Package 'CCGA'

November 4, 2016

Type Package

Title Case-control genetic association analysis incorporating non-confounding covariates

Version 1.0

Date 2016-10-27

Author Hong Zhang

Maintainer Hong Zhang <zhanghfd@fudan.edu.cn>

Depends rootSolve, maxLik, parallel, R (>= 2.14)

Description The power of disease-SNP association analysis in case-control studies can be potentially improved through adjustment of non-confounding covariates.

License Artistic License 2.0

LazyLoad yes

NeedsCompilation no

RoxygenNote 5.0.1

URL http://github.com/zhanghfd/CCGA

BugReports http://github.com/zhanghfd/CCGA/issues

R topics documented:

CCGA-package *Case-control genetic association analysis with adjustment of non-confounding covariates.*

Description

In CCGA, the power of disease-SNP association analysis in case-control studies can be potentially improved through efficiently adjustmenting non-confounding covariates, compared with the standard logistic regression method with or without covariate adjustment.

Let Y, S, G, and X denote the case-control status, stratum variable, SNP genotype coded by 0, 1, or 2, and covariate(s), respectively. The model for relating the response Y and (S,G,X) is

g(pr(Y=1)) = alpha + beta_S S + beta_G G + beta_X X,

where g(.) is a given function, which can be the logit function or the probit function. The disease prevalence is required for each stratum.

The disease prevalence(s), Hardy-Weinberg equilibrium, and independence between non-confounding covariate(s) and SNP genotype are efficiently incorporated in the retrospective likelihood function, and the nonparametric distribution of the covariate(s) is profiled out through the application of Lagrange's multipler method. The multipliers can be directly estimated using the available data, which yields the so called profile maximum likelihood estimates of the regression parameter (pMLE). Alternatively, the data-dependent multiplier(s) can be replaced with the limiting value(s) to yield a modified profile likeihood function, which results in modified profile maximum likelihood estimates (mpMLE). Theoretically, mpMLE and pMLE are asymptotically equivalent in terms of estimation efficiency, but mpMLE could be computationally much simplier and faster.

This package includes two main functions (i.e., SingleSNP and MultipleSNP) and a simulated dataset for illustration. In SingleSNP and MultipleSNP, two candidate link functions (i.e., the logit function and probit function) can be used, and both mpMLE and pMLE can be implemented. In the function MultipleSNP, multiple CPU cores can be used to speed up the analysis with UNIX-like OS.

Details

Package:	CCGA
Type:	Package
Version:	1.0
Date:	2016-10-27
License:	Artistic License 2.0

Author(s)

Hong Zhang

Maintainer: Hong Zhang <zhanghfd@fudan.edu.cn>

References

Zhang H, Chatterjee N, Rader D, Chen J. (2016) Adjustment of Non-confounding Covariates in Case-control Genetic Association Studies. Annals of Applied Statistics (revised).

`data` *A simulated dataset.*

Description

The dataset contains the information on stratum, disease status, SNP genotype, and three covariates.

Usage

```
data(data)
```

Format

A data frame containing the following variables for each of 1200 observations: status (disease status coded as 1 for case or 0 for control), stratum (a three-level stratum variable coded as 1, 2, or 2), covariate.1, covariate.2, covariate.3 (three continuous covariates), SNP.1,...,SNP.100 (100 SNP genotypes coded as 0, 1, or 2 according to the number of minor alleles).

Examples

```
data(data);
```

`MultipleSNP` *Estimation of log-ORs (SEs) and significance test p-values for multiple SNPs.*

Description

This function returns (1) log-OR estimates and the corresponding (2) standard errors and the significance test (3) p-values for multiple SNPs.

The imputs of this function include (1) disease status, (2) SNP genotype, (3) covariate(s), and (4) stratum indicator for each subject. Furthermore, the disease prevalence for each stratum is also required. Parallel computation can be used to speed up the analysis through specification of more than one CPU core. Because the function "mcapply" in the R package "parallel" is used, the core number can be greater than 1 with UNIX-like OS but it can be only 1 with Windows OS.

If a covariate is categorical, dummy variables should be constructed before using the function.

Usage

```
MultipleSNP(Gs, Y, Z, S, fs, par = NULL, link = "logit", modified=TRUE, cl.cores=1)
```

Arguments

Gs	SNP genotype matrix coded by 0, 1, or 2 according to the number of minor alleles, one colume per one SNP.
Y	Response variable (vector), which should take a value of 1 (case) or 0 (control).
Z	Covariate variable(s) (numerical vector for a single covariate or matrix for multiple covariates). Categorical variable should be coded as dummy variables.
S	Stratum variable (vector), which should take a value of 1, 2, ..., or K, where K is the number of strata.
fs	Prevalence(s) (vector). The ith entry is the prevalence for stratum i.
par	Initial regression parameters (list) including alpha, betaG, betaX, and betaS, with the default value being NULL.
link	Link function, which should be either 'logit' (default value) or 'probit'.
modified	An indicator for modifiying the profile likelihood or not. If it is TRUE (default value), then the profile likelihood will be modified; otherwise the original profile likelihood function will be used.
cl.cores	CPU cores used, with a defaul value of 1.

Examples

```
data(data);

status = data[,1];

stratum = data[,2];

covariate = data[,3:5];

Gs = data[,-(1:5)];

fs = c(0.010869, 0.000867, 0.001707);

res = MultipleSNP(Gs,Y=status,Z=covariate,S=stratum,fs=fs,cl.cores=1)
```

SingleSNP *Estimation of regression parameters and their standard errors for a single SNP.*

Description

This function returns parameter estimates and their standard errors for the parameters in the model relating disease status and (1) SNP genotype, (2) stratum variable, and (3) covariate(s). The estimated parameters include (1) regression parameters and (2) minor allele frequency. If the original profile likelihood is maximized, If the outputs also include (3) Lagrange's multipliers.

The imputs of this function include (1) disease status, (2) SNP genotype, (3) covariate(s), and (4) stratum indicator for each subject. Furthermore, the disease prevalence for each stratum is also required.

If a covariate is categorical, dummy variables should be constructed before using the function.

Usage

```
SingleSNP(Y, G, X, S, fs, par = NULL, link = "logit", modified = TRUE)
```

Arguments

Y	Response variable (vector), which should take a value of 1 (case) or 0 (control).
G	SNP genotype coded by 0, 1, or 2 according to the number of minor alleles.
X	Covariate variable(s) (numerical vector for a single covariate or matrix for multiple covariates). Categorical variable should be coded as dummy variables.
S	Stratum variable (vector), which should take a value of 1, 2, ..., or K, where K is the number of strata.
fs	Prevalence(s) (vector). The ith entry is the prevalence for stratum i.
par	Initial regression parameters (list) including alpha, betaG, betaX, and betaS, with the default value being NULL.
link	Link function, which should be either 'logit' (default value) or 'probit'.
modified	An indicator for modifiying the profile likelihood or not. If it is TRUE (default value), then the profile likelihood will be modified; otherwise the original profile likelihood function will be used.

Value

LOGIT	Standard logistic regression results including the MLEs of regression parameters and their standard errors.
mpMLE	Modified profile MLEs and their standard errors.
pMLE	The original prifile MLEs and their standard errors, available when "modified" is FALSE.

Examples

```
data(data);

status = data[,1];

stratum = data[,2];

covariate = data[,3:5];

genotype = data[,6];

fs = c(0.010869, 0.000867, 0.001707);

fit = SingleSNP(Y=status,G=genotype,X=covariate,S=stratum,fs=fs);
```

Index