
profs.nz
profs.nz[order(with(profs.nz, pubtotal), decreasing = TRUE), ]
```

rugby *Wins, Losses and Draws Between the Top 10 Rugby Teams*

Description

The number of wins, losses and draws for each of the top 10 rugby teams agains each other

Usage

```
data(rugby)
data(rugby.ties)
```

Format

The format is as two matrices.

Details

The first matrix is of the number of games won by each team against each of the other teams. The other matrix is the number of draws (ties) between each team. This data is current as of 2013-10-07.

Source

The match statistics are compiled from http://www.rugbydata.com/ on 2013-10-07 by J. T. Gray, jamsgry@gmail.com.

The top ten teams are determined by the International Rugby Board world rankings, http://www.irb.com.

Examples

```
data(rugby); data(rugby.ties)
rugby
rugby.ties
```

students.tw *Taiwanese students answer a multiple response question*

Description

This data is a subset from a survey of 49609 first-year college students in Taiwan collected in the year 2003 about their preferences for college study.

Usage

```
data(students.tw)
```

Format

A data frame with 49609 observations on the following 12 response variables. For binary variables, a "1" means yes, and "0" means no. See below for exact wording (translated from the original Chinese).

ID a numeric vector, a unique identification number for each student in the survey.

read Read Chinese and foreign classics.

t.travel Travel around Taiwan.

conference Present academic papers in conferences.

act.leader Lead large-scale activities.

team Be on a school team.

stu.leader Be a student association leader.

intern Participate internship programs.

love Fall in love.

sex Have sexual experience.

o.travel Travel around the world.

friends Make many friends.

other Other experience which is not included in the survey.

Details

This data frame is a subset of a larger data set where any student with any missing value was deleted. The remaining data set contains of 32792 students. Unfortunately, other variables such as age and sex were not made available.

Each student was asked the following multiple response question.

Question : What kind of experience do you expect to receive during the period of college study? (Select at least one response)

1. Read Chinese and foreign classics
2. Travel around Taiwan
3. Present academic papers in conferences
4. Lead large-scale activities
5. Be on a school team
6. Be a student association leader
7. Participate internship programs
8. Fall in love
9. Have sexual experience
10. Travel around the world
11. Make many friends
12. Other

Source

Originally, the data set for was downloaded from a survey center of Academia Sinica https://srda.sinica.edu.tw/news. It now seems unavailable.

References

Wang, H. and Huang, W. H. (2013) Bayesian Ranking Responses in Multiple Response Questions. *Journal of the Royal Statistical Society, Series A*, **(to appear)**.

Help from Viet Hoang Quoc is gratefully acknowledged.

Examples

```
data(students.tw)
summary(students.tw)

with(students.tw, table(love, sex))
## Not run:
plot(jitter(sex) ~ jitter(love), data = students.tw, col = "blue",
     main = "Taiwanese students")

## End(Not run)
```

---

trapO *Trout Data at the Te Whaiau Trap on Lake Otamangakau*

Description

Rainbow and brown trout trout trapped at the Te Whaiau Trap at Lake Otamangakau in the central North Island of New Zealand. The data were collected by the Department of Conservation.

Usage

```
data(trapO)
```

Format

A data frame with 1226 observations on the following 15 variables.

Date Date as a class "Date" variable.

BFTW, BMTW, RFTW, RMTW numeric vectors, the number of fish trapped daily. B/R is for brown/rainbow trout. F/M is for female/male. TW is for the Te Whaiau trap location (there was another trap just off the Tongariro River).

MinAT, MaxAT numeric vectors, daily minimum and maximum ambient temperatures in Celsius.

Rain numeric vector, daily rainfall that has been scaled between 0 (none) and 100 (flooding situation).

LevelTW numeric vector, water level of the stream that has been scaled between 0 (none) and 100 (flooding situation). In a flooding situation it is possible that some fish going upstream were not caught.

Year, Month, Day numeric vectors, extracted from `Date`.

doy a numeric vector, Julian day of year. The value 1 means 1st of January, and so on up to 365.

f.Year a factor vector, the year as a factor.

fict.Year similar to `Date` but a fictitional year is used for all the data. This allows all the data to be plotted along one calendar year.

Details

These are the daily numbers of fish trapped at the Te Whaiau trap near Lake Otamangakau, during the winter months when spawning is at its peak. These fish were all going upstream. There are two species of trout, split up by males and females, in the data set. The first is brown trout (*Salmo trutta*) and the second is rainbow trout (*Oncorhynchus mykiss*). Information on the movement patterns of brown and rainbow trout in Lake Otamangakau and Lake Te Whaiau can be found in Dedual et al. (2000).

Brown trout are more sedentary compared with rainbow trout, and spawning activities of brown trout occur between May and June whilst peak spawning of rainbow trout occurs between July and August. Furthermore, brown trout have been observed avoiding water above 19 degrees Celsius and optimum temperatures for growth are between 10–15 degrees for brown trout and 16.5–17.2 degrees for rainbow trout.

See also `lake0`.

Source

Many thanks to Dr Michel Dedual (`http://www.doc.govt.nz`) for making this data available. Help from Simeon Pattenwise is also acknowledged.

References

Dedual, M. and Maxwell, I. D. and Hayes, J. W. and Strickland, R. R. (2000). Distribution and movements of brown (*Salmo trutta*) and rainbow trout (*Oncorhynchus mykiss*) in Lake Otamangakau, central North Island, New Zealand. *New Zealand Journal of Marine and Freshwater Research*, **34**: 615–627.

Examples

```
data("trap0")
summary(trap0)
```

ugss *Undergraduate Statistics Students Lifestyle Questionnaire*

Description

About 800 students studying undergraduate statistics were asked many lifestyle questions.

Usage

```
data(ugss)
```

Format

A data frame with 804 observations on the following 29 variables.

`sex` Gender, a factor, (female or male)

`age` age in years, a numeric vector

`eyes` eye colour, a factor, (blue, brown, green, hazel or other)

`piercings` Number of body piercings, a numeric vector

`pierced` Any body piercings? a factor, (Yes or No)

`tattoos` Number of tattoos, a numeric vector

`tattooed` Any tattoos? a factor, (Yes or No)

`glasses` Wears glasses etc.? a factor, (Yes or No)

`sleep` Average number of hours of sleep per night, a numeric vector

`study` Average number of hours of study per week, a numeric vector

`tv` Average number of hours watching TV per week, a numeric vector

`movies` Number of movies seen at a cinema during the last 3 months, a numeric vector

`movies3m` Seen movies in last 3 months? a factor, (Yes or No)

`sport` Favourite sport, a factor, about 19 of them

`entertainment` Favourite entertainment, a factor, about 15 of them

`fruit` Favourite fruit a factor, about 13 of them

`income` Average income during semester per week, a numeric vector

`rent` Amount spent on rent or room and board per week, a numeric vector

`clothes` Average amount spent on clothes per month, a numeric vector

`hair` Average cost to get a hair-cut, a numeric vector

`tobacco` Average amount spent on tobacco per week, a numeric vector

`smokes` Smokes? a factor, (Yes or No)

`alcohol` Average amount spent on alcohol per week, a numeric vector

`buy.alcohol` Buys (purchases) alcohol? a factor, (Yes or No)

`sendtxt` Average number text messages sent per day, a numeric vector.

`receivetxt` Average number text messages received per day, a numeric vector.

`txts` Uses text messaging? a factor, (Yes or No)

`country` Country of birth, a factor, about 54 of them

`status` Student status, a factor, (International, NZ.Citizen, NZ.Resident)

Details

This data was collected online and anonymously in 2010. The respondents were students studying an undergraduate statistics course at a New Zealand university. Possibly there are duplicate students (due to failing and re-enrolling). All monies are in NZD. Note the data has had minimal checking. Most numerical variables tend to have measurement error, and all of them happen to be all integer-valued.

Examples

```
summary(ugss)
```

---

vtinpat *Vermont Hospital Inpatient Data*

Description

Information on inpatients discharged from hospitals in Vermont, USA, 2012.

Usage

```
data(vtinpat)
```

Format

A data frame with 52206 observations on the following 7 variables.

`hospital` a factor with levels `1` = Northwestern Medical Center, `2` = North Country Hospital and Health Center, `3` = Northeastern Vermont Regional Hospital, `4` = Copley Hospital, `5` = Fletcher Allen Health Care, `6` = Central Vermont Hospital, `8` = Rutland Regional Medical Center, `9` = Porter Medical Center, `10` = Gifford Memorial Hospital, `11` = Mount Ascutney Hospital and Health Center, `12` = Springfield Hospital, `14` = Grace Cottage Hospital, `15` = Brattleboro Memorial Hospital, `16` = Southwestern Vermont Medical Center

`admit` a factor with levels `1` = Emergency, `2` = Urgent, `3` = Elective, `4`, Newborn, `5` = Trauma

`age.group` a factor with levels `1` = Under 1, `2` = 1-17, `3` = 18-24, `4` = 25-29, `5` = 30-34, `6` = 35-39, `7` = 40-44, `8` = 45-49, `9` = 50-54, `10` = 55-59, `11` = 60-64, `12` = 65-69, `13` = 70-74, `14` = 75+

`sex` a factor with levels `1` = Male, `2` = Female

`discharge` a factor with levels `1` = To another medical facility, `2` = home, `3` = against medical advice, `4` = Died, `5` = To court or law enforcement, `6` = still a patient

`diagnosis` a factor with levels `1` = Brain And C.N.S., `2` = Eye, `3` = Ear, Nose & Throat, `4` = Respiratory, `5` = Heart & Circulatory, `6` = Digestive, `7` = Liver & Pancreas, `8` = Musculoskeletal, `9` = Skin and Breast, `10` = Endocrine, `11` = Kidney & Urinary, `12` = Male Reproductive, `13` = Female Reproductive, `14` = Pregnancy, Childbirth, `15` = Neonatal, `16` = Spleen & Blood, `17` = Lymphatic, `18` = Infection, `19` = Mental Illness, `20` = Substance Abuse, `21` = Injury, Toxic Effects, `22` = Burns, `23` = Other, `24` = Trauma, `25` = H.I.V.

`los` a numeric vector, number of days spent in hospital

Details

This data set contains details on inpatients discharged from hospitals in Vermont, USA, in 2012 as part of the Vermont Uniform Hospital Discharge Data Set. The Vermont Department of Financial Regulation administers this program and the Vermont Department of Health manages the data set.

Source

Vermont department of Health, http://healthvermont.gov/research/hospital-utilization/RECENT_PU_FILES.aspx formatted into R by J. T. Gray in mid-2014.

Examples

```
summary(vtinpat)
```

---

wffc *2008 World Fly Fishing Championships Data*

Description

Capture records of the 2008 FIPS-MOUCHE World Fly Fishing Championships held in Rotorua, New Zealand during 22–30 March 2008.

Usage

```
data(wffc)
```

Format

A data frame with 4267 observations on the following 8 variables. Each row is a recorded capture.

`length` a numeric vector; length of fish in mm.

`water` a factor with levels `Waihou`, `Waimakariri`, `Whanganui`, `Otamangakau`, `Rotoaira`. These are known as Sectors IV, V, I, II, III respectively, and are also represented by the variable `sector`.

`session` a numeric vector; a value from the set 1,2,...,6. These are time ordered, and there were two sessions per competition day.

`sector` a numeric vector; a value from the set 1,2,...,5.

`beatboat` a numeric vector; beat or boat number, a value from the set 1,2,...,19.

`comid` a numeric vector; the competitor's ID number. Uniquely identifies each competitor. These ID numbers actually correspond to their rankings in the individual level.

`iname` a character vector; the individual competitor's name.

`country` a character vector; what country the competitors represented. The countries represented were Australia (AUS), Canada (CAN), Croatia (CRO), Czech Republic (CZE), England (ENG), Finland (FIN), France (FRA), Holland (NED), Ireland (IRE), Italy (ITA), Japan (JPN), Malta (MAL), New Zealand (NZL), Poland (POL), Portugal (POR), South Africa (RSA), Slovakia (SVK), USA (USA), Wales (WAL).

Details

Details may be obtained at Yee (2010) and Yee (2014). Here is a brief summary. The three competition days were 28–30 March. Each session was fixed at 9.00am–12.00pm and 2.30–5.30pm daily. One of the sessions was a rest session. Each of 19 teams had 5 members, called A, B, C, D and E (there was a composite team, actually). The scoring system allocated 100 points to each eligible fish (minimum length was 18 cm) and 20 points for each cm of its length (rounded up to the nearest centimeter). Thus a 181mm or 190mm fish was worth 480 points. Each river was divided into 19 contiguous downstream beats labelled 1,2,...,19. Each lake was fished by 9 boats, each with two competitors except for one boat which only had one. Each competitor was randomly assigned to a beat/boat.

Competitors were ranked according to their placings at each sector-session combination, and then these placings were summed. Those with the minimum total placings were the winners, thus it was not necessarily those who had the maximum points who won. For example, in Session 1 at the Waihou River, each of the 19 competitors was ranked 1 (best) to 19 (worst) according to the point system. This is the "placing" for that session. These placings were added up over the 5 sessions to give the "total placings".

All sectors have naturally wild Rainbow trout (*Oncorhynchus mykiss*) while Lake Otamangakau and the Whanganui River also holds Brown trout (*Salmo trutta*). Only these two species were targetted. The species was not recorded electronically, however a post-analysis of the paper score sheets from the two lakes showed that, approximately, less than 5 percent were Brown trout. It may be safely assumed that all the Waihou and Waimakariri fish were Rainbow trout. The gender of the fish were also not recorded electronically, and anyway, distinguishing between male and female was very difficult for small fish.

Although species and gender data were supposed to have been collected at the time of capture the quality of these variables is rather poor and furthermore they were not recorded electronically.

Note that some fish may have been caught more than once, hence these data do not represent individual fish but rather recorded captures.

Note also that a few internal discrepancies may be found within and between the data frames `wffc`, `wffc.nc`, `wffc.indiv`, `wffc.teams`. This is due to various reasons, such as competitors being replaced by reserves when sick, fish that were included or excluded upon the local judge's decision, competitors who fished two hours instead of three by mistake, etc. The data has already been cleaned of errors and internal inconsistencies but a few may remain.

Source

This data frame was adapted from the WFFC's spreadsheet. Special thanks goes to Paul Dewar, Jill Mandeno, Ilkka Pirinen, and the other members of the Organising Committee of the 28th FIPS-

Mouche World Fly Fishing Championships for access to the data. The assistance and feedback of Colin Shepherd is gratefully acknowledged.

References

Yee, T. W. (2010) VGLMs and VGAMs: an overview for applications in fisheries research. *Fisheries Research*, **101**, 116–126.

Yee, T. W. (2014) Scoring rules, and the role of chance: analysis of the 2008 World Fly Fishing Championships. *In preparation*.

See Also

wffc.indiv, wffc.teams, wffc.nc, wffc.P1, lake0.

Examples

```
summary(wffc)
with(wffc, table(water, session))

# Obtain some simple plots
waihou <- subset(wffc, water == "Waihou")
waimak <- subset(wffc, water == "Waimakariri")
whang  <- subset(wffc, water == "Whanganui")
otam   <- subset(wffc, water == "Otamangakau")
roto   <- subset(wffc, water == "Rotoaira")
minlength <- min(wffc[, "length"])
maxlength <- max(wffc[, "length"])
nwater <- c("Waihou"    = nrow(waihou), "Waimakariri" = nrow(waimak),
            "Whanganui" = nrow(whang),  "Otamangakau" = nrow(otam),
            "Rotoaira"  = nrow(roto))
## Not run:
par(mfrow = c(2, 3), las = 1)
# Overall distribution of length
with(wffc, boxplot(length/10 ~ water, ylim = c(minlength, maxlength)/10,
                   border = "blue", main = "Length (cm)", cex.axis = 0.5))

# Overall distribution of LOG length
with(wffc, boxplot(length/10 ~ water, ylim = c(minlength, maxlength)/10,
                   border = "blue", log = "y", cex.axis = 0.5,
                   main = "Length (cm) on a log scale"))

# Overall distribution of number of captures
pie(nwater, border = "blue", main = "Proportion of captures",
    labels = names(nwater), density = 10, col = 1:length(nwater),
    angle = 85+30* 1:length(nwater))

# Overall distribution of number of captures
with(wffc, barplot(nwater, main = "Number of captures", cex.names = 0.5,
                   col = "lightblue"))

# Overall distribution of proportion of number of captures
with(wffc, barplot(nwater / sum(nwater), cex.names = 0.5, col = "lightblue",
```

```
                    main = "Proportion of captures"))
# An interesting lake
with(roto, hist(length/10, xlab = "Fish length (cm)", col = "lightblue",
                breaks = seq(18, 70, by = 3), prob = TRUE, ylim = c(0, 0.08),
                border = "blue", ylab = "", main = "Lake Rotoaira", lwd = 2))

## End(Not run)
```

`wffc.indiv` *2008 World Fly Fishing Championships (Individual results) Data*

Description

Individual competitors' results of the 2008 FIPS-MOUCHE World Fly Fishing Championships held in Rotorua, New Zealand during 22–30 March 2008.

Usage

```
data(wffc.indiv)
```

Format

A data frame with 99 observations on the following 8 variables. Some of these variable are described in `wffc`.

`totalPlacings` a numeric vector; these are the summed placings over the 5 sessions.

`points` a numeric vector.

`noofcaptures` a numeric vector.

`longest` a numeric vector.

`individual` a numeric vector; did the competitor fish in a team or as an individual? (one team was made of composite countries due to low numbers).

`country` a character vector.

`iname` a character vector.

`comid` a numeric vector.

Details

This data frame gives the individual results of the competition. See also `wffc` and `wffc.teams` for more details and links.

References

Yee, T. W. (2010) VGLMs and VGAMs: an overview for applications in fisheries research. *Fisheries Research*, **101**, 116–126.

Examples

```
summary(wffc.indiv)
```

wffc.nc *2008 World Fly Fishing Championships (Number of captures) Data*

Description

Number of captures in the 2008 FIPS-MOUCHE World Fly Fishing Championships held in Rotorua, New Zealand during 22–30 March 2008.

Usage

```
data(wffc.nc)
```

Format

A data frame with 475 observations on the following 7 variables. Most of these variable are described in wffc. Each row is sorted by sector, session and beat.

`sector` a numeric vector.

`session` a numeric vector.

`beatboat` a numeric vector.

`numbers` a numeric vector.

`comid` a numeric vector.

`iname` a character vector.

`country` a character vector.

Details

This data frame was obtained by processing wffc. The key variable is `numbers`, which is sector-session-beat specific.

Note that some fish may have been caught more than once, hence these data do not represent individual fish.

References

Yee, T. W. (2010) VGLMs and VGAMs: an overview for applications in fisheries research. *Fisheries Research*, **101**, 116–126.

See Also

DeLury, lake0.

Examples

```
xtabs( ~ sector + session, wffc.nc)
```

wffc.points *Point System for the 2008 World Fly Fishing Championships*

Description

Point system for the 2008 World Fly Fishing Championships: current and some proposals.

Usage

```
wffc.P1(length, c1 = 100, min.eligible = 0.18, ppm = 2000)
wffc.P2(length, c1 = 100, min.eligible = 0.18, ppm = 2000,
        c.quad = 12700)
wffc.P3(length, c1 = 100, min.eligible = 0.18, ppm = 2000)
wffc.P1star(length, c1 = 100, min.eligible = 0.18, ppm = 2000)
wffc.P2star(length, c1 = 100, min.eligible = 0.18, ppm = 2000,
            c.quad = 12700)
wffc.P3star(length, c1 = 100, min.eligible = 0.18, ppm = 2000)
```

Arguments

length	Length of the fish, in meters. Numeric vector.
c1	Points added to each eligible fish.
min.eligible	The 2008 WFFC regulations stipulated that the smallest eligible fish was 0.180 m, which is 180 mm.
ppm	Points per meter of length of the fish.
c.quad	Constants for the quadratic terms. The defaults mean that a fish twice the minimum legal size will award about 50 percent more points compared to wffc.P1() and wffc.P1star(). See below for examples.

Details

The official website contains a document with the official rules and regulations of the competition. The function wffc.P1() implements the current WFFC point system, and is 'discrete' in that fish lengths are rounded up to the nearest centimeter (provided it is greater or equal to min.eligible m). wffc.P1star() is a continuous version of it in that it is piecewise linear with two pieces and it is discontinuous at min.eligible.

The function wffc.P2() is a new proposal which rewards catching bigger fish. It is based on a quadratic polynomial. wffc.P2star() is a continuous version of it.

The function wffc.P3() is another new proposal which rewards catching bigger fish. Named a *cumulative linear proposal*, its slope is ppm between min.eligible and 2 * min.eligible, its slope is 2 * ppm between 2 * min.eligible and 3 * min.eligible, its slope is 3 * ppm between 3 * min.eligible and 4 * min.eligible, etc. One adds the usual c1 to each eligible fish. wffc.P3star() is a continuous version of wffc.P3().

The function `wffc.P4()` is another new proposal which rewards catching bigger fish. Named a *cumulative linear proposal*, its slope is `ppm` between `min.eligible` and `2 * min.eligible`, its slope is `2 * ppm` between `2 * min.eligible` and `1.5 * min.eligible`, its slope is `3 * ppm` between `1.5 * min.eligible` and `2 * min.eligible`, etc. One adds the usual `c1` to each eligible fish. `wffc.P4star()` is a continuous version of `wffc.P4()`.

Value

A vector with the number of points.

Note

`wffc.P2` and `wffc.P2star` may change in the future, as well as possibly `wffc.P3` and `wffc.P3star` and `wffc.P4` and `wffc.P4star`.

Author(s)

T. W. Yee.

References

Yee, T. W. (2014) Scoring rules, and the role of chance: analysis of the 2008 World Fly Fishing Championships. *In preparation*.

See Also

`wffc`.

Examples

```
## Not run:  fishlength <- seq(0.0, 0.72, by = 0.001)
plot(fishlength, wffc.P2star(fishlength), type = "l", col = "blue",
     las = 1, lty = "dashed", lwd = 2, las = 1, cex.main = 0.8,
     xlab = "Fish length (m)", ylab = "Competition points",
     main = "Current (red) and proposed (blue and green) WFFC point system")
lines(fishlength, wffc.P1star(fishlength), type = "l", col = "red", lwd = 2)
lines(fishlength, wffc.P3star(fishlength), type = "l", col = "darkgreen",
      lwd = 2, lty = "dashed")
lines(fishlength, wffc.P4star(fishlength), type = "l", col = "orange",
      lwd = 2, lty = "dashed")
abline(v = (1:4) * 0.18, lty = "dotted")
abline(h = (1:13) * wffc.P1star(0.18), lty = "dotted")
## End(Not run)

# Successive slopes:
(wffc.P1star((2:8)*0.18) - wffc.P1star((1:7)*0.18)) / (0.18 * 2000)
(wffc.P3star((2:8)*0.18) - wffc.P3star((1:7)*0.18)) / (0.18 * 2000)
(wffc.P4star((2:8)*0.18) - wffc.P4star((1:7)*0.18)) / (0.18 * 2000)

# More successive slopes:
MM2 <- 0.18 / 2
ind1 <- 2:12
```

```
(wffc.P4star((ind1)*MM2) - wffc.P4star((ind1-1)*MM2)) / (MM2 * 2000)

# About 50 percent more points:
wffc.P2    (2 * 0.18) / wffc.P1    (2 * 0.18)
wffc.P2star(2 * 0.18) / wffc.P1star(2 * 0.18)
```

`wffc.teams` *2008 World Fly Fishing Championships (Team results) Data*

Description

Team results of the 2008 FIPS-MOUCHE World Fly Fishing Championships held in Rotorua, New Zealand during 22–30 March 2008.

Usage

```
data(wffc.teams)
```

Format

A data frame with 18 observations on the following 5 variables. Some of these variable are described in `wffc`.

`country` a character vector.

`totalPlacings` a numeric vector; these are the summed placings over the 5 sessions and 5 team members.

`points` a numeric vector; see `wffc`.

`noofcaptures` a numeric vector.

`longestfish` a numeric vector.

Details

This data frame gives the team results of the competition. See also `wffc` and `wffc.indiv` for more details and links.

Examples

```
wffc.teams
```

xs.nz *Cross-sectional Data from the New Zealand Population*

Description

A cross-sectional data set of a workforce company, plus another health survey, in New Zealand during the 1990s,

Usage

```
data(xs.nz)
```

Format

A data frame with 10529 observations on the following 58 variables. For binary variables, a "1" or TRUE means yes, and "0" or FALSE means no. Also, "D" means don't know, and "-" means not applicable. The pregnancy questions were administered to women only.

regnum a numeric vector, a unique registration number. This differs from their original registration number, and the rows are sorted by their new registration number.

study1 a logical vector, Study 1 (workforce) or Study 2?

age a numeric vector, age in years.

sex a factor with levels F and M.

pulse a numeric vector, beats per minute.

sbp a numeric vector, systolic blood pressure (mm Hg).

dbp a numeric vector, diastolic blood pressure (mm Hg).

cholest a numeric vector, cholesterol (mmol/L).

height a numeric vector, in m.

weight a numeric vector, in kg.

fh.heartdisease a factor with levels 0, 1, D. Has a family history of heart disease (heart attack, angina, or had a heart bypass operation) within the immediate family (brother, sister, father or mother, blood relatives only)? Note that D means: do not know.

fh.age a factor, following from fh.heartdisease, if yes, how old was the family member when it happened (if more than one family member, give the age of the youngest person)?

fh.cancer a factor with levels 0, 1, D. Has a family history of cancer within the immediate family (blood relatives only)? Note that D means: do not know.

heartattack a numeric vector, have you ever been told by a doctor that you have had a heart attack ("coronary")?

stroke a numeric vector, have you ever been told by a doctor that you have had a stroke?

diabetes a numeric vector, have you ever been told by a doctor that you have had diabetes?

hypertension a numeric vector, have you ever been told by a doctor that you have had high blood pressure (hypertension)?

highchol a numeric vector, have you ever been told by a doctor that you have had high cholesterol?

asthma a numeric vector, have you ever been told by a doctor that you have had asthma?

cancer a numeric vector, have you ever been told by a doctor that you have had cancer?

acne a numeric vector, have you ever received treatment from a doctor for acne (pimples)?

sunburn a numeric vector, have you ever received treatment from a doctor for sunburn?

smokeever a numeric vector, have you ever smoked tailor-made or roll-you-own cigarettes once a week or more?

smokenow a numeric vector, do you smoke tailor-made or roll-you-own cigarettes now?

smokeagequit a factor, if no to smokenow, how old were you when you stopped smoking?

smokeyears a numeric vector, if yes to smokeever, for how many years altogether have you smoked tailor-made or roll-you-own cigarettes?

drinkmonth a numeric vector, do you drink alcohol once a month or more?

drinkfreqweek a numeric vector, if yes to drinkmonth, about how often do you drink alcohol (days per week)? Note: 0.25 is once a month, 0.5 is once every two weeks, 1 is once a week, 2.5 is 2-3 days a week, 4.5 is 4-5 days a week, 6.5 is 6-7 days a week.

Further note: 1 can, small bottle or handle of beer or home brew = 1 drink, 1 quart bottle of beer = 2 drinks, 1 jug of beer = 3 drinks, 1 flagon/peter of beer = 6 drinks, 1 glass of wine, sherry = 1 drink, 1 bottle of wine = 6 drinks, 1 double nip of spirits = 1 drink.

drinkweek a numeric vector, how many drinks per week, on average. This is the average daily amount of drinks multiplied by the frequency of drinking per week. See drinkfreqweek on what constitutes a 'drink'.

drinkmaxday a numeric vector, in the last three months, what is the largest number of drinks that you had on any one day? Warning: some values are considered unrealistically excessive.

pregnant a factor, have you ever been pregnant for more than 5 months?

pregfirst a factor, if yes to pregnant, how old were you when your first baby was born (or you had a miscarriage after 5 months)?

preglast a factor, how old were you when your last baby was born (or you had a miscarriage after 5 months)?

babies numeric, how many babies have you given birth to?

moody a numeric vector, does your mood often go up or down?

miserable a numeric vector, do you ever feel 'just miserable' for no reason?

hurt a numeric vector, are your feelings easily hurt?

fedup a numeric vector, do you often feel 'fed up'?

nervous a numeric vector, would you call yourself a nervous person?

worrier a numeric vector, are you a worrier?

worry a numeric vector, do you worry about awful things that might happen?

tense a numeric vector, would you call yourself tense or 'highly strung'?

embarrassed a numeric vector, do you worry too long after an embarrassing experience?

nerves a numeric vector, do you suffer from 'nerves'?

`nofriend` a numeric vector, do you have a friend or family member that you can talk to about problems or worries that you may have? The value 1 effectively means `"no"`, i.e., s/he has no friend or friends.

`depressed` a numeric vector, in your lifetime, have you ever had two weeks or more when nearly every day you felt sad or depressed?

`exervig` a numeric vector, how many hours per week would you do any vigorous activity or exercise either at work or away from work that makes you breathe hard and sweat? Values here ought be be less than 168.

`exermod` a numeric vector, how many hours per week would you do any moderate activity or exercise such as brisk walking, cycling or mowing the lawn? Values here ought be be less than 168.

`feethour` a numeric vector, on an average work day, how long would you spend on your feet, either standing or moving about?

`ethnicity` a factor with 4 levels, what ethnic group do you belong to? `European` = European (NZ European or British or other European), `Maori` = Maori, `Polynesian` = Pacific Island Polynesian, `Other` = Other (Chinese, Indian, Other).

`sleep` a numeric vector, how many hours do you usually sleep each night?

`snore` a factor with levels `0`, `1`, `D`. Do you usually snore? Note that `D` means: do not know.

`cat` a numeric vector, do you have a household pet cat?

`dog` a numeric vector, do you have a household pet dog?

`hand` a factor with levels `right` = right, `left` = left, `both` = either. Are you right-handed, left-handed, or no preference for left or right?

`numhouse` an ordered factor with 4 levels: `1` = 1, `2` = 2, `3` = 3, `4+` = four or more. how many people (including yourself) usually live in your house?

`marital` a factor with 4 levels: `single` = single, `married` = married or living with a partner, `separated` = separated or divorced, `widowed` = widowed.

`educ` an ordered factor with 4 levels: `primary` = Primary school, `secondary` = High school/secondary school, `polytechnic` = Polytechnic or similar, `university` = University. What was the highest level of education you received?

Details

The data frame is a subset of the entire data set which was collected from a confidential self-administered questionnaire administered in a large New Zealand workforce observational study conducted during 1992–3. The data were augmented by a second study consisting of retirees. The data can be considered a reasonable representation of the white male New Zealand population in the early 1990s. There were physical, lifestyle and psychological variables that were measured. The psychological variables were headed "Questions about your feelings".

Although some data cleaning was performed and logic checks conducted, anomalies remain. Some variables, of course, are subject to a lot of measurement error and bias. It is conceivable that some participants had poor reading skills!

Warning

More variables may be added in the future and these may be placed in any column position. Therefore references such as `xs.nz[, 12]` are dangerous. Also, variable names may change in the future as well as their format or internal structure, e.g., `factor` versus `numeric`.

Note

More error checking are needed for the pregnancy and smoking variables.

Source

Originally, Clinical Trials Research Unit, University of Auckland, New Zealand, http://www.ctru.auckland.ac.nz. Originally much of the error checking and formatting was performed by Stephen Vander Hoorn. Lately (2014), more changes and error checks were made to the data by James T. Gray.

References

MacMahon, S., Norton, R., Jackson, R., Mackie, M. J., Cheng, A., Vander Hoorn, S., Milne, A., McCulloch, A. (1995) Fletcher Challenge-University of Auckland Heart & Health Study: design and baseline findings. *New Zealand Medical Journal*, **108**, 499–502.

See Also

chest.nz.

Examples

```
data(xs.nz)
summary(xs.nz)
```

Index