Package 'VGAMdata'

July 21, 2025

Version 1.1-13

Date 2025-02-06

Title Data Supporting the 'VGAM' Package

Author Thomas Yee [aut, cre, cph] (ORCID:
<https://orcid.org/0000-0002-9970-3907>),
James Gray [dtc]

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

Depends R (>= 4.0.0), methods, stats, VGAM

Description Mainly data sets to accompany the VGAM package and
the book ``Vector Generalized Linear and
Additive Models: With an Implementation in R'' (Yee, 2015)
<DOI:10.1007/978-1-4939-2818-7>.
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). This package now contains some
old VGAM family functions which have been replaced by newer
ones (often because they are now special cases).

License GPL-2

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

Repository CRAN

LazyLoad yes

LazyData yes

NeedsCompilation no

Date/Publication 2025-02-06 11:20:03 UTC

Contents

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

Description

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

Details

This package mainly 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. Recently, some older or less-used **VGAM** family functions have been shifted into **VGAMdata**, and this is likely to continue in the future.

Author(s)

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

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

References

Yee, T. W. (2015). Vector Generalized Linear and Additive Models: With an Implementation in R. New York, USA: *Springer*.

See Also

VGAM-package.

Examples

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

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

`airbnb.ac` *Airbnb Accommodation in Two Sardinian Cities*

Description

The `airbnb.ac` data frame has 18159 rows and 9 columns.

Usage

```
data(airbnb.ac)
```

Format

This data frame contains the following columns:

LOS length of stay, in days.

discountWeek Multiplicative factor for the discount for booking one week: $1-$ (published weekly rate) / ($7 \times$ published nightly rate), e.g., 0.2 means a 20 percent savings off the regular price.

NumberofReviews Number of reviews on the website.

PriceAvg Average price per night per person, in Euros.

Bedrooms Number of bedrooms.

Superhost Logical. Superhost?

MinimumStay Minimum stay period, in days.

MaxGuests Maximum number of guests.

City Character, values are `"Alghero"` and `"Cagliari"`.

Details

The data frame comprises Airbnb bookings in two cities located in Sardinia, Italy. The stays were during the whole of 2016. Stays of 30 days or longer and any rows with missing variables were deleted from the original source. Variable LOS exhibits heaping at the values 7 and 14 days.

Source

The data was obtained with permission from Luca Frigau, University of Cagliari, from https://www.airbnb.com.

See Also

gaitdzeta, flamingo.

Examples

```
## Not run:
mytab <- with(subset(airbnb.ac, City == "Alghero"), table(LOS))
plot(prop.table(mytab), col = "blue", ylab = "Proportion")

## End(Not run)
```

Alap *The Laplace Distribution*

Description

Density, distribution function, quantile function and random generation for the 3-parameter asymmetric Laplace distribution with location parameter location, scale parameter scale, and asymmetry parameter kappa.

Usage

```
dalap(x, location = 0, scale = 1, tau = 0.5, kappa = sqrt(tau/(1-tau)),
      log = FALSE)
palap(q, location = 0, scale = 1, tau = 0.5, kappa = sqrt(tau/(1-tau)),
      lower.tail = TRUE, log.p = FALSE)
qalap(p, location = 0, scale = 1, tau = 0.5, kappa = sqrt(tau/(1-tau)),
      lower.tail = TRUE, log.p = FALSE)
ralap(n, location = 0, scale = 1, tau = 0.5, kappa = sqrt(tau/(1-tau)))
```

Arguments

x, q	vector of quantiles.
p	vector of probabilities.
n	number of observations. If length(n) > 1 then the length is taken to be the number required.

location	the location parameter ξ.
scale	the scale parameter σ. Must consist of positive values.
tau	the quantile parameter τ. Must consist of values in $(0, 1)$. This argument is used to specify kappa and is ignored if kappa is assigned.
kappa	the asymmetry parameter κ. Must consist of positive values.
log	if TRUE, probabilities p are given as log(p).
lower.tail, log.p	Same meaning as in pnorm or qnorm.

Details

There are many variants of asymmetric Laplace distributions (ALDs) and this one is known as *the* ALD by Kotz et al. (2001). See alaplace3, the **VGAM** family function for estimating the three parameters by maximum likelihood estimation, for formulae and details. The ALD density may be approximated by dextlogF.

Value

dalap gives the density, palap gives the distribution function, qalap gives the quantile function, and ralap generates random deviates.

Author(s)

T. W. Yee and Kai Huang

References

Kotz, S., Kozubowski, T. J. and Podgorski, K. (2001). *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*, Boston: Birkhauser.

See Also

alaplace3, dextlogF, extlogF1.

Examples

```
x <- seq(-5, 5, by = 0.01)
loc <- 0; sigma <- 1.5; kappa <- 2
## Not run:  plot(x, dalap(x, loc, sigma, kappa = kappa), type = "l",
     main = "Blue is density, orange is the CDF",
     ylim = c(0, 1), sub = "Purple are 5, 10, ..., 95 percentiles",
     las = 1, ylab = "", cex.main = 0.5, col = "blue")
abline(h = 0, col = "blue", lty = 2)
lines(qalap(seq(0.05, 0.95, by = 0.05), loc, sigma, kappa = kappa),
      dalap(qalap(seq(0.05, 0.95, by = 0.05), loc, sigma, kappa = kappa),
            loc, sigma, kappa = kappa), col="purple", lty=3, type = "h")
lines(x, palap(x, loc, sigma, kappa = kappa), type = "l", col = "orange")
abline(h = 0, lty = 2)
```

```
## End(Not run)

pp <- seq(0.05, 0.95, by = 0.05)  # Test two functions
max(abs(palap(qalap(pp, loc, sigma, kappa = kappa),
              loc, sigma, kappa = kappa) - pp))  # Should be 0
```

---

alaplace *Asymmetric Laplace Distribution Family Functions*

Description

Maximum likelihood estimation of the 1, 2 and 3-parameter asymmetric Laplace distributions (ALDs). The 2-parameter ALD may, with trepidation and lots of skill, sometimes be used as an approximation of quantile regression.

Usage

```
alaplace1(tau = NULL, llocation = "identitylink",
          ilocation = NULL, kappa = sqrt(tau/(1 - tau)), Scale.arg = 1,
          ishrinkage = 0.95, parallel.locat = TRUE  ~ 0, digt = 4,
          idf.mu = 3, zero = NULL, imethod = 1)

alaplace2(tau = NULL,  llocation = "identitylink", lscale = "loglink",
          ilocation = NULL, iscale = NULL, kappa = sqrt(tau/(1 - tau)),
          ishrinkage = 0.95,
          parallel.locat =  TRUE ~ 0,
          parallel.scale = FALSE ~ 0,
          digt = 4, idf.mu = 3, imethod = 1, zero = "scale")

alaplace3(llocation = "identitylink", lscale = "loglink",
          lkappa = "loglink", ilocation = NULL, iscale = NULL,
          ikappa = 1, imethod = 1, zero = c("scale", "kappa"))
```

Arguments

tau, kappa
: Numeric vectors with $0 < \tau < 1$ and $\kappa > 0$. Most users will only specify `tau` since the estimated location parameter corresponds to the τth regression quantile, which is easier to understand. See below for details.

llocation, lscale, lkappa
: Character. Parameter link functions for location parameter ξ, scale parameter σ, asymmetry parameter κ. See `Links` for more choices. For example, the argument `llocation` can help handle count data by restricting the quantiles to be positive (use `llocation = "loglink"`). However, `llocation` is best left alone since the theory only works properly with the identity link.

ilocation, iscale, ikappa
: Optional initial values. If given, it must be numeric and values are recycled to the appropriate length. The default is to choose the value internally.

parallel.locat, parallel.scale
: See the parallel argument of CommonVGAMffArguments. These arguments apply to the location and scale parameters. It generally only makes sense for the scale parameters to be equal, hence set parallel.scale = TRUE. Note that assigning parallel.locat the value TRUE circumvents the seriously embarrassing quantile crossing problem because all constraint matrices except for the intercept correspond to a parallelism assumption.

imethod
: Initialization method. Either the value 1, 2, 3 or 4.

idf.mu
: Degrees of freedom for the cubic smoothing spline fit applied to get an initial estimate of the location parameter. See vsmooth.spline. Used only when imethod = 3.

ishrinkage
: How much shrinkage is used when initializing ξ. The value must be between 0 and 1 inclusive, and a value of 0 means the individual response values are used, and a value of 1 means the median or mean is used. This argument is used only when imethod = 4. See CommonVGAMffArguments for more information.

Scale.arg
: The value of the scale parameter σ. This argument may be used to compute quantiles at different τ values from an existing fitted alaplace2() model (practical only if it has a single value). If the model has parallel.locat = TRUE then only the intercept need be estimated; use an offset. See below for an example.

digt
: Passed into Round as the digits argument for the tau values; used cosmetically for labelling.

zero
: See CommonVGAMffArguments for more information. Where possible, the default is to model all the σ and κ as an intercept-only term. See CommonVGAMffArguments for more information.

Details

These **VGAM** family functions implement one variant of asymmetric Laplace distributions (ALDs) suitable for quantile regression. Kotz et al. (2001) call it *the* ALD. Its density function is

$$f(y;\xi,\sigma,\kappa) = \frac{\sqrt{2}}{\sigma}\,\frac{\kappa}{1+\kappa^2}\,\exp\left(-\frac{\sqrt{2}}{\sigma\,\kappa}|y-\xi|\right)$$

for $y \leq \xi$, and

$$f(y;\xi,\sigma,\kappa) = \frac{\sqrt{2}}{\sigma}\,\frac{\kappa}{1+\kappa^2}\,\exp\left(-\frac{\sqrt{2}\,\kappa}{\sigma}|y-\xi|\right)$$

for $y > \xi$. Here, the ranges are for all real y and ξ, positive σ and positive κ. The special case $\kappa = 1$ corresponds to the (symmetric) Laplace distribution of Kotz et al. (2001). The mean is $\xi + \sigma(1/\kappa - \kappa)/\sqrt{2}$ and the variance is $\sigma^2(1+\kappa^4)/(2\kappa^2)$. The enumeration of the linear/additive predictors used for alaplace2() is the first location parameter followed by the first scale parameter, then the second location parameter followed by the second scale parameter, etc. For alaplace3(), only a vector response is handled and the last (third) linear/additive predictor is for the asymmetry parameter.

It is known that the maximum likelihood estimate of the location parameter ξ corresponds to the regression quantile estimate of the classical quantile regression approach of Koenker and Bassett (1978). An important property of the ALD is that $P(Y \leq \xi) = \tau$ where $\tau = \kappa^2/(1+\kappa^2)$ so

that $\kappa = \sqrt{\tau/(1-\tau)}$. Thus `alaplace2()` might be used as an alternative to `rq` in the **quantreg** package, although scoring is really an unsuitable algorithm for estimation here.

Both `alaplace1()` and `alaplace2()` can handle multiple responses, and the number of linear/additive predictors is dictated by the length of `tau` or `kappa`. The functions `alaplace1()` and `alaplace2()` can also handle multiple responses (i.e., a matrix response) but only with a *single-valued* `tau` or `kappa`.

Value

An object of class `"vglmff"` (see `vglmff-class`). The object is used by modelling functions such as `vglm` and `vgam`.

In the `extra` slot of the fitted object are some list components which are useful, e.g., the sample proportion of values which are less than the fitted quantile curves.

Warning

These functions are *experimental* and especially subject to change or withdrawal. The usual MLE regularity conditions do *not* hold for this distribution so that misleading inferences may result, e.g., in the `summary` and `vcov` of the object. The 1-parameter ALD can be approximated by `extlogF1` which has continuous derivatives and is recommended over `alaplace1`.

Care is needed with `tau` values which are too small, e.g., for count data with `llocation = "loglink"` and if the sample proportion of zeros is greater than `tau`.

Note

These **VGAM** family functions use Fisher scoring. Convergence may be slow and half-stepping is usual (although one can use `trace = TRUE` to see which is the best model and then use `maxit` to choose that model) due to the regularity conditions not holding. Often the iterations slowly crawl towards the solution so monitoring the convergence (set `trace = TRUE`) is highly recommended. Instead, `extlogF1` is recommended.

For large data sets it is a very good idea to keep the length of `tau`/`kappa` low to avoid large memory requirements. Then for `parallel.locat = FALSE` one can repeatedly fit a model with `alaplace1()` with one τ at a time; and for `parallel.locat = TRUE` one can refit a model with `alaplace1()` with one τ at a time but using offsets and an intercept-only model.

A second method for solving the noncrossing quantile problem is illustrated below in Example 3. This is called the *accumulative quantile method* (AQM) and details are in Yee (2015). It does not make the strong parallelism assumption.

The functions `alaplace2()` and `laplace` differ slightly in terms of the parameterizations.

Author(s)

Thomas W. Yee

References

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, **46**, 33–50.

Kotz, S., Kozubowski, T. J. and Podgorski, K. (2001). *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*, Boston: Birkhauser.

See Also

ralap, laplace, extlogF1, CommonVGAMffArguments, lms.bcn, amlnormal, sc.studentt2, simulate.vlm.

Examples

```
## Not run:
# Example 1: quantile regression with smoothing splines
set.seed(123); adata <- data.frame(x2 = sort(runif(n <- 500)))
mymu <- function(x) exp(-2 + 6*sin(2*x-0.2) / (x+0.5)^2)
adata <- transform(adata, y = rpois(n, lambda = mymu(x2)))
mytau <- c(0.25, 0.75); mydof <- 4

fit <- vgam(y ~ s(x2, df = mydof), data=adata, trace=TRUE, maxit = 900,
            alaplace2(tau = mytau, llocat = "loglink",
                      parallel.locat = FALSE))
fitp <- vgam(y ~ s(x2, df = mydof), data = adata, trace=TRUE, maxit=900,
     alaplace2(tau = mytau, llocat = "loglink", parallel.locat = TRUE))

par(las = 1); mylwd <- 1.5
with(adata, plot(x2, jitter(y, factor = 0.5), col = "orange",
                 main = "Example 1; green: parallel.locat = TRUE",
                 ylab = "y", pch = "o", cex = 0.75))
with(adata, matlines(x2, fitted(fit ), col = "blue",
                     lty = "solid", lwd = mylwd))
with(adata, matlines(x2, fitted(fitp), col = "green",
                     lty = "solid", lwd = mylwd))
finexgrid <- seq(0, 1, len = 1001)
for (ii in 1:length(mytau))
  lines(finexgrid, qpois(p = mytau[ii], lambda = mymu(finexgrid)),
        col = "blue", lwd = mylwd)
fit@extra  # Contains useful information


# Example 2: regression quantile at a new tau value from an existing fit
# Nb. regression splines are used here since it is easier.
fitp2 <- vglm(y ~ sm.bs(x2, df = mydof), data = adata, trace = TRUE,
              alaplace1(tau = mytau, llocation = "loglink",
                        parallel.locat = TRUE))

newtau <- 0.5  # Want to refit the model with this tau value
fitp3 <- vglm(y ~ 1 + offset(predict(fitp2)[, 1]),
              alaplace1(tau = newtau, llocation = "loglink"), adata)
with(adata, plot(x2, jitter(y, factor = 0.5), col = "orange",
               pch = "o", cex = 0.75, ylab = "y",
               main = "Example 2; parallel.locat = TRUE"))
with(adata, matlines(x2, fitted(fitp2), col = "blue",
                     lty = 1, lwd = mylwd))
```

```
with(adata, matlines(x2, fitted(fitp3), col = "black",
                     lty = 1, lwd = mylwd))


# Example 3: noncrossing regression quantiles using a trick: obtain
# successive solutions which are added to previous solutions; use a log
# link to ensure an increasing quantiles at any value of x.

mytau <- seq(0.2, 0.9, by = 0.1)
answer <- matrix(0, nrow(adata), length(mytau))  # Stores the quantiles
adata <- transform(adata, offsety = y*0)
usetau <- mytau
for (ii in 1:length(mytau)) {
# cat("\n\nii  = ", ii, "\n")
  adata <- transform(adata, usey = y-offsety)
  iloc <- ifelse(ii == 1, with(adata, median(y)), 1.0)  # Well-chosen!
  mydf <- ifelse(ii == 1, 5, 3)  # Maybe less smoothing will help
  fit3 <- vglm(usey ~ sm.ns(x2, df = mydf), data = adata, trace = TRUE,
            alaplace2(tau = usetau[ii], lloc = "loglink", iloc = iloc))
  answer[, ii] <- (if(ii == 1) 0 else answer[, ii-1]) + fitted(fit3)
  adata <- transform(adata, offsety = answer[, ii])
}

# Plot the results.
with(adata, plot(x2, y, col = "blue",
     main = paste("Noncrossing and nonparallel; tau  = ",
                  paste(mytau, collapse = ", "))))
with(adata, matlines(x2, answer, col = "orange", lty = 1))

# Zoom in near the origin.
with(adata, plot(x2, y, col = "blue", xlim = c(0, 0.2), ylim = 0:1,
     main = paste("Noncrossing and nonparallel; tau  = ",
                  paste(mytau, collapse = ", "))))
with(adata, matlines(x2, answer, col = "orange", lty = 1))

## End(Not run)
```

---

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)
```

---

bd.us *Births and Deaths of 348 Notable Americans*

Description

A 12 x 12 table of the Births and Deaths of 348 Notable Americans. The rows and columns are for each month.

Usage

```
data(bd.us)
```

Format

The format is: chr "bd.us"

Details

Rows denote the month of birth; columns for the month of death. These data appear as Table 1 of Phillips and Feldman (1973), who collected the data from Morris (1965). Not all of the 400 people were used because some had not died by that time and other individuals lacked the information.

Source

Phillips, D. P. and Feldman, K. A. (1973) A dip in deaths before ceremonial occasions: Some new relationships between social integration and mortality, American Sociological Review, **38**, 678–696.

Morris, R. B. (Ed.) (1965) *Four Hundred Notable Americans*. Harper and Row: New York, USA.

See Also

`rcim`.

Examples

```
print(bd.us)
sum(bd.us)
rowSums(bd.us)
colSums(bd.us)
```

---

belcap *BELCAP Dental Data*

Description

A prospective data set containing the DMFT index of children 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 persons dental health. Only the eight deciduous molars are considered, so the lowest value is 0, and the highest is 8.

Source

`https://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) ...
```

Bell *The Bell Distribution*

Description

Density, and random generation for the Topp-Leone distribution.

Usage

```
dbell(x, shape, log = FALSE)
rbell(n, shape)
```

Arguments

x, n	Same as Uniform.
shape	the (shape) parameter, which is positive.
log	Logical. If log = TRUE then the logarithm of the density is returned.

Details

See bellff, the **VGAM** family function for estimating the parameter s by maximum likelihood estimation.

Value

dbell gives the density, rbell generates random deviates. If shape is large then rbell will become computationally expensive.

See Also

bell.

Examples

```
## Not run: plot(0:15, dbell(0:15, shape = 1.5), type = "h", col = "blue")
```

---

bellff *Bell Distribution Family Function*

Description

Estimating the shape parameter of the Bell distribution by maximum likelihood estimation.

Usage

```
bellff(lshape = "loglink", zero = NULL, gshape = expm1(1.6 * ppoints(7)))
```

Arguments

lshape, zero, gshape

More information is at CommonVGAMffArguments.

Details

The Bell distribution has a probability density function that can be written

$$f(y; s) = \frac{s^y \exp(1 - e^s) B_y}{y!}$$

for $y = 0(1)\infty$ and shape parameter $0 < s$. The mean of Y is $\exp(s)s$ (returned as the fitted values). Fisher-scoring is used. This **VGAM** family function handles multiple responses.

The function bell returns the first 218 Bell numbers as finite numbers, and returns Inf when its argument has a higher value. Hence this **VGAM** family function can only handle low-value counts of less than 219.

Value

An object of class "vglmff" (see vglmff-class). The object is used by modelling functions such as vglm, and vgam.

Author(s)

T. W. Yee

References

Castellares, F. and Ferrari, S. L. P. and Lemonte, A. J. (2018). On the Bell distribution and its associated regression model for count data. *Applied Mathematical Modelling*, **56**, 172–185.

See Also

bell, dbell, poissonff.

Examples

```
bdata <- data.frame(y = rbell(1000, loglink(0.5, inverse = TRUE)))
bfit <- vglm(y ~ 1, bellff, bdata, trace = TRUE, crit = "coef")
coef(bfit, matrix = TRUE)
Coef(bfit)
```

---

bigamma.mckay *Bivariate Gamma: McKay's Distribution*

Description

Estimate the three parameters of McKay's bivariate gamma distribution by maximum likelihood estimation.

Usage

```
bigamma.mckay(lscale = "loglink", lshape1 = "loglink",
              lshape2 = "loglink", iscale = NULL, ishape1 = NULL,
              ishape2 = NULL, imethod = 1, zero = "shape")
```

Arguments

lscale, lshape1, lshape2
: Link functions applied to the (positive) parameters a, p and q respectively. See Links for more choices.

iscale, ishape1, ishape2
: Optional initial values for a, p and q respectively. The default is to compute them internally.

imethod, zero
: See CommonVGAMffArguments.

Details

One of the earliest forms of the bivariate gamma distribution has a joint probability density function given by

$$f(y_1, y_2; a, p, q) = (1/a)^{p+q} y_1^{p-1} (y_2 - y_1)^{q-1} \exp(-y_2/a)/[\Gamma(p)\Gamma(q)]$$

for $a > 0$, $p > 0$, $q > 0$ and $0 < y_1 < y_2$ (Mckay, 1934). Here, Γ is the gamma function, as in gamma. By default, the linear/additive predictors are $\eta_1 = \log(a)$, $\eta_2 = \log(p)$, $\eta_3 = \log(q)$.

The marginal distributions are gamma, with shape parameters p and $p+q$ respectively, but they have a common scale parameter a. Pearson's product-moment correlation coefficient of y_1 and y_2 is $\sqrt{p/(p+q)}$. This distribution is also known as the bivariate Pearson type III distribution. Also, $Y_2 - y_1$, conditional on $Y_1 = y_1$, has a gamma distribution with shape parameter q.

Value

An object of class `"vglmff"` (see `vglmff-class`). The object is used by modelling functions such as `vglm` and `vgam`.

Note

The response must be a two column matrix where the first column is y_1 and the second y_2. It is necessary that each element of the vectors y_1 and $y_2 - y_1$ be positive. Currently, the fitted value is a matrix with two columns; the first column has values ap for the marginal mean of y_1, while the second column has values $a(p+q)$ for the marginal mean of y_2 (all evaluated at the final iteration).

Author(s)

T. W. Yee

References

McKay, A. T. (1934). Sampling from batches. *Journal of the Royal Statistical Society—Supplement*, **1**, 207–216.

Kotz, S. and Balakrishnan, N. and Johnson, N. L. (2000). *Continuous Multivariate Distributions Volume 1: Models and Applications*, 2nd edition, New York: Wiley.

Balakrishnan, N. and Lai, C.-D. (2009). *Continuous Bivariate Distributions*, 2nd edition. New York: Springer.

See Also

`gammaff.mm`, `gamma2`.

Examples

```
shape1 <- exp(1); shape2 <- exp(2); scalepar <- exp(3)
nn <- 1000
mdata <- data.frame(y1 = rgamma(nn, shape1, scale = scalepar),
                    z2 = rgamma(nn, shape2, scale = scalepar))
mdata <- transform(mdata, y2 = y1 + z2)  # z2 \equiv Y2-y1|Y1=y1
fit <- vglm(cbind(y1, y2) ~ 1, bigamma.mckay, mdata, trace = TRUE)
coef(fit, matrix = TRUE)
Coef(fit)
vcov(fit)

colMeans(depvar(fit))  # Check moments
head(fitted(fit), 1)
```

`birds.df` *A Bird Data Set*

Description

An undocumented data set concerning birds.

Usage

```
data("birds.df")
```

Format

A data frame with 67 observations on the following 15 variables.

`Species` a character vector

`Status` a numeric vector

`Length` a numeric vector

`Mass` a numeric vector

`Range` a numeric vector

`Migr` a numeric vector

`Insect` a numeric vector

`Diet` a numeric vector

`Clutch` a numeric vector

`Broods` a numeric vector

`Wood` a numeric vector

`Upland` a numeric vector

`Water` a numeric vector

`Release` a numeric vector

`Indiv` a numeric vector

Details

This is a data set used in a book (or in a draft version of a book). Unfortunately very few details seem available concerning the background and the variables. The purpose of its inclusion in **VGAMdata** is to show that a logistic regression applied to these data, mimicking the book, has estimates suffering from the Hauck–Donner effect (HDE; `hdeff`). It is called `model.final` in the book.

Source

`http://rcompanion.org/documents/RCompanionBioStatistics.pdf`, around pages 249–.

Examples

```
data(birds.df)
str(birds.df)

model.final.copy <-   # Try mimic the book example
  vglm(Status ~ Upland + Migr + Mass + Indiv + Insect + Wood,
       binomialff, data = birds.df, trace = TRUE)
summary(model.final.copy, hde.NA = FALSE)  # Wald table
hdeff(model.final.copy)
hdeff(model.final.copy, deriv = 2)
```

`covid19.nz` *COVID-19 in New Zealand: The First Month or So*

Description

The `covid19.nz` data frame has 69 rows and 3 columns. The number of new cases is tracked daily.

Usage

```
data(covid19.nz)
```

Format

This data frame contains the following columns:

doy output from `as.Date`, is the day of year.

newcases a numeric vector, counts, the number of new cases.

Day a numeric vector, 0 for when the first case occurred; incrementally daily after that.

Details

These were collected from https://www.nzherald.co.nz/ during the first month or so after the first case.

Source

The NZ Herald states their source was Johns Hopkins University.

Examples

```
## Not run: plot(newcases ~ doy, covid19.nz, col = "blue", type = "h",
     las = 1, xlab = "Date")

## End(Not run)
```

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
```

crim.nz *New Zealand Conviction and Sentencing Data Subset 2001–2022*

Description

These data were collected from the New Zealand Ministry of Justice and comprises selected years from 2001 to 2022.

Usage

```
data("crim.nz")
```

Format

A data frame with 257600 observations on the following 7 variables.

`prison` a logical vector; `FALSE` if and only if the sentence was either `"Imprisonment"`, `"Imprisonment sentences"` or `"Life Imprisonment"`. So the other sentences were not imprisonment.

`agegp` an ordered factor with levels `17-19` < `20-24` < `25-29` < `30-39` < `40+`. The age of the offender.

`offence` a factor with levels `abduction`, `injury`, `endanger`, `fraud`, `homicide`, `drugs`, `miscoff`, `antigovt`, `weapons`, `property`, `order`, `robbery`, `sexoff`, `theft`, `burglary`. The *main* offence. In more detail, the levels correspond to `"Abduction, harassment and other offences against the person"`, `"Acts intended to cause injury"`, `"Dangerous or negligent acts endangering persons"`, `"Fraud, deception and related offences"`, `"Homicide and related offences"`, `"Illicit drug offences"`, `"Miscellaneous offences"`, `"Offences against justice procedures, gov. security and gov. operations"`, `"Prohibited and regulated weapons and explosives offences"`, `"Property damage and environmental pollution"`, `"Public order offences"`, `"Robbery, extortion and related offences"`, `"Sexual assault and related offences"`, `"Theft and related offences"`, `"Unlawful entry with intent/burglary, break and enter"`, respectively.

`gender` a factor with levels `F` and `M`.

`ethnicity` a factor with levels `Asian`, `European`, `Maori`, `Other`, `Polynesian`.

`year` a numeric vector. The calendar year when the crime occurred.

`sentence` a factor with levels `ComDet`, `ComSent`, `ComWork`, `HomeDet`, `Prison`, `PrisonSent`, `IntSupv`, `Life`, `Money`, `Other`, `PrevDet`, `Supv`. In more detail, the levels correspond to `"Community Detention"`, `"Community sentences"`, `"Community Work"`, `"Home Detention"`, `"Imprisonment"`, `"Imprisonment sentences"`, `"Intensive Supervision"`, `"Life Imprisonment"`, `"Monetary "`, `"Other"`, `"Preventive Detention"`, `"Supervision"` respectively.

Details

The data were collected in late 2023 and is described in detail in Garcia et al. (2023). Almost all the information here comes from that document. The original data comprised each year from 2001 to 2022 inclusive, however only the years 2001, 2010, 2019 and 2022 are included here for brevity.

Variables with values `"Unknown/Organisation"` were treated as missing and deleted. The `rownames` were stripped off to make the data frame much smaller.

A matching New Zealand Police data set on proceedings against offenders was also collected but not included in **VGAMdata** because of its size.

Offences are categorised using the Australian and New Zealand Standard Offence Classification (ANZSOC). There are sixteen top level 'divisions' which are further subdivided into 'subdivisions' and 'groups'. One of the sixteen main divisions (Traffic Offences) are excluded. Only the most serious offence by the offender is recorded here.

Each row contains the count of sentences received by offenders with particular demographic information in the given period, along with the ANZSOC offence code. However, since ANZSOC is a hierarchical classification, a unique offender is represented in three separate rows: one for the ANZSOC Group of the offence, one for the Subdivision of the offence, and one for the Division of the offence.

Source

https://datainfoplus.stats.govt.nz/

References

Fairness analysis and machine learning model implementation using New Zealand prosecution and conviction data. Garcia, F. and Omidyar, P. and Rodrigues, L. and Xie, J. (2023). Postgraduate Report, Computer Science Department, University of Auckland.

Only ANZSOC Divisions 01–16 were collected. More details can be found at https://www.police.govt.nz/about-us/publications-statistics/data-and-statistics/policedatanz/proceedings-offender-demographics and https://www.abs.gov.au/statistics/classifications/australian-and-new-zealand-standard-offence-classification-anzsoc/latest-release

Examples

```
data(crim.nz)
summary(crim.nz)
```

---

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)
```

`ecb06.it` *Italian Household Data for 2006 and 2014*

Description

Part of the data collected at two time points (2006 and 2014) by the Bank of Italy, as part of the European Central Banks Eurosystem collection of statistics, within the periodic sample surveys on households, businesses and selected intermediaries.

Format

Data frame with the following 20 variables:

`ID` a numeric vector, a unique identification number of the household.

`area` a factor with 5 levels, the Italian geographic area or region in which the household lives: `NW` = North-West, `NE` = North-East, `C` = Center, `S` = South, `I` = Islands. For users wanting a North–South contrast, this variable might be coded as `NC` = North and Center (NW, NE and C), `SI` = South and Islands (S and I).

`sex` a factor with 2 levels, the gender of the head householder: `M` = Male, `F` = Female.

`age` a numeric vector, age in years of the head householder.

`marital` a factor with 4 levels, marital status of the head householder: `married` = Married, `single` = Single, `separated` = Separated or divorced, `widowed` = Widowed.

`education` an ordered factor with 8 levels, the education level of the head householder: `none` = No education, `primaryschool` = Primary school, `middleschool` = Middle school, `profschool` = Professional school, `highschool` = High school, `bachelors` = Bachelors degree, `masters` = Masters degree, `higherdegree` = Higher degree.

`job` a factor with 7 levels, the type of job done by the head householder: `worker` = Worker, `employee` = Employee, `manager` = Manager, `business` = Business person, `other` = Other kind of independent job, `retired` = Retired, `unemployed` = Unemployed.

`occupants` a numeric vector, the number of people living in the same house.

`children` a numeric vector, the number of children of the head householder living with him/her.

`other.children` a numeric vector, the number of children of the head householder not living with the household.

`house.owned` a numeric vector, the ownership of the house in which the householder lives; `0` = The house is not owned, `1` = The house is owned.

`houses` a numeric vector, the number of houses owned by the family, including the one in which the family lives.

`earners` a numeric vector, the number of people in the house who have salary or some kind of earnings.

`accounts` a numeric vector, the number of bank accounts collectively owned by the household.

`ccards` a numeric vector, the number of credit cards collectively owned by the household.

`tot.income`, `dep.income`, `pens.income`, `self.income`, `cap.income` numeric vectors, the amount of income (in Euros) collectively earned by the household through different activities. The variables can be negative if the household has installments. In order, they are the total amount of income, the amount of income earned through dependent working, the amount of income earned by the household through pensions, the amount of income earned by the household through self-dependent working, the amount of income earned by the household through capital investments.

Details

The European Central Banks (ECB) Eurosystem requests and helps organize each country within the European Union to routinely collect statistical information via their central banks. These data frames are a subset from data collected by the Bank of Italy. Each year can be considered a cross-sectional study, although there are some people common to each year. Hence the data collectively can be considered partly a longitudinal study too.

Source

Data was downloaded at https://www.bancaditalia.it in May 2016 by Lucia Pilleri.

References

Supplements to the Statistical Bulletin, Sample Surveys, Household Income and Wealth in 2006, New series, Volume XVIII, Number 7–28, January 2008. Banca D'Italia, Centro Stampa, Roma, Pubbl. Mensile, https://www.bancaditalia.it.

Examples

```
data(ecb06.it); data(ecb14.it)
summary(ecb06.it)
summary(ecb14.it)
## Not run:
with(ecb14.it, table(house.owned))
with(ecb14.it, barplot(table(education), col = "lightblue"))

## 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(multiple.responses = 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)
```

`flamingo` *Flamingo Hotel in Sardinia*

Description

The `flamingo` data frame has 4871 rows and 12 columns.

Usage

```
data(flamingo)
```

Format

This data frame contains the following columns:

LOS length of stay, in days.

year year of the stay.

arrdate date of arrival.

depdate date of departure.

seasonindex seasonality index of the period of the stay, a low value means low season, a high value means very high (peak) season.

booking if the guests booked the room through the hotel's internet website (`internet`), the hotel's telephone (`direct`) or a tour operator (`agency`).

roomtype the 16 hotel's room types.

rmtype3 categorization in 3 groups of `roomtype`.

guests number of guests. Does not include any childrenzz; see the two zz `kids` variables below.

arrangement arrangement type: Bed and Breakfast (`BB`), Half Board (`HB`) and Full Board (`FB`).

kids02 number of kids between 0 and 2 years-old.

kids311 number of kids between 3 and 11 years-old.

Details

The Flamingo Hotel is located on the beach in the southern Sardinia, about 40 kilometers from Cagliari. This data concerns stays from early summer 2019 to late summer 2020 with no stays during the winter period in between. Stays longer than 30 days were deleted from the original source. Variable `LOS` exhibits heaping at the values 7 and 14 days.

Source

The data was obtained with permission from Luca Frigau, University of Cagliari.

See Also

`gaitdzeta`, `airbnb.ac`.

Examples

```
## Not run:  with(flamingo, spikeplot(LOS, col = "blue", ylab = "Proportion"))
```

genpoisson *Generalized Poisson Regression*

Description

Estimation of the two-parameter generalized Poisson distribution.

Usage

```
genpoisson(llambda = "rhobitlink", ltheta = "loglink",
           ilambda = NULL, itheta = NULL, imethod = 1,
           ishrinkage = 0.95, zero = "lambda")
```

Arguments

`llambda`, `ltheta`	Parameter link functions for λ and θ. See `Links` for more choices. The λ parameter lies at least within the interval $[-1, 1]$; see below for more details, and an alternative link is `rhobitlink`. The θ parameter is positive, therefore the default is the log link.
`ilambda`, `itheta`	Optional initial values for λ and θ. The default is to choose values internally.
`imethod`	An integer with value 1 or 2 or 3 which specifies the initialization method for the parameters. If failure to converge occurs try another value and/or else specify a value for `ilambda` and/or `itheta`.
`ishrinkage`, `zero`	See `CommonVGAMffArguments` for information.

Details

This family function is *not* recommended for use; instead try `genpoisson1` or `genpoisson2`. For underdispersion with respect to the Poisson try the GTE (generally-truncated expansion) method described by Yee and Ma (2023).

The generalized Poisson distribution has density

$$f(y) = \theta(\theta + \lambda y)^{y-1} \exp(-\theta - \lambda y)/y!$$

for $\theta > 0$ and $y = 0, 1, 2, \ldots$. Now $\max(-1, -\theta/m) \leq \lambda \leq 1$ where $m(\geq 4)$ is the greatest positive integer satisfying $\theta + m\lambda > 0$ when $\lambda < 0$ [and then $P(Y = y) = 0$ for $y > m$]. Note the complicated support for this distribution means, for some data sets, the default link for `llambda` will not always work, and some tinkering may be required to get it running.

As Consul and Famoye (2006) state on p.165, the lower limits on λ and $m \geq 4$ are imposed to ensure that there are at least 5 classes with nonzero probability when λ is negative.

An ordinary Poisson distribution corresponds to $\lambda = 0$. The mean (returned as the fitted values) is $E(Y) = \theta/(1 - \lambda)$ and the variance is $\theta/(1 - \lambda)^3$.

For more information see Consul and Famoye (2006) for a summary and Consul (1989) for full details.

Value

An object of class "vglmff" (see vglmff-class). The object is used by modelling functions such as vglm, and vgam.

Warning

Monitor convergence! This family function is fragile. Don't get confused because theta (and not lambda) here really matches more closely with lambda of dpois.

Note

This family function handles multiple responses. This distribution is potentially useful for dispersion modelling. Convergence problems may occur when lambda is very close to 0 or 1. If a failure occurs then you might want to try something like llambda = "extlogitlink(min = -0.9, max = 1)" to handle the LHS complicated constraint, and if that doesn't work, try llambda = "extlogitlink(min = -0.8, max = 1)", etc.

Author(s)

T. W. Yee. Easton Huch derived the EIM and it has been implemented in the weights slot.

References

Consul, P. C. (1989). *Generalized Poisson Distributions: Properties and Applications*. New York, USA: Marcel Dekker.

Consul, P. C. and Famoye, F. (2006). *Lagrangian Probability Distributions*, Boston, USA: Birkhauser.

Jorgensen, B. (1997). *The Theory of Dispersion Models*. London: Chapman & Hall

Yee, T. W. and Ma, C. (2024). Generally altered, inflated, truncated and deflated regression. *Statistical Science*, **39** (in press).

See Also

genpoisson1, genpoisson2, poissonff, dpois. dgenpois0, rhobitlink, extlogitlink.

Examples

```
## Not run:
gdata <- data.frame(x2 = runif(nn <- 500))  # NBD data:
gdata <- transform(gdata, y1 = rnbinom(nn, exp(1), mu = exp(2 - x2)))
fit <- vglm(y1 ~ x2, genpoisson, data = gdata, trace = TRUE)
coef(fit, matrix = TRUE)
summary(fit)
## 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

`https://oir.harvard.edu/fact-book`

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

`https://oir.harvard.edu/fact-book`

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

https://oir.harvard.edu/fact-book

References

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

See Also

hued, huie.

Examples

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

Loglap *The Log-Laplace Distribution*

Description

Density, distribution function, quantile function and random generation for the 3-parameter log-Laplace distribution with location parameter `location.ald`, scale parameter `scale.ald` (on the log scale), and asymmetry parameter `kappa`.

Usage

```
dloglap(x, location.ald = 0, scale.ald = 1,
        tau = 0.5, kappa = sqrt(tau/(1-tau)), log = FALSE)
ploglap(q, location.ald = 0, scale.ald = 1, tau = 0.5,
        kappa = sqrt(tau/(1-tau)), lower.tail = TRUE, log.p = FALSE)
qloglap(p, location.ald = 0, scale.ald = 1, tau = 0.5,
        kappa = sqrt(tau/(1-tau)), lower.tail = TRUE, log.p = FALSE)
rloglap(n, location.ald = 0, scale.ald = 1,
        tau = 0.5, kappa = sqrt(tau/(1-tau)))
```

Arguments

x, q	vector of quantiles.
p	vector of probabilities.
n	number of observations. If length(n) > 1 then the length is taken to be the number required.
location.ald, scale.ald	the location parameter ξ and the (positive) scale parameter σ, on the log scale.
tau	the quantile parameter τ. Must consist of values in $(0, 1)$. This argument is used to specify kappa and is ignored if kappa is assigned.
kappa	the asymmetry parameter κ. Must consist of positive values.
log	if TRUE, probabilities p are given as log(p).
lower.tail, log.p	Same meaning as in pnorm or qnorm.

Details

A positive random variable Y is said to have a log-Laplace distribution if $\log(Y)$ has an asymmetric Laplace distribution (ALD). There are many variants of ALDs and the one used here is described in alaplace3.

Value

dloglap gives the density, ploglap gives the distribution function, qloglap gives the quantile function, and rloglap generates random deviates.

Author(s)

T. W. Yee and Kai Huang

References

Kozubowski, T. J. and Podgorski, K. (2003). Log-Laplace distributions. *International Mathematical Journal*, **3**, 467–495.

See Also

dalap, alaplace3, loglaplace1.

Examples

```
loc <- 0; sigma <- exp(0.5); kappa <- 1
x <- seq(-0.2, 5, by = 0.01)
## Not run: plot(x, dloglap(x, loc, sigma, kappa = kappa),
     type = "l", col = "blue", ylim = c(0,1),
     main = "Blue is density, red is the CDF",
     sub = "Purple are 5,10,...,95 percentiles", las = 1, ylab = "")
abline(h = 0, col = "blue", lty = 2)
lines(qloglap(seq(0.05,0.95,by = 0.05), loc, sigma, kappa = kappa),
```

```
  dloglap(qloglap(seq(0.05,0.95,by = 0.05), loc, sigma, kappa = kappa),
              loc, sigma, kappa = kappa),
      col = "purple", lty = 3, type = "h")
lines(x, ploglap(x, loc, sigma, kappa = kappa), type = "l", col = 2)
abline(h = 0, lty = 2)

## End(Not run)
ploglap(qloglap(seq(0.05,0.95,by = 0.05), loc, sigma, kappa = kappa),
        loc, sigma, kappa = kappa)
```

---

loglaplace *Log-Laplace and Logit-Laplace Distribution Family Functions*

Description

Maximum likelihood estimation of the 1-parameter log-Laplace and the 1-parameter logit-Laplace distributions. These may be used for quantile regression for counts and proportions respectively.

Usage

```
loglaplace1(tau = NULL, llocation = "loglink",
    ilocation = NULL, kappa = sqrt(tau/(1 - tau)), Scale.arg = 1,
    ishrinkage = 0.95, parallel.locat = FALSE, digt = 4,
    idf.mu = 3, rep0 = 0.5, minquantile = 0, maxquantile = Inf,
    imethod = 1, zero = NULL)
logitlaplace1(tau = NULL, llocation = "logitlink",
    ilocation = NULL, kappa = sqrt(tau/(1 - tau)),
    Scale.arg = 1, ishrinkage = 0.95, parallel.locat = FALSE,
    digt = 4, idf.mu = 3, rep01 = 0.5, imethod = 1, zero = NULL)
```

Arguments

tau, kappa	See alaplace1.
llocation	Character. Parameter link functions for location parameter ξ. See Links for more choices. However, this argument should be left unchanged with count data because it restricts the quantiles to be positive. With proportions data llocation can be assigned a link such as logitlink, probitlink, cloglog link, etc.
ilocation	Optional initial values. If given, it must be numeric and values are recycled to the appropriate length. The default is to choose the value internally.
parallel.locat	Logical. Should the quantiles be parallel on the transformed scale (argument llocation)? Assigning this argument to TRUE circumvents the seriously embarrassing quantile crossing problem.
imethod	Initialization method. Either the value 1, 2, or ....
idf.mu, ishrinkage, Scale.arg, digt, zero	See alaplace1. See CommonVGAMffArguments for information.

`rep0`, `rep01` Numeric, positive. Replacement values for 0s and 1s respectively. For count data, values of the response whose value is 0 are replaced by `rep0`; it avoids computing `log(0)`. For proportions data values of the response whose value is 0 or 1 are replaced by `min(rangey01[1]/2, rep01/w[y< = 0])` and `max((1 + rangey01[2])/2, 1-rep01/w[y >= 1])` respectively; e.g., it avoids computing `logitlink(0)` or `logitlink(1)`. Here, `rangey01` is the 2-vector `range(y[(y > 0) & (y < 1)])` of the response.

`minquantile`, `maxquantile` Numeric. The minimum and maximum values possible in the quantiles. These argument are effectively ignored by default since `loglink` keeps all quantiles positive. However, if `llocation = logofflink(offset = 1)` then it is possible that the fitted quantiles have value 0 because `minquantile = 0`.

Details

These **VGAM** family functions implement translations of the asymmetric Laplace distribution (ALD). The resulting variants may be suitable for quantile regression for count data or sample proportions. For example, a log link applied to count data is assumed to follow an ALD. Another example is a logit link applied to proportions data so as to follow an ALD. A positive random variable Y is said to have a log-Laplace distribution if $Y = e^W$ where W has an ALD. There are many variants of ALDs and the one used here is described in `alaplace1`.

Value

An object of class `"vglmff"` (see `vglmff-class`). The object is used by modelling functions such as `vglm` and `vgam`.

In the `extra` slot of the fitted object are some list components which are useful. For example, the sample proportion of values which are less than the fitted quantile curves, which is `sum(wprior[y <= location]) / sum(wprior)` internally. Here, `wprior` are the prior weights (called `ssize` below), `y` is the response and `location` is a fitted quantile curve. This definition comes about naturally from the transformed ALD data.

Warning

The **VGAM** family function `logitlaplace1` will not handle a vector of just 0s and 1s as the response; it will only work satisfactorily if the number of trials is large.

See `alaplace1` for other warnings. Care is needed with `tau` values which are too small, e.g., for count data the sample proportion of zeros must be less than all values in `tau`. Similarly, this also holds with `logitlaplace1`, which also requires all `tau` values to be less than the sample proportion of ones.

Note

The form of input for `logitlaplace1` as response is a vector of proportions (values in $[0, 1]$) and the number of trials is entered into the `weights` argument of `vglm`/`vgam`. See Example 2 below. See `alaplace1` for other notes in general.

Author(s)

Thomas W. Yee

References

Kotz, S., Kozubowski, T. J. and Podgorski, K. (2001). *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*, Boston: Birkhauser.

Kozubowski, T. J. and Podgorski, K. (2003). Log-Laplace distributions. *International Mathematical Journal*, **3**, 467–495.

Yee, T. W. (2020). Quantile regression for counts and proportions. In preparation.

See Also

alaplace1, dloglap.

Examples

```
# Example 1: quantile regression of counts with regression splines
set.seed(123); my.k <- exp(0)
adata <- data.frame(x2 = sort(runif(n <- 500)))
mymu <- function(x) exp( 1 + 3*sin(2*x) / (x+0.5)^2)
adata <- transform(adata, y = rnbinom(n, mu = mymu(x2), my.k))
mytau <- c(0.1, 0.25, 0.5, 0.75, 0.9); mydof = 3
# halfstepping is usual:
fitp <- vglm(y ~ sm.bs(x2, df = mydof), data = adata, trace = TRUE,
            loglaplace1(tau = mytau, parallel.locat = TRUE))

## Not run:  par(las = 1)  # Plot on a log1p() scale
mylwd <- 1.5

plot(jitter(log1p(y), factor = 1.5) ~ x2, adata, col = "red",
     pch = "o", cex = 0.75,
     main = "Example 1; green=truth, blue=estimated")
with(adata, matlines(x2, log1p(fitted(fitp)), col = "blue",
                     lty = 1, lwd = mylwd))
finexgrid <- seq(0, 1, len = 201)
for (ii in 1:length(mytau))
  lines(finexgrid, col = "green", lwd = mylwd,
        log1p(qnbinom(mytau[ii], mu = mymu(finexgrid), my.k)))

## End(Not run)
fitp@extra  # Contains useful information


# Example 2: sample proportions
set.seed(123); nnn <- 1000; ssize <- 100  # ssize = 1 wont work!
adata <- data.frame(x2 = sort(runif(nnn)))
mymu <- function(x) logitlink( 1.0 + 4*x, inv = TRUE)
adata <- transform(adata, ssize = ssize,
                   y2 = rbinom(nnn, ssize, prob = mymu(x2)) / ssize)
```

```
mytau <- c(0.25, 0.50, 0.75)
fit1 <- vglm(y2 ~ sm.bs(x2, df = 3),
        logitlaplace1(tau = mytau, lloc = "cloglogLink", paral = TRUE),
        data = adata, weights = ssize, trace = TRUE)

## Not run:
# Check the solution.  Note: this is like comparing apples with oranges.
plotvgam(fit1, se = TRUE, scol = "red", lcol = "blue",
         main = "Truth = 'green'")
# Centered approximately !
linkFunctionChar <- as.character(fit1@misc$link)
adata <- transform(adata, trueFunction =
           theta2eta(theta = mymu(x2), link = linkFunctionChar))
with(adata, lines(x2, trueFunction - mean(trueFunction), col = "green"))

# Plot the data + fitted quantiles (on the original scale)
myylim <- with(adata, range(y2))
plot(y2 ~ x2, adata, col = "blue", ylim = myylim, las = 1,
     pch = ".", cex = 2.5)
with(adata, matplot(x2, fitted(fit1), add = TRUE, lwd = 3, type = "l"))
truecol <- rep(1:3, len = fit1@misc$M)  # Add the 'truth'
smallxgrid <- seq(0, 1, len = 501)
for (ii in 1:length(mytau))
  lines(smallxgrid, col = truecol[ii], lwd = 2,
        qbinom(mytau[ii], pr = mymu(smallxgrid), si = ssize) / ssize)

# Plot on the eta (== logitlink()/probitlink()/...) scale
  with(adata, matplot(x2, predict(fit1), lwd = 3, type = "l"))
# Add the 'truth'
for (ii in 1:length(mytau)) {
  true.quant <- qbinom(mytau[ii], prob = mymu(smallxgrid),
                       size = ssize) / ssize
  lines(smallxgrid, theta2eta(true.quant, link = linkFunctionChar),
        col = truecol[ii], lwd = 2)
}
## End(Not run)
```

`mbflood` *Madawask Basin Flood Data*

Description

Flood peak and flood volume in the Madawask (River) basin, Quebec, Canada

Usage

```
data(mbflood)
```

Format

A data frame with the following variables.

year Numeric.

D Numeric. Duration, in days.

V Numeric. Volume (days per cubic metre per second).

Q Numeric. Flood peak (per cubic metre per second).

Details

From Shen (2001), the basin is located in the province of Quebec, Canada, and it covers an area of 2690 square km. The data comprises 77 years daily streamflow data from 1919 to 1995 from the HYDAT CD (Environment Canada, 1998). In the basin, winter lasts for about 4 months and precipitation is mainly in the form of snowfall during this period. Spring represents the high flow season due to the contribution of spring snowmelt to river runoff and the spring flood is the annual maximum flood both in flood peak and volume. A typical flood hydrograph with its characteristic values are: flood peak Q, flood volume V, and flood duration D. These flood characteristics are determined by first identifying the flood duration: the dates of start and end of flood runoff using the approach proposed by Yue (2000). Generally, time boundaries of a flood are marked by a rise in stage and discharge from base flow (start of flood runoff) and a return to base flow (end of flood runoff).

Source

Table 1 of Yue, S. (2001). A bivariate gamma distribution for use in multivariate flood frequency analysis. *Hydrological Processes*, **15**, 1033–45.

References

Yue, S. (2000). The bivariate lognormal distribution to model a multivariate flood episode. *Hydrological Processes*, **14**, 2575–88.

Examples

```
summary(mbflood)
```

`Oalog` *One-Altered Logarithmic Distribution*

Description

Density, distribution function, quantile function and random generation for the one-altered logarithmic distribution with parameter `pobs1`.

Usage

```
doalog(x, shape, pobs1 = 0, log = FALSE)
poalog(q, shape, pobs1 = 0)
qoalog(p, shape, pobs1 = 0)
roalog(n, shape, pobs1 = 0)
```

Arguments

x, q, n, p	Same Unif.
shape, log	Same as Otlog).
pobs1	Probability of (an observed) one, called $pobs1$. The default value of pobs1 = 0 corresponds to the response having a 1-truncated logarithmic distribution.

Details

The probability function of Y is 1 with probability pobs1, else a 1-truncated logarithmic(shape) distribution.

Value

doalog gives the density and poalog gives the distribution function, qoalog gives the quantile function, and roalog generates random deviates.

Note

The argument pobs1 is recycled to the required length, and must have values which lie in the interval $[0, 1]$.

Author(s)

T. W. Yee

See Also

Gaitdlog, oalog, oilog, Otlog.

Examples

```
shape <- 0.75; pobs1 <- 0.10; x <- (-1):7
doalog(x, shape = shape, pobs1 = pobs1)
table(roalog(100, shape = shape, pobs1 = pobs1))

## Not run:  x <- 0:10
barplot(rbind(doalog(x, shape = shape, pobs1 = pobs1),
                dlog(x, shape = shape)),
        beside = TRUE, col = c("blue", "orange"), cex.main = 0.7, las = 1,
        ylab = "Probability", names.arg = as.character(x),
        main = paste("OAL(shape = ", shape, ", pobs1 = ", pobs1,
                   ") [blue] vs",  " Logarithmic(shape = ", shape,
                   ") [orange] densities", sep = ""))
```

```
## End(Not run)
```

---

oalog *One-Altered Logarithmic Distribution*

Description

Fits a one-altered logarithmic distribution based on a conditional model involving a Bernoulli distribution and a 1-truncated logarithmic distribution.

Usage

```
oalog(lpobs1 = "logitlink", lshape = "logitlink",
      type.fitted = c("mean", "shape", "pobs1", "onempobs1"),
      ipobs1 = NULL, gshape = ppoints(8), zero = NULL)
```

Arguments

`lpobs1`	Link function for the parameter p_1 or ϕ, called `pobs1` or `phi` here. See `Links` for more choices.
`lshape`	See `logff` for details.
`gshape, type.fitted`	See `CommonVGAMffArguments` and `fittedvlm` for information.
`ipobs1, zero`	See `CommonVGAMffArguments` for information.

Details

The response Y is one with probability p_1, or Y has a 1-truncated logarithmic distribution with probability $1 - p_1$. Thus $0 < p_1 < 1$, which is modelled as a function of the covariates. The one-altered logarithmic distribution differs from the one-inflated logarithmic distribution in that the former has ones coming from one source, whereas the latter has ones coming from the logarithmic distribution too. The one-inflated logarithmic distribution is implemented in the **VGAM** package. Some people call the one-altered logarithmic a *hurdle* model.

The input can be a matrix (multiple responses). By default, the two linear/additive predictors of `oalog` are $(logit(\phi), logit(s))^T$.

Value

An object of class `"vglmff"` (see `vglmff-class`). The object is used by modelling functions such as `vglm`, and `vgam`.

The `fitted.values` slot of the fitted object, which should be extracted by the generic function `fitted`, returns the mean μ (default) which is given by

$$\mu = \phi + (1 - \phi)A$$

where A is the mean of the one-truncated logarithmic distribution. If `type.fitted = "pobs1"` then p_1 is returned.

Note

This family function effectively combines binomialff and otlog into one family function.

Author(s)

T. W. Yee

See Also

Gaitdlog, Oalog, logff, oilog, CommonVGAMffArguments, simulate.vlm.

Examples

```
## Not run: odata <- data.frame(x2 = runif(nn <- 1000))
odata <- transform(odata, pobs1  = logitlink(-1 + 2*x2, inv = TRUE),
                          shape  = logitlink(-2 + 3*x2, inv = TRUE))
odata <- transform(odata, y1 = roalog(nn, shape, pobs1 = pobs1),
                          y2 = roalog(nn, shape, pobs1 = pobs1))
with(odata, table(y1))

ofit <- vglm(cbind(y1, y2) ~ x2, oalog, data = odata, trace = TRUE)
coef(ofit, matrix = TRUE)
head(fitted(ofit))
head(predict(ofit))
summary(ofit)

## End(Not run)
```

Oapospois *One-Altered Logarithmic Distribution*

Description

Density, distribution function, quantile function and random generation for the one-altered positive-Poisson distribution with parameter pobs1.

Usage

```
doapospois(x, lambda, pobs1 = 0, log = FALSE)
poapospois(q, lambda, pobs1 = 0)
qoapospois(p, lambda, pobs1 = 0)
roapospois(n, lambda, pobs1 = 0)
```

Arguments

x, q, n, p	Same Unif.
lambda, log	Same as Otpospois).
pobs1	Probability of (an observed) one, called $pobs1$. The default value of pobs1 = 0 corresponds to the response having a 1-truncated positive-Poisson distribution.

Details

The probability function of Y is 1 with probability pobs1, else a 1-truncated positive-Poisson(lambda) distribution.

Value

doapospois gives the density and poapospois gives the distribution function, qoapospois gives the quantile function, and roapospois generates random deviates.

Note

The argument pobs1 is recycled to the required length, and must have values which lie in the interval $[0, 1]$.

Author(s)

T. W. Yee

See Also

oapospoisson, Oipospois, Otpospois.

Examples

```
lambda <- 3; pobs1 <- 0.30; x <- (-1):7
doapospois(x, lambda = lambda, pobs1 = pobs1)
table(roapospois(100, lambda = lambda, pobs1 = pobs1))

## Not run:  x <- 0:10
barplot(rbind(doapospois(x, lambda = lambda, pobs1 = pobs1),
                dpospois(x, lambda = lambda)),
        beside = TRUE, col = c("blue", "orange"), cex.main = 0.7, las = 1,
        ylab = "Probability", names.arg = as.character(x),
        main = paste("OAPP(lambda = ", lambda, ", pobs1 = ", pobs1,
                   ") [blue] vs",  " PosPoisson(lambda = ", lambda,
                   ") [orange] densities", sep = ""))
## End(Not run)
```

---

oapospoisson *One-Altered Positive-Poisson Distribution*

Description

Fits a one-altered positive-Poisson distribution based on a conditional model involving a Bernoulli distribution and a 1-truncated positive-Poisson distribution.

Usage

```
oapospoisson(lpobs1 = "logitlink", llambda = "loglink",
    type.fitted = c("mean", "lambda", "pobs1", "onempobs1"),
    ipobs1 = NULL, zero = NULL)
```

Arguments

lpobs1	Link function for the parameter p_1 or ϕ, called pobs1 or phi here. See Links for more choices.
llambda	See pospoisson for details.
type.fitted	See CommonVGAMffArguments and fittedvlm for information.
ipobs1, zero	See CommonVGAMffArguments for information.

Details

The response Y is one with probability p_1, or Y has a 1-truncated positive-Poisson distribution with probability $1 - p_1$. Thus $0 < p_1 < 1$, which is modelled as a function of the covariates. The one-altered positive-Poisson distribution differs from the one-inflated positive-Poisson distribution in that the former has ones coming from one source, whereas the latter has ones coming from the positive-Poisson distribution too. The one-inflated positive-Poisson distribution is implemented in the **VGAM** package. Some people call the one-altered positive-Poisson a *hurdle* model.

The input can be a matrix (multiple responses). By default, the two linear/additive predictors of oapospoisson are $(logit(\phi), log(\lambda))^T$.

Value

An object of class "vglmff" (see vglmff-class). The object is used by modelling functions such as vglm, and vgam.

The fitted.values slot of the fitted object, which should be extracted by the generic function fitted, returns the mean μ (default) which is given by

$$\mu = \phi + (1 - \phi)A$$

where A is the mean of the one-truncated positive-Poisson distribution. If type.fitted = "pobs1" then p_1 is returned.

Note

This family function effectively combines binomialff and otpospoisson into one family function.

Author(s)

T. W. Yee

See Also

Oapospois, pospoisson, oipospoisson, CommonVGAMffArguments, simulate.vlm.

Examples

```
## Not run: odata <- data.frame(x2 = runif(nn <- 1000))
odata <- transform(odata, pobs1  = logitlink(-1 + 2*x2, inv = TRUE),
                          lambda =   loglink( 1 + 1*x2, inv = TRUE))
odata <- transform(odata, y1 = roapospois(nn, lambda, pobs1 = pobs1),
                          y2 = roapospois(nn, lambda, pobs1 = pobs1))
with(odata, table(y1))

ofit <- vglm(cbind(y1, y2) ~ x2, oapospoisson, odata, trace = TRUE)
coef(ofit, matrix = TRUE)
head(fitted(ofit))
head(predict(ofit))
summary(ofit)

## End(Not run)
```

Oazeta *One-Altered Logarithmic Distribution*

Description

Density, distribution function, quantile function and random generation for the one-altered zeta distribution with parameter pobs1.

Usage

```
doazeta(x, shape, pobs1 = 0, log = FALSE)
poazeta(q, shape, pobs1 = 0)
qoazeta(p, shape, pobs1 = 0)
roazeta(n, shape, pobs1 = 0)
```

Arguments

x, q, n, p	Same Unif.
shape, log	Same as Otzeta).
pobs1	Probability of (an observed) one, called $pobs1$. The default value of pobs1 = 0 corresponds to the response having a 1-truncated zeta distribution.

Details

The probability function of Y is 1 with probability pobs1, else a 1-truncated zeta distribution.

Value

doazeta gives the density and poazeta gives the distribution function, qoazeta gives the quantile function, and roazeta generates random deviates.

Note

The argument pobs1 is recycled to the required length, and must have values which lie in the interval $[0, 1]$.

Author(s)

T. W. Yee

See Also

oazeta, Oizeta, Otzeta, zeta.

Examples

```
shape <- 1.1; pobs1 <- 0.10; x <- (-1):7
doazeta(x, shape = shape, pobs1 = pobs1)
table(roazeta(100, shape = shape, pobs1 = pobs1))

## Not run:  x <- 0:10
barplot(rbind(doazeta(x, shape = shape, pobs1 = pobs1),
                dzeta(x, shape = shape)),
        beside = TRUE, col = c("blue", "orange"), cex.main = 0.7, las = 1,
        ylab = "Probability", names.arg = as.character(x),
        main = paste("OAZ(shape = ", shape, ", pobs1 = ", pobs1,
                   ") [blue] vs",  " zeta(shape = ", shape,
                   ") [orange] densities", sep = ""))
## End(Not run)
```

---

oazeta *One-Altered Zeta Distribution*

Description

Fits a one-altered zeta distribution based on a conditional model involving a Bernoulli distribution and a 1-truncated zeta distribution.

Usage

```
oazeta(lpobs1 = "logitlink", lshape = "loglink",
       type.fitted = c("mean", "shape", "pobs1", "onempobs1"),
       gshape = exp((-4:3)/4), ishape = NULL, ipobs1 = NULL, zero = NULL)
```

Arguments

lpobs1	Link function for the parameter p_1 or ϕ, called pobs1 or phi here. See Links for more choices.
lshape	See zeta for details.
type.fitted	See CommonVGAMffArguments and fittedvlm for information.
gshape, ishape, ipobs1, zero	See CommonVGAMffArguments for information.

Details

The response Y is one with probability p_1, or Y has a 1-truncated zeta distribution with probability $1 - p_1$. Thus $0 < p_1 < 1$, which is modelled as a function of the covariates. The one-altered zeta distribution differs from the one-inflated zeta distribution in that the former has ones coming from one source, whereas the latter has ones coming from the zeta distribution too. The one-inflated zeta distribution is implemented in the **VGAM** package. Some people call the one-altered zeta a *hurdle* model.

The input can be a matrix (multiple responses). By default, the two linear/additive predictors of `oazeta` are $(logit(\phi), log(shape))^T$.

Value

An object of class `"vglmff"` (see `vglmff-class`). The object is used by modelling functions such as `vglm`, and `vgam`.

The `fitted.values` slot of the fitted object, which should be extracted by the generic function `fitted`, returns the mean μ (default) which is given by

$$\mu = \phi + (1 - \phi)A$$

where A is the mean of the one-truncated zeta distribution. If `type.fitted = "pobs1"` then p_1 is returned.

Note

This family function effectively combines `binomialff` and `otzeta` into one family function.

Author(s)

T. W. Yee

See Also

`Oazeta`, `zetaff`, `oizeta`, `otzeta`, `CommonVGAMffArguments`, `simulate.vlm`.

Examples

```
## Not run: odata <- data.frame(x2 = runif(nn <- 1000))
odata <- transform(odata, pobs1 = logitlink(-1 + 2*x2, inverse = TRUE),
                          shape =  loglink( 1 + 1*x2, inverse = TRUE))
odata <- transform(odata, y1 = roazeta(nn, shape = shape, pobs1 = pobs1),
                          y2 = roazeta(nn, shape = shape, pobs1 = pobs1))
with(odata, table(y1))

ofit <- vglm(cbind(y1, y2) ~ x2, oazeta, data = odata, trace = TRUE)
coef(ofit, matrix = TRUE)
head(fitted(ofit))
head(predict(ofit))
summary(ofit)

## End(Not run)
```

`Oilog` *One-Inflated Logarithmic Distribution*

Description

Density, distribution function, quantile function and random generation for the one-inflated logarithmic distribution with parameter `pstr1`.

Usage

```
doilog(x, shape, pstr1 = 0, log = FALSE)
poilog(q, shape, pstr1 = 0)
qoilog(p, shape, pstr1 = 0)
roilog(n, shape, pstr1 = 0)
```

Arguments

`x`, `q`, `p`, `n`	Same as `Uniform`.
`shape`	Vector of parameters that lie in $(0, 1)$.
`pstr1`	Probability of a structural one (i.e., ignoring the logarithmic distribution), called ϕ. The default value of $\phi = 0$ corresponds to the response having an ordinary logarithmic distribution.
`log`	Same as `Uniform`.

Details

The probability function of Y is 1 with probability ϕ, and $Logarithmic(prob)$ with probability $1 - \phi$. Thus

$$P(Y = 1) = \phi + (1 - \phi)P(W = 1)$$

where W is distributed as a $Logarithmic(shape)$ random variable. The **VGAM** family function `oilog` estimates ϕ by MLE.

Value

`doilog` gives the density, `poilog` gives the distribution function, `qoilog` gives the quantile function, and `roilog` generates random deviates.

Note

The argument `pstr1` is recycled to the required length, and usually has values which lie in the interval $[0, 1]$.

These functions actually allow for the *zero-deflated logarithmic* distribution. Here, `pstr1` is also permitted to lie in the interval `[-dlog(1, shape) / (1 - dlog(1, shape)), 0]`. The resulting probability of a unit count is *less than* the nominal logarithmic value, and the use of `pstr1` to stand for the probability of a structural 1 loses its meaning. When `pstr1` equals `-dlog(1, shape) / (1 - dlog(1, shape))` this corresponds to the 1-truncated logarithmic distribution.

Author(s)

T. W. Yee

See Also

Gaitdlog, oilog, rlog, logff, Otlog.

Examples

```
shape <- 0.5; pstr1 <- 0.3; x <- (-1):7
(ii <- doilog(x, shape, pstr1 = pstr1))
max(abs(poilog(1:200, shape) -
  cumsum(shape^(1:200) / (-(1:200) * log1p(-shape))))  # Should be 0

## Not run:  x <- 0:10
par(mfrow = c(2, 1))  # One-Inflated logarithmic
barplot(rbind(doilog(x, shape, pstr1 = pstr1), dlog(x, shape)),
   beside = TRUE, col = c("blue", "orange"),
   main = paste0("OILogff(", shape, ", pstr1 = ", pstr1,
                 ") (blue) vs Logff(", shape, ") (orange)"),
   names.arg = as.character(x))

deflat.limit <- -dlog(1, shape) / plog(1, shape, lower.tail = FALSE)
newpstr1 <- round(deflat.limit, 3) + 0.001  # Near the boundary
barplot(rbind(doilog(x, shape, pstr1 = newpstr1),
               dlog(x, shape)),
   beside = TRUE, col = c("blue","orange"),
   main = paste0("ODLogff(", shape, ", pstr1 = ", newpstr1,
                 ") (blue) vs Logff(", shape, ") (orange)"),
   names.arg = as.character(x))
## End(Not run)
```

---

oilog *One-inflated Logarithmic Distribution Family Function*

Description

Fits a 1-inflated logarithmic distribution.

Usage

```
oilog(lpstr1 = "logitlink", lshape = "logitlink",
      type.fitted = c("mean", "shape", "pobs1", "pstr1",
      "onempstr1"), ishape = NULL, gpstr1 = ppoints(8), gshape =
      ppoints(8), zero = NULL)
```

Arguments

`lpstr1`, `lshape` Link functions. For `lpstr1`: the same idea as `zipoisson` except it applies to a structural 1.

`gpstr1`, `gshape`, `ishape`
For initial values. See `CommonVGAMffArguments` for information.

`type.fitted`, `zero`
See `CommonVGAMffArguments` for information.

Details

The 1-inflated logarithmic distribution is a mixture distribution of the logarithmic distribution with some probability of obtaining a (structural) 1. Thus there are two sources for obtaining the value 1. This distribution is written here in a way that retains a similar notation to the one-inflated positive-Poisson, i.e., the probability $P[Y = 1]$ involves another parameter ϕ. See `oipospoisson`.

This family function can handle multiple responses.

Value

An object of class `"vglmff"` (see `vglmff-class`). The object is used by modelling functions such as `vglm`, `rrvglm` and `vgam`.

Author(s)

Thomas W. Yee

See Also

`Gaitdlog`, `Oilog`, `logff`, `Oizeta`.

Examples

```
## Not run:  odata <- data.frame(x2 = runif(nn <- 1000))  # Artificial data
odata <- transform(odata, pstr1 = logitlink(-1 + x2, inv = TRUE),
                          shape = 0.5)
odata <- transform(odata, y1 = roilog(nn, shape, pstr1 = pstr1))
with(odata, table(y1))
fit1 <- vglm(y1 ~ x2, oilog(zero = "shape"), odata, trace = TRUE)
coef(fit1, matrix = TRUE)

## End(Not run)
```

`Oiposbinom` *One-Inflated Positive Binomial Distribution*

Description

Density, distribution function, quantile function and random generation for the one-inflated positive binomial distribution with parameter `pstr1`.

Usage

```
doiposbinom(x, size, prob, pstr1 = 0, log = FALSE)
poiposbinom(q, size, prob, pstr1 = 0)
qoiposbinom(p, size, prob, pstr1 = 0)
roiposbinom(n, size, prob, pstr1 = 0)
```

Arguments

`x`, `p`, `q`, `n`	Same as `Posbinom`.
`size`, `prob`	Same as `Posbinom`.
`pstr1`	Probability of a structural one (i.e., ignoring the positive binomial distribution), called ϕ. The default value of $\phi = 0$ corresponds to the response having a positive binomial distribution. However, `pstr1` can also be negative, in which case it ceases its interpretation as a probability, and this is known as *one-deflation*.
`log`	Logical. Return the logarithm of the answer?

Details

The probability function of Y is 1 with probability ϕ, and $PosBinomial(size, prob)$ with probability $1 - \phi$. Thus

$$P(Y = 1) = \phi + (1 - \phi)P(W = 1)$$

where W is distributed as a positive $binomial(size, prob)$ random variable.

Value

`doiposbinom` gives the density, `poiposbinom` gives the distribution function, `qoiposbinom` gives the quantile function, and `roiposbinom` generates random deviates.

Note

The argument `pstr1` is recycled to the required length, and usually has values which lie in the interval $[0, 1]$. These functions actually allow for the *zero-deflated binomial* distribution. Here, `pstr1` is also permitted to lie in the interval $[-A, 0]$ for some positive quantity A. The resulting probability of a unit value is *less than* the nominal positive binomial value, and the use of `pstr1` to stand for the probability of a structural 1 loses its meaning. If `pstr1` equals A then this corresponds to the 0- and 1-truncated binomial distribution.

Author(s)

T. W. Yee

See Also

posbinomial, dbinom, binomialff.

Examples

```
size <- 10; prob <- 0.2; pstr1 <- 0.4; x <- (-1):size
(ii <- doiposbinom(x, size, prob, pstr1 = pstr1))
table(roiposbinom(100, size, prob, pstr1 = pstr1))
round(doiposbinom(x, size, prob, pstr1 = pstr1) * 100)  # Similar?

## Not run:  x <- 0:size
par(mfrow = c(2, 1))  # One-Inflated Positive Binomial
barplot(rbind(doiposbinom(x, size, prob, pstr1 = pstr1),
              dposbinom(x, size, prob)),
   beside = TRUE, col = c("blue", "orange"),
   main = paste0("OIPB(", size, ",", prob, ", pstr1 = ", pstr1,
     ") (blue) vs PosBinomial(", size, ",", prob, ") (orange)"),
   names.arg = as.character(x))

# Zero-deflated Pos Binomial:
def.limit <- -dposbinom(1, size, prob) / (1 - dposbinom(1, size, prob))
def.limit <- size * prob / (1 + (size-1) * prob-1 / (1-prob)^(size-1))
newpstr1 <- round(def.limit, 3) + 0.001  # A little from the boundary
barplot(rbind(doiposbinom(x, size, prob, pstr1 = newpstr1),
               dposbinom(x, size, prob)),
   beside = TRUE, col = c("blue","orange"),
   main = paste0("ODPB(", size, ",", prob, ", pstr1 = ", newpstr1,
      ") (blue) vs PosBinomial(", size, ",", prob, ") (orange)"),
   names.arg = as.character(x))
## End(Not run)
```

`oiposbinomial` *One-Inflated Positive Binomial Distribution Family Function*

Description

Fits a one-inflated positive binomial distribution by maximum likelihood estimation.

Usage

```
oiposbinomial(lpstr1 = "logitlink", lprob = "logitlink",
    type.fitted = c("mean", "prob", "pobs1", "pstr1", "onempstr1"),
    iprob = NULL, gpstr1 = ppoints(9), gprob  = ppoints(9),
    multiple.responses = FALSE, zero = NULL)
```

Arguments

`lpstr1`, `lprob`	Link functions for the parameter ϕ and the positive binomial probability μ parameter. See `Links` for more choices. See `CommonVGAMffArguments` also. For the one-*deflated* model see below.
`type.fitted`	See `CommonVGAMffArguments` and `fittedvlm`.
`iprob`, `gpstr1`, `gprob`	For initial values; see `CommonVGAMffArguments`.
`multiple.responses`	Logical. See `binomialff` and `posbinomial`.
`zero`	See `CommonVGAMffArguments` for information.

Details

These functions are based on

$$P(Y = y) = \phi + (1 - \phi)N\mu(1 - \mu)^N/(1 - (1 - \mu)^N),$$

for $y = 1/N$, and

$$P(Y = y) = (1 - \phi)\binom{N}{Ny}\mu^{Ny}(1 - \mu)^{N(1-y)}/(1 - (1 - \mu)^N).$$

for $y = 2/N, \ldots, 1$. That is, the response is a sample proportion out of N trials, and the argument `size` in `roiposbinom` is N here. Ideally $N > 2$ is needed. The parameter ϕ is the probability of a structural one, and it satisfies $0 < \phi < 1$ (usually). The mean of Y is $E(Y) = \phi + (1 - \phi)\mu/(1 - (1 - \mu)^N)$ and these are returned as the default fitted values. By default, the two linear/additive predictors for `oiposbinomial()` are $(logit(\phi), logit(\mu))^T$.

Value

An object of class `"vglmff"` (see `vglmff-class`). The object is used by modelling functions such as `vglm` and `vgam`.

Note

The response variable should have one of the formats described by `binomialff`, e.g., a factor or two column matrix or a vector of sample proportions with the `weights` argument specifying the values of N.

To work well, one ideally needs large values of N and μ much greater than 0, i.e., the larger N and μ are, the better. If $N = 1$ then the model is unidentifiable since the number of parameters is excessive.

Estimated probabilities of a structural one and an observed one are returned, as in `zipoisson`.

The one-*deflated* positive binomial distribution might be fitted by setting `lpstr1 = "identitylink"`, albeit, not entirely reliably. See `zipoisson` for information that can be applied here.

Author(s)

T. W. Yee

See Also

roiposbinom, posbinomial, binomialff, rbinom.

Examples

```
size <- 10  # Number of trials; N in the notation above
nn <- 200
odata <- data.frame(pstr1  = logitlink( 0, inv = TRUE),  # 0.50
             mubin1 = logitlink(-1, inv = TRUE),  # Binomial mean
             svec   = rep(size, length = nn),
             x2     = runif(nn))
odata <- transform(odata,
                   mubin2 = logitlink(-1 + x2, inv = TRUE))
odata <- transform(odata,
             y1 = roiposbinom(nn, svec, pr = mubin1, pstr1 = pstr1),
             y2 = roiposbinom(nn, svec, pr = mubin2, pstr1 = pstr1))
with(odata, table(y1))
fit1 <- vglm(y1 / svec ~  1, oiposbinomial, data = odata,
             weights = svec, trace = TRUE, crit = "coef")
fit2 <- vglm(y2 / svec ~ x2, oiposbinomial, data = odata,
             weights = svec, trace = TRUE)

coef(fit1, matrix = TRUE)
Coef(fit1)  # Useful for intercept-only models
head(fitted(fit1, type = "pobs1"))  # Estimate of P(Y = 1)
head(fitted(fit1))
with(odata, mean(y1))  # Compare this with fitted(fit1)
summary(fit1)
```

Oipospois *One-Inflated Positive Poisson Distribution*

Description

Density, distribution function, quantile function and random generation for the one-inflated positive Poisson distribution with parameter pstr1.

Usage

```
doipospois(x, lambda, pstr1 = 0, log = FALSE)
poipospois(q, lambda, pstr1 = 0)
qoipospois(p, lambda, pstr1 = 0)
roipospois(n, lambda, pstr1 = 0)
```

Arguments

x, p, q, n	Same as Pospois.
lambda	Vector of positive means.

pstr1 — Probability of a structural one (i.e., ignoring the positive Poisson distribution), called ϕ. The default value of $\phi = 0$ corresponds to the response having a positive Poisson distribution.

log — Logical. Return the logarithm of the answer?

Details

The probability function of Y is 1 with probability ϕ, and $PosPoisson(\lambda)$ with probability $1 - \phi$. Thus

$$P(Y = 1) = \phi + (1 - \phi)P(W = 1)$$

where W is distributed as a positive $Poisson(\lambda)$ random variate.

Value

doipospois gives the density, poipospois gives the distribution function, qoipospois gives the quantile function, and roipospois generates random deviates.

Note

The argument pstr1 is recycled to the required length, and usually has values which lie in the interval $[0, 1]$.

These functions actually allow for the *zero-deflated Poisson* distribution. Here, pstr1 is also permitted to lie in the interval [-lambda / (expm1(lambda) - lambda), 0]. The resulting probability of a unit count is *less than* the nominal positive Poisson value, and the use of pstr1 to stand for the probability of a structural 1 loses its meaning. When pstr1 equals -lambda / (expm1(lambda) - lambda) this corresponds to the 0- and 1-truncated Poisson distribution.

Author(s)

T. W. Yee

See Also

Pospois, oapospoisson, oipospoisson, otpospoisson, pospoisson, dpois, poissonff.

Examples

```
lambda <- 3; pstr1 <- 0.2; x <- (-1):7
(ii <- doipospois(x, lambda, pstr1 = pstr1))
table(roipospois(100, lambda, pstr1 = pstr1))
round(doipospois(1:10, lambda, pstr1 = pstr1) * 100)  # Similar?

## Not run:  x <- 0:10
par(mfrow = c(2, 1))  # One-Inflated Positive Poisson
barplot(rbind(doipospois(x, lambda, pstr1 = pstr1),
              dpospois(x, lambda)),
   beside = TRUE, col = c("blue", "orange"),
   main = paste0("OIPP(", lambda, ", pstr1 = ", pstr1,
                 ") (blue) vs PosPoisson(", lambda, ") (orange)"),
   names.arg = as.character(x))
```

```
# 0-deflated Pos Poisson:
deflat.limit <- -lambda / (expm1(lambda) - lambda)
newpstr1 <- round(deflat.limit, 3) + 0.001  # Near the boundary
barplot(rbind(doipospois(x, lambda, pstr1 = newpstr1),
              dpospois(x, lambda)),
   beside = TRUE, col = c("blue","orange"),
   main = paste0("ODPP(", lambda, ", pstr1 = ", newpstr1,
                 ") (blue) vs PosPoisson(", lambda, ") (orange)"),
   names.arg = as.character(x))
## End(Not run)
```

---

oipospoisson *One-inflated Positive Poisson Distribution Family Function*

Description

Fits a 1-inflated positive Poisson distribution.

Usage

```
oipospoisson(lpstr1 = "logitlink", llambda = "loglink",
   type.fitted = c("mean", "lambda", "pobs1", "pstr1", "onempstr1"),
   ilambda = NULL, gpstr1 = (1:19)/20, gprobs.y = (1:19)/20,
   imethod = 1, zero = NULL)
```

Arguments

`lpstr1`, `llambda` For `lpstr1`: the same idea as `zipoisson` except it applies to a structural 1.

`ilambda`, `gpstr1`, `gprobs.y`, `imethod`
For initial values. See `CommonVGAMffArguments` for information.

`type.fitted`, `zero`
See `CommonVGAMffArguments` for information.

Details

The 1-inflated positive Poisson distribution is a mixture distribution of the positive (0-truncated) Poisson distribution with some probability of obtaining a (structural) 1. Thus there are two sources for obtaining the value 1. It is similar to a zero-inflated Poisson model, except the Poisson is replaced by a positive Poisson and the 0 is replaced by 1. This distribution is written here in a way that retains a similar notation to the zero-inflated Poisson, i.e., the probability $P[Y = 1]$ involves another parameter ϕ. See `zipoisson`.

This family function can handle multiple responses.

Value

An object of class `"vglmff"` (see `vglmff-class`). The object is used by modelling functions such as `vglm`, `rrvglm` and `vgam`.

Warning

Under- or over-flow may occur if the data is ill-conditioned.

Author(s)

Thomas W. Yee

See Also

Oipospois, pospoisson, oapospoisson, otpospoisson, zipoisson, poissonff, simulate.vlm.

Examples

```
## Not run: set.seed(1)
pdata <- data.frame(x2 = runif(nn <- 1000))  # Artificial data
pdata <- transform(pdata, pstr1 = 0.5, lambda = exp(3 - x2))
pdata <- transform(pdata, y1 = roipospois(nn, lambda, pstr1 = pstr1))
with(pdata, table(y1))
fit1 <- vglm(y1 ~ x2, oipospoisson, data = pdata, trace = TRUE)
coef(fit1, matrix = TRUE)

## End(Not run)
```

Oizeta *One-Inflated Zeta Distribution*

Description

Density, distribution function, quantile function and random generation for the one-inflated zeta distribution with parameter pstr1.

Usage

```
doizeta(x, shape, pstr1 = 0, log = FALSE)
poizeta(q, shape, pstr1 = 0)
qoizeta(p, shape, pstr1 = 0)
roizeta(n, shape, pstr1 = 0)
```

Arguments

x, q, p, n	Same as Uniform.
shape	Vector of positive shape parameters.
pstr1	Probability of a structural one (i.e., ignoring the zeta distribution), called ϕ. The default value of $\phi = 0$ corresponds to the response having an ordinary zeta distribution.
log	Same as Uniform.

Details

The probability function of Y is 1 with probability ϕ, and $Zeta(shape)$ with probability $1 - \phi$. Thus

$$P(Y = 1) = \phi + (1 - \phi)P(W = 1)$$

where W is distributed as a $zeta(shape)$ random variable.

Value

`doizeta` gives the density, `poizeta` gives the distribution function, `qoizeta` gives the quantile function, and `roizeta` generates random deviates.

Note

The argument `pstr1` is recycled to the required length, and usually has values which lie in the interval $[0, 1]$.

These functions actually allow for the *zero-deflated zeta* distribution. Here, `pstr1` is also permitted to lie in the interval `[-dzeta(1, shape) / (1 - dzeta(1, shape)), 0]`. The resulting probability of a unit count is *less than* the nominal zeta value, and the use of `pstr1` to stand for the probability of a structural 1 loses its meaning. When `pstr1` equals `-dzeta(1, shape) / (1 - dzeta(1, shape))` this corresponds to the 1-truncated zeta distribution.

Author(s)

T. W. Yee

See Also

`Zeta`, `zetaff`. `Otzeta`,

Examples

```
shape <- 1.5; pstr1 <- 0.3; x <- (-1):7
(ii <- doizeta(x, shape, pstr1 = pstr1))
max(abs(poizeta(1:200, shape) -
        cumsum(1/(1:200)^(1+shape)) / zeta(shape+1)))  # Should be 0

## Not run:  x <- 0:10
par(mfrow = c(2, 1))  # One-Inflated zeta
barplot(rbind(doizeta(x, shape, pstr1 = pstr1), dzeta(x, shape)),
   beside = TRUE, col = c("blue", "orange"),
   main = paste0("OIZeta(", shape, ", pstr1 = ", pstr1,
                 ") (blue) vs Zeta(", shape, ") (orange)"),
   names.arg = as.character(x))

deflat.limit <- -dzeta(1, shape) / pzeta(1, shape, lower.tail = FALSE)
newpstr1 <- round(deflat.limit, 3) + 0.001  # Near the boundary
barplot(rbind(doizeta(x, shape, pstr1 = newpstr1),
               dzeta(x, shape)),
   beside = TRUE, col = c("blue","orange"),
   main = paste0("ODZeta(", shape, ", pstr1 = ", newpstr1,
```

```
                ") (blue) vs Zeta(", shape, ") (orange)"),
   names.arg = as.character(x))
## End(Not run)
```

---

oizeta *One-inflated Zeta Distribution Family Function*

Description

Fits a 1-inflated zeta distribution.

Usage

```
oizeta(lpstr1 = "logitlink", lshape = "loglink",
  type.fitted = c("mean", "shape", "pobs1", "pstr1", "onempstr1"),
  ishape = NULL, gpstr1 = ppoints(8), gshape = exp((-3:3) / 4),
  zero = NULL)
```

Arguments

`lpstr1`, `lshape` For `lpstr1`: the same idea as `zipoisson` except it applies to a structural 1.

`gpstr1`, `gshape`, `ishape`
For initial values. See `CommonVGAMffArguments` for information.

`type.fitted`, `zero`
See `CommonVGAMffArguments` for information.

Details

The 1-inflated zeta distribution is a mixture distribution of the zeta distribution with some probability of obtaining a (structural) 1. Thus there are two sources for obtaining the value 1. This distribution is written here in a way that retains a similar notation to the zero-inflated Poisson, i.e., the probability $P[Y = 1]$ involves another parameter ϕ. See `zipoisson`.

This family function can handle multiple responses.

Value

An object of class `"vglmff"` (see `vglmff-class`). The object is used by modelling functions such as `vglm`, `rrvglm` and `vgam`.

Warning

Under- or over-flow may occur if the data is ill-conditioned. Lots of data is needed to estimate the parameters accurately. Usually, probably the shape parameter is best modelled as intercept-only.

Author(s)

Thomas W. Yee

See Also

Oizeta, zetaff, oazeta, otzeta, diffzeta, zeta, Oizipf.

Examples

```
## Not run:  odata <- data.frame(x2 = runif(nn <- 1000))  # Artificial data
odata <- transform(odata, pstr1 = logitlink(-1 + x2, inv = TRUE),
                          shape = exp(-0.5))
odata <- transform(odata, y1 = roizeta(nn, shape, pstr1 = pstr1))
with(odata, table(y1))
fit1 <- vglm(y1 ~ x2, oizeta(zero = "shape"), odata, trace = TRUE)
coef(fit1, matrix = TRUE)

## End(Not run)
```

Oizipf *One-Inflated Zipf Distribution*

Description

Density, distribution function, quantile function and random generation for the one-inflated Zipf distribution with parameter pstr1.

Usage

```
doizipf(x, N, shape, pstr1 = 0, log = FALSE)
poizipf(q, N, shape, pstr1 = 0)
qoizipf(p, N, shape, pstr1 = 0)
roizipf(n, N, shape, pstr1 = 0)
```

Arguments

x, q, p, n	Same as Uniform.
N, shape	See Zipf.
pstr1	Probability of a structural one (i.e., ignoring the Zipf distribution), called ϕ. The default value of $\phi = 0$ corresponds to the response having an ordinary Zipf distribution.
log	Same as Uniform.

Details

The probability function of Y is 1 with probability ϕ, and $Zipf(N, s)$ with probability $1 - \phi$. Thus

$$P(Y = 1) = \phi + (1 - \phi)P(W = 1)$$

where W is distributed as a $Zipf(N, s)$ random variable. The **VGAM** family function oizeta estimates the two parameters of this model by Fisher scoring.

Value

doizipf gives the density, poizipf gives the distribution function, qoizipf gives the quantile function, and roizipf generates random deviates.

Note

The argument pstr1 is recycled to the required length, and usually has values which lie in the interval $[0, 1]$.

These functions actually allow for the *zero-deflated Zipf* distribution. Here, pstr1 is also permitted to lie in the interval [-dzipf(1, N, s) / (1 - dzipf(1, N, s)), 0]. The resulting probability of a unit count is *less than* the nominal zipf value, and the use of pstr1 to stand for the probability of a structural 1 loses its meaning. When pstr1 equals -dzipf(1, N, s) / (1 - dzipf(1, N, s)) this corresponds to the 1-truncated zipf distribution.

Author(s)

T. W. Yee

See Also

oizeta. Zipf, zipf, Oizeta.

Examples

```
N <- 10; shape <- 1.5; pstr1 <- 0.3; x <- (-1):N
(ii <- doizipf(x, N, shape, pstr1 = pstr1))

## Not run:  x <- 0:10
par(mfrow = c(2, 1))  # One-Inflated zipf
barplot(rbind(doizipf(x, N, shape, pstr1 = pstr1),
              dzipf(x, N, shape)),
   beside = TRUE, col = c("blue", "orange"),
   main = paste0("OIZipf(", N, ", ", shape, ", pstr1 = ", pstr1,
                 ") (blue) vs Zipf(", N, ", ", shape, ") (orange)"),
   names.arg = as.character(x))

deflat.limit <- -dzipf(1, N, shape) / (1 - dzipf(1, N, shape))
newpstr1 <- round(deflat.limit, 3) + 0.001  # Near the boundary
barplot(rbind(doizipf(x, N, shape, pstr1 = newpstr1),
              dzipf(x, N, shape)),
   beside = TRUE, col = c("blue", "orange"),
   main = paste0("ODZipf(", N, ", ", shape, ", pstr1 = ", newpstr1,
                 ") (blue) vs Zipf(", N, ", ", shape, ") (orange)"),
   names.arg = as.character(x))
## End(Not run)
```

oizipf *One-inflated Zipf Distribution Family Function*

Description

Fits a 1-inflated Zipf distribution.

Usage

```
oizipf(N = NULL, lpstr1 = "logitlink", lshape = "loglink",
  type.fitted = c("mean", "shape", "pobs1", "pstr1", "onempstr1"),
  ishape = NULL, gpstr1 = ppoints(8), gshape = exp((-3:3) / 4),
  zero = NULL)
```

Arguments

N	Same as zipf.
lpstr1, lshape	For lpstr1: the same idea as zipoisson except it applies to a structural 1.
gpstr1, gshape, ishape	For initial values. See CommonVGAMffArguments for information.
type.fitted, zero	See CommonVGAMffArguments for information.

Details

The 1-inflated Zipf distribution is a mixture distribution of the Zipf distribution with some probability of obtaining a (structural) 1. Thus there are two sources for obtaining the value 1. This distribution is written here in a way that retains a similar notation to the zero-inflated Poisson, i.e., the probability $P[Y = 1]$ involves another parameter ϕ. See zipoisson.

This family function can handle multiple responses.

Value

An object of class "vglmff" (see vglmff-class). The object is used by modelling functions such as vglm, rrvglm and vgam.

Warning

Under- or over-flow may occur if the data is ill-conditioned. Lots of data is needed to estimate the parameters accurately. Usually, probably the shape parameter is best modelled as intercept-only.

Author(s)

Thomas W. Yee

See Also

Oizipf. zipf, Oizeta.

Examples

```
## Not run:  odata <- data.frame(x2 = runif(nn <- 1000))  # Artificial data
odata <- transform(odata, pstr1 = logitlink(-1 + x2, inv = TRUE),
                          myN   = 10,
                          shape = exp(-0.5))
odata <- transform(odata, y1 = roizipf(nn, N = myN, shape, pstr1 = pstr1))
with(odata, table(y1))
fit1 <- vglm(y1 ~ x2, oizipf(zero = "shape"), odata, trace = TRUE)
coef(fit1, matrix = TRUE)

## End(Not run)
```

`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; more recently it has changed to https://www.theguardian.com/sport/series/london-2012-olympics-data.

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)
```

Otlog *One-truncated Logarithmic Distribution*

Description

Density, distribution function, quantile function, and random generation for the one-truncated logarithmic distribution.

Usage

```
dotlog(x, shape, log = FALSE)
potlog(q, shape, log.p = FALSE)
qotlog(p, shape)
rotlog(n, shape)
```

Arguments

x, q	Vector of quantiles. For the density, it should be a vector with integer values > 1 in order for the probabilities to be positive.
p	vector of probabilities.
n	number of observations. Same as in `runif`.

shape	The parameter value c described in in logff. Here it is called shape because $0 < c < 1$ is the range.
log, log.p	Logical. If log.p = TRUE then all probabilities p are given as log(p).

Details

The one-truncated logarithmic distribution is a logarithmic distribution but with the probability of a one being zero. The other probabilities are scaled to add to unity. Some more details are given in logff.

Value

dotlog gives the density, potlog gives the distribution function, qotlog gives the quantile function, and rotlog generates random deviates.

Note

Given some response data, the **VGAM** family function otlog estimates the parameter shape. Function potlog() suffers from the problems that plog sometimes has.

Author(s)

T. W. Yee

See Also

Gaitdlog, otlog, rlog, Oilog.

Examples

```
dotlog(1:20, 0.5)
rotlog(20, 0.5)

## Not run:  shape <- 0.8; x <- 1:10
plot(x, dotlog(x, shape = shape), type = "h", ylim = 0:1,
     sub = "shape=0.8", las = 1, col = "blue", ylab = "Probability",
     main = "1-truncated logarithmic distn: blue=PMF; orange=CDF")
lines(x+0.1, potlog(x, shape), col = "orange", lty = 3, type = "h")
## End(Not run)
```

---

otlog *One-truncated Logarithmic Distribution*

Description

Estimating the (single) parameter of the 1-truncated logarithmic distribution.

Usage

```
otlog(lshape = "logitlink", gshape = ppoints(8), zero = NULL)
```

Arguments

lshape, gshape, zero
: Same as logff. See CommonVGAMffArguments for information.

Details

The 1-truncated logarithmic distribution is a logarithmic distribution but with the probability of a one being zero. The other probabilities are scaled to add to unity. Some more details can be found at logff. Multiple responses are permitted.

Value

An object of class "vglmff" (see vglmff-class). The object is used by modelling functions such as vglm, and vgam.

Author(s)

T. W. Yee

See Also

Gaitdlog, Otlog, logff, oalog, oilog, simulate.vlm.

Examples

```
## Not run:
odata <- data.frame(y1 = rotlog(1000, shape = logitlink(1/3, inv = TRUE)))
ofit <- vglm(y1 ~ 1, otlog, data = odata, trace = TRUE, crit = "c")
coef(ofit, matrix = TRUE)
Coef(ofit)
with(odata,
    hist(y1, shape = TRUE, breaks = seq(0.5, max(y1) + 0.5, by = 1),
         border = "blue"))
x <- seq(1, with(odata, max(y1)), by = 1)
with(odata, lines(x, dotlog(x, Coef(ofit)[1]), col = "orange",
                  type = "h", lwd = 2))
## End(Not run)
```

Otpospois *One-truncated Positive-Poisson Distribution*

Description

Density, distribution function, quantile function, and random generation for the one-truncated positive-Poisson distribution.

Usage

```
dotpospois(x, lambda, log = FALSE)
potpospois(q, lambda, log.p = FALSE)
qotpospois(p, lambda)
rotpospois(n, lambda)
```

Arguments

x, q, p, n	Same as Pospois.
lambda, log, log.p	Same as Pospois.

Details

The one-truncated positive-Poisson is a Poisson distribution but with the probability of a one and a zero being 0. That is, its support is 2, 3, ... The other probabilities are scaled to add to unity. Some more details are given in pospoisson.

Value

dotpospois gives the density, potpospois gives the distribution function, qotpospois gives the quantile function, and rotpospois generates random deviates.

Note

Given some response data, the **VGAM** family function otpospoisson estimates the parameter lambda.

Author(s)

T. W. Yee

See Also

otpospoisson, Pospois, Oipospois.

Examples

```
dotpospois(1:20, 0.5)
rotpospois(20, 0.5)

## Not run:  lambda <- 4; x <- 1:10
plot(x, dotpospois(x, lambda = lambda), type = "h", ylim = 0:1,
     sub = "lambda=4", las = 1, col = "blue", ylab = "Probability",
     main = "1-truncated positive-Poisson distn: blue=PMF; orange=CDF")
lines(x+0.1, potpospois(x, lambda), col = "orange", lty=3, type = "h")
## End(Not run)
```

`otpospoisson` *One-truncated Poisson Distribution*

Description

Estimating the (single) parameter of the 1-truncated positive Poisson distribution.

Usage

```
otpospoisson(llambda = "loglink",
   type.fitted = c("mean", "lambda", "prob0", "prob1"),
   ilambda = NULL, imethod = 1, zero = NULL)
```

Arguments

`llambda`, `type.fitted`, `ilambda`
: Same as `pospoisson`.

`imethod`, `zero`
: Same as `pospoisson`. See `CommonVGAMffArguments` for information.

Details

The 1-truncated positive Poisson distribution has support on 2, 3, ... It is a Poisson distribution but with the probability of a one or zero being 0. The other probabilities are scaled to add to unity. Some more details can be found at `pospoisson`. Multiple responses are permitted.

Value

An object of class `"vglmff"` (see `vglmff-class`). The object is used by modelling functions such as `vglm`, and `vgam`.

Author(s)

T. W. Yee

See Also

`Otpospois`, `oipospoisson`, `simulate.vlm`.

Examples

```
## Not run:
odata <- data.frame(y1 = rotpospois(1000, lambda = loglink(1, inv = TRUE)))
ofit <- vglm(y1 ~ 1, otpospoisson, data = odata, trace = TRUE, crit = "c")
coef(ofit, matrix = TRUE)
Coef(ofit)
with(odata,
    hist(y1, prob = TRUE, breaks = seq(0.5, max(y1) + 0.5, by = 1),
         border = "blue"))
x <- seq(1, with(odata, max(y1)), by = 1)
with(odata, lines(x, dotpospois(x, Coef(ofit)[1]), col = "orange",
                  type = "h", lwd = 2))
## End(Not run)
```

Otzeta *One-truncated Zeta Distribution*

Description

Density, distribution function, quantile function, and random generation for the one-truncated zeta distribution.

Usage

```
dotzeta(x, shape, log = FALSE)
potzeta(q, shape, log.p = FALSE)
qotzeta(p, shape)
rotzeta(n, shape)
```

Arguments

x, q, p, n	Same as in runif.
shape	The positive shape parameter described in in zetaff. Here it is called shape because it is positive.
log, log.p	Same as in runif.

Details

The one-truncated zeta distribution is a zeta distribution but with the probability of a one being zero. The other probabilities are scaled to add to unity. Some more details are given in zetaff.

Value

dotzeta gives the density, potzeta gives the distribution function, qotzeta gives the quantile function, and rotzeta generates random deviates.

Note

Given some response data, the **VGAM** family function otzeta estimates the parameter shape.

Author(s)

T. W. Yee

See Also

Otzeta, zetaff, Oizeta.

Examples

```
dotzeta(1:20, 0.5)
rotzeta(20, 0.5)

## Not run:  shape <- 0.8; x <- 1:10
plot(x, dotzeta(x, shape = shape), type = "h", ylim = 0:1,
     sub = "shape=0.8", las = 1, col = "blue", ylab = "Probability",
     main = "1-truncated zeta distn: blue=PMF; orange=CDF")
lines(x + 0.1, potzeta(x, shape), col = "orange", lty = 3, type = "h")
## End(Not run)
```

---

otzeta *One-truncated Zeta Distribution Family Function*

Description

Estimates the parameter of the 1-truncated zeta distribution.

Usage

```
otzeta(lshape = "loglink", ishape = NULL,
       gshape = exp((-4:3)/4), zero = NULL)
```

Arguments

lshape, ishape, gshape, zero
: Same as zetaff. See CommonVGAMffArguments for information.

Details

The 1-truncated zeta distribution is the ordinary zeta distribution but with the probability of one being 0. Thus the other probabilities are scaled up (i.e., divided by $1 - P[Y = 1]$). The mean is returned by default as the fitted values. More details can be found at zetaff. Multiple responses are handled.

Value

An object of class "vglmff" (see vglmff-class). The object is used by modelling functions such as vglm, and vgam.

Author(s)

T. W. Yee

See Also

Otzeta, zetaff, oizeta, diffzeta, zeta, dzeta, hzeta, zipf.

Examples

```
## Not run: odata <- data.frame(x2 = runif(nn <- 1000))  # Artificial data
odata <- transform(odata, shape = loglink(-0.25 + x2, inverse = TRUE))
odata <- transform(odata, y1 = rotzeta(nn, shape))
with(odata, table(y1))
ofit <- vglm(y1 ~ x2, otzeta, data = odata, trace = TRUE, crit = "coef")
coef(ofit, matrix = TRUE)

## End(Not run)
```

---

pirates1 *Personal data of the executed pirates associated with Bartholomew Roberts*

Description

The age, names and habitation of 52 pirates who were found guilty of piracy and executed, after the ships associated with Bartholomew Roberts were captured.

Usage

```
data(pirates1)
```

Format

A data frame with the following 3 variables.

`age` a numeric vector, their age in years at the time of trial. Bartholomew Roberts himself was 39 years old at his death.

`name` character.

`habitation` character.

Details

According to Wiki, in February 1722 Captain Ogle was sent by the British Government to find and capture the notorious pirate Bartholomew Roberts (real name: John Roberts, but also known later as Black Bart). When his warship caught up with the *Royal Fortune* he attacked and Bartholomew Roberts was the first to fall, followed by 2 others. The remaining pirates surrendered soon afterwards. A total of 272 men were captured, and of these, 65 were black, and they were sold into

slavery. The remainder were taken to Cape Coast Castle, apart from those who died on the voyage back. The trial was held in April, 1722, and 54 were condemned to death, of whom 52 were hanged and two were reprieved. Of those executed, their personal data (name, age, habitation) were recorded.

Source

Pages 248–249 of Johnson, Captain Charles, (1955) (Editor: Arthur L. Hayward). *A General History of the Robberies and Murders of the Most Notorious Pirates*, London: Routledge and Kegan Paul Ltd. This edition was first published in 1926. The earliest manuscript of the book dates back to 1724.

This data was entered into R by Lucia Pilleri.

See Also

pirates2.

Examples

```
summary(pirates1)
```

`pirates2` *Personal data of the crew of the ship Ranger, associated with the pirate Edward Low*

Description

A data frame containing the age, name, birthplace and verdict of 35 members of a pirate ship associated with Edward Low, who were taken to trial on 10th to 12th July, 1723.

Usage

```
data(pirates2)
```

Format

The variables `age` and `name` are analagous to `pirates1`. The variable `guilty` is binary and 1 means yes, 0 means not guilty. Guilty crew members were executed except for two: John Brown and Patrick Cunningham; they were respited for one year and recommended to the King's favour.

Details

Starting on the 10th July, 1723, the crew of the *Ranger* were judged. The captain of the ship was Charles Harris, and this ship was one of two pirate ships under Captain Edward Low. Their personal data (name, age, place of birth) and verdicts are recorded in Johnson (1955). This data was constructed from pp.295–296 of that book and includes those who were not found guilty (and therefore were not executed). The execution of the 25 men were performed on 19 July near Newport, Rhode Island, USA. The notorious pirate Edward Low himself was brought to trial in 1724 under different circumstances and was hanged in Martinique.

Source

Same as pirates1. This data was entered into R by Lucia Pilleri.

See Also

pirates1.

Examples

```
summary(pirates2)
```

Posbinom *Positive-Binomial Distribution*

Description

Density, distribution function, quantile function and random generation for the positive-binomial distribution.

Usage

```
dposbinom(x, size, prob, log = FALSE)
pposbinom(q, size, prob)
qposbinom(p, size, prob)
rposbinom(n, size, prob)
```

Arguments

x, q	vector of quantiles.
p	vector of probabilities.
n	number of observations. Fed into runif.
size	number of trials. It is the N symbol in the formula given in posbinomial and should be positive.
prob	probability of success on each trial. Should be in $(0, 1)$.
log	See dbinom.

Details

The positive-binomial distribution is a binomial distribution but with the probability of a zero being zero. The other probabilities are scaled to add to unity. The mean therefore is

$$\mu/(1-(1-\mu)^N)$$

where μ is the argument prob above. As μ increases, the positive-binomial and binomial distributions become more similar. Unlike similar functions for the binomial distribution, a zero value of prob is not permitted here.

Value

dposbinom gives the density, pposbinom gives the distribution function, qposbinom gives the quantile function, and rposbinom generates random deviates.

Note

These functions are or are likely to be deprecated. Use Gaitdbinom instead.

For dposbinom(), if arguments size or prob equal 0 then a NaN is returned.

The family function posbinomial estimates the parameters by maximum likelihood estimation.

Author(s)

T. W. Yee.

See Also

posbinomial, dposbern, Gaitdbinom, zabinomial, zibinomial, Binomial.

Examples

```
prob <- 0.2; size <- 10
table(y <- rposbinom(n = 1000, size, prob))
mean(y)  # Sample mean
size * prob / (1 - (1 - prob)^size)  # Population mean

(ii <- dposbinom(0:size, size, prob))
cumsum(ii) - pposbinom(0:size, size, prob)  # Should be 0s
table(rposbinom(100, size, prob))

table(qposbinom(runif(1000), size, prob))
round(dposbinom(1:10, size, prob) * 1000)  # Should be similar

## Not run:  barplot(rbind(dposbinom(x = 0:size, size, prob),
                           dbinom(x = 0:size, size, prob)),
        beside = TRUE, col = c("blue", "green"),
        main = paste("Positive-binomial(", size, ",",
                     prob, ") (blue) vs",
        " Binomial(", size, ",", prob, ") (green)", sep = ""),
        names.arg = as.character(0:size), las = 1)
## End(Not run)

# Simulated data example
nn <- 1000; sizeval1 <- 10; sizeval2 <- 20
pdata <- data.frame(x2 = seq(0, 1, length = nn))
pdata <- transform(pdata, prob1  = logitlink(-2 + 2 * x2, inv = TRUE),
                          prob2  = logitlink(-1 + 1 * x2, inv = TRUE),
                          sizev1 = rep(sizeval1, len = nn),
                          sizev2 = rep(sizeval2, len = nn))
pdata <- transform(pdata, y1 = rposbinom(nn, sizev1, prob = prob1),
                          y2 = rposbinom(nn, sizev2, prob = prob2))
with(pdata, table(y1))
```

```
with(pdata, table(y2))
# Multiple responses
fit2 <- vglm(cbind(y1, y2) ~ x2, posbinomial(multip = TRUE),
             trace = TRUE, pdata, weight = cbind(sizev1, sizev2))
coef(fit2, matrix = TRUE)
```

Posnegbin *Positive-Negative Binomial Distribution*

Description

Density, distribution function, quantile function and random generation for the positive-negative binomial distribution.

Usage

```
dposnegbin(x, size, prob = NULL, munb = NULL, log = FALSE)
pposnegbin(q, size, prob = NULL, munb = NULL,
           lower.tail = TRUE, log.p = FALSE)
qposnegbin(p, size, prob = NULL, munb = NULL)
rposnegbin(n, size, prob = NULL, munb = NULL)
```

Arguments

x, q	vector of quantiles.
p	vector of probabilities.
n	number of observations. Fed into runif.
size, prob, munb, log	Same arguments as that of an ordinary negative binomial distribution (see dnbinom). Some arguments have been renamed slightly. Short vectors are recycled. The parameter 1/size is known as a dispersion parameter; as size approaches infinity, the negative binomial distribution approaches a Poisson distribution. Note that prob must lie in $(0, 1)$, otherwise a NaN is returned.
log.p, lower.tail	Same arguments as that of an ordinary negative binomial distribution (see pnbinom).

Details

The positive-negative binomial distribution is a negative binomial distribution but with the probability of a zero being zero. The other probabilities are scaled to add to unity. The mean therefore is

$$\mu/(1 - p(0))$$

where μ the mean of an ordinary negative binomial distribution.

Value

dposnegbin gives the density, pposnegbin gives the distribution function, qposnegbin gives the quantile function, and rposnegbin generates n random deviates.

Note

These functions are or are likely to be deprecated. Use Gaitdnbinom instead.

Author(s)

T. W. Yee

References

Welsh, A. H., Cunningham, R. B., Donnelly, C. F. and Lindenmayer, D. B. (1996). Modelling the abundances of rare species: statistical models for counts with extra zeros. *Ecological Modelling*, **88**, 297–308.

See Also

Gaitdnbinom, posnegbinomial, zanegbinomial, zinegbinomial, rnbinom.

Examples

```
munb <- 5; size <- 4; n <- 1000
table(y <- rposnegbin(n, munb = munb, size = size))
mean(y)  # Sample mean
munb / (1 - (size / (size + munb))^size)  # Population mean
munb / pnbinom(0, mu = munb, size, lower.tail = FALSE)  # Same

x <- (-1):17
(ii <- dposnegbin(x, munb = munb, size = size))
max(abs(cumsum(ii) - pposnegbin(x, munb = munb, size)))  # 0?

## Not run:  x <- 0:10
barplot(rbind(dposnegbin(x, munb = munb, size = size),
                dnbinom(x, mu   = munb, size = size)),
   beside = TRUE, col = c("blue","green"),
   main = paste0("dposnegbin(munb = ", munb, ", size = ", size,
                ") (blue) vs dnbinom(mu = ", munb,
                ", size = ", size, ") (green)"),
   names.arg = as.character(x))
## End(Not run)

# Another test for pposnegbin()
nn <- 5000
mytab <- cumsum(table(rposnegbin(nn, munb = munb, size))) / nn
myans <- pposnegbin(sort(as.numeric(names(mytab))), munb = munb, size)
max(abs(mytab - myans))  # Should be 0
```

Pospois *Positive-Poisson Distribution*

Description

Density, distribution function, quantile function and random generation for the positive-Poisson distribution.

Usage

```
dpospois(x, lambda, log = FALSE)
ppospois(q, lambda)
qpospois(p, lambda)
rpospois(n, lambda)
```

Arguments

x, q	vector of quantiles.
p	vector of probabilities.
n	number of observations. Fed into runif.
lambda	vector of positive means (of an ordinary Poisson distribution). Short vectors are recycled.
log	logical.

Details

The positive-Poisson distribution is a Poisson distribution but with the probability of a zero being zero. The other probabilities are scaled to add to unity. The mean therefore is

$$\lambda/(1-\exp(-\lambda)).$$

As λ increases, the positive-Poisson and Poisson distributions become more similar. Unlike similar functions for the Poisson distribution, a zero value of lambda returns a NaN.

Value

dpospois gives the density, ppospois gives the distribution function, qpospois gives the quantile function, and rpospois generates random deviates.

Note

These functions are or are likely to be deprecated. Use Gaitdpois instead.

The family function pospoisson estimates λ by maximum likelihood estimation.

Author(s)

T. W. Yee

See Also

Gaitdpois, pospoisson, zapoisson, zipoisson, rpois.

Examples

```
lambda <- 2; y = rpospois(n = 1000, lambda)
table(y)
mean(y)  # Sample mean
lambda / (1 - exp(-lambda))  # Population mean

(ii <- dpospois(0:7, lambda))
cumsum(ii) - ppospois(0:7, lambda)  # Should be 0s
table(rpospois(100, lambda))

table(qpospois(runif(1000), lambda))
round(dpospois(1:10, lambda) * 1000)  # Should be similar

## Not run:  x <- 0:7
barplot(rbind(dpospois(x, lambda), dpois(x, lambda)),
        beside = TRUE, col = c("blue", "orange"),
        main = paste("Positive Poisson(", lambda, ") (blue) vs",
        " Poisson(", lambda, ") (orange)", sep = ""),
        names.arg = as.character(x), las = 1, lwd = 2)
## 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. This site may have become stale.

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.

`pub1stAuthor` a numeric vector, the total number of publications which are first authored by the person.

`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, C. S. Withers.

Source

The data was compiled from the equivalent of https://mathscinet.ams.org/mathscinet/publications-search by J. T. Gray in April 2014.

Examples

```
profs.nz
profs.nz[order(with(profs.nz, pubtotal), decreasing = TRUE), ]
## Not run:
plot(pub1stAuthor / pubtotal ~ pubtotal,
     main = "Professors of Statistics in NZ",
     xlab = "Number of publications in MathSciNet",
     ylab = "Proportion of first-authored papers",
     data = profs.nz, col = "blue", las = 1, type = "n")
with(profs.nz, text(pubtotal, y = pub1stAuthor / pubtotal,
     labels = initials, col = "blue", las = 1))

## End(Not run)
```

---

rar.df *Rock and Roll and Life Expectancy*

Description

Ages at death in rock and roll.

Usage

```
data(rar.df)
```

Format

A data frame with the following variables.

age Numeric. Age when the person died, in years. A handful of values are approximate or missing.

yod Numeric. The year that the person died.

Details

This type of data is very subjective because many definitions cannot ever be finalized. The Wiki website was written by an anonymous person and the data collected informally and there is no quality assurance. Of course, it could be debated as whether a particular person should be included. And there will be many people who should be there but are not. Nevertheless, it can be considered a fun data set that should be of interest to anybody liking modern music or the history of such in musicology.

In the source, some peoples' ages were listed as a decade, e.g., age was `40s` or `50--51`. These were either replaced by `NA` or the mean of the endpoints was taken.

Source

The lists at https://en.wikipedia.org/wiki/List_of_deaths_in_rock_and_roll. were downloaded and edited.

Examples

```
summary(rar.df)
```

---

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, https://www.world.rugby/.

Examples

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

`SardiniaHotels` *Data from hotels in Sardinia, Italy*

Description

This data set contains information and satisfaction scores appearing on the TripAdvisor website between the years 2008 and 2016 regarding hotels in Sardinia, Italy.

The satisfaction data refer to the reputation of hotel located along Sardinian coasts, as expressed by clients with respect to different services (e.g., breakfast, restaurant, swimming pool) offered by the hotel.

Usage

```
data(SardiniaHotels)
```

Format

A data frame with 518 rows and 43 columns (variables). Each row refers to a single hotel.

The following variables are included in the dataset:

`municipality` a factor, the municipality where the hotel is located.

`stars` an ordered factor with levels:

- `1OR2stars` for 1 star or 2 star hotels,
- `3stars` 3 star hotels,
- `residence`,
- `4stars`, 4 star hotels,
- `5starsORresort`, 5 star hotels or resorts.

`area` a factor with levels related to the area of the Sardinian coast where each single hotel is located: `AlgheroSassari`, `CagliariVillasimius`, `CostaSmeralda`, `DorgaliOrosei`, `Gallura`, `NurraAnglona`, `Ogliastra`, `Olbia`, `OristanoBosa`, `PulaChia`, `Sarrabus`, `Sulcis`.

`seaLocation` a factor with levels `yes` (if the hotel is located close to the sea) and `no` (otherwise).

`excellent` a numeric vector, the number of people that expressed the highest level of satisfaction.

`good` a numeric vector, the number of people that expressed a good level of satisfaction.

`average` a numeric vector, the number of people that expressed an average level of satisfaction.

`bad` a numeric vector, the number of people that expressed a bad level of satisfaction.

`poor` a numeric vector, the number of people that expressed the lowest level of satisfaction.

`family` a numeric vector, the number of people travelling with family.

`couple` a numeric vector, the number of people travelling with their partner.

`single` a numeric vector, the number of people travelling alone.

`business` a numeric vector, the number of people travelling for work.

`MarMay` a numeric vector, the number of people travelling during the period March to May.

`JunAug` a numeric vector, the number of people travelling during the period June to August.

`SepNov` a numeric vector, the number of people travelling during the period September to November.

`DecFeb` a numeric vector, the number of people travelling during the period December to February.

`location` a numeric vector, the satisfaction score expressed by tourists towards the location.

`sleepQuality` a numeric vector, the satisfaction score expressed by tourists towards the sleep quality.

`room` a numeric vector, the satisfaction score expressed by tourists towards the comfort and quality of the room.

`services` a numeric vector, the satisfaction score expressed by tourists towards the quality of the services.

`priceQualityRate` a numeric vector, the satisfaction score expressed by tourists towards ratio between price and quality.

`cleaning` a numeric vector, the satisfaction score expressed by tourists towards level of room and hotel cleaning.

`bt1` a factor with levels `breakfast`, `cleaning`, `location`, `overall`, `price`, `restaurant`, `room`, `services`, `staff`, `structure` and `Wi-Fi`.

It expresses the 1st most used word in reviews for a hotel.

`ratebt1` a factor with levels `-1` (if the satisfaction score espressed in `bt1` is prevalently negative) and `1` (if the satisfaction score espressed in `bt1` is prevalently positive).

`bt2` a factor with levels `breakfast`, `cleaning`, `location`, `overall`, `price`, `restaurant`, `room`, `services`, `staff`, `structure` and `Wi-Fi`.

It expresses the 2nd most used word in reviews for a hotel.

`ratebt2` a factor with levels `-1` (if the satisfaction score espressed in `bt2` is prevalently negative) and `1` (if the satisfaction score espressed in `bt2` is prevalently positive).

`bt3` similar to `bt1` and `bt2`, but with a corresponding different ranking.

`bt4` similar to `bt1` and `bt2`, but with a corresponding different ranking.

`bt5` similar to `bt1` and `bt2`, but with a corresponding different ranking.

`bt6` similar to `bt1` and `bt2`, but with a corresponding different ranking.

`bt7` similar to `bt1` and `bt2`, but with a corresponding different ranking.

`bt8` similar to `bt1` and `bt2`, but with a corresponding different ranking.

`bt9` similar to `bt1` and `bt2`, but with a corresponding different ranking.

`bt10` similar to `bt1` and `bt2`, but with a corresponding different ranking.

`ratebt3` similar to `ratebt1` and `ratebt2`, but with a corresponding different ranking.

`ratebt4` similar to `ratebt1` and `ratebt2`, but with a corresponding different ranking.

`ratebt5` similar to `ratebt1` and `ratebt2`, but with a corresponding different ranking.

`ratebt6` similar to `ratebt1` and `ratebt2`, but with a corresponding different ranking.

`ratebt7` similar to `ratebt1` and `ratebt2`, but with a corresponding different ranking.

`ratebt8` similar to `ratebt1` and `ratebt2`, but with a corresponding different ranking.

`ratebt9` similar to `ratebt1` and `ratebt2`, but with a corresponding different ranking.

`ratebt10` similar to `ratebt1` and `ratebt2`, but with a corresponding different ranking.

Details

These data were manually collected during March–June 2016 by students of the class of "Statistics for Turism" at the University of Cagliari, Italy (Bachelor's degree in Tourism Economics and Managment), under the supervision of Prof. Claudio Conversano and Dr. Giulia Contu.

Many of the variables fall into several natural groups, e.g., [`municipality`, `stars`, `area`, `seaLocation`]; [`excellent`, `good`, `average`, `bad`, `poor`]; [`MarMay`, `JunAug`, `SepNov`, `DecFeb`]; [`family`, `couple`, `single`, `business`]; [`location`,...`cleaning`]; [`bt1`,...`bt10`]; [`ratebt1`,...`ratebt10`].

Source

TripAdvisor, `https://www.tripadvisor.it/`.

Examples

```
data(SardiniaHotels)
summary(SardiniaHotels)
```

---

smqP *Smoking Data in the NHANES 2017–2020 Cycle*

Description

Selected variables mainly focussed on the smoking questionnaire of the National Health and Nutrition Examination Survey collected during 1.5 cycles just prior to Covid-19.

Usage

```
data(smqP)
```

Format

A data frame with the following variables.

SEQN Identifier for individuals that can be used to merge in other data sets from the same cycle.

TC30 Numeric. Tobacco consumption: average number of cigarettes/day during past 30 days. Aka `SMD650` with some preprocessing.

SCA Smoking cessation age (SCA). Is missing for those who have not quit. Computing this involved variables SMQ050Q, SMQ050U and age. The variable SMQ050Q is for How long since (you) quit smoking cigarettes? and SMQ050U are the units (e.g., years, months, days). The variable SMQ050Q is right-censored at 50 years (66666 means 50 or more years) and for such people SCA was set to NA.

TCquit Tobacco consumption: number of cigarettes smoked per day when quit. Aka SMD057 with some preprocessing.

SIA Smoking initiation age (SIA): age when individuals started smoking cigarettes regularly. Aka SMD030 with some preprocessing.

age, gender Age (RIDAGEYR) when surveyed; the value 80 is right-censored, i.e., subjects 80 years or older were recorded as 80. Gender is RIAGENDR .

race, race2 Race (RIDRETH1) and a binary simplification of race ("Non-Hispanic White" versus "Others").

educ, marital Education (DMDEDUC2) and marital status (DMDMARTZ). Both variables were not collected for those aged 12–19, hence those are NAs.

poverty_ratio Ratio of family income to poverty (INDFMPIR).

meals_fastfood_pizza, diet, readytoeat_foods For example, for the first one, this is the number of meals from a fast food or pizza place (DBD900) during the last 7 days, where the value 5555 means more than 21 meals per week, the value 7777 means the person refused to answer, and the value 9999 means the person didn't know. The other variables may be described later in more detail.

frozen_meals To be described later.

people_fam_smoking, people_home_smoking Numeric. For people_fam_smoking, this is Number of people who live here smoke tobacco? (SMD460). For this, 0 is: No one in household is a smoker; 1 is: 1 household member is a smoker; 2 is: 2 or more household members are smokers.

For people_home_smoking, this is the number of people who smoke inside this home? (SMD470). For this, 0 is: No one smokes inside the house; 1 is: 1 household member smokes inside the house; 2 is: 2 or more household members smoke inside the house.

use_cigarettes, use_pipes, use_cigars Use during the last 5 days of smoking variants. The codes are SMQ690A, SMQ690B, SMQ690C.

use_hookah, use_ecigarettes The codes are SMQ690G, SMQ690H.

use_snuff, use_otherSmokeless The codes are SMQ690E, SMQ690K.

passiveSmoke_job, passiveSmoke_rest Binary 0 (no) or 1 (yes) measuring exposure to passive smoke at certain places in the past 7 days, e.g., While you were working at a job or business outside of the home, did someone else smoke cigarettes or other tobacco products indoors? The codes are SMQ858, SMQ862.

passiveSmoke_bar, passiveSmoke_car See above. The codes are SMQ868, SMQ872.

passiveSmoke_home, passiveSmoke_other See above. The codes are SMQ876, SMQ880.

passiveSmoke_ecigarettes See above. The code is SMQ940.

Details

The National Health and Nutrition Examination Survey (NHANES) is a well-known longitudinal study located in USA (a country just north of Mexico). This data frame shares a selection of variables mainly to do with the smoking questionnaire (codeword: `SMQ`); some demographic and anthropometric variables might also be included and/or added later. The significance of `P` is that the 2019–2020 cycle was not completed due to Covid-19, hence this data concerns 2017–2020 and is a merging of the 2017–2018 cycle (data codenamed `J`) with further data collected just prior to the pandemic.

The original data has been preprocessed and/or simplified. For example, `"Don't know"` and `"Refused"` usually have been converted to a `NA`.

This data frame is subject to change, especially with the addition of new variables.

Source

The URL https://wwwn.cdc.gov/Nchs/Nhanes/ provides an entry point from which many data sets may be downloaded. Comprehensive documentation is available there too.

The data was downloaded in late 2022 by Luca Frigau, University of Cagliari, and some subsequent edits were made by Thomas Yee.

Examples

```
summary(smqP)
```

---

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)
```

`T101` *Alcohol, Cigarette, and Marijuana Use, by Gender and Race*

Description

Alcohol, cigarette, and marijuana use, by gender and race, in 2276 senior high school students (Table 10.1 of Agresti, 2013).

Usage

```
data(T101)
```

Format

The complete data frame contains 32 observations (rows) and the following columns:

alc Alcohol use: 1 is yes, 0 is no.

cig Cigarette use: 1 is yes, 0 is no.

marj Marijuana use: 1 is yes, 0 is no.

race Factor: `White` or `Other`.

gender Factor: `F` or `M`.

count Frequency. Data frame `T101` has the 0 counts in `tb101` deleted.

Source

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

References

Agresti, A. (2013) *Categorical Data Analysis*, Third Edition. Hoboken, NJ, USA: John Wiley & Sons.

Examples

```
summary(T101)
```

tb101 *Alcohol, Cigarette, and Marijuana Use, by Gender and Race*

Description

Alcohol, cigarette, and marijuana use, by gender and race, in 2276 senior high school students (Table 10.1 of Agresti, 2013).

Usage

```
data(tb101)
```

Format

The complete data frame contains 32 observations (rows) and the following columns:

alc Alcohol use: 1 is yes, 0 is no.

cig Cigarette use: 1 is yes, 0 is no.

marj Marijuana use: 1 is yes, 0 is no.

race Factor: `White` or `Other`.

gender Factor: `F` or `M`.

count Frequency. Data frame `T101` has the 0 counts in `tb101` deleted.

Source

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

References

Agresti, A. (2013) *Categorical Data Analysis*, Third Edition. Hoboken, NJ, USA: John Wiley & Sons.

Examples

```
summary(tb101)
```

Tikuv *A Short-tailed Symmetric Distribution*

Description

Density, cumulative distribution function, quantile function and random generation for the short-tailed symmetric distribution of Tiku and Vaughan (1999).

Usage

```
dtikuv(x, d, mean = 0, sigma = 1, log = FALSE)
ptikuv(q, d, mean = 0, sigma = 1, lower.tail = TRUE, log.p = FALSE)
qtikuv(p, d, mean = 0, sigma = 1, lower.tail = TRUE, log.p = FALSE, ...)
rtikuv(n, d, mean = 0, sigma = 1, Smallno = 1.0e-6)
```

Arguments

x, q	vector of quantiles.
p	vector of probabilities.
n	number of observations. Same as in runif.
d, mean, sigma	arguments for the parameters of the distribution. See tikuv for more details. For rtikuv, arguments mean and sigma must be of length 1.
Smallno	Numeric, a small value used by the rejection method for determining the lower and upper limits of the distribution. That is, ptikuv(L) < Smallno and ptikuv(U) > 1-Smallno where L and U are the lower and upper limits respectively.
...	Arguments that can be passed into uniroot.
log	Logical. If log = TRUE then the logarithm of the density is returned.
lower.tail, log.p	Same meaning as in pnorm or qnorm.

Details

See tikuv for more details.

Value

dtikuv gives the density, ptikuv gives the cumulative distribution function, qtikuv gives the quantile function, and rtikuv generates random deviates.

Author(s)

T. W. Yee and Kai Huang

See Also

tikuv.

Examples

```
## Not run:  par(mfrow = c(2, 1))
x <- seq(-5, 5, len = 401)
plot(x, dnorm(x), type = "l", col = "black", ylab = "", las = 1,
     main = "Black is standard normal, others are dtikuv(x, d)")
lines(x, dtikuv(x, d = -10), col = "orange")
lines(x, dtikuv(x, d = -1 ), col = "blue")
lines(x, dtikuv(x, d =  1 ), col = "green")
legend("topleft", col = c("orange","blue","green"), lty = rep(1, len = 3),
       legend = paste("d =", c(-10, -1, 1)))

plot(x, pnorm(x), type = "l", col = "black", ylab = "", las = 1,
     main = "Black is standard normal, others are ptikuv(x, d)")
lines(x, ptikuv(x, d = -10), col = "orange")
lines(x, ptikuv(x, d = -1 ), col = "blue")
lines(x, ptikuv(x, d =  1 ), col = "green")
legend("topleft", col = c("orange","blue","green"), lty = rep(1, len = 3),
       legend = paste("d =", c(-10, -1, 1)))
## End(Not run)

probs <- seq(0.1, 0.9, by = 0.1)
ptikuv(qtikuv(p = probs, d =  1), d = 1) - probs  # Should be all 0
```

---

tikuv *Short-tailed Symmetric Distribution Family Function*

Description

Fits the short-tailed symmetric distribution of Tiku and Vaughan (1999).

Usage

```
tikuv(d, lmean = "identitylink", lsigma = "loglink",
      isigma = NULL, zero = "sigma")
```

Arguments

d	The d parameter. It must be a single numeric value less than 2. Then $h = 2 - d > 0$ is another parameter.
lmean, lsigma	Link functions for the mean and standard deviation parameters of the usual univariate normal distribution (see **Details** below). They are μ and σ respectively. See Links for more choices.
isigma	Optional initial value for σ. A NULL means a value is computed internally.
zero	A vector specifying which linear/additive predictors are modelled as intercept-only. The values can be from the set {1,2}, corresponding respectively to μ, σ. If zero = NULL then all linear/additive predictors are modelled as a linear combination of the explanatory variables. For many data sets having zero = 2 is a good idea. See CommonVGAMffArguments for information.

Details

The short-tailed symmetric distribution of Tiku and Vaughan (1999) has a probability density function that can be written

$$f(y) = \frac{K}{\sqrt{2\pi}\sigma}\left[1 + \frac{1}{2h}\left(\frac{y-\mu}{\sigma}\right)^2\right]^2 \exp\left(-\frac{1}{2}(y-\mu)^2/\sigma^2\right)$$

where $h = 2 - d > 0$, K is a function of h, $-\infty < y < \infty$, $\sigma > 0$. The mean of Y is $E(Y) = \mu$ and this is returned as the fitted values.

Value

An object of class `"vglmff"` (see `vglmff-class`). The object is used by modelling functions such as `vglm`, and `vgam`.

Warning

Under- or over-flow may occur if the data is ill-conditioned, e.g., when d is very close to 2 or approaches `-Inf`.

Note

The density function is the product of a univariate normal density and a polynomial in the response y. The distribution is bimodal if $d > 0$, else is unimodal. A normal distribution arises as the limit as d approaches $-\infty$, i.e., as h approaches ∞. Fisher scoring is implemented. After fitting the value of `d` is stored in `@misc` with component name `d`.

Author(s)

Thomas W. Yee

References

Akkaya, A. D. and Tiku, M. L. (2008). Short-tailed distributions and inliers. *Test*, **17**, 282–296.

Tiku, M. L. and Vaughan, D. C. (1999). A family of short-tailed symmetric distributions. *Technical report, McMaster University, Canada*.

See Also

`dtikuv`, `uninormal`.

Examples

```
m <- 1.0; sigma <- exp(0.5)
tdata <- data.frame(y = rtikuv(1000, d = 1, m = m, s = sigma))
tdata <- transform(tdata, sy = sort(y))
fit <- vglm(y ~ 1, tikuv(d = 1), data = tdata, trace = TRUE)
coef(fit, matrix = TRUE)
(Cfit <- Coef(fit))
with(tdata, mean(y))
```

```
## Not run:  with(tdata, hist(y, prob = TRUE))
lines(dtikuv(sy, d = 1, m = Cfit[1], s = Cfit[2]) ~ sy,
      data = tdata, col = "orange")
## End(Not run)
```

---

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

Description

Rainbow and brown 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)
```

---

triangle *Triangle Distribution Family Function*

Description

Estimating the parameter of the triangle distribution by maximum likelihood estimation.

Usage

```
triangle(lower = 0, upper = 1,
         link = "extlogitlink(min = 0, max = 1)", itheta = NULL)
```

Arguments

lower, upper	lower and upper limits of the distribution. Must be finite. Called A and B respectively below.
link	Parameter link function applied to the parameter θ, which lies in (A, B). See Links for more choices. The default constrains the estimate to lie in the interval.
itheta	Optional initial value for the parameter. The default is to compute the value internally.

Details

The triangle distribution has a probability density function that consists of two lines joined at θ, which is the location of the mode. The lines intersect the $y = 0$ axis at A and B. Here, Fisher scoring is used.

On fitting, the `extra` slot has components called `lower` and `upper` which contains the values of the above arguments (recycled to the right length). The fitted values are the mean of the distribution, which is $(A + B + \theta)/3$.

Value

An object of class `"vglmff"` (see `vglmff-class`). The object is used by modelling functions such as `vglm` and `vgam`.

Warning

The MLE regularity conditions do not hold for this distribution (e.g., the first derivative evaluated at the mode does not exist because it is not continuous) so that misleading inferences may result, e.g., in the `summary` and `vcov` of the object. Additionally, convergence to the MLE often appears to fail.

Note

The response must contain values in (A, B). For most data sets (especially small ones) it is very common for half-stepping to occur.

Arguments `lower` and `upper` and `link` must match. For example, setting `lower = 0.2` and `upper = 4` and `link = "extlogitlink(min = 0.2, max = 4.1)"` will result in an error. Ideally `link = "extlogitlink(min = lower, max = upper)"` ought to work but it does not (yet)! Minimal error checking is done for this deficiency.

Author(s)

T. W. Yee

References

Kotz, S. and van Dorp, J. R. (2004). Beyond Beta: Other Continuous Families of Distributions with Bounded Support and Applications. Chapter 1. World Scientific: Singapore.

Nguyen, H. D. and McLachlan, G. J. (2016). Maximum likelihood estimation of triangular and polygon distributions. *Computational Statistics and Data Analysis*, **102**, 23–36.

See Also

`Triangle`, `Topple`, `simulate.vlm`.

Examples

```
## Not run:
# Example 1
tdata <- data.frame(y = rtriangle(n <- 3000, theta = 3/4))
fit <- vglm(y ~ 1, triangle(link = "identitylink"), tdata,
            trace = TRUE)
coef(fit, matrix = TRUE)
Coef(fit)
head(fit@extra$lower)
head(fitted(fit))
with(tdata, mean(y))

# Example 2; Kotz and van Dorp (2004), p.14
rdata <- data.frame(y = c(0.1,0.25,0.3,0.4,0.45, 0.6, 0.75, 0.8))
fit <- vglm(y ~ 1, triangle(link = "identitylink"), rdata,
           trace = TRUE, crit = "coef", maxit = 1000)
Coef(fit)  # The MLE is the 3rd order statistic, which is 0.3.
fit <- vglm(y ~ 1, triangle(link = "identitylink"), rdata,
           trace = TRUE, crit = "coef", maxit = 1001)
Coef(fit)  # The MLE is the 3rd order statistic, which is 0.3.

## End(Not run)
```

---

tube10 *London Underground (Tube) Passenger Counts for November 2010*

Description

The data set contains counts of the number of passengers entering London Underground stations (also known as *the Tube*) on a typical day of November 2010 in quarter-hour time blocks.

Usage

```
data("tube10")
```

Format

A data frame with 100 observations on the following 270 variables.

`ActonTown` a numeric vector

`Aldgate` a numeric vector

`AldgateEast` a numeric vector

`Alperton` a numeric vector

`Amersham` a numeric vector

`Angel` a numeric vector

`Archway` a numeric vector

ArnosGrove a numeric vector

Arsenal a numeric vector

BakerStreet a numeric vector

Balham a numeric vector

BankAndMonument a numeric vector

Barbican a numeric vector

Barking a numeric vector

Barkingside a numeric vector

BaronsCourt a numeric vector

Bayswater a numeric vector

Becontree a numeric vector

BelsizePark a numeric vector

Bermondsey a numeric vector

BethnalGreen a numeric vector

Blackfriars a numeric vector

BlackhorseRoad a numeric vector

BondStreet a numeric vector

Borough a numeric vector

BostonManor a numeric vector

BoundsGreen a numeric vector

BowRoad a numeric vector

BrentCross a numeric vector

Brixton a numeric vector

BromleyByBow a numeric vector

BuckhurstHill a numeric vector

BurntOak a numeric vector

CaledonianRoad a numeric vector

CamdenTown a numeric vector

CanadaWater a numeric vector

CanaryWharf a numeric vector

CanningTown a numeric vector

CannonStreet a numeric vector

CanonsPark a numeric vector

ChalfontAndLatimer a numeric vector

ChalkFarm a numeric vector

ChanceryLane a numeric vector

CharingCross a numeric vector

Chesham a numeric vector

Chigwell a numeric vector

ChiswickPark a numeric vector

Chorleywood a numeric vector

ClaphamCommon a numeric vector

ClaphamNorth a numeric vector

ClaphamSouth a numeric vector

Cockfosters a numeric vector

Colindale a numeric vector

ColliersWood a numeric vector

CoventGarden a numeric vector

Croxley a numeric vector

DagenhamEast a numeric vector

DagenhamHeathway a numeric vector

Debden a numeric vector

DollisHill a numeric vector

EalingBroadway a numeric vector

EalingCommon a numeric vector

EarlsCourt a numeric vector

EastActon a numeric vector

EastFinchley a numeric vector

EastHam a numeric vector

EastPutney a numeric vector

Eastcote a numeric vector

Edgware a numeric vector

EdgwareRoadBak a numeric vector

EdgwareRoadCir a numeric vector

ElephantAndCastle a numeric vector

ElmPark a numeric vector

Embankment a numeric vector

Epping a numeric vector

Euston a numeric vector

EustonSquare a numeric vector

Fairlop a numeric vector

Farringdon a numeric vector

FinchleyCentral a numeric vector

FinchleyRoad a numeric vector

FinsburyPark a numeric vector

FulhamBroadway a numeric vector

GantsHill a numeric vector

GloucesterRoad a numeric vector

GoldersGreen a numeric vector

GoldhawkRoad a numeric vector

GoodgeStreet a numeric vector

GrangeHill a numeric vector

GreatPortlandStreet a numeric vector

GreenPark a numeric vector

Greenford a numeric vector

Gunnersbury a numeric vector

Hainault a numeric vector

HammersmithDis a numeric vector

HammersmithHC a numeric vector

Hampstead a numeric vector

HangerLane a numeric vector

Harlesden a numeric vector

HarrowAndWealdstone a numeric vector

HarrowOnTheHill a numeric vector

HattonCross a numeric vector

HeathrowTerminals123 a numeric vector

HeathrowTerminal4 a numeric vector

HeathrowTerminal5 a numeric vector

HendonCentral a numeric vector

HighBarnet a numeric vector

HighStreetKensington a numeric vector

HighburyAndIslington a numeric vector

Highgate a numeric vector

Hillingdon a numeric vector

Holborn a numeric vector

HollandPark a numeric vector

HollowayRoad a numeric vector

Hornchurch a numeric vector

HounslowCentral a numeric vector

HounslowEast a numeric vector

HounslowWest a numeric vector

HydeParkCorner a numeric vector

Ickenham a numeric vector

Kennington a numeric vector

KensalGreen a numeric vector

KensingtonOlympia a numeric vector

KentishTown a numeric vector

Kenton a numeric vector

KewGardens a numeric vector

Kilburn a numeric vector

KilburnPark a numeric vector

KingsCrossStPancras a numeric vector

Kingsbury a numeric vector

Knightsbridge a numeric vector

LadbrokeGrove a numeric vector

LambethNorth a numeric vector

LancasterGate a numeric vector

LatimerRoad a numeric vector

LeicesterSquare a numeric vector

Leyton a numeric vector

Leytonstone a numeric vector

LiverpoolStreet a numeric vector

LondonBridge a numeric vector

Loughton a numeric vector

MaidaVale a numeric vector

ManorHouse a numeric vector

MansionHouse a numeric vector

MarbleArch a numeric vector

Marylebone a numeric vector

MileEnd a numeric vector

MillHillEast a numeric vector

MoorPark a numeric vector

Moorgate a numeric vector

Morden a numeric vector

MorningtonCrescent a numeric vector

Neasden a numeric vector

NewburyPark a numeric vector

NorthActon a numeric vector

NorthEaling a numeric vector

NorthGreenwich a numeric vector

NorthHarrow a numeric vector

NorthWembley a numeric vector

Northfields a numeric vector

Northolt a numeric vector

NorthwickPark a numeric vector

Northwood a numeric vector

NorthwoodHills a numeric vector

NottingHillGate a numeric vector

Oakwood a numeric vector

OldStreet a numeric vector

Osterley a numeric vector

Oval a numeric vector

OxfordCircus a numeric vector

Paddington a numeric vector

ParkRoyal a numeric vector

ParsonsGreen a numeric vector

Perivale a numeric vector

PiccadillyCircus a numeric vector

Pimlico a numeric vector

Pinner a numeric vector

Plaistow a numeric vector

PrestonRoad a numeric vector

PutneyBridge a numeric vector

QueensPark a numeric vector

Queensbury a numeric vector

Queensway a numeric vector

RavenscourtPark a numeric vector

RaynersLane a numeric vector

Redbridge a numeric vector

RegentsPark a numeric vector

Richmond a numeric vector

Rickmansworth a numeric vector

RodingValley a numeric vector

RoyalOak a numeric vector

Ruislip a numeric vector

RuislipGardens a numeric vector

RuislipManor a numeric vector

RussellSquare a numeric vector

SevenSisters a numeric vector

ShepherdsBushCen a numeric vector

ShepherdsBushHC a numeric vector

SloaneSquare a numeric vector

Snaresbrook a numeric vector

SouthEaling a numeric vector

SouthHarrow a numeric vector

SouthKensington a numeric vector

SouthKenton a numeric vector

SouthRuislip a numeric vector

SouthWimbledon a numeric vector

SouthWoodford a numeric vector

Southfields a numeric vector

Southgate a numeric vector

Southwark a numeric vector

StJamessPark a numeric vector

StJohnsWood a numeric vector

StPauls a numeric vector

StamfordBrook a numeric vector

Stanmore a numeric vector

StepneyGreen a numeric vector

Stockwell a numeric vector

StonebridgePark a numeric vector

Stratford a numeric vector

SudburyHill a numeric vector

SudburyTown a numeric vector

SwissCottage a numeric vector

Temple a numeric vector

TheydonBois a numeric vector

TootingBec a numeric vector

TootingBroadway a numeric vector

TottenhamCourtRoad a numeric vector

TottenhamHale a numeric vector

TotteridgeAndWhetstone a numeric vector

`TowerHill` a numeric vector
`TufnellPark` a numeric vector
`TurnhamGreen` a numeric vector
`TurnpikeLane` a numeric vector
`Upminster` a numeric vector
`UpminsterBridge` a numeric vector
`Upney` a numeric vector
`UptonPark` a numeric vector
`Uxbridge` a numeric vector
`Vauxhall` a numeric vector
`Victoria` a numeric vector
`WalthamstowCentral` a numeric vector
`Wanstead` a numeric vector
`WarrenStreet` a numeric vector
`WarwickAvenue` a numeric vector
`Waterloo` a numeric vector
`Watford` a numeric vector
`WembleyCentral` a numeric vector
`WembleyPark` a numeric vector
`WestActon` a numeric vector
`WestBrompton` a numeric vector
`WestFinchley` a numeric vector
`WestHam` a numeric vector
`WestHampstead` a numeric vector
`WestHarrow` a numeric vector
`WestKensington` a numeric vector
`WestRuislip` a numeric vector
`WestbournePark` a numeric vector
`Westminster` a numeric vector
`WhiteCity` a numeric vector
`Whitechapel` a numeric vector
`WillesdenGreen` a numeric vector
`WillesdenJunction` a numeric vector
`Wimbledon` a numeric vector
`WimbledonPark` a numeric vector
`WoodGreen` a numeric vector
`WoodLane` a numeric vector
`Woodford` a numeric vector
`WoodsidePark` a numeric vector
`Total` a numeric vector; the total over all stations.
`mins24` a numeric vector; minutes on a 24 hour clock, e.g., 0 is midnight, 120 is 2am.

Details

Each cell contains the number of passengers entering a station during a quarter-hour period of time on a typical day during November 2010. The column names of the data frame are the station names and the most of the rows are the start time of each quarter-hour time block given in 24 hour time, e.g., 2215 means 10:15pm to 10:29pm. The last four rows are `"Total"`, `"A.M.Peak"`, `"Interpeak"`, `"P.M.Peak"`.

The data is adjusted to remove the effect of abnormal circumstances that many affect passeger numbers such as industrial action.

Source

The data came from the UK Government Transport for London website https://www.tfl.gov.uk. Downloaded in December 2013 and formatted into R by J. T. Gray (and slightly edited by T. W. Yee).

Examples

```
## Not run:
data(tube10)
waterloo <- tube10[1:(4*24), "Waterloo"]  # Omit the totals and the peaks
barplot(unlist(waterloo))
barplot(log10(1 + unlist(waterloo)), col = "lightblue",
        ylab = "log10(1+.)", las = 1)

## End(Not run)
```

---

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, https://www.healthvermont.gov/stats/systems 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. Ideally these should be converted to a factor.

`beatboat` a numeric vector; beat or boat number, a value from the set 1,2,...,19. Ideally these should be converted to a factor. For a river though, they are contiguous whereas on a lake it is less so.

`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. *Journal of Quantitative Analysis in Sports*. **10**, 397–409.

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. *Journal of Quantitative Analysis in Sports*. **10**, 397–409.

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) & proposed (blue&green) WFFC point system")
lines(fishlength, wffc.P1star(fishlength), col = "red", lwd = 2)
lines(fishlength, wffc.P3star(fishlength), col = "darkgreen",
      lwd = 2, lty = "dashed")
lines(fishlength, wffc.P4star(fishlength), 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 64 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?

smokepassive a numeric vector, on average, how many hours each week (at work and at home) would you spend near someone who is smoking? (put "0" if none)

smokeever a numeric vector, have you ever smoked tailor-made or roll-you-own cigarettes once a week or more? A 1 means yes and 0 means no.

smokenow a numeric vector, do you smoke tailor-made or roll-you-own cigarettes now? A 1 means yes and 0 means no.

smokeagequit a factor, if no to smokenow, how old were you when you stopped smoking? Using as.numeric(as.character(smokeagequit)) will work for those values which are not as.character(smokeagequit) == "-".

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

smoketailormade a numeric vector, how many tailor-made cigarettes do you smoke each day?

smokeweekpack a numeric vector, how many packets of *roll-your-own* tobacco do you use each week? (put "0" if none)

smokepacketsize a numeric vector, what size packets of *roll-your-own* tobacco do you usually buy? ("0" means don't smoke *roll-your-owns*, else 25g or 30g or 35g or 50g)

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.

eggs a numeric vector, how many eggs do you eat a week (raw, boiled, scrambled, poached, or in quiche)?

chocbiscuits a numeric vector, how many chocolate biscuits do you usually eat in a week?

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! In particular, the smoking variables contain a small percentage of conflicting values, and when `NA`s are taken into account then there would be several different ways the data might be cleaned. If `smokeever == 0` then strictly speaking, only `smokepassive` is the other variable—the other smoking variables should either be `NA` or `0`.

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)
```

yip88 *Zero-Inflated Poisson Distribution (Yip (1988) algorithm)*

Description

Fits a zero-inflated Poisson distribution based on Yip (1988).

Usage

```
yip88(link = "loglink", n.arg = NULL, imethod = 1)
```

Arguments

link	Link function for the usual λ parameter. See Links for more choices.
n.arg	The total number of observations in the data set. Needed when the response variable has all the zeros deleted from it, so that the number of zeros can be determined.
imethod	Details at CommonVGAMffArguments.

Details

The method implemented here, Yip (1988), maximizes a *conditional* likelihood. Consequently, the methodology used here deletes the zeros from the data set, and is thus related to the positive Poisson distribution (where $P(Y = 0) = 0$).

The probability function of Y is 0 with probability ϕ, and Poisson(λ) with probability $1 - \phi$. Thus

$$P(Y = 0) = \phi + (1 - \phi)P(W = 0)$$

where W is Poisson(λ). The mean, $(1 - \phi)\lambda$, can be obtained by the extractor function fitted applied to the object.

This family function treats ϕ as a scalar. If you want to model both ϕ and λ as a function of covariates, try zipoisson.

Value

An object of class "vglmff" (see vglmff-class). The object is used by modelling functions such as vglm, rrvglm and vgam.

Warning

Under- or over-flow may occur if the data is ill-conditioned. Yip (1988) only considered ϕ being a scalar and not modelled as a function of covariates. To get around this limitation, try zipoisson.

Inference obtained from summary.vglm and summary.vgam may or may not be correct. In particular, the p-values, standard errors and degrees of freedom may need adjustment. Use simulation on artificial data to check that these are reasonable.

Note

The data may be inputted in two ways. The first is when the response is a vector of positive values, with the argument `n` in `yip88` specifying the total number of observations. The second is simply include all the data in the response. In this case, the zeros are trimmed off during the computation, and the `x` and `y` slots of the object, if assigned, will reflect this.

The estimate of ϕ is placed in the `misc` slot as `@misc$pstr0`. However, this estimate is computed only for intercept models, i.e., the formula is of the form `y ~ 1`.

Author(s)

Thomas W. Yee

References

Yip, P. (1988). Inference about the mean of a Poisson distribution in the presence of a nuisance parameter. *The Australian Journal of Statistics*, **30**, 299–306.

Angers, J-F. and Biswas, A. (2003). A Bayesian analysis of zero-inflated generalized Poisson model. *Computational Statistics & Data Analysis*, **42**, 37–46.

See Also

`zipoisson`, `Zipois`, `zapoisson`, `pospoisson`, `poissonff`, `dzipois`.

Examples

```
phi <- 0.35; lambda <- 2  # Generate some artificial data
y <- rzipois(n <- 1000, lambda, phi)
table(y)

# Two equivalent ways of fitting the same model
fit1 <- vglm(y ~ 1, yip88(n = length(y)), subset = y > 0)
fit2 <- vglm(y ~ 1, yip88, trace = TRUE, crit = "coef")
(true.mean <- (1-phi) * lambda)
mean(y)
head(fitted(fit1))
fit1@misc$pstr0  # The estimate of phi

# Compare the ZIP with the positive Poisson distribution
pp <- vglm(y ~ 1, pospoisson, subset = y > 0, crit = "c")
coef(pp)
Coef(pp)
coef(fit1) - coef(pp)            # Same
head(fitted(fit1) - fitted(pp))  # Different

# Another example (Angers and Biswas, 2003) ----------------------
abdata <- data.frame(y = 0:7, w = c(182, 41, 12, 2, 2, 0, 0, 1))
abdata <- subset(abdata, w > 0)

yy <- with(abdata, rep(y, w))
fit3 <- vglm(yy ~ 1, yip88(n = length(yy)), subset = yy > 0)
```

```
fit3@misc$pstr0  # phi estimate (they get 0.5154 with SE 0.0707)
coef(fit3, matrix = TRUE)
Coef(fit3)  # Estimate of lambda (they get 0.6997 with SE 0.1520)
head(fitted(fit3))
mean(yy)  # Compare this with fitted(fit3)
```

Index