Package 'LGEWIS'

September 1, 2017

Type Package

Title Tests for Genetic Association/Gene-Environment Interaction in Longitudinal Studies

Version 1.1

Date 2017-8-30

Author Zihuai He, Seunggeun Lee, Bhramar Mukherjee, Min Zhang

Maintainer Zihuai He <zihuai@umich.edu>

Description Functions for genome-wide association studies (GWAS)/gene-environment-wide interaction studies (GEWIS) with longitudinal outcomes and exposures. He et al. (2017) ``Set-Based Tests for Gene-Environment Interaction in Longitudinal Studies" and He et al. (2017) ``Rare-variant association tests in longitudinal studies, with an application to the Multi-Ethnic Study of Atherosclerosis (MESA)".

License GPL-3

Depends CompQuadForm, SKAT, geeM, splines, mvtnorm

NeedsCompilation no

Repository CRAN

Date/Publication 2017-09-01 13:30:52 UTC

R topics documented:

GA.prelim   *The preliminary data management for GA (tests for genetic association)*

Description

Before testing a specific region using a generalized score type test, this function does the preliminary data management, such as fitting the model under the null hypothesis.

Usage

```
GA.prelim(Y,time,X=NULL,corstr="exchangeable")
```

Arguments

Y	The outcome variable, an n*1 matrix where n is the total number of observations
time	An n*2 matrix describing how the observations are measured. The first column is the subject id. The second column is the measured exam (1,2,3,etc.).
X	An n*p covariates matrix where p is the total number of covariates.
corstr	The working correlation as specified in 'geeglm'. The following are permitted: "independence", "exchangeable", "ar1".

Value

It returns a list used for function GA.test().

Examples

```
library(LGEWIS)

# Load data example
# Y: outcomes, n by 1 matrix where n is the total number of observations
# X: covariates, n by p matrix
# time: describe longitudinal structure, n by 2 matrix
# G: genotype matrix, m by q matrix where m is the total number of subjects

data(LGEWIS.example)
Y<-LGEWIS.example$Y
X<-LGEWIS.example$X
time<-LGEWIS.example$time
G<-LGEWIS.example$G

# Preliminary data management
result.prelim<-GA.prelim(Y,time,X=X)
```

GA.SSD.All *Genetic association tests for multiple regions/genes using SSD format files*

Description

Test the association between an quantitative outcome and multiple region/genes using SSD format files.

Usage

```
GA.SSD.All(SSD.INFO, result.prelim, Gsub.id=NULL, B=5000, ...)
```

Arguments

SSD.INFO   SSD format information file, output of function "Open_SSD". The sets are defined by this file.

result.prelim   Output of function "GA.prelim()".

Gsub.id   The subject id corresponding to the genotype matrix, an m dimensional vector. This is in order to match the phenotype and genotype matrix. The default is NULL, where the order is assumed to be matched with Y, X, E and time.

B   Number of Bootstrap replicates. The default is 5000.

...   Other options of the generalized score type test. Defined same as in function "GA.test()".

Details

Please see SKAT vignettes for using SSD files.

Value

results   Results of the set based analysis. First column contains the set ID; Second column (second and third columns when the MinP test is compared) contains the p-values; Last column contains the number of tested SNPs.

results.single Results of the single variant analysis for all variants in the sets. First column contains the regions' names; Second column is the variants' names; Third column contains the minor allele frequencies; Last column contains the p.values.

GA.SSD.OneSet_SetIndex

Genetic association tests for a single region/gene using SSD format files

Description

Test the genetic association between an quantitative outcome and one region/gene using SSD format files.

Usage

```
GA.SSD.OneSet_SetIndex(SSD.INFO, SetIndex, result.prelim, Gsub.id=NULL,
B=5000, ...)
```

Arguments

SSD.INFO   SSD format information file, output of function "Open_SSD". The sets are de-
                fined by this file.

SetIndex   Set index. From 1 to the total number of sets.

result.prelim   Output of function "GA.prelim()".

Gsub.id   The subject id corresponding to the genotype matrix, an m dimensional vector.
                This is in order to match the phenotype and genotype matrix. The default is
                NULL, where the order is assumed to be matched with Y, X, E and time.

B   Number of Bootstrap replicates. The default is 5000.

...   Other options of the generalized score type test. Defined same as in function
                "GA.test()".

Details

Please see SKAT vignettes for using SSD files.

Value

p.value   P-value of the set based generalized score type test.

p.single   P-values of the incorporated single SNP analyses

n.marker   number of tested SNPs in the SNP set.

GA.test | *Test the association between an quantitative outcome variable and a region/gene by a generalized score type test.*

Description

Once the preliminary work is done using "GA.prelim()", this function tests a specifc region/gene. Single SNP analyses are also incorporated.

Usage

```
GA.test(result.prelim,G,Gsub.id=NULL,weights='beta',B=5000, B.coef=NULL,
impute.method='fixed')
```

Arguments

result.prelim   The output of function "GEI.prelim()"

G   Genetic variants in the target region/gene, an m*q matrix where m is the subject ID and q is the total number of genetic variables. Note that the number of rows in Z should be same as the number of subjects.

Gsub.id   The subject id corresponding to the genotype matrix, an m dimensional vector. This is in order to match the phenotype and genotype matrix. The default is NULL, where the order is assumed to be matched with Y, X and time.

weights   Can be a numeric vector of weights for genetic variants (The length should be same as the number of genetic variants in the set.), or pre-determined weights: "beta" (beta weights as in SKAT paper), "rare" (restrited to MAF<0.01), "common" (restrited to MAF>0.01). The default is NULL, where the flat weights are applied.

B   Number of Bootstrap replicates. The default is 5000.

B.coef   Direct import of Bootstrap coefficients, an m by B matrix. This is in order to efficiently implement the Bootstrap step. The default is NULL.

impute.method   Choose the imputation method when there is missing genotype. Can be "random", "fixed" or "bestguess". Given the estimated allele frequency, "random" simulates the genotype from binomial distribution; "fixed" uses the genotype expectation; "Best guess" uses the genotype with highest probability.

Value

p.value   P-value of the set based generalized score type test.

p.single   P-values of the incorporated single SNP analyses

n.marker   number of heterozygous SNPs in the SNP set.

Examples

```
## GA.prelim does the preliminary data management.
# Input: Y, time, X (covariates)
## GA.test tests a region.
# Input: G (genetic variants) and result of GEI.prelim

library(LGEWIS)

# Load data example
# Y: outcomes, n by 1 matrix where n is the total number of observations
# X: covariates, n by p matrix
# time: describe longitudinal structure, n by 2 matrix
# G: genotype matrix, m by q matrix where m is the total number of subjects

data(LGEWIS.example)
Y<-LGEWIS.example$Y
X<-LGEWIS.example$X
time<-LGEWIS.example$time
G<-LGEWIS.example$G

# Preliminary data management
result.prelim<-GA.prelim(Y,time,X=X)

# test with 5000 bootstrap replicates
result<-GA.test(result.prelim,G,B=5000)
```

GEI.prelim	*The preliminary data management for GEI (tests for gene-environment interaction)*

Description

Before testing a specific region using a generalized score type test, this function does the preliminary data management, such as pareparing spline basis functions for E etc..

Usage

```
GEI.prelim(Y,time,E,X=NULL,E.method='ns',E.df=floor(sqrt(length(unique(time[,1])))),
corstr="exchangeable")
```

Arguments

Y	The outcome variable, an n*1 matrix where n is the total number of observations
time	An n*2 matrix describing how the observations are measured. The first column is the subject id. The second column is the measured exam (1,2,3,etc.).
E	An n*1 environmental exposure.
X	An n*p covariates matrix where p is the total number of covariates.

E.method	The method of sieves for the main effect of E. It can be "ns" for natural cubic spline sieves; "bs" for B-spline sieves; "ps" for polynomial sieves. The default is "ns".
E.df	Model complexity for the method of sieves, i.e., number of basis functions. The default is sqrt(m).
corstr	The working correlation as specified in 'geeglm'. The following are permitted: "independence", "exchangeable", "ar1".

Value

It returns a list used for function GEI.test().

Examples

```
library(LGEWIS)

# Load data example
# Y: outcomes, n by 1 matrix where n is the total number of observations
# X: covariates, n by p matrix
# E: environmental exposure, n by 1 matrix
# time: describe longitudinal structure, n by 2 matrix
# G: genotype matrix, m by q matrix where m is the total number of subjects

data(LGEWIS.example)
Y<-LGEWIS.example$Y
X<-LGEWIS.example$X
E<-LGEWIS.example$E
time<-LGEWIS.example$time
G<-LGEWIS.example$G

# Preliminary data management
result.prelim<-GEI.prelim(Y,time,E,X=X)
```

GEI.SSD.All	*Gene-environment interaction tests for multiple regions/genes using SSD format files*

Description

Test the interaction between an environmental exposure and multiple region/genes on a quantitative outcome using SSD format files.

Usage

```
GEI.SSD.All(SSD.INFO, result.prelim, Gsub.id=NULL, MinP.adjust=NULL, ...)
```

Arguments

`SSD.INFO`	SSD format information file, output of function "Open_SSD". The sets are defined by this file.
`result.prelim`	Output of function "GEI.prelim()".
`Gsub.id`	The subject id corresponding to the genotype matrix, an m dimensional vector. This is in order to match the phenotype and genotype matrix. The default is NULL, where the order is assumed to be matched with Y, X, E and time.
`MinP.adjust`	If the users would like to compare with the MinP test, this parameter specify the adjustment thereshold as in Gao, et al. (2008) "A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms". Values from 0 to 1 are permitted. The default is NULL, i.e., no comparison. The value suggested by Gao, et al. (2008) is 0.95.
`...`	Other options of the generalized score type test. Defined same as in function "GEI.test()".

Details

Please see SKAT vignettes for using SSD files.

Value

`results`	Results of the set based analysis. First column contains the set ID; Second column (second and third columns when the MinP test is compared) contains the p-values; Last column contains the number of tested SNPs.
`results.single`	Results of the single variant analysis for all variants in the sets. First column contains the regions' names; Second column is the variants' names; Third column contains the minor allele frequencies; Last column contains the p.values.

`GEI.SSD.OneSet_SetIndex`

Gene-environment interaction tests for a single region/gene using SSD format files

---

Description

Test the interaction between an environmental exposure and one region/gene on a quantitative outcome using SSD format files.

Usage

```
GEI.SSD.OneSet_SetIndex(SSD.INFO, SetIndex, result.prelim, Gsub.id=NULL,
MinP.adjust=NULL, ...)
```

Arguments

`SSD.INFO`	SSD format information file, output of function "Open_SSD". The sets are defined by this file.
`SetIndex`	Set index. From 1 to the total number of sets.
`result.prelim`	Output of function "GEI.prelim()".
`Gsub.id`	The subject id corresponding to the genotype matrix, an m dimensional vector. This is in order to match the phenotype and genotype matrix. The default is NULL, where the order is assumed to be matched with Y, X, E and time.
`MinP.adjust`	If the users would like to compare with the MinP test, this parameter specify the adjustment thereshold as in Gao, et al. (2008) "A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms". Values from 0 to 1 are permitted. The default is NULL, i.e., no comparison. The value suggested by Gao, et al. (2008) is 0.95.
`...`	Other options of the generalized score type test. Defined same as in function "GEI.test()".

Details

Please see SKAT vignettes for using SSD files.

Value

`p.value`	P-value of the set based generalized score type test.
`p.single`	P-values of the incorporated single SNP analyses
`p.MinP`	P-value of the MinP test.
`n.marker`	number of tested SNPs in the SNP set.
`E.df`	number of tested SNPs in the SNP set.
`G.df`	number of tested SNPs in the SNP set.

`GEI.test`	*Test the interaction between an environemental exposure and a region/gene by a generalized score type test.*

Description

Once the preliminary work is done using "GEI.prelim()", this function tests a specifc region/gene. Single SNP analyses are also incorporated.

Usage

```
GEI.test(result.prelim,G,Gsub.id=NULL,G.method='wPCA',G.df=floor(sqrt(nrow(G))),
bootstrap=NULL,MinP.adjust=NULL,impute.method='fixed')
```

Arguments

result.prelim	The output of function "GEI.prelim()"
G	Genetic variants in the target region/gene, an m*q matrix where m is the subject ID and q is the total number of genetic variables. Note that the number of rows in Z should be same as the number of subjects.
Gsub.id	The subject id corresponding to the genotype matrix, an m dimensional vector. This is in order to match the phenotype and genotype matrix. The default is NULL, where the order is assumed to be matched with Y, X and time.
G.method	The dimension reduction method for main effect adjustment of G. The following are permitted: "wPCA" for weighted principal component analysis; "PCA" for principal component analysis; "R2" for ordering the principal components by their R-squares. The dimension reduction method is in order to analyze large regions, i.e., the number of variants is close to or larger than the number of subjects. The default is "wPCA".
G.df	Number of components selected by the dimension reduction method. The default is sqrt(m).
bootstrap	Whether to use bootstrap for small sample size adjustement. This is recommended when the number of subjects is small, or the set contains rare variants. The default is NULL, but a suggested number is 10000 when it is needed.
MinP.adjust	If the users would like to compare with the MinP test, this parameter specify the adjustment thereshold as in Gao, et al. (2008) "A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms". Values from 0 to 1 are permitted. The default is NULL, i.e., no comparison. The value suggested by Gao, et al. (2008) is 0.95.
impute.method	Choose the imputation method when there is missing genotype. Can be "random", "fixed" or "bestguess". Given the estimated allele frequency, "random" simulates the genotype from binomial distribution; "fixed" uses the genotype expectation; "Best guess" uses the genotype with highest probability.

Value

p.value	P-value of the set based generalized score type test.
p.single	P-values of the incorporated single SNP analyses
p.MinP	P-value of the MinP test.
n.marker	number of heterozygous SNPs in the SNP set.
E.df	number of tested SNPs in the SNP set.
G.df	number of tested SNPs in the SNP set.

Examples

```
## GEI.prelim does the preliminary data management.
# Input: Y, time, E, X (covariates)
## GEI.test tests a region.
# Input: G (genetic variants) and result of GEI.prelim
```

```
library(LGEWIS)

# Load data example
# Y: outcomes, n by 1 matrix where n is the total number of observations
# X: covariates, n by p matrix
# E: environmental exposure, n by 1 matrix
# time: describe longitudinal structure, n by 2 matrix
# G: genotype matrix, m by q matrix where m is the total number of subjects

data(LGEWIS.example)
Y<-LGEWIS.example$Y
X<-LGEWIS.example$X
E<-LGEWIS.example$E
time<-LGEWIS.example$time
G<-LGEWIS.example$G

# Preliminary data management
result.prelim<-GEI.prelim(Y,time,E,X=X)

# test without the MinP test
result<-GEI.test(result.prelim,G,MinP.adjust=NULL)

# test with the MinP test
result<-GEI.test(result.prelim,G,MinP.adjust=0.95)

# test with the MinP test and the small sample adjustment
result<-GEI.test(result.prelim,G,MinP.adjust=0.95,bootstrap=1000)
```

LGEWIS.example	*Data example for LGEWIS (tests for genetic association or gene-environment interaction)*

Description

Example data for LGEWIS.

Usage

```
data(LGEWIS.example)
```

Format

LGEWIS.example contains the following objects:

G a numeric genotype matrix of 10 individuals and 20 SNPs. Each row represents a different individual, and each column represents a different SNP marker.

Y a numeric vector of continuous phenotypes of 10 individuals with 4 repeated measurements.

time a numeric matrix. The first column is the subject ID and the second column is the measured exam.

X a numeric matrix of 1 covariates.

E a numeric vector of environmental exposure.

Index