cluster  Identify clusters

Description
This is a special function used in the drm or depratio syntax. It identifies correlated groups of observations, and is used on the right hand side of a formula.

Usage
cluster(x)

Arguments
x A character, factor, or numeric variable.

Details
The function’s only action is semantic, to mark a variable as the cluster indicator.

Value
x

See Also
drm, depratio, Time

depratio(formula, data, subset, ord = 2, boot.ci = FALSE, n.boot = NULL, 
       ci.width=0.95)

Description
Calculates the observed values of the adjacent dependence ratios from the data.

Usage
depratio(formula, data, subset, ord = 2, boot.ci = FALSE, n.boot = NULL, 
          ci.width=0.95)
Arguments

formula the syntax is of form $y \sim \text{cluster}(id) + \text{Time}(time)$, where id denotes the cluster indicator, and Time denotes the order along which the adjacent dependence ratios will be calculated.

data optional data frame containing the variables in the formula

subset an optional vector specifying a subset of observations from the data

ord order of the dependence ratios to be calculated. The default is 2

boot.ci logical argument specifying whether bootstrap confidence intervals will be calculated for the empirical dependence ratio estimates

n.boot number of bootstrap replicates

CI.width width of the confidence interval. Default is 0.95

Value

An object of class depratio. Generic functions print and plot are also available.

An object of class depratio is a list containing at least the following two components:

tau matrix of the observed dependence ratios

freq matrix of the frequencies of events for the numerator of the observed dependence ratios

See Also

drm, cluster, Time

Examples

```r
## calculate and plot the observed 2nd order dependence ratios
## for the marijuana data:
data(marijuana)
dr.male <- depratio(y~cluster(id)+Time(age), data=marijuana,
                   subset=sex=="male")
dr.male
plot(dr.male)

## confirm that the 1st order Markov assumption is adequate
## for the madras data:
data(madras)
dr2 <- depratio(symptom~cluster(id)+Time(month), data=madras)
dr3 <- depratio(symptom~cluster(id)+Time(month), ord=3, data=madras)
dr <- rbind(dr2$tau[-length(dr2$tau)]*dr2$tau[-1], dr3$tau)
matplot(1:ncol(dr), t(dr))
```
**Description**

drm fits a combined regression and association model for longitudinal or otherwise clustered categorical responses using dependence ratio as a measure of the association.

**Usage**

drm(formula, family=binomial, data=sys.parent(), weights, offset, subset=NULL, na.action, start=NULL, link="cum", dep="I", Ncond=TRUE, Lclass=2, dropout=FALSE, drop.x=NULL, save.profiles=TRUE, pmatrix=NULL, print.level=2, iterlim=200, ...)

**Arguments**

- **formula**: a formula expression as for other regression models. In addition the cluster term has to be specified in the expression by cluster() and if using temporal association structure the temporal term has to be specified by Time(). See examples below and the documentation of lm and formula for further details.
- **family**: a description of the link function to be used in the model for a binary response. Default is logit link. See family for details. For an ordinal response, link is defined for the cumulative probabilities when link-argument is set to "cum". See link below.
- **data**: an optional data frame containing the variables in the model.
- **weights**: an optional vector of weights to be used in the fitting process. Only equal weights within cluster are allowed.
- **offset**: this can be used to specify an a priori known component to be included in the linear predictor during fitting.
- **subset**: an optional vector specifying a subset of observations to be used in the fitting process.
- **na.action**: a function which indicates what should happen when the data contain NAs. The default is na.include after which the analysis assumes missing data mechanism at random (MAR) if dropout=FALSE, and not at random if dropout = TRUE. See dropout below.
- **start**: an optional vector of starting values for the parameters. By default, the starting values are estimated from glm-procedure assuming independence.
- **link**: this can be used to specify alternative link functions for nominal and ordinal responses. By default "cum", after which the link is specified through family = binomial(link=?) for the cumulative probabilities. Alternative links include adjacent category logit "acl" and baseline category logit "bcl" (baseline category being the last category). For "bcl", the regression parameters are estimated for each logit level. For a binary response, this argument is ignored.
dep
dep defines the association structure. The default is independence "I". Other singular options are for the exchangeable association: Necessary factor "N", Latent categorical factor "L", Latent Beta-distributed propensity "B" (binary response), Latent Dirichlet-distributed propensities "D" (multicategorical response), and for the temporal association: first order Markov "M", and second order Markov "M2" (binary response). By default, Markov structure for the adjacent 2-way dependence ratios is assumed to be stationary. Superpositions of these structures can be imposed, such as "NL", "NB","ND", "NM", "LM", "NLM","NM2". See [3-7] for further details. Parameter restrictions, covariates and functional forms for the association parameters can also be specified. In that case the dep-argument must be a list. See examples below. For the interpretation of the association parameters, see the documentation of the support function getass.drm.

Ncond
logical argument defining whether the regression model is marginal or conditional when the association is "N". The default is true, i.e. the regression estimation is conditional on \( N = 1 \). If covariates are used for the "N"-association, it is advisable to set Ncond=FALSE, since otherwise the interpretation of the regression parameters is less clear.

Lclass
Number of latent classes in the population when the association is "L". Default is 2. Available only for binary response. Note that in the current implementation, the conditional probabilities are not calculated for Lclass>2. For checking the validity of the model, the user needs to check whether the estimated conditional probabilities fall within 0 and 1. See example in getass.drm for parameter interpretation and how to calculate the conditional probabilities.

dropout
logical argument. For monotone missing patterns in longitudinal studies, this argument allows to impose a selection model (see [8] for details) on top of regression and association model to investigate the sensitivity of the results due to missingness. The model formula notation is: \( \text{logit}(hz(\text{dropNcurII})) = (\text{InterceptId}+\text{responseNcur}+\text{responseNprev}) \), where response.Ncur denotes the effect of current, possibly missing response value and response.Nprev denotes the effect of previous response value. MCAR, MAR and MNAR-models can be specified by imposing restrictions on selection model parameters in dep-argument as for the association parameters. See dep above and examples below. If the response is a factor, the effect of the factor levels are estimated contrasting to the lowest level.

drop.x
an optional covariate vector for the selection model. The covariate’s previous value (notation: covariate.Nprev) is used in the selection model.

save.profiles
logical argument defining whether the fitted values of all possible profiles is saved. If FALSE, only the indicator vector (-1 for a negative, 1 for a positive profile) over all units will be saved. If the cluster size is large, using save.profiles=TRUE may result in a very large object.

pmatrix
a character object specifying the name of the matrix for all possible profiles, created using profiles.drm. If the cluster size is large, this speeds up the estimation in case several models are fitted. See examples below.

print.level
level of printing during numerical optimisation. The default is 2. See nlm for further details.

iterlim
maximum iteration limit for the numerical maximisation. See nlm for further details.
... other arguments passed to \texttt{nlm}, e.g. controlling the convergence. See \texttt{nlm} for further details.

Details

\texttt{drm} gives maximum likelihood estimates for the combined regression and association model by decomposing a joint probability of responses in a cluster to univariate marginal or cumulative probabilities and dependence ratios of all orders. See [1] and [5] for further details. The dimensionality of the association part is reduced by imposing a model for the association structure with \texttt{dep}-argument. See \texttt{getass.drm} and [3-7] for details. Furthermore, a selection model can be added on top of regression and association model. See examples below and [5] and [8] for details.

Value

\texttt{drm} returns an object of class \texttt{drm}. The function \texttt{summary} (i.e., \texttt{summary.drm}) can be used to obtain or print a summary of the results. The generic accessor function \texttt{coefficients} can be used to extract coefficients.

An object of class \texttt{drm} is a list containing at least the following components:

- \texttt{coefficients}: a named vector of regression, and possibly association and selection model coefficients.
- \texttt{cov.scaled}: a variance-covariance matrix of the parameter estimates.
- \texttt{fitted.marginals}: the fitted values for the univariate means, obtained by transforming the linear predictors by the inverse of the link function.
- \texttt{fitted.conditionals}: in case of \texttt{"L"}-structure, the fitted values for the conditional univariate means, otherwise NULL. Not yet implemented for \texttt{lclass} > 2; see also \texttt{getass.drm}.
- \texttt{fitted.profiles}: the fitted response profile probabilities within each cluster, calculated by using the maximum likelihood estimates from the model. See also \texttt{save.profiles} above. Note that within each cluster, the order of the responses is by Time for Markov structures, and for exchangeable structures with missing values, by response value, with missing values (NA) last.
- \texttt{deviance}: minus twice the maximised log-likelihood.
- \texttt{aic}: An Information Criterion: minus twice the maximised log-likelihood plus twice the number of coefficients. Not available if the likelihood is weighted with the dropout probabilities.
- \texttt{niter}: the number of iterations that \texttt{nlm} used.
- \texttt{code}: convergence code from \texttt{nlm}. See \texttt{nlm} for details.
- \texttt{call}: the matched call.
- \texttt{terms}: the ‘terms’ object used.

WARNING

The maximum likelihood estimates may sometimes lead to negative fitted probabilities. In this case, both generic print-methods warn about this. In this case, the model is considered to be wrongly specified and model specification should be changed.
Author(s)

Jukka Jokinen, <jukka.jokinen@helsinki.fi>

References


See Also

`getass.drm, nlm, cluster, Time profiles.drm, depratio`

Examples

```
### Examples for binary responses

#### Wheeze among Steubenville (see [3]):
#### Latent Beta-distributed propensity

data(wheeze)
fit1 <- drm(wheeze~I(age>9)+smoking+cluster(id),data=wheeze,dep="B", print=0)

#### Obesity among Muscatine children (see [2]):
#### Analysis for completers: M2 for girls

data(obese)
fit2 <- drm(obese~age+cluster(id)+Time(year), subset=sex=="female",
            dep="M2",data=obese)

#### Not run:
#### Muscatine children continued (see [3]):
#### LM for boys and girls separately

data(obese)
fit3 <- drm(obese~age+cluster(id)+Time(age), subset=sex=="male",
            dep="LM",data=obese)
```
fit4 <- drm(obese~age+cluster(id)+Time(age), subset=sex=="female",
            dep="LM", data=obese)

## End(Not run)

## Examples for ordinal responses

## Movie critic example (see [6]):
## Latent Dirichlet propensities with baseline category link.
data(movie)

options(contrasts=c("contr.treatment","contr.treatment"))
fit5 <- drm(y~critic+cluster(movie), data=movie, dep="D", link="bcl")

## Longitudinal dataset on teenage marijuana use (see [6]):
## Superposition of structures N, L and M for the girls.
data(marijuana)

fit6 <- drm(y~age+cluster(id)+Time(age), data=marijuana,
            subset=sex=="female", dep=list("NLM", ~kappa1==1,
            ~kappa2==0, ~tau12==1, ~tau21==1, ~tau11==tau22))

## Parameter restrictions with functions using M-structure for the boys.
## Plot the second order dependence ratios:
plot(depratio(y~cluster(id)+Time(age), data=marijuana,
             subset=sex=="male"))

## fit the model in [6]:
fit7 <- drm(y~age+cluster(id)+Time(age), data=marijuana,
            subset=sex=="male", dep=list("M",
            tau12=function(a=1,b=0) a+b*c(0:3),
            tau21=function(a=1,b=0) a+b*c(0:3))

## Not run:

## Covariates for the association (see [7]):
data(madras)

## plot empirical 2nd order dependence ratios with bootstrap CI's
tau.madras <- depratio(symptom~cluster(id)+Time(month), data=madras,
                      boot.ci = TRUE, n.boot = 1000)
plot(tau.madras, log="y", ylim=c(1,40), plot.ci=TRUE)

## create matrix for profiles:
W.madras <- profiles.drm(n.categories=2, n.repetitions=12, "M")

## create four-level covariate, combining age and sex:
madras$age.sex <- factor(paste(madras$age, madras$sex, sep="."))

## fit the model in [7], Section 4:
fit8 <- drm(symptom~age+sex+month+month:age+month:sex+cluster(id)+Time(month),
            data=madras, Ncond=FALSE, save.profiles=FALSE, pmatrix="W.madras")
### Internal drm functions

#### Description

Internal drm functions
Details

These are internal drm functions and are not to be called by the user.

---

**getass.drm**

Support function for drm: specification of the association model

---

Description

A support function called by drm that parses from the argument dep the covariates and functional forms for the association model. This function should not be used directly by the user.

Details

In the argument dep, the user needs to specify the covariates and functional forms on the association parameters. The following list describes the notation and concise interpretation of the association parameters. For more details, see the reference below.

- nu
  For "N"-structure: the proportion of susceptibles in the population, i.e. proportion of units that can have a value greater than the smallest category. Covariates can be specified. See examples below.

- nu1, nu2, ..., nuk
  For "L"-structure: the proportion in the population in the latent class 1, 2, ..., k. The proportion in the latent class 0 can be obtained by calculating \(1 - nu1 - nu2 - \ldots\). Note that for binary responses, the number of latent classes can be specified with an argument lclass (Default 2). Covariates can be specified. See examples below.

- kappa
  (binary response) For "L"-structure with lclass=2 (default), the success probability in the latent class 0 divided by the success probability in the latent class 1, i.e. \(kappa = \Pr(Y = 1|L = 0)/\Pr(Y = 1|L = 1)\). Covariates can be specified; examples below.

- kappa0, kappa1, ..., kappak-1
  (binary response) For "L"-structure with lclass=k+1: the success probabilities in the latent class 0, 1, ..., k-1 divided by the success probability in the latent class k. For example, \(kappa0 = \Pr(Y = 1|L = 0)/\Pr(Y = 1|L = k)\). Covariates can be specified. See examples below.

- kappa1, kappa2, ..., kappak-1
  (multicategorical response with k levels) For "L"-structure with two latent classes (i.e. lclass=2): the category probabilities in categories 1, 2, ..., k-1 in the latent class 0 divided by the corresponding probabilities in the latent class 1. For example, \(kappa2 = \Pr(Y = 2|L = 0)/\Pr(Y = 2|L = 1)\). The smallest response value is regarded as the baseline, denoted by 0

- xi1, xi0
  (binary response) For "B"-structure: the shape parameters of the Beta-distribution (sometimes also noted as xi1=p and xi0=q). Covariates can be specified. See examples below.

- xi0, xi1, ..., xik
  (multicategorical response) For "D"-structure: the shape parameters of the Dirichlet distribution

- tau
  (binary responses) For "M"-structure: adjacent second order dependence ratio. If the number of repeated measurements is greater than two, the adjacent tau's are assumed to be equal (i.e. stationarity of the dependence ratios). In order to specify equalities or functional forms (i.e. non-stationary overlapping dependence ratios), see examples below.
• tau12, tau13, tau123 (binary responses) For "M2"-structure: adjacent second order dependence ratio (tau12), adjacent third order dependence ratio (tau123) and the second order dependence ratio between first and third response (tau13). If the number of repeated measurements is greater than three, the tau's are assumed to be equal (i.e. stationarity of the dependence ratios). Equalities and functional forms can be specified. See examples below.

• tau11, tau12, ..., tau21, tau22, ..., taukk (multicategorical responses) For "M"-structure: adjacent second order dependence ratios for categories 1, 2, ..., k, where the smallest response value is regarded as the baseline, denoted by 0. Equalities and functional forms can be specified as for the binary "M"-structures above. See examples below.

References

See Also
drm

Examples

```r
### Example of functional forms:
# non-stationary second order Markov structure
# initial values of the dependence ratios are set to 1.
# Not run:
data(wheeze)
assoc <- list("M2",
              tau12 ~ function(a78=1, a89=1, a910=1)c(a78, a89, a910),
              tau123 ~ function(a78=1, a8910=1)c(a78, a8910),
              tau13 ~ function(a79=1, a810=1)c(a79, a810))
fit1 <- drm(wheeze~i(age>YI+smoking+cluster(idI+time(ageI, data=wheeze, dep=assoc, print=0)

### Example of other parameter restrictions:
# fixing parameters to a known value: -tau12==1, -tau21==1
# setting parameters to equal: -tau11==tau22
data(marijuana)
assoc <- list("M", -tau12==1, -tau21==1, -tau11==tau22)
fit2 <- drm(y~age+cluster(idI)+Time(age), data=marijuana,
           subset=sex=="female", dep=assoc, print=0)

# setting all parameters to equal:
assoc <- list("M", -tau11==tau12, -tau11==tau21, -tau11==tau22)
fit3 <- drm(y~age+cluster(idI)+Time(age), data=marijuana,
           subset=sex=="female", dep=assoc, print=0)

### End(Not run)
### Example of covariates for the association parameters:
```
allow the probabilities within the latent class vary by sex. Note: covariate needs to be a factor.
data(obese)
assoc <- list("L", kappa ~ kappa:factor(sex))

fit4 <- drm(obese~age+cluster(id)+Time(age), data=obese,
          dep=assoc, print=0)

### Example how to derive conditional probabilities from marginals
### Fit a model with three latent classes:
data(wheeze)
latent3 <- drm(wheeze~I(age>9)+smoking+cluster(id), data=wheeze,
              dep="L", lclass=3, print=0)

### calculate conditional probabilities:
### pr(Y|L=2) = pr(Y=1)/(nu2+kappa1*nu1+kappa0*(1-nu1-nu2))
### pr(Y|L=1) = kappa1*pr(Y|L=2)
### pr(Y|L=0) = kappa0*pr(Y|L=2)
est <- coef(latent3)
psi2 <- latent3$fitted.marginals/
   (est["nu2"]+est["kappa1"]*est["nu1"]+
est["kappa0"]*(1-est["nu1"]-est["nu2"]))
psi1 <- psi2*est["kappa1"]
psi0 <- psi2*est["kappa0"]

### check the model validity, i.e. require that 0 < psi_i < 1:
range(cbind(psi0, psi1, psi2))

---

### madras

**Madras Longitudinal Schizophrenia Study**

**Description**

Monthly records of presence/abscence of psychiatric symptom ‘thought disorder’ of 86 patients over the first year after initial hospitalisation for disease.

**Usage**

data(madras)

**Format**

A data frame with 1032 observations on the following 5 variables.

- **id**: identification number of a patient
- **symptom**: response ‘thought disorder’: 0 = absent, 1 = present
- **month**: month since hospitalisation
- **age**: age indicator: 0 = less than 20 years, 1 = 20 or over
- **sex**: sex indicator: 0 = male, 1 = female
Source

http://faculty.washington.edu/heagerty/Books/AnalysisLongitudinal/datasets.html

References


marijuana Longitudinal data set on teenage marijuana use

Description

This data set is from US National Youth Survey reporting a repeated questionnaire on marijuana use of 237 teenagers yearly from ages 13 to 17 on an ordinal scale.

Usage

data(marijuana)

Format

This data frame contains the following columns:

- **id** identification number of a teenager.
- **y** marijuana response, 1: never; 2: less than once a month; 3: more than once a month.
- **sex** a factor with levels female male
- **age** age as a factor with levels 13 14 15 16 17

Source


References

### Movie reviews from a Variety magazine

**Description**

This dataset contains reviews of 93 movies by four critics in the Variety magazine from April 1995 through March 1997.

**Usage**

```r
data(movie)
```

**Format**

This data frame contains the following columns:

- `movie` Identification number of a movie.
- `y` review response, 1: Positive; 2: Mixture of positive and negative; 3: Negative.
- `critic` an ordered factor: critics with levels medved < siskel < ebert < lyons

**Source**


**References**


### Obesity among Muscatine children

**Description**

Data of the completers concerning the presence or absence of obesity in five age cohorts in Muscatine, Ohio, measured repeatedly in -77, -79 and -81.

**Usage**

```r
data(obese)
```
profiles.drm

Format

A data frame with 5310 observations on the following 5 variables.

- **id**: identification number of a child
- **obese**: a binary indicator of obesity
- **sex**: a factor with levels male female
- **year**: year of measurement
- **age**: a factor indicating age cohorts from (5, 7] to (17, 19]

Source


http://www.biostat.harvard.edu/~fitzmaur/ala/obesity.txt

References


---

**profiles.drm**

*Creates a support matrix for ‘drm’ for profile probability calculation*

Description

Creates a support matrix for the calculation of all possible profile probabilities using the ML estimates obtained from **drm**.

Usage

`profiles.drm(n.categories, n.repetitions, structure = "exchangeable")`

Arguments

- **n.categories**: Number of response variable categories
- **n.repetitions**: Number of response repetitions within a cluster
- **structure**: Association structure indicator of the model to be fitted: "exchangeable" if the structure is exchangeable, otherwise either "M" or "M2"

Details

Creation of this support object is useful if the cluster size is large and if several models are fitted to the same dataset. Avoids creating the same matrix over and over again for each model fit. See examples for **drm**.
Value

If structure="exchangeable", returns a matrix with row and column size equal to n.categories to the power of n.repetitions. For temporal structures "M" and "M2", the result is an array.

See Also

drm

Examples

## use marijuana-data
data(marijuana)

## create matrix for profiles:
W.mari <- profiles.drm(3, 5, "M")

## fit the model using W.mari:
fit <- drm(y~age+cluster(id)+Time(age), data=marijuana, 
  subset=sex=="male", pmatrix="W.mari", save.profiles=FALSE, 
  dep=list("M", tau12=function(a=1, b=0) a+b*c(0:3), 
  tau21=function(a=1, b=0) a+b*c(0:3)), print=0)

## a speedy way of saving the profile probabilities to the object:
fit <- drm(y~age+cluster(id)+Time(age), data=marijuana, 
  subset=sex=="male", pmatrix="W.mari", save.profiles=TRUE, 
  dep=list("M", tau12=function(a=1, b=0) a+b*c(0:3), 
  tau21=function(a=1, b=0) a+b*c(0:3)), print=0, 
  start=coef(fit))

---

**Time**

<table>
<thead>
<tr>
<th>Time</th>
<th>Identify ordering of the responses</th>
</tr>
</thead>
</table>

Description

This is a special function used in the drm or depratio syntax. It identifies the ordering in correlated groups of observations, and is used on the right hand side of a formula.

Usage

Time(x)

Arguments

x A factor, or numeric variable.

Details

The function’s only action is semantic, to mark a variable as the Time indicator.
wheeze

See Also
drm, depratio, cluster

Examples

```
formula(wheeze~cluster(id)+Time(age))
```

---

| wheeze | Wheeze among Steubenville children |

Description

wheeze is a longitudinal dataset reporting the presence and absence of wheeze at ages 7, 8, 9 and 10 among 537 Steubenville children.

Usage

data(wheeze)

Format

This data frame contains the following columns:

- `id` identification number of a child.
- `wheeze` a binary indicator of the presence of wheeze.
- `age` a numeric vector of age from 7 to 10.
- `smoking` a binary indicator of mother's smoking status at baseline.

Source


References

Index

*Topic **arith**
depratio, 2

*Topic **array**
profiles.drm, 15

*Topic **datasets**
madras, 12
marijuana, 13
movie, 14
obese, 14
wheeze, 17

*Topic **misc**
cluster, 2
   drm-internal, 9
   Time, 16

*Topic **models**
drm, 4

*Topic **utilities**
getass.drm, 10

cluster, 2, 3, 7, 17

depratio, 2, 2, 7, 17
drm, 2, 3, 4, 10, 11, 15–17
drm-internal, 9

getass.drm, 5–7, 10
getpath.drm(drm-internal), 9

kronecker.drm(drm-internal), 9
kroneckerd.drm(drm-internal), 9

loglikd.drm(drm-internal), 9
logliks.drm(drm-internal), 9

madras, 12
marijuana, 13
movie, 14

na.include(drm-internal), 9
nlm, 5–7

obese, 14
plot.depratio(depratio), 2
print.depratio(depratio), 2
print.drm(drm), 4
print.summary.drm(drm), 4
profiles.drm, 5, 7, 15

strata(drm-internal), 9
summary.drm(drm), 4

Time, 2, 3, 7, 16
untangle.specials(drm-internal), 9
wheeze, 17