Package 'riskclustr'

May 21, 2018

Type Package

Title Functions to study etiologic heterogeneity

Version 0.5.1

Description Functions related to the study of etiologic heterogeneity both across disease subtypes and across individual tumor markers.

Depends R (>= 3.1.0)

License GPL-2

Encoding UTF-8

LazyData true

Imports aod, mlogit, gtools

RoxygenNote 6.0.1

Author Emily Zabor [aut, cre]

Maintainer Emily Zabor <zabore@mskcc.org>

R topics documented:

`dest` *Estimate the incremental explained risk variation*

Description

`dest` estimates the incremental explained risk variation across a set of pre-specified disease subtypes in a case-control study

Usage

```
dest(formula, cls, M, data)
```

Arguments

formula	an mFormula() model formula for a polytmous logistic regression model to be fit with mlogit() using the appropriate variable names from the data of interest, see Examples
cls	the name of the subtype variable in the data, where 0 indicates control subjects, should be supplied in quotes, e.g. cls = "class"
M	the number of subtypes. This should not include controls, but only the number of subtypes among case subjects. For M>=2.
data	the name of the dataframe that contains the relevant variables

References

Begg, C. B., Zabor, E. C., Bernstein, J. L., Bernstein, L., Press, M. F., & Seshan, V. E. (2013). A conceptual and methodological framework for investigating etiologic heterogeneity. Stat Med, 32(29), 5039-5052. doi: 10.1002/sim.5902

Examples

```
# generate data - 4 classes and 2 risk factors
class <- rep(c(0, 1, 2, 3, 4), times = c(1000, 250, 250, 250, 250))
x <- matrix(rnorm(2000 * 2), 2000, 2) +
model.matrix(~factor(class))[, -1] %*%
t(matrix(c(1.5, 0, 0.75, 0.25, 0.25, 0.75, 0, 1.5), ncol = 4))
df <- data.frame(class = class, x1 = x[, 1], x2 = x[, 2])

# specify the model formula
library(mlogit)
mform <- mFormula(class ~ 1 | x1 + x2)

dest(mform, "class", 4, df)
```

---

dstarest *Estimate the incremental explained risk variation in case-only study*

Description

dstarest estimates the incremental explained risk variation across a set of pre-specified disease subtypes in a case-only study

Usage

```
dstarest(formula, cls, M, data)
```

Arguments

formula	an mFormula() model formula for a polytmous logistic regression model to be fit with mlogit() using the appropriate variable names from the data of interest, see Examples
cls	the name of the subtype variable in the data with values 1 through M. Should be supplied in quotes, e.g. cls = "class"
M	the number of subtypes among case subjects. For M>=2.
data	the name of the dataframe that contains the relevant variables

References

Begg, C. B., Seshan, V. E., Zabor, E. C., Furberg, H., Arora, A., Shen, R., . . . Hsieh, J. J. (2014). Genomic investigation of etiologic heterogeneity: methodologic challenges. BMC Med Res Methodol, 14, 138. doi: 10.1186/1471-2288-14-138

Examples

```
# generate data - 4 classes and 2 risk factors
set.seed(20180521)
class <- rep(c(0, 1, 2, 3, 4), times = c(1000, 250, 250, 250, 250))
x <- matrix(rnorm(2000 * 2), 2000, 2) +
model.matrix(~factor(class))[, -1] %*%
t(matrix(c(1.5, 0, 0.75, 0.25, 0.25, 0.75, 0, 1.5), ncol = 4))
df <- data.frame(class = class, x1 = x[, 1], x2 = x[, 2])

# remove the controls - this is a case only analysis example
df <- df[df$class != 0, ]

library(mlogit)
mform <- mFormula(class ~ 1 | x1 + x2)

dstarest(mform, "class", 4, df)
```

---

fstat_bin *Function to calculate the test statistic and proportion of variation explained using MDMR*

Description

fstat Computes the test statistic and proportion of variation explained using multivariate distance matrix regression. For binary tumor marker data. Uses asymmetric binary distance metric.

Usage

```
fstat_bin(Y, X)
```

Arguments

Y	an NxK matrix of tumor marker data
X	an NxP matrix of risk factor data

Value

returns a list containing f, the test statistic, and R, the proportion of variation explained

Author(s)

Emily C Zabor <zabore@mskcc.org>

References

Zapala MA, Schork NJ. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. PNAS 2006; 103(51):19430-35.

`ksq` *Estimate the overall risk heterogeneity*

Description

`ksq` estimates the overall risk heterogeneity in a two-class system (cases versus controls, class 2 versus class 1)

Usage

```
ksq(formula, cls, data)
```

Arguments

`formula`	a `formula()` object for use in `glm()`
`cls`	the name of the subtype variable in the data, where 0 indicates control subjects, or the reference class level), should be supplied in quotes, e.g. cls = "class". Must be 0/1.
`data`	the name of the dataframe that contains the relevant variables

References

Begg, C. B., Zabor, E. C., Bernstein, J. L., Bernstein, L., Press, M. F., & Seshan, V. E. (2013). A conceptual and methodological framework for investigating etiologic heterogeneity. Stat Med, 32(29), 5039-5052. doi: 10.1002/sim.5902

Examples

```
class <- rep(c(0, 1), times = c(1000, 500))
x <- matrix(rnorm(1500 * 2), 1500, 2) +
model.matrix(~factor(class))[, -1] * 1.5
df <- data.frame(class = class, x1 = x[, 1], x2 = x[, 2])
ksq(formula = formula(class ~ x1 + x2), cls = "class", data = df)
```

minmc — *Calculate the minimum misclassification rate in a cross-tabulation*

Description

minmc provides a way of reconciling the arbitrary class labels that result from k-means clustering to obtain the minimum misclassification rate of a k-means clustering result to a true class solution, after aligning the class labels

Usage

```
minmc(tab)
```

Arguments

tab	a cross-tabulation table of a class solution from k-means clustering to the true class solution

Author(s)

Emily C Zabor <zabore@mskcc.org>

Examples

```
# Example where there is perfect alignment of class results, but the class
# labels do not correspond and need to be aligned
trueclass <- rep(c(1, 2, 3, 4), times = 20)
kclass <- rep(c(3, 2, 1, 4), times = 20)
table(kclass, trueclass)
minmc(table(kclass, trueclass))
```

plrsub — *Conduct an analysis of etiologic heterogeneity for pre-defined subtypes using polytomous logistic regression*

Description

plrsub takes a vector of class labels for pre-specified subtypes and a list of risk factors and returns results related to the question of whether each risk factor differs across levels of the disease subtypes

Input is a dataframe that contains the risk factors of interest and a variable containing numeric class labels that is 0 for control subjects. Risk factors can be either binary or continuous. For categorical risk factors, a reference level should be selected and then indicator variables for each remaining level of the risk factor should be created. Categorical risk factors entered as is will be treated as ordinal. Class labels for the cases can be specified as a vector. The multinomial logistic regression model is fit using the mlogit function from the mlogit package.

Usage

```
plrsub(cls, m, rf, data)
```

Arguments

cls	the name of the variable in the data that contains numeric class labels. This should be 0 for all controls. Argument must be supplied in quotes.
m	is the number of subtypes
rf	a list of the names of the binary or continuous risk factors. For binary risk factors the lowest level will be used as the reference level.
data	the name of the dataframe that contains the tumor markers and risk factors.

Value

Returns a list.

beta is a matrix containing the estimated beta parameters with a column for each risk factor and a row for each disease subtype.

beta_se contains the associated standard errors.

eh_pval is a vector of p-values for testing whether each risk factor differs across levels of the disease subtype.

or_ci_p is a dataframe with a odds ratio (95 factor/subtype combination, as well as a column of etiologic heterogeneity p-values.

beta_se_p is a dataframe with the estimated beta parameters (SE) for each risk factor/subtype combination, as well as a column of etiologic heterogeneity p-values.

Author(s)

Emily C Zabor <zabore@mskcc.org>

Examples

```
# generate the data - 4 classes and 2 risk factors
class <- rep(c(0, 1, 2, 3, 4), times = c(1000, 250, 250, 250, 250))
x <- matrix(rnorm(2000 * 2), 2000, 2) +
model.matrix(~factor(class))[, -1] %*%
t(matrix(c(1.5, 0, 0.75, 0.25, 0.25, 0.75, 0, 1.5), ncol = 4))
df <- data.frame(class = class, x1 = x[, 1], x2 = x[, 2])

plrsub("class", 4, c("x1", "x2"), df)
```

---

plrtm *Conduct an analysis of etiologic heterogeneity for individual tumor markers using polytomous logistic regression*

Description

plrtm takes a list of individual tumor markers and a list of risk factors and returns results related to the question of whether each risk factor differs across levels of the disease subtypes and the question of whether each risk factor differs across levels of each individual tumor marker of which the disease subtypes are comprised.

Input is a dataframe that contains the individual tumor markers, the risk factors of interest, and an indicator of case or control status. The tumor markers must be binary and must have levels 0 or 1 for cases. The tumor markers should be left missing for control subjects. For categorical tumor markers, a reference level should be selected and then indicator variables for each remaining level of the tumor marker should be created. For continuous tumor markers, categories should be formed and then indicator variables can be constructed as in the case of categorical tumor markers. Risk factors can be either binary or continuous. For categorical risk factors, a reference level should be selected and then indicator variables for each remaining level of the risk factor should be created. Categorical risk factors entered as is will be treated as ordinal.

Usage

```
plrtm(tm, rf, case, data)
```

Arguments

`tm`	a list of the names of the binary tumor markers. Each must have levels 0 or 1 for case subjects. This value will be missing for all control subjects.
`rf`	a list of the names of the binary or continuous risk factors. For binary risk factors the lowest level will be used as the reference level.
`case`	denotes the variable that contains each subject's status as a case or control. This value should be 1 for cases and 0 for controls. Argument must be supplied in quotes.
`data`	the name of the dataframe that contains the tumor markers and risk factors.

Value

Returns a list.

`beta` is a matrix containing the estimated beta parameters with a column for each risk factor and a row for each disease subtype.

`beta_se` contains the associated standard errors.

`eh_pval` is a vector of p-values for testing whether each risk factor differs across levels of the disease subtype.

`gamma` is a matrix containing the estimated gamma parameters, obtained as linear combinations of the beta parameters with a column for each risk factor and a row for each tumor marker.

`gamma_se` contains the associated standard errors.

`gamma_p` is a matrix of p-values for testing whether each risk factor differs across levels of each tumor marker, with a column for each risk factor and a row for each tumor marker.

`or_ci_p` is a dataframe with a odds ratio (95 factor/subtype combination, as well as a column of etiologic heterogeneity p-values.

`beta_se_p` is a dataframe with the estimated beta parameters (SE) for each risk factor/subtype combination, as well as a column of etiologic heterogeneity p-values.

`gamma_se_p` is a dataframe with estimates of the gamma tumor marker effects (SE) and their associated p-values.

Author(s)

Emily C Zabor <zabore@mskcc.org>

Examples

```
# generate the data - 2 tumor markers and 2 risk factors
class <- rep(c(0, 1, 2, 3, 4), times = c(1000, 250, 250, 250, 250))
tm1 <- c(rep(0, 1000), rep(c(0, 1, 0, 1), each = 250))
tm2 <- c(rep(0, 1000), rep(c(0, 1), each = 500))
x <- matrix(rnorm(2000 * 2), 2000, 2) +
model.matrix(~factor(class))[, -1] %*%
t(matrix(c(1.5, 0, 0.75, 0.25, 0.25, 0.75, 0, 1.5), ncol = 4))
df <- data.frame(class = class, x1 = x[, 1], x2 = x[, 2], tm1 = tm1, tm2 = tm2)
df$caseind <- ifelse(df$class == 0, 0, 1)

plrtm(c("tm1", "tm2"), c("x1", "x2"), "caseind", df)
```

`trueDsim` *Estimate the true population D*

Description

Adapted from code written by Venkat Seshan and made generalizable to any number of subtypes, any number of risk factors, and any risk factor means for the cases

Usage

```
trueDsim(N, M, pi, P, mu_m)
```

Arguments

`N`	population sample size
`M`	number of disease subtypes
`pi`	vector of control and subtype prevalences i.e. c(pi0, pi1, pi2, pi3) for controls and 3 subtypes
`P`	number of risk factors
`mu_m`	P x M matrix of risk factor means for the cases, by default all risk factors have mean 0 for control subjects

Author(s)

Emily C Zabor <zabore@mskcc.org>

Index