Package 'iPRSue'

August 13, 2025

Title Individual Polygenic Risk Score Uncertainty Estimation

Version 1.0.0

Description Provides tools for estimating uncertainty in individual polygenic risk scores (PRSs) using both simulation-based and analytical methods. These methods help quantify variability in PRS estimates for both binary and quantitative traits.

URL https://github.com/DoviniJ/iPRSue

License GPL-3

Encoding UTF-8

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.2

Depends R (>= 2.10)

Imports data.table,
bigstatsr,
logistf,
stats

LazyData true

R topics documented:

`BLUE_estimates_BT` *BLUE_estimates_BT function*

Description

Estimates individual-level polygenic risk scores (PRS) with uncertainty using a frequentist approach for binary traits. This implementation applies Firth's bias-reduced logistic regression on the discovery sample, computes the coefficient covariance matrix, and uses the delta method to derive PRS variance and confidence intervals.

Usage

```
BLUE_estimates_BT(
  discovery_pheno,
  discovery_geno_mat,
  target_pheno,
  target_geno_mat,
  significance_level = 0.05,
  max_iterations = 100
)
```

Arguments

`discovery_pheno`	Character. Path to the phenotype file for the discovery dataset. Assumes no header and that the binary trait is in the third column.
`discovery_geno_mat`	Character. Path to the genotype matrix file for the discovery dataset. Assumes no header.
`target_pheno`	Character. Path to the phenotype file for the target dataset. Assumes no header and individual IDs in the second column.
`target_geno_mat`	Character. Path to the genotype matrix file for the target dataset. Assumes no header.
`significance_level`	Numeric. Significance level for confidence intervals (e.g., 0.05 for 95% CI). Default is 0.05.
`max_iterations`	Integer. Maximum number of iterations allowed in Firth logistic regression. Default is 100.

Details

The function fits a Firth logistic regression model using the `logistf` package to reduce small-sample bias in the discovery set. It extracts SNP effect estimates and their covariance matrix, and propagates this uncertainty through to the individual-level PRS in the target dataset via the delta method. Confidence intervals are derived assuming normality.

Missing or non-estimable coefficients and variances are set to zero.

Value

A data frame with the following columns:

IID Individual identifier (from the target phenotype file).

PRS Estimated polygenic risk score for each individual.

Variance Estimated variance of the PRS.

Lower_Limit Lower bound of the confidence interval.

Upper_Limit Upper bound of the confidence interval.

Examples

```
## Not run:
results <- BLUE_estimates_BT(discovery_pheno = "Bpd.txt",
                                    discovery_geno_mat = "Gd.txt",
                                    target_pheno = "Bpt.txt",
                                    target_geno_mat = "Gt.txt",
                                    significance_level = 0.05,
                                    max_iterations = 100)
head(results)

## End(Not run)
```

`BLUE_estimates_QT` *BLUE_estimates_QT function*

Description

Estimates individual-level polygenic risk scores (PRS) with uncertainty using a frequentist approach for quantitative traits. This implementation fits a multiple linear regression model in the discovery dataset, computes the coefficient covariance matrix, and applies the delta method to propagate uncertainty to the target dataset.

Usage

```
BLUE_estimates_QT(
  discovery_pheno,
  discovery_geno_mat,
  target_pheno,
  target_geno_mat,
  significance_level = 0.05
)
```

Arguments

`discovery_pheno`
: Character. Path to the phenotype file for the discovery dataset. Assumes no header and that the quantitative trait is in the third column.

`discovery_geno_mat`
: Character. Path to the genotype matrix file for the discovery dataset. Assumes no header.

`target_pheno`
: Character. Path to the phenotype file for the target dataset. Assumes no header and individual IDs in the second column.

`target_geno_mat`
: Character. Path to the genotype matrix file for the target dataset. Assumes no header.

`significance_level`
: Numeric. Significance level for confidence intervals (e.g., 0.05 for 95% CI). Default is 0.05.

Details

The function fits a multiple linear regression model (`lm`) using the discovery data. The estimated SNP effects and their covariance matrix are used to compute PRS and associated uncertainty for each individual in the target dataset. Confidence intervals are constructed using the normal approximation.

Missing or non-estimable coefficients and variances are set to zero.

Value

A data frame with the following columns:

IID Individual identifier (from the target phenotype file).

PRS Estimated polygenic risk score for each individual.

Variance Estimated variance of the PRS.

Lower_Limit Lower bound of the confidence interval.

Upper_Limit Upper bound of the confidence interval.

Examples

```
## Not run:
results <- BLUE_estimates_QT(discovery_pheno = "Qpd.txt",
                                    discovery_geno_mat = "Gd.txt",
                                    target_pheno = "Qpt.txt",
                                    target_geno_mat = "Gt.txt",
                                    significance_level = 0.05)
head(results)

## End(Not run)
```

`GWAS_BT` *GWAS_BT function*

Description

Performs genome-wide association analysis for a binary trait using logistic regression. It reads a phenotype file and a genotype matrix, and estimates the SNP effect sizes and standard errors.

Usage

```
GWAS_BT(discovery_pheno, discovery_geno_mat)
```

Arguments

`discovery_pheno`
: A character string specifying the path to the phenotype file. The file should have no header and contain individual IDs, and the third column should contain the binary trait (0/1).

`discovery_geno_mat`
: A character string specifying the path to the genotype matrix file. The file should have no header and contain numeric genotype data (e.g., 0, 1, 2) for each SNP.

Details

The function uses logistic regression (`glm` with `binomial(link="logit")`) to regress the binary phenotype on each SNP individually. The output includes only the regression coefficient and standard error for each SNP.

Value

A data frame with two columns:

beta Estimated effect size (log odds) for each SNP.

se Standard error of the estimated effect size.

Examples

```
## Not run:
# Phenotype file: 3rd column must contain binary outcome (0/1)
# Genotype file: SNPs in columns, rows correspond to individuals
# Run GWAS on a binary trait with discovery phenotype and genotype files
gwas_results <- GWAS_BT(discovery_pheno = "Bpd.txt", discovery_geno_mat = "Gd.txt")
head(gwas_results)

## End(Not run)
```

`GWAS_QT` *GWAS_QT function*

Description

Performs univariate linear regression for each SNP to estimate effect sizes and standard errors in a genome-wide association study (GWAS) for a quantitative trait.

Usage

```
GWAS_QT(discovery_pheno, discovery_geno_mat)
```

Arguments

`discovery_pheno`
: Character. Path to the phenotype file. This file should be tab- or space-delimited, with no header, and the quantitative phenotype should be located in the third column.

`discovery_geno_mat`
: Character. Path to the genotype matrix file. This file should also be delimited with no header, and each column corresponds to a SNP (e.g., encoded as 0, 1, 2).

Details

The function uses linear regression (`lm`) to regress the quantitative phenotype on each SNP separately. The phenotype is standardized prior to analysis. No covariates are included in the model. The genotype matrix and phenotype vector are assumed to be ordered consistently.

Value

A data frame with two columns:

beta Estimated effect size from linear regression.

se Standard error of the effect size estimate.

Examples

```
## Not run:
# Example usage:
# Phenotype file: 3rd column must contain a continuous outcome
# Genotype file: SNPs in columns, rows correspond to individuals
results <- GWAS_QT(discovery_pheno = "Qpd.txt", discovery_geno_mat = "Gd.txt")
head(results)

## End(Not run)
```

`iPRSue_estimates_BT` *iPRSue_estimates_BT function*

Description

Computes individual-level polygenic risk scores (PRS) with uncertainty estimates using a simulation-based approach for binary traits. This implementation follows the iPRSue framework, simulating multiple PRSs by sampling from the GWAS effect size distribution and deriving individual-level confidence intervals.

Usage

```
iPRSue_estimates_BT(
  gwas,
  target_pheno,
  target_geno_mat,
  no_of_PRSs = 500,
  significance_level = 0.05,
  seed = NULL
)
```

Arguments

`gwas`	A data frame with GWAS summary statistics for binary traits. Must contain `beta` and `se` columns representing estimated SNP effect sizes and their standard errors.
`target_pheno`	Character. Path to the target phenotype file. Assumes no header and individual IDs in the second column.
`target_geno_mat`	Character. Path to the genotype matrix of target individuals. No header is expected; columns correspond to SNPs.
`no_of_PRSs`	Integer. Number of simulations used to construct PRS uncertainty intervals. Default is 500.

`significance_level`
: Numeric. Significance level for confidence intervals (e.g., 0.05 gives 95% CI). Default is 0.05.

`seed`
: Integer or NULL. Random seed for reproducibility. If NULL, results may vary across runs. Default is NULL.

Details

For each SNP, the function simulates `no_of_PRSs` effect sizes from a normal distribution defined by its GWAS beta and SE. These sampled betas are multiplied by the genotype matrix to generate PRS replicates for each individual. Confidence intervals are then calculated using the specified significance level.

This function is designed for binary traits and should be used with GWAS summary statistics obtained from logistic regression.

Value

A data frame containing the following columns:

IID Individual identifier (from target phenotype file).

PRS Mean of simulated PRSs for each individual.

Variance Variance across simulated PRSs.

Lower_Limit Lower bound of the confidence interval.

Upper_Limit Upper bound of the confidence interval.

Examples

```
## Not run:
# Step 1: Run GWAS on binary trait
gwas_res <- GWAS_BT("Bpd.txt", "Gd.txt")

# Step 2: Estimate individual PRS with uncertainty
prs_estimates <- iPRSue_estimates_BT(gwas = gwas_res,
                                     target_pheno = "Bpt.txt",
                                     target_geno_mat = "Gt.txt",
                                     no_of_PRSs = 500,
                                     significance_level = 0.05,
                                     seed = 123)
head(prs_estimates)

## End(Not run)
```

`iPRSue_estimates_QT` *iPRSue_estimates_QT function*

Description

Computes individual-level polygenic risk scores (PRS) with uncertainty estimates using a simulation-based approach for quantitative traits. This implementation follows the iPRSue framework, simulating multiple PRSs by sampling from the GWAS effect size distribution and deriving individual-level confidence intervals.

Usage

```
iPRSue_estimates_QT(
  gwas,
  target_pheno,
  target_geno_mat,
  no_of_PRSs = 500,
  significance_level = 0.05,
  seed = NULL
)
```

Arguments

gwas	A data frame with GWAS summary statistics for a quantitative trait. Must contain beta and se columns representing estimated SNP effect sizes and their standard errors.
target_pheno	Character. Path to the target phenotype file. Assumes no header and individual IDs in the second column.
target_geno_mat	Character. Path to the genotype matrix of target individuals. No header is expected; columns correspond to SNPs.
no_of_PRSs	Integer. Number of simulations used to construct PRS uncertainty intervals. Default is 500.
significance_level	Numeric. Significance level for confidence intervals (e.g., 0.05 gives 95% CI). Default is 0.05.
seed	Integer or NULL. Random seed for reproducibility. If NULL, results may vary across runs. Default is NULL.

Details

For each SNP, the function simulates no_of_PRSs effect sizes from a normal distribution defined by its GWAS beta and SE. These sampled betas are multiplied by the genotype matrix to generate PRS replicates for each individual. Confidence intervals are then calculated using the specified significance level.

This function is designed for quantitative traits and should be used with GWAS summary statistics obtained from linear regression.

Value

A data frame containing the following columns:

IID Individual identifier (from target phenotype file).

PRS Mean of simulated PRSs for each individual.

Variance Variance across simulated PRSs.

Lower_Limit Lower bound of the confidence interval.

Upper_Limit Upper bound of the confidence interval.

Examples

```
## Not run:
# Step 1: Run GWAS on quantitative trait
gwas_res <- GWAS_QT("Qpd.txt", "Gd.txt")

# Step 2: Estimate individual PRS with uncertainty
prs_estimates <- iPRSue_estimates_QT(gwas = gwas_res,
                                     target_pheno = "Qpt.txt",
                                     target_geno_mat = "Gt.txt",
                                     no_of_PRSs = 500,
                                     significance_level = 0.05,
                                     seed = 123)
head(prs_estimates)

## End(Not run)
```

Index