Package 'ethet'

October 24, 2017

Type Package

Title Functions to study etiologic heterogeneity

Version 0.1.1

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 nnet, aod, mlogit

RoxygenNote 6.0.1

Author Emily Zabor [aut, cre]

Maintainer Emily Zabor <zabore@mskcc.org>

R topics documented:

`calcmm` *Calculate the minimum misclassification rate in a cross-tabulation*

Description

This function provides a way of reconciling the arbitrary class labels that result from k-means clustering to obtain the misclassification rate of the k-means clustering result to the true class solution

Usage

```
calcmm(tab)
```

Arguments

tab	a cross-tabulation table of the simulation-based class solution that resulted in maximum D to the true class solution

Author(s)

Emily C Zabor <zabore@mskcc.org>

---

ehCalcD *Calculate the D metric to measure the extent of etiologic heterogeneity in a case-control study*

Description

Calculate the D metric to measure the extent of etiologic heterogeneity in a case-control study

Usage

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

Arguments

data	the name of the dataframe that contains the relevant variables
cls	the names of the subtype variable in the data, should be supplied in quotes, e.g. cls = "class"
M	the number of subtypes. This could should not include controls, but only the number of subtypes among case subjects.
formula	an mFormula() model formula for a polytmous logistic regression model to be fit with mlogit() using the appropriate variable names from the data as interest e.g. formula = mFormula(class ~ 1 | x1 + x2) for a model with subtype variable class and two individual-specific predictors x1 and x2

---

ehpoly *Conduct an analysis of etiologic heterogeneity using polytomous logistic regression*

Description

ehpoly 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

```
ehpoly(tm, rf, case, df)
```

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.
df	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>

---

ehpoly2 *Conduct an analysis of etiologic heterogeneity using polytomous logistic regression*

Description

ehpoly2 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.

Usage

```
ehpoly2(cls, m, rf, df)
```

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.
df	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>

---

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.

---

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