Package 'EHR'

October 20, 2017

Version 0.1-3

Date 2017-10-19

Title Electronic Health Record (EHR) Data Processing and Analysis Tool

Author Leena Choi [aut, cre], Cole Beck [aut]

Maintainer Leena Choi <naturechoi@gmail.com>

Description Process and analyze Electronic Health Record (EHR) data. Frequency and contingency tables for many binary outcomes and a binary exposure variable can be generated more efficiently. Phenome Wide Association Study (PheWAS) performed using EHR data can be analyzed using three commonly used statistical analysis methods: Firth's penalized-likelihood logistic regression; logistic regression with likelihood ratio test; conventional logistic regression with Wald test.

Depends R (>= 2.10)

License GPL (>= 3)

Imports stats, utils, logistf

Suggests glmnet

NeedsCompilation no

Repository CRAN

Date/Publication 2017-10-20 10:33:53 UTC

R topics documented:

EHR-package *Electronic Health Record (EHR) Data Processing and Analysis Tool*

Description

Process and analyze Electronic Health Record (EHR) Data. Implement three statistical methods for Phenome Wide Association Study (PheWAS).

Details

Contingency tables for many binary outcomes (e.g., phenotypes) and a binary covariate (e.g., exposure) can be efficiently generated by `zeroOneTable`, and three commonly used statistical methods to analyze data for PheWAS can be implement by `analysisPheWAS`.

Author(s)

Leena Choi <naturechoi@gmail.com> and Cole Beck <cole.beck@Vanderbilt.Edu>

Maintainer: Leena Choi <naturechoi@gmail.com>

analysisPheWAS *Statistical Analysis for PheWAS*

Description

Implement three commonly used statistical methods to analyze data for Phenome Wide Association Study (PheWAS)

Usage

```
analysisPheWAS(method=c('firth','glm', 'lr'),
               adjust=c('PS','demo','PS.demo','none'), Exposure, PS,
               demographics, phenotypes, data)
```

Arguments

`method`	define the statistical analysis method from 'firth', 'glm', and 'lr'. 'firth': Firth's penalized-likelihood logistic regression; 'glm': logistic regression with Wald test, 'lr': logistic regression with likelihood ratio test.
`adjust`	define the adjustment method from 'PS','demo','PS.demo', and 'none'. 'PS': adjustment of PS only; 'demo': adjustment of demographics only; 'PS.demo': adjustment of PS and demographics; 'none': no adjustment.
`Exposure`	define the variable name of exposure variable.
`PS`	define the variable name of propensity score.
`demographics`	define the list of demographic variables.
`phenotypes`	define the list of phenotypes that need to be analyzed.
`data`	define the data.

Details

Implements three commonly used statistical methods to analyze the associations between exposure (e.g., drug exposure, genotypes) and various phenotypes in PheWAS. Firth's penalized-likelihood logistic regression is the default method to avoid the problem of separation in logistic regression, which is often a problem when analyzing sparse binary outcomes and exposure. Logistic regression with likelihood ratio test and conventional logistic regression with Wald test can be also performed.

Value

`estimate`	the estimate of log odds ratio.
`stdError`	the standard error.
`statistic`	the test statistic.
`pvalue`	the p-value.

Author(s)

Leena Choi <naturechoi@gmail.com> and Cole Beck <cole.beck@Vanderbilt.Edu>

Examples

```
## use small datasets to run this example
data(dataPheWASsmall)
## make dd.base with subset of covariates from baseline data (dd.baseline.small)
## or select covariates with upper code as shown below
upper.code.list <- unique(sub("[.][^.]*(.).*", "", colnames(dd.baseline.small)) )
upper.code.list <- intersect(upper.code.list, colnames(dd.baseline.small))
dd.base <- dd.baseline.small[, upper.code.list]
## perfrom regularized logistic regression to obtain propensity score (PS)
## to adjust for potential confounders at baseline
phenos <- setdiff(colnames(dd.base), c('id', 'exposure'))
data.x <- as.matrix(dd.base[, phenos])
glmnet.fit <- glmnet::cv.glmnet(x=data.x, y=dd.base[,'exposure'],
                                family="binomial", standardize=TRUE,
                                alpha=0.1)
dd.base$PS <- c(predict(glmnet.fit, data.x, s='lambda.min'))
data.ps <- dd.base[,c('id', 'PS')]
dd.all.ps <- merge(data.ps, dd.small, by='id')
demographics <- c('age', 'race', 'gender')
phenotypeList <- setdiff(colnames(dd.small), c('id','exposure','age','race','gender'))
## run with a subset of phenotypeList to get quicker results
phenotypeList.sub <- sample(phenotypeList, 5)
results.sub <- analysisPheWAS(method='firth', adjust='PS', Exposure='exposure',
PS='PS', demographics=demographics,
phenotypes=phenotypeList.sub, data=dd.all.ps)
## run with the full list of phenotype outcomes (i.e., phenotypeList)

results <- analysisPheWAS(method='firth', adjust='PS',Exposure='exposure',
                          PS='PS', demographics=demographics,
                          phenotypes=phenotypeList, data=dd.all.ps)
```

dd *dd*

Description

Simulated outcome data example from Phenome Wide Association Study (PheWAS) that examines associations between drug exposure and various phenotypes at follow-up after the drug exposure. The dataset includes 1505 variables: subject identification number ('id'), drug exposure ('exposure'), 3 demographic variables ('age', 'race', 'gender'), and 1500 phenotypes.

Usage

```
dd
```

Format

A data frame with 10000 observations on 1505 variables.

Examples

```
data(dataPheWAS)
```

dd.baseline *dd.baseline*

Description

Simulated baseline data example from a Phenome Wide Association Study (PheWAS) obtained at baseline before drug exposure. The dataset includes 1505 variables: subject identification number ('id'), drug exposure ('exposure'), 3 demographic variables ('age', 'race', 'gender'), and 1500 phenotypes.

Usage

```
dd.baseline
```

Format

A data frame with 10000 observations on 1505 variables.

Examples

```
data(dataPheWAS)
```

dd.baseline.small *dd.baseline.small*

Description

A smaller subset of baseline data example, dd.baseline. The dataset includes 55 variables: subject identification number ('id'), drug exposure ('exposure'), 3 demographic variables ('age', 'race', 'gender'), and 50 phenotypes.

Usage

```
dd.baseline.small
```

Format

A data frame with 2000 observations on 55 variables.

Examples

```
data(dataPheWASsmall)
```

dd.small *dd.small*

Description

A smaller subset of outcome data example, 'dd'. The dataset includes 55 variables: subject identification number ('id'), drug exposure ('exposure'), 3 demographic variables ('age', 'race', 'gender'), and 50 phenotypes.

Usage

```
dd.small
```

Format

A data frame with 2000 observations on 55 variables.

Examples

```
data(dataPheWASsmall)
```

Logistf *Firth's penalized-likelihood logistic regression with more decimal places of p-value than* logistf *function in the R package* **logistf**

Description

Adapted from logistf in the R package **logistf**, this is the same as logistf except that it provides more decimal places of p-value that would be useful for Genome-Wide Association Study (GWAS) or Phenome Wide Association Study (PheWAS).

Usage

```
Logistf(formula = attr(data, "formula"), data = sys.parent(), pl = TRUE,
alpha = 0.05, control, plcontrol, firth = TRUE, init, weights,
plconf = NULL, dataout = TRUE, ...)
```

Arguments

formula	a formula object, with the response on the left of the operator, and the model terms on the right. The response must be a vector with 0 and 1 or FALSE and TRUE for the outcome, where the higher value (1 or TRUE) is modeled. It is possible to include contrasts, interactions, nested effects, cubic or polynomial splines and all S features as well, e.g. `Y ~ X1*X2 + ns(X3, df=4)`. From version 1.10, you may also include offset() terms.
data	a data.frame where the variables named in the formula can be found, i. e. the variables containing the binary response and the covariates.
pl	specifies if confidence intervals and tests should be based on the profile penalized log likelihood (`pl=TRUE`, the default) or on the Wald method (`pl=FALSE`).
alpha	the significance level (1-α the confidence level, 0.05 as default).
control	Controls Newton-Raphson iteration. Default is `control= logistf.control(maxstep, maxit, maxhs, lconv, gconv, xconv)`
plcontrol	Controls Newton-Raphson iteration for the estimation of the profile likelihood confidence intervals. Default is `plcontrol=logistpl.control(maxstep, maxit, maxhs, lconv, xconv, ortho, pr)`
firth	use of Firth's penalized maximum likelihood (`firth=TRUE`, default) or the standard maximum likelihood method (`firth=FALSE`) for the logistic regression. Note that by specifying `pl=TRUE` and `firth=FALSE` (and probably a lower number of iterations) one obtains profile likelihood confidence intervals for maximum likelihood logistic regression parameters.
init	specifies the initial values of the coefficients for the fitting algorithm.
weights	specifies case weights. Each line of the input data set is multiplied by the corresponding element of `weights`.
plconf	specifies the variables (as vector of their indices) for which profile likelihood confidence intervals should be computed. Default is to compute for all variables.

dataout	If TRUE, copies the data set to the output object.
...	Further arguments to be passed to logistf.

Value

same as logistf except for providing more decimal places of p-value.

Author(s)

Leena Choi <naturechoi@gmail.com> and Cole Beck <cole.beck@Vanderbilt.Edu>

References

same as those provided in the R package **logistf**.

Examples

```
data(dataPheWAS)
fit <- Logistf(X264.3 ~ exposure + age + race + gender, data=dd)
summary(fit)
```

---

zeroOneTable *Make Zero One Contingency Tables*

Description

Make contingency tables for many binary outcomes and a binary covariate

Usage

```
zeroOneTable(EXPOSURE, phenotype)
```

Arguments

EXPOSURE	binary covariate (e.g., exposure).
phenotype	binary outcome (e.g., phenotype).

Details

Generates frequency and contingency tables for many binary outcomes (e.g., large number of phenotypes) and a binary covariate (e.g., drug exposure, genotypes) more efficiently.

Value

t00	frequency for non-exposed group and non-case outcome.
t01	frequency for non-exposed group and case outcome.
t10	frequency for exposed group and non-case outcome.
t11	frequency for exposed group and case outcome.

Author(s)

Leena Choi <naturechoi@gmail.com> and Cole Beck <cole.beck@Vanderbilt.Edu>

Examples

```
## full example data
data(dataPheWAS)
demo.covariates <- c('id','exposure','age','race','gender')
phenotypeList <- setdiff(colnames(dd), demo.covariates)
tablePhenotype <- matrix(NA, ncol=4, nrow=length(phenotypeList),
dimnames=list(phenotypeList, c("n.nocase.nonexp", "n.case.nonexp",
"n.nocase.exp", "n.case.exp")))
for(i in seq_along(phenotypeList)) {
    tablePhenotype[i, ] <- zeroOneTable(dd[, 'exposure'], dd[, phenotypeList[i]])
}
```

Index