Package 'riskclustr'

July 23, 2025

Type Package

Title Functions to Study Etiologic Heterogeneity

Version 0.4.1

Description A collection of functions related to the study of etiologic heterogeneity both across disease subtypes and across individual disease markers. The included functions allow one to quantify the extent of etiologic heterogeneity in the context of a case-control study, and provide p-values to test for etiologic heterogeneity across individual risk factors. Begg CB, Zabor EC, Bernstein JL, Bernstein L, Press MF, Seshan VE (2013) <doi:10.1002/sim.5902>.

Depends R (>= 4.0)

License GPL-2

URL https://www.emilyzabor.com/riskclustr/, https://github.com/zabore/riskclustr

BugReports https://github.com/zabore/riskclustr/issues

Encoding UTF-8

Imports mlogit, stringr, Matrix

Language en-US

LazyData true

RoxygenNote 7.2.3

VignetteBuilder knitr

Suggests testthat, covr, rmarkdown, dplyr, knitr, usethis, spelling

NeedsCompilation no

Author Emily C. Zabor [aut, cre]

Maintainer Emily C. Zabor <zabore2@ccf.org>

Repository CRAN

Date/Publication 2024-01-18 20:10:02 UTC

Contents

`d` *Estimate the incremental explained risk variation in a case-control study*

Description

`d` estimates the incremental explained risk variation across a set of pre-specified disease subtypes in a case-control study. This function takes the name of the disease subtype variable, the number of disease subtypes, a list of risk factors, and a wide dataset, and does the needed transformation on the dataset to get the correct format. Then the polytomous logistic regression model is fit using `mlogit`, and D is calculated based on the resulting risk predictions.

Usage

```
d(label, M, factors, data)
```

Arguments

`label`	the name of the subtype variable in the data. This should be a numeric variable with values 0 through M, where 0 indicates control subjects. Must be supplied in quotes, e.g. `label = "subtype"`. quotes.
`M`	is the number of subtypes. For M>=2.
`factors`	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. e.g. `factors = list("age", "sex", "race")`.
`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

```
d(
  label = "subtype",
  M = 4,
  factors = list("x1", "x2", "x3"),
  data = subtype_data
)
```

`dstar` *Estimate the incremental explained risk variation in a case-only study*

Description

`dstar` estimates the incremental explained risk variation across a set of pre-specified disease subtypes in a case-only study. The highest frequency level of label is used as the reference level, for stability. This function takes the name of the disease subtype variable, the number of disease subtypes, a list of risk factors, and a wide case-only dataset, and does the needed transformation on the dataset to get the correct format. Then the polytomous logistic regression model is fit using `mlogit`, and D* is calculated based on the resulting risk predictions.

Usage

```
dstar(label, M, factors, data)
```

Arguments

`label`	the name of the subtype variable in the data. This should be a numeric variable with values 0 through M, where 0 indicates control subjects. Must be supplied in quotes, e.g. `label = "subtype"`. quotes.
`M`	is the number of subtypes. For M>=2.
`factors`	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. e.g. `factors = list("age", "sex", "race")`.
`data`	the name of the case-only 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.

Examples

```
# Exclude controls from data as this is a case-only calculation
dstar(
  label = "subtype",
  M = 4,
  factors = list("x1", "x2", "x3"),
  data = subtype_data[subtype_data$subtype > 0, ]
)
```

`eh_test_marker` *Test for etiologic heterogeneity of risk factors according to individual disease markers in a case-control study*

Description

`eh_test_marker` takes a list of individual disease markers, a list of risk factors, a variable name denoting case versus control status, and a dataframe, 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 disease marker of which the disease subtypes are comprised. Input is a dataframe that contains the individual disease markers, the risk factors of interest, and an indicator of case or control status. The disease markers must be binary and must have levels 0 or 1 for cases. The disease markers should be left missing for control subjects. For categorical disease markers, a reference level should be selected and then indicator variables for each remaining level of the disease marker should be created. 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.

Usage

```
eh_test_marker(markers, factors, case, data, digits = 2)
```

Arguments

`markers`	a list of the names of the binary disease markers. Each must have levels 0 or 1 for case subjects. This value will be missing for all control subjects. e.g. `markers = list("marker1", "marker2")`
`factors`	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. e.g. `factors = list("age", "sex", "race")`
`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, e.g. `case = "status"`.
`data`	the name of the dataframe that contains the relevant variables.
`digits`	the number of digits to round the odds ratios and associated confidence intervals, and the estimates and associated standard errors. Defaults to 2.

Value

Returns a list.

beta is a matrix containing the raw estimates from the polytomous logistic regression model fit with mlogit with a row for each risk factor and a column for each disease subtype.

beta_se is a matrix containing the raw standard errors from the polytomous logistic regression model fit with mlogit with a row for each risk factor and a column for each disease subtype.

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

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

gamma_se is a matrix containing the estimated disease marker standard errors, obtained based on a transformation of the beta standard errors, with a row for each risk factor and a column for each disease marker.

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

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

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

gamma_se_p is a dataframe with disease marker estimates (SE) and their associated p-values.

Author(s)

Emily C Zabor <zabore@mskcc.org>

Examples

```
# Run for two binary tumor markers, which will combine to form four subtypes
eh_test_marker(
  markers = list("marker1", "marker2"),
  factors = list("x1", "x2", "x3"),
  case = "case",
  data = subtype_data,
  digits = 2
)
```

---

eh_test_subtype *Test for etiologic heterogeneity of risk factors according to disease subtypes in a case-control study*

Description

eh_test_subtype takes the name of the variable containing the pre-specified subtype labels, the number of subtypes, a list of risk factors, and the name of the dataframe 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. The multinomial logistic regression model is fit using mlogit.

Usage

```
eh_test_subtype(label, M, factors, data, digits = 2)
```

Arguments

label	the name of the subtype variable in the data. This should be a numeric variable with values 0 through M, where 0 indicates control subjects. Must be supplied in quotes, e.g. label = "subtype".
M	is the number of subtypes. For M>=2.
factors	a list of the names of the binary or continuous risk factors. For binary or categorical risk factors the lowest level will be used as the reference level. e.g. factors = list("age", "sex", "race").
data	the name of the dataframe that contains the relevant variables.
digits	the number of digits to round the odds ratios and associated confidence intervals, and the estimates and associated standard errors. Defaults to 2.

Value

Returns a list.

beta is a matrix containing the raw estimates from the polytomous logistic regression model fit with mlogit with a row for each risk factor and a column for each disease subtype.

beta_se is a matrix containing the raw standard errors from the polytomous logistic regression model fit with mlogit with a row for each risk factor and a column for each disease subtype.

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

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

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

var_covar contains the variance-covariance matrix associated with the model estimates contained in beta.

Author(s)

Emily C Zabor <zabore@mskcc.org>

Examples

```
eh_test_subtype(
  label = "subtype",
  M = 4,
  factors = list("x1", "x2", "x3"),
  data = subtype_data,
  digits = 2
)
```

`optimal_kmeans_d` *Obtain optimal D solution based on k-means clustering of disease marker data in a case-control study*

Description

`optimal_kmeans_d` applies k-means clustering using the `kmeans` function with many random starts. The D value is then calculated for the cluster solution at each random start using the `d` function, and the cluster solution that maximizes D is returned, along with the corresponding value of D. In this way the optimally etiologically heterogeneous subtype solution can be identified from possibly high-dimensional disease marker data.

Usage

```
optimal_kmeans_d(markers, M, factors, case, data, nstart = 100, seed = NULL)
```

Arguments

`markers`	a vector of the names of the disease markers. These markers should be of a type that is suitable for use with `kmeans` clustering. All markers will be missing for control subjects. e.g. `markers = c("marker1", "marker2")`
`M`	is the number of clusters to identify using `kmeans` clustering. For M>=2.
`factors`	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. e.g. `factors = list("age", "sex", "race")`
`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, e.g. `case = "status"`.
`data`	the name of the dataframe that contains the relevant variables.
`nstart`	the number of random starts to use with `kmeans` clustering. Defaults to 100.
`seed`	an integer argument passed to `set.seed`. Default is NULL. Recommended to set in order to obtain reproducible results.

Value

Returns a list

`optimal_d` The D value for the optimal D solution

`optimal_d_data` The original data frame supplied through the `data` argument, with a column called `optimal_d_label` added for the optimal D subtype label. This has the subtype assignment for cases, and is 0 for all controls.

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.

Examples

```
# Cluster 30 disease markers to identify the optimally
# etiologically heterogeneous 3-subtype solution
res <- optimal_kmeans_d(
  markers = c(paste0("y", seq(1:30))),
  M = 3,
  factors = list("x1", "x2", "x3"),
  case = "case",
  data = subtype_data,
  nstart = 100,
  seed = 81110224
)

# Look at the value of D for the optimal D solution
res[["optimal_d"]]

# Look at a table of the optimal D solution
table(res[["optimal_d_data"]]$optimal_d_label)
```

---

posthoc_factor_test *Post-hoc test to obtain overall p-value for a factor variable used in a* `eh_test_subtype` *fit.*

Description

`posthoc_factor_test` takes a `eh_test_subtype` fit and returns an overall p-value for a specified factor variable.

Usage

```
posthoc_factor_test(fit, factor, nlevels)
```

Arguments

fit	the resulting eh_test_subtype fit.
factor	is the name of the factor variable of interest, supplied in quotes, e.g. factor = "race". Only supports a single factor.
nlevels	is the number of levels the factor variable in factor has.

Value

Returns a list.

pval is a formatted p-value.

pval_raw is the raw, unformatted p-value.

Author(s)

Emily C Zabor <zabore@mskcc.org>

---

subtype_data *Simulated subtype data*

Description

A dataset containing 2000 patients: 1200 cases and 800 controls. There are four subtypes, and both numeric and character subtype labels. The subtypes are formed by cross-classification of two binary disease markers, disease marker 1 and disease marker 2. There are three risk factors, two continuous and one binary. One of the continuous risk factors and the binary risk factor are related to the disease subtypes. There are also 30 continuous tumor markers, 20 of which are related to the subtypes and 10 of which represent noise, which could be used in a clustering analysis.

Usage

```
subtype_data
```

Format

A data frame with 2000 rows–one row per patient

case Indicator of case control status, 1 for cases and 0 for controls

subtype Numeric subtype label, 0 for control subjects

subtype_name Character subtype label

marker1 Disease marker 1

marker2 Disease marker 2

x1 Continuous risk factor 1

x2 Continuous risk factor 2

x3 Binary risk factor

y1 Continuous tumor marker 1

y2 Continuous tumor marker 2

y3 Continuous tumor marker 3

y4 Continuous tumor marker 4

y5 Continuous tumor marker 5

y6 Continuous tumor marker 6

y7 Continuous tumor marker 7

y8 Continuous tumor marker 8

y9 Continuous tumor marker 9

y10 Continuous tumor marker 10

y11 Continuous tumor marker 11

y12 Continuous tumor marker 12

y13 Continuous tumor marker 13

y14 Continuous tumor marker 14

y15 Continuous tumor marker 15

y16 Continuous tumor marker 16

y17 Continuous tumor marker 17

y18 Continuous tumor marker 18

y19 Continuous tumor marker 19

y20 Continuous tumor marker 20

y21 Continuous tumor marker 21

y22 Continuous tumor marker 22

y23 Continuous tumor marker 23

y24 Continuous tumor marker 24

y25 Continuous tumor marker 25

y26 Continuous tumor marker 26

y27 Continuous tumor marker 27

y28 Continuous tumor marker 28

y29 Continuous tumor marker 29

y30 Continuous tumor marker 30

Index