Package: GRAPPLE (via r-universe)

September 23, 2024

Title R Package for MR Framework GRAPPLE

Version 0.2.2

Description Fitting and diagnosing two-sample summary data Mendelian randomization with heterogeneous instruments.

Depends R (>= 3.4.0),

License GPL (>=2)

Encoding UTF-8

LazyData true

Imports stats, ggplot2, nortest, data.table, haploR, ggrepel, dplyr, tools

RoxygenNote 7.1.1

Repository https://mrcieu.r-universe.dev

RemoteUrl https://github.com/jingshuw/GRAPPLE

RemoteRef HEAD

RemoteSha 317e837340a29129c52c9d9b22f5f8bae72e27d0

Contents

computeQ *Compute the conditional Q-statistic for assessing instrument strength*

Description

Compute the conditional Q-statistic for assessing instrument strength

Usage

```
computeQ(dat.list, p.thres = NULL)
```

Arguments

dat.list	Object returned from getInput
p.thres	The p-value threshold for SNP selection. The SNPs whose selection_pvals are less than p.thres are selected. The default value is NULL, which is to include all SNPs in data. If it is not NULL, then data should have a column selection_pvals that stores the selection p-value of each SNP.

Value

The conditional Q-statistics, degrees of freedom, and the corresponding p-values

References

Eleanor Sanderson, George Davey Smith, Frank Windmeijer, Jack Bowden, An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings, International Journal of Epidemiology, Volume 48, Issue 3, June 2019, Pages 713–727, https://doi.org/10.1093/ije/dyy262.

findModes *Use the multiple modes of the robust profile likelihood function to find out multiple pathways in MR and their marker SNPs.*

Description

THis function can be run only for k = 1 when there is only one risk factor

Usage

```
findModes(
  data,
  p.thres = NULL,
  marker.data = NULL,
  marker.p.thres = NULL,
  mode.lmts = c(-5, 5),
  cor.mat = NULL,
  loss.function = c("tukey", "huber", "l2"),
  k.findmodes = switch(loss.function[1], l2 = NA, huber = 1.345, tukey = 3),
  include.thres = 1,
  exclude.thres = 2,
  map.marker = T,
  ldThres = 0.9,
  npoints = 10000
)
```

Arguments

`data`	A data frame containing the information of the selected genetic instruments. One can simply take the `data` element from the output of function `getInput`, or provide their own data frame. The required columns include `SNP` for the SNP rsID, the columns `gamma_exp1` to `gamma_expk` for the estimated effect sizes of the SNPs on risk factors `1` to `k`, the columns `se_exp1` to `se_expk` for their standard deviations, the columns `gamma_out1` to `gamma_outm` for the estimated effect sizes of the SNPs on diseases `1` to `m`,the columns `se_out1` to `se_outm` for their standard deviations.
`p.thres`	The p-value threshold for SNP selection. The SNPs whose `selection_pvals` are less than `p.thres` are selected. The default value is NULL, which is to include all SNPs in `data`. If it is not NULL, then `data` should have a column `selection_pvals` that stores the selection p-value of each SNP.
`marker.data`	A data frame containing the information of candidate marker genes. Default is NULL, which sets `marker.data` to `data`. Another choice is to use the `marker.data` element in the output of `getInput`.
`marker.p.thres`	P-value threshold for marker SNP selection. See `p.thres`
`cor.mat`	Either NULL or a `k + 1` by `k + 1` symmetric matrix. The shared correlation matrix for `(b_exp[j], b_out[j])` across SNP `j`. Used only when `input.list` is NULL and the default value is NULL, for the identity matrix.
`loss.function`	Loss function used, one of "tukey", "huber" or "l2". Default is "tukey", which is robust to outlier SNPs with large pleiotropic effects
`k.findmodes`	Tuning parameters of the loss function, for loss "l2", it is NA, for loss "huber", default is 1.345 and for loss "tukey", default is 3.
`include.thres`	Absolute value upper threshold of the standardized test statistics of one SNP on one mode for the SNP to be included as a marker for that mode, default is 1.4
`exclude.thres`	Absolute value lower threshold of the standardized test statistics of one SNP on other modes for the SNP to be included as a marker for that mode, default is `qnorm(0.975)`

map.marker	Whether map each marker to the earist gene or not. Default is TRUE if multiple markers are found. It is always FALSE if there is just one mode.
ldThres	the parameter passed to the queryhaploReg function. Increase to 1 when there is a "timeout" error.
npoints	Number of equally spaced points chosen for grid search of modes within the range mode.lmts.
mode_lmts	The range of beta that the modes are searched from. Default is c(-5, 5)

Value

A list containing the following elements:

fun	The profile likelihood function with argument beta
modes	The position of modes. Only include modes where marker genes can be detected
p	The profile likelihood plot with gene markers when there are multiple modes. The range of the x.axis depends on the distance between the maximum mode and minimum mode when there are multiple modes.
markers	A data frame of marker information
raw.modes	All modes of the profile likelihood function within the range of mode_lmts
supp_gwas	More information about the markers.

---

getInput *Preprocess GWAS summary statistics datasets*

Description

This function has GWAS summary statistics data files as inputs, perform genetic instrument selection and return matrices that are ready to use for GRAPPLE

Usage

```
getInput(
  sel.files,
  exp.files,
  out.files,
  plink_refdat,
  max.p.thres = 0.01,
  cal.cor = T,
  p.thres.cor = 0.5,
  get.marker.candidates = T,
  marker.p.thres = 1e-05,
  marker.p.source = "exposure",
  clump_r2 = 0.001,
  clump_r2_formarkers = 0.05,
  plink_exe = NULL
)
```

Arguments

sel.files	A vector of the GWAS summary statistics file names for the risk factors SNP selection. Each GWAS file is a ".csv" or ".txt" file containing a data frame that at least has a column "SNP" for the SNP ids and "pval" for the p-values. The length of sel.files are not required to be the same as that of exp.files and the order of the files do not matter, while we strongly suggest having one selection file for each risk factor.
exp.files	A vector of length k of the GWAS summary statistics file names of the k risk factors for getting the effect sizes and standard deviations. Each GWAS file should have a column "SNP" for the SNP ids, "beta" for the effect sizes, "se" for the standard deviation, "effect_allele" for the effect allele and "other_allele" for the other allele of the SNP.
out.files	The GWAS summary statistics file name for the disease data, can be a vector of length m to allow preprocessing m diseases simultaneously. Each GWAS file should have a column "SNP" for the SNP ids, "beta" for the effect sizes, "se" for the standard deviation, "effect_allele" for the effect allele and "other_allele" for the other allele of the SNP.
plink_refdat	The reference genotype files (.bed, .bim, .fam) for clumping using PLINK (loaded with –bfile).
max.p.thres	The upper threshold of the selection p-values for a SNP to be selected before clumping. It only requires that at least one of the p-values of the risk factors of the SNPs to be below the threshold. Default is 0.01.
cal.cor	Whether calculate the (k + 1) by (k + 1) correlation matrix between the k risk factors and the outcome. The default is TRUE
p.thres.cor	The lower threshold of the p-values for a SNP to be used in calculating the correlation matrix. It only select SNPs whose p-values are above the threshold for all risk factors. Default is 0.5.
get.marker.candidates	Whether getting SNPs which are used for mode marker selection. Only applies to cases where the number of risk factors k = 1. Default is TRUE for k = 1.
marker.p.thres	P-value threshold of p-values in the exposure files for mode markers. Default is 1e-5.
marker.p.source	source of p-values of mode markers, a string of either "exposure" or "selection". Default is "exposure" for obtaining more markers.
clump_r2	The clumping r2 threshold in PLINK for genetic instrument selection. Default is set to 0.001 for selection of independent SNPs.
clump_r2_formarkers	The clumping r2 threshold in PLINK. Default is set to 0.05 for selection of candidates for the marker SNPs.
plink_exe	The name of the plink exe. Default is NULL, which uses "plink". For users with Linux systems, one may want to have a different name, like "./plink" depending on where they install plink

Value

A list of selected summary statistics, which include

`data` A data frame of size `p * (3 + 2k + 2m + 1)` for the effect sizes of `p` number of selected independent SNPs (instruments) on `k` risk factors (exposures). The first three columns include the SNP rsID, the effect allele and other allele after harmonizing, the next `2k` columns are the estimated effect sizes and standard deviations for the `k` risk factors stored in `exp.files`, the next `2m` columns are the estimated effect sizes and standard deviations for the `m` diseases stored in `exp.files` and the the last columns are the selection p-values obtained from `sel.files`

`marker.data` A data frame for marker candidate SNPs, which has the same columns as `data`

.

`cor.mat` The estimated `(k + m)` by `(k + m)` correlation matrix between the `k` risk factors and the disease (outcome) shared by SNPs. The last column is for the outcome trait.

`grappleRobustEst` *Robust multivariate MR estimation*

Description

The main function of GRAPPLE to estimate causal effects of risk factors `beta` under a random effect model of the pleiotropic effects.

Usage

```
grappleRobustEst(
  data,
  p.thres = NULL,
  cor.mat = NULL,
  tau2 = NULL,
  loss.function = c("tukey", "huber", "l2"),
  k = switch(loss.function[1], l2 = NA, huber = 1.345, tukey = 4.685),
  niter = 20,
  tol = .Machine$double.eps^0.5,
  opt.method = "L-BFGS-B",
  diagnosis = T,
  plot.it = T
)
```

Arguments

data	A data frame containing the information of the selected genetic instruments. One can simply take the data element from the output of function getInput, or provide their own data frame. The required columns include SNP for the SNP rsID, the columns gamma_exp1 to gamma_expk for the estimated effect sizes of the SNPs on risk factors 1 to k, the columns se_exp1 to se_expk for their standard deviations, the columns gamma_out1 to gamma_outm for the estimated effect sizes of the SNPs on diseases 1 to m,the columns se_out1 to se_outm for their standard deviations.
p.thres	The p-value threshold for SNP selection. The SNPs whose selection_pvals are less than p.thres are selected. The default value is NULL, which is to include all SNPs in data. If it is not NULL, then data should have a column selection_pvals that stores the selection p-value of each SNP.
cor.mat	Either NULL or a k + 1 by k + 1 symmetric matrix. The shared correlation matrix for (b_exp[j], b_out[j]) across SNP j. Used only when input.list is NULL and the default value is NULL, for the identity matrix.
tau2	The dispersion parameter. The default value is NULL, which is to be estimated by the function
loss.function	Loss function used, one of "tukey", "huber" or "l2". Default is "tukey", which is robust to outlier SNPs with large pleiotropic effects
k	Tuning parameters of the loss function, for loss "l2", it is NA, for loss "huber", default is 1.345 and for loss "tukey", default is 4.685
niter	Number of maximum iterations allowed for optimization. Default is 20
tol	Tolerance for convergence, default is the square root of the smallest positive floating number depending on the machine R is running on
opt.method	the optimization used, which is one of choices the R function optim accepts. Default value is "L-BFGS-B".
diagnosis	Run diagnosis analysis based on the residuals or not, default is FALSE
plot.it	Whether show the QQ-plot or not if diagnosis if performed. Default is TRUE.

Value

A list with elements

beta.hat	Point estimates of beta
tau2.hat	Point estimates of the pleiotropic effect variance tau2 if the argument tau2 is set to NULL
beta.variance	Estimated covariance matrix of beta.hat
tau2.se	Estimated standard deviation of tau2.hat
beta.p.vaue	A vector of p-values where the kth element is the p-value for whether beta_k = 0
std.resid	Returned if diagnosis is TRUE. A vector of standardized residuals of each SNP

qqDiagnosis *QQ-plot diagnosis and outlier detection from standardized residuals*

Description

QQ-plot diagnosis and outlier detection from standardized residuals

Usage

```
qqDiagnosis(
  std.residuals,
  outlier.quantile = 0.1/length(std.residuals),
  plot.it = T
)
```

Arguments

std.residuals	A vector of standardized residuals, can be the output element std.resid from function grappleRobustEst
outlier.quantile	The quantile threshold for outliers.
plot.it	Whether show the QQ-plot or not if diagnosis if performed. Default is TRUE.

Value

A list with two elements

p	The QQ plot
outliers	The data frame for the detected outliers

rho.tukey *Tukey's beweight loss function and its derivatives*

Description

Tukey's beweight loss function and its derivatives

Usage

```
rho.tukey(r, k = 4.685, deriv = 0)
```

Arguments

r	Function value
k	Tuning parameter, default value is 4.685
deriv	The derivative function to calculate. 0 is the Tukey's loss function, 1 is for the first derivative and 2 for the second derivative function

Value

The value of the corresponding function at r

---

robustLossFixtau *Calculate the robustified profile likelihood function*

Description

Calculate the robustified profile likelihood function

Usage

```
robustLossFixtau(
  data = NULL,
  b_exp = NULL,
  b_out = NULL,
  se_exp = NULL,
  se_out = NULL,
  cor.mat = NULL,
  loss.function = c("tukey", "huber", "l2"),
  k.findmodes = switch(loss.function[1], l2 = NA, huber = 1.345, tukey = 3)
)
```

Arguments

data	A data frame containing the information of the selected genetic instruments. One can simply take the data element from the output of function getInput, or provide their own data frame. The required columns include SNP for the SNP rsID, the columns gamma_exp1 to gamma_expk for the estimated effect sizes of the SNPs on risk factors 1 to k, the columns se_exp1 to se_expk for their standard deviations, the columns gamma_out1 to gamma_outm for the estimated effect sizes of the SNPs on diseases 1 to m,the columns se_out1 to se_outm for their standard deviations.
cor.mat	Either NULL or a k + 1 by k + 1 symmetric matrix. The shared correlation matrix for (b_exp[j], b_out[j]) across SNP j. Used only when input.list is NULL and the default value is NULL, for the identity matrix.
loss.function	Loss function used, one of "tukey", "huber" or "l2". Default is "tukey", which is robust to outlier SNPs with large pleiotropic effects
k.findmodes	Tuning parameters of the loss function, for loss "l2", it is NA, for loss "huber", default is 1.345 and for loss "tukey", default is 3.

Value

The robustified profile likelihood function

Index