Package 'SPCompute'

September 5, 2022

Type Package

Title Compute Power or Sample Size for GWAS with Covariate Effect

Version 1.0.2

Author Ziang Zhang, Lei Sun

Maintainer Ziang Zhang <aguero.zhang@mail.utoronto.ca>

Description Fast computation of the required sample size or the achieved power, for GWAS studies with different types of covariate effects and different types of covariate-gene dependency structure. For the detailed description of the methodology, see Zhang (2022) ``Power and Sample Size Computation for Genetic Association Studies of Binary Traits: Accounting for Covariate Effects" <arXiv:2203.15641>.

License GPL (>= 3)

Imports Matrix, stats

Suggests knitr, rmarkdown, testthat

VignetteBuilder knitr

Encoding UTF-8

RoxygenNote 7.1.1

NeedsCompilation no

Repository CRAN

Date/Publication 2022-09-05 21:00:02 UTC

R topics documented:

check_parameters *Check if the parameter list contains all the parameters required for the computation.*

Description

Check if the parameter list contains all the parameters required for the computation.

Usage

```
check_parameters(parameters, response, covariate)
```

Arguments

parameters	A list of parameters that contains all the required parameters in the model. If response is "binary", this list needs to contain "preva" which denotes the prevalence of the disease (or case to control ratio for case-control sampling). If response is continuous, the list needs to contain "TraitSD" and "TraitMean" which represent the standard deviation and mean of the continuous trait. #' If covariate is not "none", a parameter "gammaG" needs to be defined to capture the dependence between the SNP and the covariate (through linear regression model if covariate is continuous, and logistic model if covariate is binary). If covariate is "binary", list needs to contains "pE" that defines the frequency of the covariate. If it is continuous, list needs to contain "muE" and "sigmaE" to define #' its mean and standard deviation. The MAF is defined as "pG", with HWE assumed to hold.
response	A string of either "binary" or "continuous", indicating the type of response/trait variable in the model.
covariate	A string of either "binary", "continuous" or "none" indicating the type of covariate E in the model.

Value

TRUE or FALSE if all the parameters are correctly defined.

Examples

```
parameters <- list(TraitMean = 0.3, TraitSD = 1, pG = 0.2, betaG = log(1.1),
betaE = log(1.1), muE = 0, sigmaE = 3, gammaG = log(2.1))

SPCompute::check_parameters(parameters, "continuous", "continuous")
```

`Compute_Power` *Compute the Power of an association study, at a given sample size.*

Description

Compute the Power of an association study, at a given sample size.

Usage

```
Compute_Power(
  parameters,
  n,
  response = "binary",
  covariate = "binary",
  mode = "additive",
  alpha = 0.05,
  seed = 123,
  LargePowerApproxi = FALSE,
  searchSizeGamma0 = 100,
  searchSizeBeta0 = 100,
  B = 10000,
  method = "semi-sim"
)
```

Arguments

`parameters`	A list of parameters that contains all the required parameters in the model. If response is "binary", this list needs to contain "prev" which denotes the prevalence of the disease (or case to control ratio for case-control sampling). If response is continuous, the list needs to contain "traitSD" and "traitMean" which represent the standard deviation and mean of the continuous trait. If covariate is not "none", a parameter "gammaG" needs to be defined to capture the dependence between the SNP and the covariate (through linear regression model if covariate is continuous, and logistic model if covariate is binary). If covariate is "binary", list needs to contains "pE" that defines the frequency of the covariate. If it is continuous, list needs to contain "muE" and "sigmaE" to define its mean and standard deviation. The MAF is defined as "pG", with HWE assumed to hold.
`n`	An integer number that indicates the sample size.
`response`	A string of either "binary" or "continuous", indicating the type of response/trait variable in the model, by default is "binary"
`covariate`	A string of either "binary", "continuous" or "none" indicating the type of covariate E in the model, by default is "binary".
`mode`	A string of either "additive", "dominant" or "recessive", indicating the genetic mode, by default is "additive".
`alpha`	A numeric value that denotes the significance level used in the study, by default is 0.05.

`seed` An integer number that indicates the seed used for the simulation to compute the approximate fisher information matrix, by default is 123.

`LargePowerApproxi` TRUE or FALSE indicates whether to use the large power approximation formula.

`searchSizeGamma0` The interval radius for the numerical search of gamma0, by default is 8. Setting to higher values may solve some numerical problems at the cost of longer runtime.

`searchSizeBeta0` The interval radius for the numerical search of beta0, by default is 8. Setting to higher values may solve some numerical problems at the cost of longer runtime.

`B` An integer number that indicates the number of simulated sample to approximate the fisher information matrix, by default is 10000 (Should only be changed if computation uses semi-simulation method).

`method` An character that is either "semi-sim" (default) or "expand" using the idea of representative dataset. This specifies the method being used to compute the power/sample size when the trait is binary using logistic regression. The default method will be faster for large sample size computation.

Value

The power that can be achieved at the given sample size.

Examples

```
parameters <- list(TraitMean = 0.3, TraitSD = 1, pG = 0.2, betaG = log(1.1),
betaE = log(1.1), muE = 0, sigmaE = 3, gammaG = log(2.1))

Compute_Power(parameters, n = 1000, response = "continuous",
covariate = "continuous", method = "semi-sim")
```

Compute_Power_multi *Compute the Power of an association study at a given sample size, accommodating more than one covariates, using the Semi-Simulation method.*

Description

Compute the Power of an association study at a given sample size, accommodating more than one covariates, using the Semi-Simulation method.

Usage

```
Compute_Power_multi(
  parameters,
  n,
  response = "binary",
  covariate,
  mode = "additive",
  alpha = 0.05,
  seed = 123,
  searchSizeBeta0 = 8,
  searchSizeGamma0 = 8,
  LargePowerApproxi = FALSE,
  B = 10000
)
```

Arguments

parameters	A list of parameters that contains all the required parameters in the model. If response is "binary", this list needs to contain "prev" which denotes the prevalence of the disease (or case to control ratio for case-control sampling). If response is continuous, the list needs to contain "traitSD" and "traitMean" which represent the standard deviation and mean of the continuous trait. If covariate is not "none", a parameter "gammaG" needs to be defined to capture the dependence between the SNP and the covariate (through linear regression model if covariate is continuous, and logistic model if covariate is binary). If covariate is "binary", list needs to contains "pE" that defines the frequency of the covariate. If it is continuous, list needs to contain "muE" and "sigmaE" to define its mean and standard deviation. The MAF is defined as "pG", with HWE assumed to hold.
n	An integer number that indicates the sample size.
response	A string of either "binary" or "continuous", indicating the type of response/trait variable in the model, by default is "binary"
covariate	A vector of length two with elements being either c("binary", "continuous"),c("binary", "binary") or c("continuous", "continuous"), indicating the type of covariate E in the model.
mode	A string of either "additive", "dominant" or "recessive", indicating the genetic mode, by default is "additive".
alpha	A numeric value that denotes the significance level used in the study, by default is 0.05.
seed	An integer number that indicates the seed used for the simulation to compute the approximate fisher information matrix, by default is 123.
searchSizeBeta0	The interval radius for the numerical search of beta0, by default is 8. Setting to higher values may solve some numerical problems at the cost of longer runtime.
searchSizeGamma0	The interval radius for the numerical search of gamma0, by default is 8. Setting to higher values may solve some numerical problems at the cost of longer runtime.

LargePowerApproxi
: TRUE or FALSE indicates whether to use the large power approximation formula.

B
: An integer number that indicates the number of simulated sample to approximate the fisher information matrix, by default is 10000 (Should only be changed if computation uses semi-simulation method).

Value

The power that can be achieved at the given sample size.

Examples

```
parameters <- list(TraitMean = 0.3, TraitSD = 1, pG = 0.2, betaG = log(1.1),
betaE = c(log(1.1), log(1.2)),
muE = 0, sigmaE = 3, gammaG = c(log(2.1), log(2.2)), pE = 0.4)
SPCompute::Compute_Power_multi(parameters, n = 1000, response = "continuous",
covariate = c("binary", "continuous"))
```

Compute_Size *Compute the sample size of an association study, to achieve a target power.*

Description

Compute the sample size of an association study, to achieve a target power.

Usage

```
Compute_Size(
  parameters,
  PowerAim,
  response = "binary",
  covariate = "binary",
  mode = "additive",
  alpha = 0.05,
  seed = 123,
  LargePowerApproxi = FALSE,
  searchSizeGamma0 = 100,
  searchSizeBeta0 = 100,
  B = 10000,
  method = "semi-sim",
  lower.lim.n = 1000,
  upper.lim.n = 8e+05
)
```

Arguments

parameters	A list of parameters that contains all the required parameters in the model. If response is "binary", this list needs to contain "prev" which denotes the prevalence of the disease (or case to control ratio for case-control sampling). If response is continuous, the list needs to contain "traitSD" and "traitMean" which represent the standard deviation and mean of the continuous trait. If covariate is not "none", a parameter "gammaG" needs to be defined to capture the dependence between the SNP and the covariate (through linear regression model if covariate is continuous, and logistic model if covariate is binary). If covariate is "binary", list needs to contains "pE" that defines the frequency of the covariate. If it is continuous, list needs to contain "muE" and "sigmaE" to define its mean and standard deviation. The MAF is defined as "pG", with HWE assumed to hold.
PowerAim	An numeric value between 0 and 1 that indicates the aimed power of the study.
response	A string of either "binary" or "continuous", indicating the type of response/trait variable in the model, by default is "binary"
covariate	A string of either "binary", "continuous" or "none" indicating the type of covariate E in the model, by default is "binary".
mode	A string of either "additive", "dominant" or "recessive", indicating the genetic mode, by default is "additive".
alpha	A numeric value that denotes the significance level used in the study, by default is 0.05.
seed	An integer number that indicates the seed used for the simulation to compute the approximate fisher information matrix, by default is 123.
LargePowerApproxi	TRUE or FALSE indicates whether to use the large power approximation formula.
searchSizeGamma0	The interval radius for the numerical search of gamma0, by default is 8. Setting to higher values may solve some numerical problems at the cost of longer runtime.
searchSizeBeta0	The interval radius for the numerical search of beta0, by default is 8. Setting to higher values may solve some numerical problems at the cost of longer runtime.
B	An integer number that indicates the number of simulated sample to approximate the fisher information matrix, by default is 10000 (Should only be changed if computation uses semi-simulation method).
method	An character that is either "semi-sim" (default) or "expand" using the idea of representative dataset. This specifies the method being used to compute the power/sample size when the trait is binary using logistic regression. The default method will be faster for large sample size computation.
lower.lim.n	An integer number that indicates the smallest sample size to be considered, only for "expand" method.
upper.lim.n	An integer number that indicates the largest sample size to be considered.

Value

The required sample size.

Examples

```
parameters <- list(TraitMean = 0.3, TraitSD = 1, pG = 0.2, betaG = log(1.1),
betaE = log(1.1), muE = 0, sigmaE = 3, gammaG = log(2.1))

Compute_Size(parameters, PowerAim = 0.8, response = "continuous",
covariate = "continuous", method = "semi-sim")
```

`Compute_Size_multi` *Compute the sample size of an association study to achieve a target power for multiple E's, using semi-sim.*

Description

Compute the sample size of an association study to achieve a target power for multiple E's, using semi-sim.

Usage

```
Compute_Size_multi(
  parameters,
  PowerAim,
  response = "binary",
  covariate,
  mode = "additive",
  alpha = 0.05,
  seed = 123,
  LargePowerApproxi = FALSE,
  searchSizeGamma0 = 100,
  searchSizeBeta0 = 100,
  B = 10000,
  upper.lim.n = 8e+05
)
```

Arguments

`parameters`	A list of parameters that contains all the required parameters in the model. If response is "binary", this list needs to contain "prev" which denotes the prevalence of the disease (or case to control ratio for case-control sampling). If response is continuous, the list needs to contain "traitSD" and "traitMean" which represent the standard deviation and mean of the continuous trait. If covariate is not "none", a parameter "gammaG" needs to be defined to capture the dependence between the SNP and the covariate (through linear regression model if covariate is continuous, and logistic model if covariate is binary). If covariate is "binary",

list needs to contains "pE" that defines the frequency of the covariate. If it is continuous, list needs to contain "muE" and "sigmaE" to define its mean and standard deviation. The MAF is defined as "pG", with HWE assumed to hold.

`PowerAim` An numeric value between 0 and 1 that indicates the aimed power of the study.

`response` A string of either "binary" or "continuous", indicating the type of response/trait variable in the model, by default is "binary"

`covariate` Same as in Compute_Power_multi.

`mode` A string of either "additive", "dominant" or "recessive", indicating the genetic mode, by default is "additive".

`alpha` A numeric value that denotes the significance level used in the study, by default is 0.05.

`seed` An integer number that indicates the seed used for the simulation to compute the approximate fisher information matrix, by default is 123.

`LargePowerApproxi`
TRUE or FALSE indicates whether to use the large power approximation formula.

`searchSizeGamma0`
The interval radius for the numerical search of gamma0, by default is 8. Setting to higher values may solve some numerical problems at the cost of longer runtime.

`searchSizeBeta0`
The interval radius for the numerical search of beta0, by default is 8. Setting to higher values may solve some numerical problems at the cost of longer runtime.

`B` An integer number that indicates the number of simulated sample to approximate the fisher information matrix, by default is 10000 (Should only be changed if computation uses semi-simulation method).

`upper.lim.n` An integer number that indicates the largest sample size to be considered.

Value

The required sample size.

Examples

```
parameters <- list(TraitMean = 0.3, TraitSD = 1, pG = 0.2,
betaG = log(1.1), betaE = c(log(1.1), log(1.2)),
muE = 0, sigmaE = 3, gammaG = c(log(2.1), log(2.2)), pE = 0.4)
SPCompute::Compute_Size_multi(parameters, PowerAim = 0.8,
response = "continuous", covariate = c("binary", "continuous"))
```

`convert_preva_to_intercept` *Convert the prevalence value to the intercept value beta0.*

Description

Convert the prevalence value to the intercept value beta0.

Usage

```
convert_preva_to_intercept(
  parameters,
  mode = "additive",
  covariate = "binary",
  seed = 123,
  B = 10000,
  searchSizeBeta0 = 8,
  searchSizeGamma0 = 8
)
```

Arguments

`parameters`	A list of parameters that contains all the required parameters in the model. If response is "binary", this list needs to contain "prev" which denotes the prevalence of the disease (or case to control ratio for case-control sampling). If response is continuous, the list needs to contain "traitSD" and "traitMean" which represent the standard deviation and mean of the continuous trait. If covariate is not "none", a parameter "gammaG" needs to be defined to capture the dependence between the SNP and the covariate (through linear regression model if covariate is continuous, and logistic model if covariate is binary). If covariate is "binary", list needs to contains "pE" that defines the frequency of the covariate. If it is continuous, list needs to contain "muE" and "sigmaE" to define its mean and standard deviation. The MAF is defined as "pG", with HWE assumed to hold.
`mode`	A string of either "additive", "dominant" or "recessive", indicating the genetic mode, by default is "additive".
`covariate`	A string of either "binary", "continuous" or "none" indicating the type of covariate E in the model, by default is "binary".
`seed`	An integer number that indicates the seed used for the simulation if needed, by default is 123.
`B`	An integer number that indicates the number of simulated sample to use if needed, by default is 10000.
`searchSizeBeta0`	The interval radius for the numerical search of beta0, by default is 8. Setting to higher values may solve some numerical problems at the cost of longer runtime.

searchSizeGamma0
: The interval radius for the numerical search of gamma0, by default is 8. Setting to higher values may solve some numerical problems at the cost of longer runtime.

Value

The corresponding gamma0, beta0 and residual variance of E (if applicable).

Examples

```
convert_preva_to_intercept(parameters = list(preva = 0.2, betaG = 0.6, betaE = c(0.9),
gammaG = c(0.2), muE = c(0), sigmaE = c(1), pG = 0.3), covariate = "continuous")
```

Index