Package 'evian'

July 18, 2018

Type Package

Title Evidential Analysis of Genetic Association Data

Version 2.0.0

Depends R (>= 3.2.5), ProfileLikelihood, sandwich, foreach, doParallel

Author Dr. Lisa J Strug <lisa.strug@utoronto.ca>;
Dr. Zeynep Baskurt <zeynep.baskurt@sickkids.ca>;
Bowei Xiao <bowei.xiao@sickkids.ca>;
Ted Chiang <theodorechiang@gmail.com>

Maintainer Bowei Xiao <bowei.xiao@sickkids.ca>

Description Evidential regression analysis for dichotomous and quantitative outcome data. The following references described the methods in this package:
Strug, L. J., Hodge, S. E., Chiang, T., Pal, D. K., Corey, P. N., & Rohde, C. (2010) <doi:10.1038/ejhg.2010.47>.
Strug, L. J., & Hodge, S. E. (2006) <doi:10.1159/000094709>.
Royall, R. (1997) <ISBN:0-412-04411-0>.

License GPL (>= 2)

Encoding UTF-8

LazyData true

RoxygenNote 6.0.1

NeedsCompilation no

Repository CRAN

Date/Publication 2018-07-18 13:10:06 UTC

R topics documented:

`adjustModel` *Genotype coding adjusement*

Description

This is a helper function that adjusts the genotye coding scheme based on the genetic model specified.

Usage

```
adjustModel(data_nomiss,model)
```

Arguments

`data_nomiss`	a data frame that contains the phenotype, covariates and genotype columns. Genotypes need to be coded as 0/1/2.
`model`	The model specified. Must be a string from one of the following: `additive`, `dominant`, `recessive`, or `overdominance`

Details

`adjustModel` is an interior function that adjusts the genotype coding based on the genetic model specified. The coding scheme for different genetic models can be found in `calculateLinearMLE`.

Value

This function returns the data frame with the same columns but changed genotype coding based on the genetic model specified.

`calculateLinearMLE` *Profile likelihood calculation for quantitative outcome data using linear regression models*

Description

This is the function that calculates profileLikelihood for a single SNP. The main function `evian_linear` calls this function repeatedly to obtain results for multiple SNPs.

Usage

```
calculateLinearMLE(snp, formula_tofit, model, data, bim, lolim, hilim, m, bse,
k, robust, family)
```

Arguments

`snp`	a string specifying the SNP of interests to be calculated.
`formula_tofit`	a formula object of the genetic model. The model should be formatted as `y~nuisance parameters`. The parameter of interest should not be included here.
`model`	a string specifying the mode of inheritance parameterization: `additive, dominant, recessive, or overdominance`. See details.
`data`	data frame; read from the argument `data` in the main function `evian_linear`. It should contain the SNP ID specified in the `snp` argument as a column name.
`bim`	data frame; read from from the argument `bim` in the main function `evian_linear`. Provides allele information (base pair, effect/reference alleles) for the SNP of interest.
`lolim`	numeric; the lower limit for the grid or the minimum value of the regression parameter β used to calculate the likelihood function.
`hilim`	numeric; the upper limit for the grid or the maximum value of the regression parameter β used to calculate the likelihood funciton.
`m`	numeric; the density of the grid at which to compute the standardized likelihood function. A beta grid is defined as the grid of values for the SNP parameter used to evaluate the likelihood function.
`bse`	numeric; the number of beta standard errors to utilize in constraining the beta grid limits. Beta grid is evaluated at β +/- `bse`*s.e.
`k`	numeric or numeric vector; The strength of evidence criterion k. Reads from the input of `kcutoff` from the main `evian_linear` or `evian_logit` function
`robust`	logical; if `TRUE`, then a robust adjustment is applied to the likelihood function to account for the cluster nature in the data. See robust_forCluster.
`family`	the link function for `glm`. Currently only linear ('gaussian') and logistic ('binomial') are supported. This is currently auto-filled by the main `evian_linear` function.

Details

`calculateLinearMLE` calculates the profile likelihood for a single SNP. A proper grid range is first established for β then the standardized profile likelihood is evaluated at each of the `m` cuts uniformly spread across the grid. Based on the standardized profile likelihood, the MLE for β is computed as well as the likelihood intervals for each value of `k` provided.
For different genetic models, their coding schemes are shown as below:

```
  Additive
AA  0
AB  1
BB  2

  Dominant
AA  0
AB  1
BB  1

  Recessive
AA  0
AB  0
BB  1

Overdominance model
       A  D
   AA  0  0
   AB  1  1
   BB  2  0
```

Specifically for the overdominance model, the column of interest is the D column.

Value

This function outputs a list containg 4 elements that can be directly accessed using '`$`' operator.

`theta` numeric vector; the `m` β values used to estimate the standardized profile likelihood.

`profile.lik.norm` numeric vector; the corresponding `m` standardized profile likelihood value at each of the β values in `theta`. If `robust=TRUE`, then the values will be adjusted by the robust factor.

`k_cutoff` numeric vector; It specifies which k-cutoff had been used in the calculation, ordered from the smallest k to the largest k.

`SummaryStats` data frame; contains the summary statistics of the profile likelihood calculation. It contains the following columns:

- `mle`: the estimates for SNP effect (odds ratio for logistic, and β for linear) with respect to the effective allele
- `maxlr`: maximum likelihood ratio in the beta grid defined by `lolim` and `hilim`

- AF: allele frequency for the effective allele
- SNP: SNP ID
- bp: base pair position from the bim input
- effect, ref: the effective allele and the other allele from the bim input
- robustFactor: robust factor calculated, set to 1 if robust=FALSE.
- lo_1, hi_1, lo_2, hi_2...: the lower and upper bound of the likelihood intervals for the kth cut-off in k_cutoff.

Note

When lolim or hilim are NOT defined, then the boundaries of the beta grid will be determined by the default bse=5, or a bse defined by the user. Otherwise, the user can define the exact beta grid boundaries using lolim and hilim.

In some cases the beta grid (using bse or lolim,hilim) may need to be increased substantially (bse as large as 15) if covariates are present in the formula. This is automatically dealt by the current function, but contributes to longer computation time to find the appropriate ranges. Estimation may become inaccurate with large number of correlated covariates, which is a known limitation of profile likelihoods.

Author(s)

Dr. Lisa J Strug <lisa.strug@utoronto.ca>

References

Strug, L. J., Hodge, S. E., Chiang, T., Pal, D. K., Corey, P. N., & Rohde, C. (2010). A pure likelihood approach to the analysis of genetic association data: an alternative to Bayesian and frequentist analysis. Eur J Hum Genet, 18(8), 933-941. doi:10.1038/ejhg.2010.47

Strug, L. J., & Hodge, S. E. (2006). An alternative foundation for the planning and evaluation of linkage analysis. I. Decoupling "error probabilities" from "measures of evidence". Hum Hered, 61(3), 166-188. doi:10.1159/000094709

Royall, R. (1997). Statistical Evidence: A Likelihood Paradigm. London, Chapman and Hall.

---

calculateLogitMLE *Profile likelihood calculation for binary outcome data using logistic regression models*

Description

This is the function that calculates profileLikelihood for a single SNP. The main function evian_logit calls this function repeatedly to obtain results for multiple SNPs.

Usage

```
calculateLogitMLE(snp,formula_tofit,model,data,bim,lolim,hilim,m,bse,k,robust,family)
```

Arguments

snp	a string specifyingthe SNP to be calculated.
formula_tofit	a formula object of the genetic model. The model should be formatted as y~nuisance parameters. The parameter of interest should not be included here.
model	a string specifying the mode of inheritance parameterization: additive, dominant, recessive, or overdominance. See calculateLinearMLE for details.
data	data frame; read from the argument data in the main function evian_logit. It should contain the SNP ID specified in the snp argument as a column name.
bim	data frame; read from from the argument bim in the main function evian_logit. Provides allele information (base pair, effect/reference alleles) for the SNP of interest.
lolim	numeric; the lower limit for the grid or the minimum value of the regression parameter β used to calculate the likelihood function.
hilim	numeric; the upper limit for the grid or the maximum value of the regression parameter β used to calculate the likelihood funciton.
m	numeric; the density of the grid at which to compute the standardized likelihood function. A beta grid is defined as the grid of values for the SNP parameter used to evaluate the likelihood function.
bse	numeric; the number of beta standard errors to utilize in constraining the beta grid limits. Beta grid is evaluated at β +/- bse*s.e.
k	numeric or numeric vector; The strength of evidence criterion k. Reads from the input of kcutoff from the main function
robust	logical; if TRUE, then a robust adjustment is applied to the likelihood function to account for the cluster nature in the data. See robust_forCluster.
family	the link function for glm. Currently only linear ('gaussian') and logistic ('binomial') are supported. This is currently auto-filled by the main evian_logit function.

Details

calculateLogitMLE calculates the profile likelihood for a single SNP. A proper grid range is first established for β then the standardized profile likelihood is evaluated at each of the m cuts uniformly spread across the grid. The choice of grid range can be either specified through lolim and hilim arguments (default) or can be specified through providing a numerical bse argument. This will enable the grid search feature as in evian_linear. After the standardized profile likelihood is evalualted, the MLE for β is computed as well as the likelihood intervals for each value of k provided.

Note when the model is dominant or recessive, the odds ratio cannot simply be inverted to obtain its reciprocal model, otherwise the heteozygous genotype group will be misclassified. Instead the genotype coding is swapped so that it codes with respect to the risk allele prior to running the analysis so odds ratio are properly calculated.

Value

This function outputs a list containg 4 elements that can be directly accessed using $ operator.

`theta` numeric vector; the `m` odds ratios used to estimate the standardized profile likelihood.

`profile.lik.norm` numeric vector; the corresponding `m` standardized profile likelihood value at each of the odds ratios in `theta`. If `robust=TRUE`, then the values will be adjusted by the robust factor.

`k_cutoff` numeric vector; It specifies which k-cutoff had been used in the calculation, ordered from the smallest k to the largest k.

`SummaryStats` data frame; contains contains the summary statistics of the profile likelihood calculation:

- `mle`: the estimates for SNP effect (odds Ratio for logistic, and β for linear) with respect to the effective allele
- `maxlr`: maximum likelihood ratio in the beta grid defined by `lolim` and `hilim`
- `AF`: allele frequency for the effective allele
- `SNP`: SNP ID
- `bp`: base pair position from the `bim` input
- `effect`, `ref`: the effective allele and the other allele from the `bim` input
- `robustFactor`: robust factor calculated, set to 1 if `robust=FALSE`.
- `lo_1`, `hi_1`, `lo_2`, `hi_2`...: the lower and upper bound of the likelihood intervals for the kth cut-off in `k_cutoff`.

Note

As mentioned in the linear case, finding a proper beta grid range can be time-consuming. For logistic regression specifically, we assume that most of the odds ratios are between 0.025 to 4, and we assign those values as default for the lower and upper bound. If the grid search feature used in the linear case is preferred, specify a value for `bse` and leave `lolim` and `hilim` unchanged.

Author(s)

Dr. Lisa J Strug <lisa.strug@utoronto.ca>

References

Strug, L. J., Hodge, S. E., Chiang, T., Pal, D. K., Corey, P. N., & Rohde, C. (2010). A pure likelihood approach to the analysis of genetic association data: an alternative to Bayesian and frequentist analysis. Eur J Hum Genet, 18(8), 933-941. doi:10.1038/ejhg.2010.47

Strug, L. J., & Hodge, S. E. (2006). An alternative foundation for the planning and evaluation of linkage analysis. I. Decoupling "error probabilities" from "measures of evidence". Hum Hered, 61(3), 166-188. doi:10.1159/000094709

Royall, R. (1997). Statistical Evidence: A Likelihood Paradigm. London, Chapman and Hall.

densityPlot *Plot profile likelihood density for a single SNP.*

Description

This function plots the density distribution for a single SNP calculated from the evian_linear or evian_logit functions.

Usage

```
densityPlot(dList, snpName, kcut = NULL, pl = 'linear', xlim = NULL,
color = c('red','orange','green','blue'), round = 2)
```

Arguments

dList	a row-combined list, output from evian_linear or evian_logit.
snpName	a string specifying the SNP to be plotted.
kcut	numeric; the cut-off to be plotted. If kcut=NULL, all values of k in the kcutoff will be plotted.
pl	a string specifying the y-axis for the plot. The y-axis will be plotted as 'Odds Ratio' if pl is specified as logit, 'Beta' otherwise.
xlim	graphical parameter used in plot function
color	color of the likelihood interval lines from smallest to largest. For instance, c('red','green') for LIs of k=c(8,32) means that the 1/8 interval will be plotted as red, and 1/32 will be plotted as green.
round	numeric; number of digits displayed on the plot.

Details

This function takes output from evian_linear or evian_logit as input. It will plot the density of the estimated standardized profile likelihood for the SNP of interest. Some basic summary statistics will be included on the plot too.

Author(s)

Dr. Lisa J Strug <lisa.strug@utoronto.ca>

References

Strug, L. J., Hodge, S. E., Chiang, T., Pal, D. K., Corey, P. N., & Rohde, C. (2010). A pure likelihood approach to the analysis of genetic association data: an alternative to Bayesian and frequentist analysis. Eur J Hum Genet, 18(8), 933-941. doi:10.1038/ejhg.2010.47

Royall, R. (1997). Statistical Evidence: A Likelihood Paradigm. London, Chapman and Hall.

Examples

```
data(eviandata_linear)
data(evianmap_linear)


rst1=evian_linear(data=eviandata_linear, bim=evianmap_linear, xcols=10:ncol(eviandata_linear),
ycol=6, covariateCol=c(5,7:9), robust=FALSE, model="additive", m=1000,
kcutoff = c(32,100), multiThread=1)

# Plot the density for rs912
densityPlot(dList=rst1,snpName='rs912')
```

---

eviandata_binary *Example dataset with a binary outcome.*

Description

This dataset included the genotypic and phenotypic information of 250 individuals in proper formats. This dataset is used together with evianmap_binary to illustrate how to use the main function evian_logit.

Usage

```
data(eviandata_binary)
```

Details

This is an example dataset for evian_logit function. It contained 250 individuals, and for each of the individuals, their genotype at 30 SNPs and a binary outcome (PHENOTYPE; coded as 0/1) were stored. Three additional covariates (age, weight, city) were provided as well. Specifically, our function can incorporate with related individuals, some of these individuals in the dataset are correlated with others, and are specified through the FID column.

---

eviandata_linear *Example dataset with a quantative outcome.*

Description

This dataset included the genotypic and phenotypic information of 1444 individuals in proper formats. This dataset is used together with evianmap_linear to illustrate how to use the main function evian_linear.

Usage

```
data(eviandata_linear)
```

Details

This is an example dataset for evian_linear function. It contained 1444 individuals, and for each of the individuals, their genotype at 10 SNPs and a continuous outcome (Y_norma) were stored. Three additional covariates (Fev, BMI_group, Age_group) were provided as well. Specifically, our function can incorporate with related individuals, some of these individuals in the dataset are correlated with others, and are specified through the FID column.

---

evianmap_binary *Example map data for eviandata_binary.*

Description

This dataset stored the corresponding SNP information from eviandata_binary.

Usage

```
data(evianmap_linear)
```

Details

This is the corresponding map file for eviandata_binary. Specifically it stores the chromosome, base pair, and two alleles for the 30 SNPs listed in eviandata_binary in the same order.

---

evianmap_linear *Example map data for eviandata_linear.*

Description

This dataset stored the corresponding SNP information from eviandata_linear.

Usage

```
data(evianmap_linear)
```

Details

This is the corresponding map file for eviandata_linear. Specifically it stores the chromosome, base pair, and two alleles for the 10 SNPs listed in eviandata_linear in the same order.

`evian_linear` *Evidential analysis for quantitative outcome data using linear regression models*

Description

Calculates the likelihood intervals for genetic association of a quantitative trait in a genomic region of interest. Covariates can be accommodated.

Usage

```
evian_linear(data, bim, xcols = NULL, ycol = NULL, covariateCol = NULL, formula = NULL,
  robust = FALSE, model = 'additive', m = 200, bse = 5, lolim = NULL, hilim = NULL,
   kcutoff = c(8,32,100,1000), multiThread = 1)
```

Arguments

`data`	a data frame includes a column for the response variable, one or multiple columns of genotype data (coded as `0, 1, 2,` or `NA`), and optionally columns for covariates. Headers are assumed. If the data is from related individuals, an additional column named `'FID'` needs to be included to specify the related structure. Using the PLINK toolkit with option `--recodeA` can produce the file in the required format and is recommended.
`bim`	a data frame with six columns representing chromosome, SNP ID, physical distance, base pair position, effective allele, and reference allele. i.e. data from a file in PLINK binary format (bim). No header is assumed, but the ordering of the columns must follow the standard bim file format.
`ycol`	numeric; column index in the `data` data frame for the column representing the response variable.
`xcols`	numeric vector; the column range in the `data` where genotype information is stored. Note that although a range of X is required, only one SNP at a time is calculated.
`covariateCol`	numeric or numeric vector; optional argument specifying which columns represent covariates. If left as `NULL`, no covariates will be included and the model `Y~snp` will be used.
`formula`	string; this is an alternative way of specifying model rather than using `xcols` and `ycol` arguments. This model follows the same format as the `glm` function (e.g. `Y~snp1+age+sex`). Note that in the case where multiple SNPs are included, only one SNP will be calculated at a time (e.g. given `Y~snp1+snp2`, the function will estimate `Y~snp1` and `Y~snp2` separately). The function can automatically identify snps with rsID as proper Xs, and would treat all other predictors as covariates.
`robust`	logical; default `FALSE`. If `TRUE`, then a robust adjustment is applied to the likelihood function to account for clustering in the data; See robust_forCluster.

model	a string that specifies the mode of inheritance parameterization: additive, dominant, recessive, or overdominance. Default additive.
m	numeric; the density of the grid at which to compute the standardized likelihood function. A beta grid is defined as the grid of values for the SNP parameter used to evaluate the likelihood function.
bse	numeric; the number of beta standard errors to utilize in constraining the beta grid limits. Beta grid is evaluated at β +/- bse*s.e.
lolim	numeric; the lower limit for the grid or the minimum value of the regression parameter β used to calculate the likelihood function.
hilim	numeric; the upper limit for the grid or the maximum value of the regression parameter β used to calculate the likelihood funciton.
kcutoff	numeric or numeric vector; default = c(8,32,100,1000). The strength of evidence criterion k. The function will calculate the 1/k standardized likelihood intervals for each value provided.
multiThread	numeric; number of threads to use for parallel computing.

Details

evian_linear is the main function called to calculate the 1/k likelihood intervals for the additive, dominant, recessive, or overdominance genotypic models. This function calls calculateLinearMLE in parallel to calculate the likelihood for each SNP. The calculation details can be found in calculateLinearMLE.

The input for the data and bim arguments can be obtained from the PLINK files; data is expected to follow PLINK format when run with the --recodeA option and bim can be obtained directly from a PLINK binary format file. Note if covariates are to be included, it is expected that the covariates are appended to the data file with a header for each covariate.

The statistical model can be specified in two ways. Column index can be provided through the xcols, ycol, and covariateCol arguments or through the formula argument, which can accept a formula specified as the formula argument in the R glm function. We recommend using xcols, ycol, and covariateCol arguments in most scenarios as this is relatively easier to input and it works for all the cases that we have considered so far. The alternative formula argument is not able to detect non-rsID variants as parameters of interests, and is only suggested in the scenario where only a few variants are of interest and all of their rsID are known in advance. In this case, using this argument can save the time to search for the column index values corresponding to those SNPs.

Parallel computing is avaliable through the use of the multiThread argument. This parallelization uses the foreach and doMC packages and will typically reduce computation time significantly. Due to this dependency, parallelization is not available on Windows OS as foreach and doMC are not supported on Windows.

Value

This function outputs the row-combined the results from calculateLinearMLE for each of the SNPs included in the data/bim files. The exact output for each SNP can be found in the calculateLinearMLE documentation.

Note

When lolim/hilim is NOT defined, then the boundaries of the beta grid will be determined by the default bse=5, or a bse defined by the user. Otherwise, the user can define the exact beta grid boundaries using lolim/hilim.

In some cases the beta grid (using bse or lolim/hilim) may need to be increased substantially (bse as large as 15) if covariates are present in the formula. This is automatically dealt by the current function, but contribute to longer computation time to find the appropriate ranges. Estimation may become inaccurate with large number of correlated covariates, which is a known limitation of profile likelihoods.

See Also

calculateLinearMLE

Examples

```
data(eviandata_linear)
data(evianmap_linear)


rst1=evian_linear(data=eviandata_linear, bim=evianmap_linear, xcols=10:ncol(eviandata_linear),
ycol=6, covariateCol=c(5,7:9), robust=FALSE, model="additive",
m=1000, kcutoff = c(32,100), multiThread=1)

#Alternatively you can use the formula argument to run the same model as above

rst2=evian_linear(data=eviandata_linear, bim=evianmap_linear,
formula='Y_norma~Fev+SEX+Age_group+BMI_group+rs141+rs912+rs573+rs414+rs635+
rs356+rs877+rs168+rs449+rs580', robust=FALSE,
model="additive", m=1000, kcutoff = c(32,100), multiThread=1)
```

---

evian_logit *Evidential analysis for binary outcome data using logistic regression models*

Description

Calculates the likelihood intervals for genetic association of a binary trait in a genomic region of interest. Covariates can be accommodated.

Usage

```
evian_logit(data, bim, xcols = NULL, ycol = NULL, covariateCol = NULL, formula = NULL,
  robust = FALSE, model = "additive", m = 200, bse = NULL, lolim = log(0.025),
  hilim = log(4), kcutoff = c(8,32,100,1000), multiThread = 1)
```

Arguments

`data`	a data frame includes a column for the response variable, one or multiple columns of genotype data (coded as `0, 1, 2,` or `NA`), and optionally columns for covariates. Headers are assumed. If the data is from related individuals, an additional column named `'FID'` needs to be included to specify the related structure. Using the PLINK toolkit with option `--recodeA` can produce the file in the required format and is recommended. Note specifically for binary outcome, we assumed that the case/control status were coded as `0/1` rather than `1/2` as how PLINK codes them.
`bim`	data frame; six columns representing chromosome, SNP ID, physical distance, base pair position, effective allele, and reference allele. i.e. data from a file in PLINK binary format (bim). No header is assumed, but the order must follow the standard bim file format.
`ycol`	numeric; column index in the `data` data frame for the column representing the response variable.
`xcols`	numeric vector; the column range in the `data` where genotype information is stored. Note that although a range of X is required, only one SNP at a time is calculated.
`covariateCol`	numeric or numeric vector; optional argument specifying which columns represent covariates. If left as `NULL`, no covariates will be included and the model `Y~snp` will be used.
`formula`	string; this is an alternative way of specifying model rather than using `xcols` and `ycol` arguments. This model follows the same formate as the `glm` function (e.g. `Y~snp1+age+sex`). Note that in the case where multiple SNPs are included, only one SNP will be calculated at a time (e.g. given `Y~snp1+snp2`, the function will estimate `Y~snp1` and `Y~snp2` separately). The function can automatically identify snps with rsID as proper Xs, and would treat all other predictors as covariates.
`robust`	logical; default `FALSE`. If `TRUE`, then a robust adjustment is applied to the likelihood function to account for clustering in the data; See robust_forCluster.
`model`	string; specifies the mode of inheritance parameterization: `additive, dominant, recessive,` or `overdominance`. Default `additive`.
`m`	numeric; the density of the grid at which to compute the standardized likelihood function. A beta grid is defined as the grid of values for the SNP parameter used to evaluate the likelihood function.
`bse`	numeric; the number of beta standard errors to utilize in constraining the beta grid limits. Beta grid is evaluated at β +/- `bse`*s.e. The default for this option is NULL, meaning to use `lolim` and `hilim` values specified. See details `calculateLogitMLE`.
`lolim`	numeric; the lower limit for the grid or the minimum value of the regression parameter β used to calculate the likelihood function.
`hilim`	numeric; the upper limit for the grid or the maximum value of the regression parameter β used to calculate the likelihood funciton.

kcutoff	numeric or numeric vector; default = c(8,32,100,1000). The strength of evidence criterion k. The function will calculate the 1/k standardized likelihood intervals for each value provided.
multiThread	numeric; number of threads to use for parallel computing.

Details

evian_logistic is the main function called to calculate the 1/k likelihood intervals for the additive, dominant, recessive, or overdominance genotypic models when a binary phenotype is presented. This function calls calculateLogitMLE in parallel to calculate the likelihood for each SNP. The calculation details can be found in calculateLogitMLE.

The inputs for the data and bim arguments can be obtained from the PLINK files; data is expected to follow PLINK format when run with the --recodeA option and bim can be obtained directly from a PLINK binary format file. Note if covariates are to be included, it is expected that the covariates are appended to the data file with a header for each covariate.

The statistical model can be specified in two ways. Column index can be provided through the xcols, ycol, and covariateCol arguments or through the formula argument, which can accept a formula specified as the formula argument in the R glm function. We recommend using xcols, ycol, and covariateCol arguments in most scenarios as this is relatively easier to input and it works for all the cases that we have considered so far. The alternative formula argument is not able to detect non-rsID variants as parameters of interests, and is only suggested in the scenario where only a few variants are of interest and all of their rsID are known in advance. In this case, using this argument can save the time to search for the column index values corresponding to those SNPs.

Parallel computing is avaliable through the use of the multiThread argument. This parallelization uses the foreach and doMC packages and will typically reduce computation time significantly. Due to this dependency, parallelization is not available on Windows OS as foreach and doMC are not supported on Windows.

Value

This function outputs the row-combined the results from calculateLogitMLE for each of the SNPs included in the data/bim files. The exact output for each SNP can be found in the calculateLogitMLE documentation.

Note

In some cases the beta grid (using bse or lolim/hilim) may need to be increased substantially (bse as large as 15) if covariates are present in the formula. This is automatically dealt by the current function, but contribute to longer computation time to find the appropriate ranges. Estimation may become inaccurate with large number of correlated covariates, which is a known limitation of profile likelihoods.

See Also

calculateLogitMLE

Examples

```
data(eviandata_binary)
data(evianmap_binary)


rst1=evian_logit(data=eviandata_binary, bim=evianmap_binary, xcols=10:19, ycol=6, robust=FALSE,
model="additive", m=1000, kcutoff = c(32,100), multiThread=1)

#Alternatively you can use the formula argument to run the same model as above

rst2=evian_logit(data=eviandata_binary, bim=evianmap_binary, formula='PHENOTYPE~rs461+rs462+rs477+
rs479+rs491+rs492+rs504+rs509+rs519+rs542', robust=FALSE, model="additive",
m=1000, kcutoff = c(32,100), multiThread=1)
```

---

expandBound *A recursive function that expands the grid search for MLE.*

Description

This is an internal function that finds the proper boundary of the grid.

Usage

```
expandBound(data,bse,parameters,formula,m,k,family)
```

Arguments

data	a data frame inputted from the main function.
bse	numeric. The number of beta standard errors to utilize in constraining the beta grid limits. Passed down from argument bse in the main evian_linear or evian_logit function.
parameters	a numeric vector of length 3 providing the starting values for the search. This is obtained from the getGridBound function. The three numeric values in the vector should represent the beta estimates, s.e., and the correction factor respectively. Details can be found in getGridBound.
formula	a formula specifying the response and possible covariates to keep in the output dataframe. This is directly obtained from evian_linear or the evian_logit function.
k	numeric vector. The strength of evidence criterion k. Passed down from argument kcutoff in the main evian_linear or evian_logit function.
m	numeric. The density of the grid at which to compute the standardized likelihood function. Passed down from argument m in the main evian_linear or evian_logit function.
family	a string representing the link function for ProfileLikelihood::ProfileLikelihood.glm. Currently only supports linear ('gaussian') and logistic ('binary').

Details

Even though the initial grid bound calculated from `getGridBound` works for most of the data, there can be cases where `bse` needs to be increased in order to observe all the Likelihood Intervals (LIs) specified from the main function in the range `kcutoff` calculated. In this case, our approach is to check whether the current grid range includes the largest LIs. The function will expand the grid range by increasing `bse` by 1 if it is not included. This step will be running recursively until the largest LIs are included in the current grid range.

Value

This function returns a numeric vector of length two representing the optimal lower and upper bounds for the grid on which the later functions will search for MLE.

`getGridBound` *Obtain the range of the grid where MLE will be searched at*

Description

This is an internal function that provides the range where the profileLikelihood function would search for MLE.

Usage

```
getGridBound(formula,data,bse,k,m,family,robust)
```

Arguments

`formula`	a formula specifying the response and possible covariates to keep in the output dataframe. This is directly obtained from `evian_linear` or the `evian_logit` function.
`data`	a data frame inputted from the output of `subsetData`.
`bse`	numeric. The number of beta standard errors to utilize in constraining the beta grid limits. Passed down from argument `bse` in the main `evian_linear` or `evian_logit` function.
`k`	numeric vector. The strength of evidence criterion k. Passed down from argument `kcutoff` in the main `evian_linear` or `evian_logit` function.
`m`	numeric. The density of the grid at which to compute the standardized likelihood function. Passed down from argument `m` in the main `evian_linear` or `evian_logit` function.
`family`	a string representing the link function for `ProfileLikelihood::ProfileLikelihood.glm`. Currently only supports linear ('gaussian') and logistic ('binary').
`robust`	A numeric value, robust correction factor.

Details

getGridBound is an interior function that searches for the proper grid range that would be used to search for MLE. This is done through two steps: First, it finds a starting grid range by fitting a (generalized) linear model to obtain the estimate and s.e. of the beta. Then the starting grid range can be defined as mean +- bse*s.e. . In the case where robust correction is needed, the grid will be defined as mean +- bse*s.e./correction factor. Then the function determines an optimal grid range by using expandBound function.

Value

This function returns a numeric vector of length 2 that represents the lower and upper bounds of the grid for the MLE search.

---

multiLine_plot *Plot methods for multiple likelihood intervals in a genomic region.*

Description

This function plots the likelihood intervals (LIs) for all SNPs calculated using either evian_linear or evian_logit.

Usage

```
multiLine_plot(bpstart = 0, bpend = 1000000000, dList, title = "My model",
showmaxlr = 3, kcut = NULL, pl = 'linear', ylim = c(-0.5,10),
color = c('violet','green','red','blue'), markSNP = NULL, round = 2)
```

Arguments

bpstart,bpend	numeric; indicating the range of base pairs to be plotted. From bpstart to bpend.
dList	a row-combined list, output from evian_linear or evian_logit.
title	string; title of plot
showmaxlr	numeric; number of top SNPs to display on the graph. Default = 3. SNPs are chosen by their maximum likelihood ratio values.
kcut	numeric; the cut-off to be plotted. If kcut=NULL, all intervals will be plotted.
pl	a string specifying the y-axis for the plot. The y-axis will be plotted as 'Odds Ratio' if pl is specified as logit, 'Beta' otherwise.
markSNP	vector of strings; indicates which SNPs to be marked on the plot. By default it will mark all SNPs that are significant at the smallest cut-off.
round	numeric; number of digits displayed on the plot.
ylim	graphical parameter used in plot function
color	color of the likelihood interval lines from smallest to largest. For instance, c('red','green') for LIs of k=c(8,32) means that the 1/8 interval will be plotted as red, and 1/32 will be plotted as green.

Details

This function takes output from `evian_linear` or `evian_logit` as input. It will plot the likelihood intervals for each of the SNPs analyzed. If 1/k interval is significant then it will be colored by the specified color and will remain grey if the interval is not significant.

Author(s)

Dr. Lisa J Strug <lisa.strug@utoronto.ca>

References

Strug, L. J., Hodge, S. E., Chiang, T., Pal, D. K., Corey, P. N., & Rohde, C. (2010). A pure likelihood approach to the analysis of genetic association data: an alternative to Bayesian and frequentist analysis. Eur J Hum Genet, 18(8), 933-941. doi:10.1038/ejhg.2010.47

Examples

```
data(eviandata_linear)
data(evianmap_linear)


rst1=evian_linear(data=eviandata_linear, bim=evianmap_linear, xcols=10:ncol(eviandata_linear),
ycol=6, covariateCol=c(5,7:9), robust=FALSE, model="additive",
m=1000, kcutoff = c(32,100), multiThread=1)


# Plot the LIs for all 10 SNPs
multiLine_plot(dList=rst1)
```

---

robust_forCluster *Robust adjustment function*

Description

The robust function computes an adjustment that is applied to the likelihood function to account for the cluster nature of the data.

Usage

```
robust_forCluster(formula, data, family)
```

Arguments

formula	a formula specifying the response and possible covariates to keep in the output data frame. This is directly obtained from the evian_linear or evian_logistic function.
data	data frame; from the output of subsetData.
family	string; the link function for glm. Currently this only supported linear and logistic.

Details

The robust function is called from within evian_logit or evian_linear functions. It computes a robust adjustment factor that is applied to the likelihood function to account for the cluster nature of the data. The family ID column (FID) specifies the clusters. The robust adjustment factor is the ratio of the regular variance estimator of the maximum likelihood estimate (MLE) to the sandwich variance estimator of the MLE, where the 'meat' of the sandwich variance estimator is corrected for clustering in the data (Blume et.al, 2007). If the data is not clustered (i.e. the observations are independent) then the adjustment factor can still be applied to make the working model robust to possible model misspecifications (Royall and Tsou, 2003).

Value

A numeric constant.

Author(s)

Zeynep Baskurt <zeynep.baskurt@sickkids.ca>

References

Blume, J. D., Su, L., Remigio, M. O., & McGarvey, S. T. (2007). Statistical evidence for GLM regression parameters: A robust likelihood approach. Statistics in Medicine, 26, 2919-2936.

Royall, R. , Tsou, T. S. (2003). Interpreting statistical evidence by using imperfect models: robust adjusted likelihood functions. J Roy Stat Soc B; 65: 391-404.

---

subsetData *interior subsetting function*

Description

This is an internal function that subsets the SNP column with matching name and removes rows with missing observations.

Usage

```
subsetData(snp,formula_tofit,data)
```

Arguments

snp	a string specifying the SNP of interests. The SNP ID must exist in data.
formula_tofit	a formula object of the genetic model. This is directly obtained from evian_linear or evian_logit function.
data	a data frame inputted from the main function. Should contain the SNP ID snp as one of the column names.

Details

subsetData is an interior function that subsets the full dataset into a smaller set containing only one specific SNP by the snp option. It will then remove any rows with missing values.

Value

This function returned a dataframe containing phenotype, covariates in their original column names as in the full dataset, and a column called X representing the genotype information for the SNP chosen. The column names are essetial.

Index