Package 'avengeme'

July 4, 2017

Title Analysis of polygenic scoring methods

Version 1.0

Description AVENGEME (Additive Variance Explained and Number of Genetic Effects Method of Estimation) is a package for the analysis of the polygenic scoring method. Polygenic scores access a large proportion of the heritability of a trait by aggregating genetic effects across thousands of markers into a single composite score. The markers may not be individually associated with the trait at standard statistical significance levels. However, the composite score may be associated, indicating a substantial polygenic component within the score.

Depends R (>= 3.3.1)

License GPL-3

Encoding UTF-8

LazyData true

RoxygenNote 6.0.1

Author Frank Dudbridge [aut, cre]

Maintainer Frank Dudbridge <frank.dudbridge@leicester.ac.uk>

Imports mvtnorm

Suggests shiny

R topics documented:

`avengemeShiny` *Launch Shiny app for avengeme*

Description

Launches a Shiny graphical user interface for the avengeme package.

Usage

```
avengemeShiny()
```

Details

Uses the `estimatePolygenicModel` function to estimate all parameters that are not explicitly set in the graphical interface. If all parameters are set, calls the `polygenescore` function instead. Parameters for combined genetic and environmental risk scores are not yet implemented.

Author(s)

Frank Dudbridge

References

Palla L and Dudbridge F (2015) A fast method using polygenic scores to estimate the variance explained by genome-wide marker panels and the proportion of variants affecting a trait. Am J Hum Genet 97:250-259

See Also

`estimatePolygenicModel`, `polygenescore`

---

estimatePolygenicModel

Estimate polygenic model

Description

Estimates the parameters of an underlying genetic model from the results of association tests of a polygenic score.

Usage

```
estimatePolygenicModel(p, nsnp, n, vg = c(NA, NA), cov12 = NA, pi0 = c(NA,
  NA), pupper = 1, nested = TRUE, weighted = TRUE, binary = c(FALSE,
  FALSE), prevalence = c(0.1, 0.1), sampling = prevalence, lambdaS = c(NA,
  NA), shrinkage = FALSE, logrisk = FALSE, option = 0, boot = 0,
  bidirectional = FALSE, initial = c(), fixvg2pi02 = FALSE)
```

Arguments

`p`	Vector of P-values or Z-statistics for polygenic scores tested in the target data. Automatically detects Z-statistics if some entries of p are greater than 1 or less than 0.
`nsnp`	Number of independent markers in the polygenic score.
`n`	Vector with two elements, giving the total sizes of the training and target samples. In case/control studies, n is the sum of the number of cases and number of controls. If only one element of n is given, the training and target samples are assumed to be the same size. No default - a value must be given

`vg` Proportion of variance explained by genetic effects in training sample. By default, the variance explained is the same in the target sample; otherwise vg should be a vector with two elements for the training and target samples respectively.

`cov12` Covariance between genetic effect sizes in the two samples. If the effects are fully correlated then cov12<=sqrt(vg1). If the effects are identical then cov12=vg1 (default).

`pi0` Proportion of markers with no effect on the training trait. By default, the proportion is the same for the target trait; otherwise pi0 should be a vector with two elements for the training and target samples respectively.

`pupper` Vector of p-value thresholds for selecting markers from training sample. First element is the lower bound of the first interval, second element is the upper bound of the first interval, third element is the upper bound of the second interval, etc.

`nested` TRUE if the p-value intervals are nested, that is they have the same lower bound, which is the first element of pupper. If false, lower bound of the second interval is the upper bound of the first and so on.

`weighted` TRUE if estimated effect sizes are used as weights in forming the polygenic score. If false, an unweighted score is used, which is the sum of risk alleles carried.

`binary` TRUE if the training trait is binary. By default, the target trait is binary if the training trait is; otherwise binary should be a vector with two elements for the training and target samples respectively.

`prevalence` For a binary trait, prevalence in the training sample, By default, prevalence is the same in the target sample. Otherwise, prevalence should be a vector with two elements for the training and target samples respectively.

`sampling` For a binary trait, case/control sampling fraction in the training sample. By default, sampling equals the prevalence, as in a cohort study. If the sampling fraction is different in the target sample, sampling should be a vector with two elements for the training and target samples respectively.

`lambdaS` Sibling relative recurrence risk in training sample, can be specified instead of vg1.

`shrinkage` TRUE if effect sizes are to be shrunk to BLUPs.

`logrisk` TRUE if binary trait arises from log-risk model rather than liability threshold.

`option` Parameter used in method development. Default 0, fits the model by maximum likelihood for Z statistics. 1 and 2 fit the model by least squares to chisq and Z statistics respectively. 3 fits by maximum likelihood for chisq statistics.

`boot` Number of bootstrap replicates to estimate approximate confidence intervals. If boot==0 (default), an analytic interval is calculated using profile likelihood. if boot>0, a bootstrap interval is estimated. These intervals assume that the input P-values are independent; this assumption is generally untrue and the interval will be slightly smaller than it should be.

`bidirectional` TRUE if p also contains results when exchanging the role of training and target samples. In this case, vg and pi0 can also be estimated in the target sample. The input vector p should now be twice as long with the list of P-values for training/target followed by the list for target/training.

`initial` Specify starting values for numerical maximisation of the likelihood. The number of elements must equal the number of estimated parameters, and follows the

order vg[1], vg[2], pi0[1], pi0[2], cov12, for those parameters that are actually being estimated. Default 0.5 for all parameters.

`fixvg2pi02` TRUE if the same genetic model is assumed for the training and target samples. This fixes the target variance and the covariance to both equal the variance explained in the training sample, vg1. Also fixes the proportion of null markers in the target sample to equal the proportion in the training sample.

Details

The input is a vector of P-values or (signed) Z-statistics from the association test of the polygenic score in the target sample. P-values are assumed if all the values are in (0,1), otherwise Z-statistics are assumed. Each P-value corresponds to the association test of a polygenic score consisting of SNPs with training sample P-values in a specific interval. Up to five parameters can be estimated: vg[1], cov12, vg[2], pi0[1], pi0[2]. The number of input P-values must be greater than or equal to the number of estimated parameters, otherwise an error message is returned. Any combination of parameters can be estimated. A parameter will be estimated if its input value is unspecified or `NA`.

Value

A list with elements corresponding to the estimated genetic model. Values fixed at input are returned unchanged with a degenerate confidence interval. Each element is a vector consisting of the point estimate followed by its lower and upper 95% confidence limit.

- `vg` Variance explained in the training trait. If bidirectional estimation is selected, `vg` is a matrix with two rows corresponding to the training and target samples respectively.
- `cov12` Covariance between genetic effects in the two samples.
- `pi0` Proportion of markers with no effect on the training trait. If bidirectional estimation is selected, `pi0` is a matrix with two rows corresponding to the training and target samples respectively.
- `logLikelihood` Maximised log-likelihood at the fitted model.
- `error` Error message, if any.

Author(s)

Frank Dudbridge

References

Palla L and Dudbridge F (2015) A fast method using polygenic scores to estimate the variance explained by genome-wide marker panels and the proportion of variants affecting a trait. Am J Hum Genet 97:250-259

Examples

```
# Schizophrenia PGC2 study, rightmost column of table 5 in Palla & Dudbridge (2015)
# P-values from supplementary table 6 of Schizophrenia Working Group (2014)
# Other parameters as in table 1 of Palla & Dudbridge (2015)
pupper=c(0,5e-8,1e-6,1e-4,1e-3,0.01,0.05,0.1,0.2,0.5,1)
p=c(9.85087e-24, 4.44037e-36, 2.08048e-71, 8.0594e-103, 2.0587e-138,
    1.4131e-164,5.8954e-166,3.75e-164,7.9488e-159,2.3286e-157)
estimatePolygenicModel(p,103125,c(77195,5120),pupper=pupper,nested=TRUE,binary=TRUE,
prevalence=0.01,sampling=c(0.425,0.515),fixvg2pi02=TRUE)
# $vg
```

```
# [1] 0.2449328 0.2352049 0.2547500
#
# $cov12
# [1] 0.2449328 0.2352049 0.2547500
#
# $pi0
# [1] 0.8520102 0.8354152 0.8669681
#
# $logLikelihood
# [1] 30.1139
#
# $error
# [1] ""

# Genetic covariance between bipolar disorder and schizophrenia
# Table 6 in Palla & Dudbridge (2015)
# Nagelkerke R2 SCZ-BPD from table S5 of Cross Disorder Group (2013)
R2N=c(.0044,.0065,.015,.023,.024,.025,.024,.024,.025,.025)
n1=11922
p1=6664/n1 # sampling fraction
# Convert to observed scale R2
R2O=R2N*(1-p1^(2*p1)*(1-p1)^(2*(1-p1)))
# Convert to chisq statistics
X2=n1*R2O/(1-R2O)
# Now the same for BPD-SCZ
R2N=c(0.002,0.0048,0.012,0.017,0.021,0.021,0.022,0.021,0.021,0.021)
n2=17012
p2=9032/n2
R2O=R2N*(1-p2^(2*p2)*(1-p2)^(2*(1-p2)))
X2=c(n2*R2O/(1-R2O),X2)
# Perform bidirectional estimation with Z-scores as the first argument
# Small difference from published result due to minor bug fixes
pupper=c(0,.0001,.001,.01,.05,.1,.2,.3,.4,.5,1)
estimatePolygenicModel(sqrt(X2),nsnp=83884,n=c(n1,n2),binary=TRUE,pupper=pupper,
prevalence=c(0.01,0.01),sampling=c(p1,p2),bidirectional=TRUE)
# $vg
# vg      vgLo      vgHi
# [1,] 0.8800473 0.1784087 0.9999528
# [2,] 0.6339373 0.1258082 0.9999382
#
# $cov12
# [1] 0.2092269 0.1895850 0.2226666
#
# $pi0
# pi0     pi0Lo     pi0Hi
# [1,] 0.7297216 0.6453690 0.9138123
# [2,] 0.7237033 0.5936525 0.9127537
#
# $logLikelihood
# [1] 19.55162
#
# $error
# [1] ""
```

`plotAccuracy` *Plot predictive accuracy*

Description

Plots the AUC or the R2 as a function of training sample size.

Usage

```
plotAccuracy(xlim = c(1, 500), ylim = 0, nsnp = 1e+05, vg1 = 0,
  pi0 = 0, cov12 = NA, fix = TRUE, binary = FALSE, prevalence = 0,
  sampling = 0.5, r2gx = 0, corgx = 0, r2xy = 0,
  adjustedEffects = FALSE, plot = TRUE, col = "black", breakeven = 0.5,
  lty = 1)
```

Arguments

`xlim`	Vector of 2 elements, giving the range of sample size to display on the x-axis, in 1000s. For binary traits this is the number of cases.
`ylim`	Range of AUC/R2 to display on y-axis.
`nsnp`	Number of independent SNPs in the gene score.
`vg1`	Proportion of variance explained by genetic effects in the training sample.
`pi0`	Proportion of markers with no effect on the training trait.
`cov12`	Covariance between genetic effect sizes in the two samples. If the effects are fully correlated then cov12<=sqrt(vg1). If the effects are identical then cov12=vg1 (default).
`fix`	TRUE if the same genetic model is assumed for the training and target samples.
`binary`	TRUE if the training trait is binary. By default, the target trait is binary if the training trait is; otherwise binary should be a vector with two elements for the training and target samples respectively.
`prevalence`	For a binary trait, prevalence in the training sample, By default, prevalence is the same in the target sample. Otherwise, prevalence should be a vector with two elements for the training and target samples respectively.
`sampling`	For a binary trait, case/control sampling fraction in the training sample. By default, sampling equals the prevalence, as in a cohort study. If the sampling fraction is different in the target sample, sampling should be a vector with two elements for the training and target samples respectively.
`r2gx`	Proportion of variance in environmental risk score explained by genetic effects in training sample.
`corgx`	Genetic correlation between environmental risk score and training trait.
`r2xy`	Proportion of variance in training trait explained by environmental risk score.
`adjustedEffects`	TRUE if polygenic and envrionmental scores are combined as a weighted sum. If FALSE, the scores are combined as an unweighted sum even if they are correlated.
`plot`	TRUE is a new plot is to be drawn, otherwise draw lines on the existing plot.
`col`	Colour in which to plot.
`breakeven`	Value of AUC/R2 for which the minimum sample size will be estimated.
`lty`	Line type parameter for R plots.

Details

AUC is plotted for binary traits, R2 for quantitative traits. At each point, the p-value threshold is identified for selecting markers into the polygenic score, such that the AUC or R2 is maximised.

Value

A list with the following elements:

- `limit` Value of AUC/R2 at the maximum sample size plotted.
- `breakeven` Sample size at which the AUC/R2 exceeds the value specified by the breakeven parameter.
- `plimit` Optimal P-value threshold at the maximum sample size plotted.

Author(s)

Frank Dudbridge

References

Dudbridge F (2013) Power and predictive accuracy of polygenic risk scores. PLoS Genet 9:e1003348

Examples

```
# Breast cancer with 90% null markers, from figure 3 in Dudbridge (2013)
plotAccuracy(vg1=0.44/2,pi0=0.90,fix=TRUE,binary=TRUE,prevalence=0.036)
```

`polygenescore` *Calculate power and predictive accuracy of a polygenic score*

Description

Calculates measures of association for a polygenic score derived from a training sample to predict traits in a target sample.

Usage

```
polygenescore(nsnp, n, vg1 = 0, cov12 = vg1, pi0 = 0, pupper = c(0, 1),
  nested = TRUE, weighted = TRUE, binary = c(FALSE, FALSE),
  prevalence = c(0.1, 0.1), sampling = prevalence, lambdaS = NA,
  shrinkage = FALSE, logrisk = FALSE, alpha = 0.05, r2gx = 0,
  corgx = 0, r2xy = 0, adjustedEffects = FALSE, riskthresh = 0.1)
```

Arguments

`nsnp`	Number of independent markers in the polygenic score.
`n`	Vector with two elements, giving the total sizes of the training and target samples. In case/control studies, n is the sum of the number of cases and number of controls. If only one element of n is given, the training and target samples are assumed to be the same size. No default - a value must be given
`vg1`	Proportion of variance explained by genetic effects in the training sample.

`cov12`	Covariance between genetic effect sizes in the two samples. If the effects are fully correlated then cov12<=sqrt(vg1). If the effects are identical then cov12=vg1 (default).
`pi0`	Proportion of markers with no effect on the training trait.
`pupper`	Vector of p-value thresholds for selecting markers from training sample. First element is the lower bound of the first interval, second element is the upper bound of the first interval, third element is the upper bound of the second interval, etc.
`nested`	TRUE if the p-value intervals are nested, that is they have the same lower bound, which is the first element of pupper. If false, lower bound of the second interval is the upper bound of the first and so on.
`weighted`	TRUE if estimated effect sizes are used as weights in forming the polygenic score. If false, an unweighted score is used, which is the sum of risk alleles carried.
`binary`	TRUE if the training trait is binary. By default, the target trait is binary if the training trait is; otherwise binary should be a vector with two elements for the training and target samples respectively.
`prevalence`	For a binary trait, prevalence in the training sample, By default, prevalence is the same in the target sample. Otherwise, prevalence should be a vector with two elements for the training and target samples respectively.
`sampling`	For a binary trait, case/control sampling fraction in the training sample. By default, sampling equals the prevalence, as in a cohort study. If the sampling fraction is different in the target sample, sampling should be a vector with two elements for the training and target samples respectively.
`lambdaS`	Sibling relative recurrence risk in training sample, can be specified instead of vg1.
`shrinkage`	TRUE if effect sizes are to be shrunk to BLUPs.
`logrisk`	TRUE if binary trait arises from log-risk model rather than liability threshold.
`alpha`	Significance level for testing association of the polygenic score in the target sample.
`r2gx`	Proportion of variance in environmental risk score explained by genetic effects in training sample.
`corgx`	Genetic correlation between environmental risk score and training trait.
`r2xy`	Proportion of variance in training trait explained by environmental risk score.
`adjustedEffects`	TRUE if polygenic and envrionmental scores are combined as a weighted sum. If FALSE, the scores are combined as an unweighted sum even if they are correlated.
`riskthresh`	Absolute risk threshold for calculating net reclassification index.

Details

The following setup is assumed. Two independent samples of genotypes are available; this could be one sample of data split into two subsets. One sample is termed the training sample, the other the target sample. Traits are measured in each sample; different traits could be measured in training and target samples. Subjects are assumed to be unrelated, and genotypes assumed to be independent. In practice we recommend LD-clumping methods, such as the –clump option in PLINK, to ensure weak dependence between markers; in this case the methods are almost unbiased if an r2 threshold of 0.1 is used. Markers with P-values within a fixed range are selected from the training sample, and then used to construct a polygenic score for each subject in the target sample. The score can be tested for association to the target trait, or used to predict individual trait values in the target sample.

Value

A list with elements containing quantities describing the association of the polygenic score with the target trait:

- `R2` Squared correlation between polygenic score and target trait.
- `NCP` Non-centrality parameter of the chisq test of association between polygenic score and target trait.
- `p` Expected P-value of the chisq test of association between polygenic score and target trait.
- `power` Power of the chisq test of association between polygenic score and target trait.
- `FDR` Expected proportion of false positives among selected markers.
- `AUC` For binary traits, area under ROC curve.
- `MSE` For quantitative traits, mean square error between target trait and polygenic score.
- `NRI` Net reclassification improvement in cases, controls, and combined.
- `IDI` Integrated discrimination improvement.
- `error` Error message, if any.

Author(s)

Frank Dudbridge

References

Dudbridge F (2013) Power and predictive accuracy of polygenic risk scores. PLoS Genet 9:e1003348

Dudbridge F, Pashayan N, Yang J. Predictive accuracy of combined genetic and environmental risk scores. Submitted.

Examples

```
# P-value for ISC schizophrenia score associated with schizophrenia in MGS-EA
# See page 3, column 2, paragraph 3 of Dudbridge (2013)
polygenescore(74062,n=c(3322+3587,2687+2656),vg1=0.269,pi0=0.99,binary=TRUE,
sampling=c(3322/6909,2687/5343),pupper=c(0,0.5),prevalence=.01)$p
# [1] 1.029771e-28

# Power for ISC schizophrenia score associated with bipolar disorder in WTCCC
# See page 4, column 2, paragraph 2 of Dudbridge (2013)
polygenescore(74062,c(3322+3587,1829+2935),vg1=0.287,cov12=0.28*0.287,binary=TRUE,
sampling=c(3322/6909,1829/4764),pupper=c(0,0.5),prevalence=.01)$power
# [1] 0.8042843

# Power for cross validation study of Framingham risk score
# See page 6, column 1, paragraph 1 of Dudbridge (2013)
polygenescore(100000,c(1575,175),vg1=1,pupper=c(0,0.1,0.2,0.3,0.4,0.5),
nested=FALSE)$power
# [1] 0.19723400 0.11733175 0.09195134 0.07733049 0.06771049

# Net reclassification index for cardiovascular disease with QRISK-2 and 53 SNPs
# See table 3, row 1, columns 5-6 of Dudbridge et al (submitted)
# results vary due to stochastic evaluation of multivariate normal probabilities
polygenescore(nsnp=1e5,n=63746+130681,vg1=0.3,pi0=0.8,binary=TRUE,
prevalence=0.15,sampling=63746/194427,pupper=c(0,5e-8),
r2gx=0.3,r2xy=0.052,corgx=0.1,riskthresh=0.1,adjustedEffects=TRUE)$NRI
# [1] -0.006042718  0.015266759  0.009224041
```

`sampleSizeForGeneScore` *Sample size calculations for polygenic scores*

Description

Calculates the size of training sample to achieve a given AUC, R2 or power in the target sample.

Usage

```
sampleSizeForGeneScore(targetQuantity, targetValue, nsnp, n2 = NA, vg1 = 0,
  cov12 = vg1, pi0 = 0, weighted = TRUE, binary = FALSE,
  prevalence = 0.1, sampling = prevalence, lambdaS = NA,
  shrinkage = FALSE, logrisk = FALSE, alpha = 0.05, r2gx = 0,
  corgx = 0, r2xy = 0, adjustedEffects = FALSE)
```

Arguments

`targetQuantity`	Either "AUC", "R2" or "power" (case insensitive).
`targetValue`	The value of the targetQuantity for which to calculate sample size.
`nsnp`	Number of independent markers in the polygenic score.
`n2`	Target sample size. Only relevant when targetQuantity is "power". By default set equal to the training sample size.
`vg1`	Proportion of variance explained by genetic effects in the training sample.
`cov12`	Covariance between genetic effect sizes in the two samples. If the effects are fully correlated then cov12<=sqrt(vg1). If the effects are identical then cov12=vg1 (default).
`pi0`	Proportion of markers with no effect on the training trait.
`weighted`	TRUE if estimated effect sizes are used as weights in forming the polygenic score. If false, an unweighted score is used, which is the sum of risk alleles carried.
`binary`	TRUE if the training trait is binary. By default, the target trait is binary if the training trait is; otherwise binary should be a vector with two elements for the training and target samples respectively.
`prevalence`	For a binary trait, prevalence in the training sample, By default, prevalence is the same in the target sample. Otherwise, prevalence should be a vector with two elements for the training and target samples respectively.
`sampling`	For a binary trait, case/control sampling fraction in the training sample. By default, sampling equals the prevalence, as in a cohort study. If the sampling fraction is different in the target sample, sampling should be a vector with two elements for the training and target samples respectively.
`lambdaS`	Sibling relative recurrence risk in training sample, can be specified instead of vg1.
`shrinkage`	TRUE if effect sizes are to be shrunk to BLUPs.
`logrisk`	TRUE if binary trait arises from log-risk model rather than liability threshold.

`alpha` Significance level for testing association of the polygenic score in the target sample.

`r2gx` Proportion of variance in environmental risk score explained by genetic effects in training sample.

`corgx` Genetic correlation between environmental risk score and training trait.

`r2xy` Proportion of variance in training trait explained by environmental risk score.

`adjustedEffects` TRUE if polygenic and envrionmental scores are combined as a weighted sum. If FALSE, the scores are combined as an unweighted sum even if they are correlated.

Details

The sample size is estimated by numerical optimisation. For each possible sample size, the P-value threshold is identified for selecting markers into the polygenic score, such that targetQuantity is maximised.

Value

A list with the following elements:

- `n` Required sample size for the training sample. This is the total sample size: to obtain the number of cases, multiply by the sampling fraction.
- `p` P-value threshold for selecting markers into the polygenic score, such that the target value is achieved with the minimum sample size.
- `max` Maximum targetQuantity possible if the training sample size were increased to infinity (actually 1e10).

Author(s)

Frank Dudbridge

References

Dudbridge F (2013) Power and predictive accuracy of polygenic risk scores. PLoS Genet 9:e1003348

Examples

```
# AUC= 0.75 in breast cancer.  See Table 4, row 4, column 3 in Dudbridge (2013).
sampleSizeForGeneScore("AUC",0.75,nsnp=100000,vg1=0.44/2,pi0=0.90,binary=TRUE,
prevalence=0.036,sampling=0.5)
# $n
# [1] 313981.4
#
# $p
# [1] 0.007500909
#
# $max
# [1] 0.788842
#
# Number of cases
sampleSizeForGeneScore("AUC",0.75,nsnp=100000,vg1=0.44/2,pi0=0.90,binary=TRUE,
prevalence=0.036,sampling=0.5)$n/2
# [1] 156990.7
```

Index