Package 'SMAT'

January 29, 2013

Type Package

Title Scaled Multiple-phenotype Association Test

Version 0.98

Date 2013-01-26

Author Lin Li, Ph.D.; Elizabeth D. Schifano, Ph.D.

Maintainer Lin Li <linli@hsph.harvard.edu>; Elizabeth D. Schifano <elizabeth.schifano@uconn.edu>

Description SMAT is an R package for performing the Scaled Multiple-phenotype Association Test in cohort or case-control designs to assess common effect of a single nucleotide polymorphism (SNP) on multiple (positively correlated) continuous outcomes measuring the same underlying trait.

R topics documented:

SMAT-package 2
simdat_caco 2
simdat_co 3
SMAT 3
SMAT.plink 6

Index **9**

SMAT-package *Scaled Multiple-phenotype Association Test*

Description

SMAT is an R package for performing the Scaled Multiple-phenotype Association Test in cohort or case-control designs to assess common effect of a single nucleotide polymorphism (SNP) on multiple (positively correlated) continuous outcomes measuring the same underlying trait.

Details

Two main functions are SMAT() and SMAT.plink(). The latter is a wrapper function of SMAT(), making it easy to conduct genome-wide association scans.

Package:	SMAT
Type:	Package
Version:	0.98
Date:	2013-01-26
License:	GPL-3

Author(s)

Lin Li, Ph.D.; Elizabeth D. Schifano, Ph.D.
Maintainer: Lin Li <linli@hsph.harvard.edu>; Elizabeth D. Schifano <elizabeth.schifano@uconn.edu>

References

Schifano, E.D., Li, L., Christiani, D.C., and Lin, X. (2012) Genome-wide Association Analysis for Multiple Continuous Secondary Phenotypes. (in revision)
Roy, J., Lin, X., and Ryan, L. (2003). Scaled Marginal Models For Multiple Continuous Outcomes. Biostatistics , 4, 371-384.

simdat_caco *A simulated case-control dataset*

Description

Multiple secondary phenotypes, covariates, SNP genotypes, and weights for a simulated case-control study.

Usage

```
data(simdat_caco)
```

Format

$ y: num [1:1426, 1:4] 4 phenotypes (column) for each individual (row)
$ x: num [1:1426, 1:4] 4 covariates (column) for each individual (row). The first covariate is the intercept.
$ s int [1:1426] SNP genotypes for 1426 individuals.
$ w: num [1:1426] weights for 1426 individuals.

Examples

```
data(simdat_caco)
```

simdat_co *A simulated control-only dataset.*

Description

Multiple phenotypes, covariates, and SNP genotypes from a simulated control-only data.

Usage

```
data(simdat_co)
```

Format

The format is:
List of 3
$ y: num [1:343, 1:4] 4 phenotypes (column) for each individual (row).
$ x: num [1:343, 1:3] 3 covariates (column) for each individual (row). The first covariate (column) is the intercept.
$ s: int [1:343] SNP genotypes of a SNP for 343 individual.

Examples

```
data(simdat_co)
```

SMAT *Scaled Multiple-phenotype Association Test*

Description

SMAT is an R package for performing the Scaled Multiple-phenotype Association Test in cohort or case-control designs to assess common effect of a SNP on multiple (positively correlated) continuous outcomes measuring the same underlying trait.

Usage

```
SMAT(y, x, s, w = NULL, status = NULL, prev = NULL, working = "unstr", conv = 1e-04,
maxiter = 500)
```

Arguments

`y` A n-by-M matrix. Rows correspond to n individuals, and columns contain values of M phenotypes.

`x` A n-by-p matrix. Rows correspond to n individuals, and columns contain values of p covariates (including the intercept if applicable). Note that an intercept column should be provided if the intercept is present in the model.

`s` A vector of length n, containing SNP genotypes of n individuals. Currently only a single SNP is considered.

`w` A vector of length n, containing the weights for n individuals. This vector should be provided for analyzing secondary phenotypes in case-control studies, if the user wants to specify the weights. If not provided, SMAT will check status and prev for calculating the weights within SMAT. If none of w, status and prev are provided (e.g. in control-only studies), equal weights will be used.

`status` A vector of length n, containing the case-control status for n individuals. If the user wants to let SMAT calculate the weights automatically for analyzing secondary phenotypes in case-control studies, this vector should be provided and w should not be specified. Disease prevalence prev should also be provided. If none of w, status and prev are provided (e.g. in control-only studies), equal weights will be used. Note that status only takes values of 1 (case), 0 (control) or NA (missing).

`prev` A scalar, corresponding to the disease prevalence. This together with status should be provided if the user wants to let SMAT calculate the weights. See description of status above for more details. If not provided (e.g. in control-only studies), equal weights will be used.

`working` A string being one of "unstr", "ind", and "exch". This string specifies the choice of the working correlation matrix: unstructured, independence, or exchangeable. The default is "unstr".

`conv` A numeric value to specify the convergence threshold.

`maxiter` An integer specifying the maximum number of iterations.

Details

Consider M phenotypes for n individuals, with p covariates and genotypes of a SNP. The phenotypes are expected to be positively correlated. The function will first check the pairwise correlation of each pair of phenotypes. A warning message will be returned in the beginning if any pair of phenotypes are not positively correlated. If not, consider a transformation (i.e., change in sign) so that the phenotypes will be positively correlated. Otherwise, it is not recommended to use this function in this situation.

The coefficients for the covariates can vary among different phenotypes, and their values are stored in a p-by-M matrix (beta), with each column corresponding to a phenotype. The coefficient for the common effect of a SNP on the phenotypes is stored in a scalar (alpha).

Three ways of specifying the weights are provided: via w if the user wants to specify the weights; via status and prev if the user wants to let SMAT calculate the weights automatically, which is helpful if there are missing values in the data; if none of w, status and prev are provided, equal weights will be used.

Value

`beta` A p-by-M matrix. The estimates of covariate coefficients, each column for a phenotype.

`se.beta` A p-by-M matrix. The estimates of standard error of beta, each column for a phenotype.

`alpha` A scalar. The estimate of the common effect coefficient for a SNP.

`se.alpha` A scalar. The estimate of standard error of alpha.

`sigma2` A vector of length M. The estimates of standard error of the M phenotypes.

`se.sigma2` A vector of length M. The estimates of stand error of sigma2.

`pvalue` A scalar. The 1-DF common effect test (SMAT) p-value for a SNP.

`pvalue.common` A scalar. The p-value of the test for the common effect (homogeneity) assumption.

`is.conv` A boolean value. The indicator whether SMAT converges before the maximum iteration number (maxiter) is reached.

Author(s)

Lin Li, Ph.D.; Elizabeth D. Schifano, Ph.D.

References

Elizabeth D. Schifano, Lin Li, David C. Christiani, and Xihong Lin (2012) Genome-wide Association Analysis for Multiple Continuous Secondary Phenotypes (in revision)

Examples

```
###### a simulated example of analyzing case-control data ######
# load a simulated case-control dataset
data(simdat_caco)
# run SMAT with unstructured working matrix
out = with(simdat_caco, SMAT(y, x, s, w, working = "unstr"))
# 1-DF common effect test p-value for a SNP
print(out$pvalue)
# check common effect assumption:
print(out$pvalue.common)

###### a simulated example of analyzing control-only data ######
# load a simulated control-only dataset
data(simdat_co)
# run SMAT with unstructured working matrix
out = with(simdat_co, SMAT(y, x, s, working = "unstr"))
# 1-DF common effect test p-value for a SNP
```

```
print(out$pvalue)
# check common effect assumption:
print(out$pvalue.common)
```

SMAT.plink *Scaled Multiple-phenotype Association Test with PLINK binary files as inputs*

Description

SMAT.plink is a wrapper function for SMAT that can easily read PLINK binary files as inputs of genotypes. This function can be used for association scans of genome-wide association studies.

Usage

```
SMAT.plink(file, file.pheno, col.pheno, col.covar=NULL, col.w=NULL, col.status=NULL,
prev=NULL, working="unstr", conv=1e-4, maxiter=500, verbose=FALSE)
```

Arguments

`file` A character string. The file name of the PLINK binary files. For example, a file="sim20" indicates that the files "sim20.bed", "sim20.bim", "sim20.fam" are to be used as inputs of SNP genotypes These files must exist, otherwise the program will stop with an error message. The binary files must be in PLINK v1.00 format and SNP-major. For more information, please refer to the PLINK documentation.

`file.pheno` A character string. The file name of the phenotype file, which follows the format of PLINK phenotype files except for missing values. A header of column names needs to be specified at the first line, with "FID" and "IID" as the first two column names. Both phenotypes and covariates should be included in this file. Missing values should be replaced by NA.

`col.pheno` A numeric vector. This vector specifies the columns in the phenotype file that should be treated as the phenotypes. Note that a length of at least 2 is considered.

`col.covar` A numeric vector. This vector specifies the columns in the phenotype file that should be treated as the covariates. An intercept will be automatically included as a covariate. If no values are given, the covariate is solely the intercept.

`col.w` A scalar. This scalar specifies the column in the phenotype file that should be treated as the weights for n individuals. This vector should be provided for analyzing secondary phenotypes in case-control studies, if the user wants to specify the weights. If not provided, the user can also specify col.status and prev to let SMAT calculate the weights automatically. If none of them are provided (e.g. in control-only studies), equal weights will be used.

`col.status` A scalar. This scalar specifies the column in the phenotype file that should be treated as the case-control status for n individuals. This vector should be provided for analyzing secondary phenotypes in case-control studies, if the user wants to let SMAT calculate the weights automatically (in this situation, col.w should be NULL). If none of col.w, col.status and prev are provided (e.g. in control-only studies), equal weights will be used. Note that the column for the case-control status should only take values of 1 (case), 0 (control) or NA (missing).

`prev` A scalar. This is the disease prevalence, which should be provided if col.status is provided. See more details in the descriptions of w and col.status.

`working` A string being one of "unstr", "ind", and "exch". This string specifies the choice of the working correlation matrix: unstructured, independence, or exchangeable. The default is "unstr".

`conv` A numeric value to specify the convergence threshold.

`maxiter` An integer specifying the maximum number of iterations.

`verbose` Logical. If true, the index of each SNP analyzed will be printed out in the association scans. This helps the user to know the progress.

Details

This function is a wrapper function of SMAT that takes PLINK binary format files as inputs. For other formats, please use SMAT directly and prepare the inputs as needed.
A warning message will be returned in the beginning if any pair of phenotypes are not positively correlated. If not, consider a transformation (i.e., change in sign) so that the phenotypes will be positively correlated. Otherwise, it is not recommended to use this function in this situation.

Value

`pvalues` A data frame. The .bim information with two columns appended. One column contains to the 1-DF common effect test (SMAT) p-values, and the other column contains the p-values of the test for the common effect (homogeneity) assumption.

`fits` A vector of lists. Each elment is a list returned by SMAT for a SNP. Please check SMAT for more information of the returned values.

Author(s)

Lin Li, Ph.D.; Elizabeth D. Schifano, Ph.D.

References

Elizabeth D. Schifano, Lin Li, David C. Christiani, and Xihong Lin (2012) Genome-wide Association Analysis for Multiple Continuous Secondary Phenotypes (in revision)

Examples

```
## Not run:
# The PLINK binary data "sim20" (sim20.bed, sim20.bim, sim20.fam) contain the genotype
#  information of 20 SNPs for 2000 individuals from a cohort
# The phenotype file "pheno.txt" contains the phenotypes (columns 3-7) and the covariates
#  (columns 8-9) for the 2000 individuals
# call SMAT is very easy with these files as inputs
```

```
fits=SMAT.plink(file="sim20",file.pheno="pheno.txt",col.pheno=3:7,col.covar=8:9,verbose=T)
# Output p-values
print(fits$pvalues)

## End(Not run)
```

Index

∗Topic **datasets**
- simdat_caco, 2
- simdat_co, 3

simdat_caco, 2
simdat_co, 3
SMAT, 3
SMAT-package, 2
SMAT.plink, 6