AVENGEME

Additive Variance Explained and Number of Genetic Effects Method of Estimation

Frank Dudbridge, July 2015

AVENGEME is a set of R functions for the analysis of the polygenic scoring method. In brief, this approach aims to 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 are generally not individually associated with the trait, at least at standard statistical significance levels. However, the composite score may be associated, indicating a substantial "polygenic" component within the score.

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 with an r^2 threshold of 0.1.

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.

The general formula for the polygenic score is

$$\hat{S}_i = \sum_j \hat{\beta}_j x_{ij}$$

where $\hat{\beta}_j$ is the estimated effect size (beta coefficient,or log odds ratio) of marker *j* in the training sample, x_{ij} is the coded genotype (typically, number of minor alleles) of subject *i* at marker *j* in the target sample, and the sum is over markers *j* whose P-values in the training sample are within the specified range.

The theory underlying the methods in AVENGEME is set out in the following papers

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

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

Running AVENGEME

AVENGEME can be run from an R console, or via a Shiny graphical interface.

To run within the console, unzip the avengeme.zip file and type source(“avengeme.R”), making sure that the current working directory contains the avengeme.R file (or giving the full pathname as the argument to source).

To run under Shiny, install the package with install.packages(“shiny”) and load it into the current R session with library(shiny). Unzip the avengeme.zip file into a sub-directory called “avengeme” of the working directory, and type runApp(“avengeme”) into the R terminal.

Three functions are provided:

polygenescore: calculates power and predictive accuracy of a polygenic score, given the specified genetic model parameters

sampleSizeForGeneScore: calculates the required size of the training sample to achieve a given area under ROC curve (AUC), squared correlation (R^2), or power of association test in the target sample

estimatePolygenicModel: given the results of a test of association between polygenic score and target trait, estimates the genetic model under which these results are most likely

polygenescore

polygenescore implements the theory in Dudbridge (2013), and supercedes the previous software associated with that paper, distributed as polygenescore.R

Arguments

nsnp number of markers genotyped in both training and target samples. Markers are assumed to be independent (PLINK clumping with $r^2<0.1$ works well). No default – a value must be given

n vector with 2 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 trait variance explained by genetic effects in the training sample. For binary traits, vg1 is on the liability scale. Default = 0

cov12 covariance between genetic effects in the training and target samples. Default = vg1

pi0 proportion of markers with no effect on the training trait. Default = 0, meaning that all markers have an effect in the training sample

pupper vector of *P*-value thresholds for selecting markers into the polygenic score. The first element is the lower bound of the first interval, and the second element is the upper bound of the first interval. The third element is the upper bound of the second interval, and so on. Each interval defines a different polygenic score which can be tested in the target sample. So the length of pupper is the number of intervals plus one. Default = c(0,1), meaning that all markers are selected into a single polygenic score.

nested if true, intervals defined by pupper are assumed to be nested, ie they have the same lower bound, given by the first element of pupper. If false, the intervals are assumed to be adjacent so that the second element of pupper is both the upper bound of the first interval and the lower bound of the second interval, and so on. Default =T

weighted

if true, $\hat{\beta}_j$ is assumed to be used as in the definition of the score. If false, $\hat{\beta}_j$ is assumed to be replaced by ±1 according to the sign of $\hat{\beta}_j$. Default = T

binary vector with 2 elements. First element is T if training trait is binary, second element is T if target trait is binary. If only one element is given, target trait is assumed to be of the same type as the training trait. Default =c(F,F)

prevalence

vector with 2 elements corresponding to the population prevalence of the training and target traits, if they are binary. If only the target trait is binary, the first element is ignored. If only one element is given, and both traits are binary, target prevalence is assumed to equal the training prevalence. Default = c(0.1,0.1) sampling

vector with 2 elements corresponding to the case/control sampling fractions in the training and target samples, if they are binary. If only the target trait is binary, the first element is ignored. If only one element is given, and both traits are binary, target sampling fraction is assumed to equal the training sampling fraction. Default = prevalence, as in a cohort study

lambdaS

sibling recurrence risk, an alternative parameterisation to vg1. Not required if vg1 is specified, but if both are specified, lambdaS overrides vg1

shrinkage

if T, assumes that $\hat{\beta}_j$ are obtained by shrinkage estimation in a ridge regression model. Default = F

logrisk if T, assumes that effects are normally distributed on the log-risk scale rather than the liability scale. Default = F

alpha type-1 error rate for calculating power. Default = 0.05

<u>Value</u>

A list with the following elements

R2 squared correlations between polygenic scores and target trait. Each element of R2 corresponds to a P-value selection interval as defined by the input parameter pupper

NCP non-centrality parameters of χ^2 tests of association between polygenic scores and target trait.

p expected *P*-values of tests of association between polygenic scores and target trait

power power of association tests between polygenic scores and target trait

AUC if target trait is binary, expected areas under ROC curve for prediction

MSE if target trait is not binary, expected mean square errors of prediction

error error message

sampleSizeForGeneScore

Arguments

targetQuantity

must be “AUC”, “R2” or “Power” depending on which quantity the sample size is intended to be calculated for. No default – a value must be given

targetValue

the value of AUC, R^2 or power for which the minimum sample size is required

nsnp number of markers genotyped in both training and target samples. Markers are assumed to be independent (PLINK clumping with r^2<0.1 works well). No default – a value must be given

n2 total size of the target sample. In case/control studies, n2 is the sum of the number of cases and number of controls. No default – a value must be given

vg1 proportion of trait variance explained by genetic effects in the training sample. For binary traits, vg1 is on the liability scale. Default = 0

cov12 covariance between genetic effects in the training and target samples. Default = vg1

pi0 proportion of markers with no effect on the training trait. Default = 0, meaning that all markers have an effect in the training sample

weighted

if true, $\hat{\beta}_j$ is assumed to be used as in the definition of the score. If false, $\hat{\beta}_j$ is assumed to be replaced by ± 1 according to the sign of $\hat{\beta}_j$. Default = T

binary vector with 2 elements. First element is T if training trait is binary, second element is T if target trait is binary. If only one element is given, target trait is assumed to be of the same type as the training trait. Default =c(F,F)

prevalence

vector with 2 elements corresponding to the population prevalence of the training and target traits, if they are binary. If only the target trait is binary, the first element is ignored. If only one element is given, and both traits are binary, target prevalence is assumed to equal the training prevalence. Default = c(0.1,0.1)

sampling

vector with 2 elements corresponding to the case/control sampling fractions in the training and target samples, if they are binary. If only the target trait is binary, the first element is ignored. If only one element is given, and both traits are binary, target sampling fraction is assumed to equal the training sampling fraction. Default = prevalence, as in a cohort study

lambdaS

sibling recurrence risk, an alternative parameterisation to vg1. Not required if vg1 is specified, but if both are specified, lambdaS overrides vg1 shrinkage

if T, assumes that $\hat{\beta}_j$ are obtained by shrinkage estimation in a ridge regression model. Default = F

logrisk if T, assumes that effects are normally distributed on the log-risk scale rather than the liability scale. Default = F

alpha type-1 error rate for calculating power. Default = 0.05

Value

A list with the following elements

n minimum training sample size

p *P*-value selection threshold in the training sample for which the required AUC/R^2/power is achieved at this sample size. The selection interval is assumed to be (0,p).

max value of AUC/R^2/power which would be achieved with an infinitely large training sample (in fact, a sample size of 10^{10}). This indicates the limiting potential of the polygenic scoring method under the given genetic model.

estimatePolygenicModel

estimatePolygenicModel implements the method described in Palla & Dudbridge (2015).

Arguments

p vector of two-sided *P*-values from association tests of polygenic scores with the target trait. Each *P*-value corresponds to a polygenic score from a selection interval defined by the pupper argument. *Z*-scores can be provided instead of *P*-values, the advantage being that these can be negative, allowing negative genetic correlations between the training and target traits to be detected. The function automatically assumes *P*-values if all elements of p are between 0 and 1; if any values are negative or greater than 1, it assumes *Z*-scores.

nsnp number of markers genotyped in both training and target samples. Markers are assumed to be independent (PLINK clumping with $r^2<0.1$ works well). No default – a value must be given

n vector with 2 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 trait variance explained by genetic effects in the training sample. To estimate this proportion, specify NA. For binary traits, vg is on the liability scale. Default = NA

cov12 covariance between genetic effects in the training and target samples. To estimate this covariance, specify NA. Default = NA

pi0 proportion of markers with no effect on the training trait. To estimate this proportion, specify NA. Default = NA

pupper vector of *P*-value thresholds for selecting markers into the polygenic score. The first element is the lower bound of the first interval, and the second element is the upper bound of the first interval. The third element is the upper bound of the second interval, and so on. Each interval defines a different polygenic score which can be tested in the target sample. So the length of pupper is the number of intervals plus one. Default = c(0,1), meaning that all markers are selected into a single polygenic score.

nested if true, intervals defined by pupper are assumed to be nested, ie they have the same lower bound, given by the first element of pupper. If false, the intervals are assumed to be adjacent so that the second element of pupper is both the upper bound of the first interval and the lower bound of the second interval, and so on. Default =T

weighted

if true, $\hat{\beta}_j$ is assumed to be used as in the definition of the score. If false, $\hat{\beta}_j$ is assumed to be replaced by ±1 according to the sign of $\hat{\beta}_j$. Default = T

binary vector with 2 elements. First element is T if training trait is binary, second element is T if target trait is binary. If only one element is given, target trait is assumed to be of the same type as the training trait. Default =c(F,F) prevalence

vector with 2 elements corresponding to the population prevalence of the training and target traits, if they are binary. If only the target trait is binary, the first element is ignored. If only one element is given, and both traits are binary, target prevalence is assumed to equal the training prevalence. Default = c(0.1,0.1)

sampling

vector with 2 elements corresponding to the case/control sampling fractions in the training and target samples, if they are binary. If only the target trait is binary, the first element is ignored. If only one element is given, and both traits are binary, target sampling fraction is assumed to equal the training sampling fraction. Default = prevalence, as in a cohort study

lambdaS

sibling recurrence risk, an alternative parameterisation to vg. Not required if vg is specified, but if both are specified, lambdaS overrides vg

shrinkage

if T, assumes that $\hat{\beta}_j$ are obtained by shrinkage estimation in a ridge regression model. Default = F

logrisk if T, assumes that effects are normally distributed on the log-risk scale rather than the liability scale. Default = F

option specifies different ways of estimating parameters. Option=0 using maximum likelihood, as in Palla & Dudbridge (2015). Other methods were used in development and are retained for reference. Option=1 uses minimum squared error of the non-centrality parameter of the χ^2 tests; option=2 uses minimum squared error of the non-centrality parameter of the Z tests; option=3 uses maximum likelihood based on non-central χ^2 distributions rather than normal distributions as in Palla & Dudbridge (2015). Default = 0

boot number of bootstrap simulations used to estimate 95% confidence intervals for estimated parameters. A parametric bootstrap is used in which Z-scores are simulated from the distributions defined by the estimated parameters. Z-scores are simulated independently in each interval defined by pupper. For each simulated set of Z-scores, parameters are re-estimated and the 2.5% and 97.5% quantiles of their empirical distributions used to define the bootstrap confidence intervals. If boot = 0, analytic confidence intervals are calculated using profile likelihood, as described in Palla & Dudbridge (2015). Default = 0

bidirectional

if T, bidirectional estimation is performed in which the roles of training and target samples are reversed, allowing vg and pi0 to be estimated in both samples simultaneously. In this case, argument p should contain twice as many P-values as the number of intervals defined by pupper, with the results for the original training/target direction coming first, followed by those for the reversed direction. Also, arguments vg and pi0 should now be vectors with 2 elements; if only one value is given for either argument, the second element will be equated to the first. Default =F initial vector of initial values for the estimated parameters when numerically maximising 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. So if we wish to estimate pi0[1] and cov12, then initial will have two elements corresponding to the initial values of pi0[1] and cov12. Default = 0.5 for all parameters, which tends to work well but it is worth trying other realistic values for the parameters of interest to ensure that the likelihood has indeed been maximised

fixvg2pi02

if T, constrains cov12=vg[1], and if bidirectional=T also constrains vg[2]=vg[1] and pi0[2]=pi0[1]. Thus if fixvg2pi02=T, at most two parameters can be estimated, equivalent to assuming the same genetic model in both training and target samples. Can be abbreviated to fix. Default = F

Value

A list with the following elements

vg if bidirectional=F, a vector with 3 elements consisting of the estimated vg and its lower and upper 95% confidence limits. If the value of vg is fixed in the function call, this is returned again in vg. If bidirectional=T, a matrix with two rows corresponding to the training and target samples. Each row is a vector with 3 elements as above.

cov12 a vector with 3 elements consisting of the estimated cov12 and its lower and upper 95% confidence limits. If the value of cov12 is fixed in the function call, this is returned again in cov12.

pi0 if bidirectional=F, a vector with 3 elements consisting of the estimated pi0 and its lower and upper 95% confidence limits. If the value of pi0 is fixed in the function call, this is returned again in pi0. If bidirectional=T, a matrix with two rows corresponding to the training and target samples. Each row is a vector with 3 elements as above.

logLikelihood

value of the maximised log-likelihood

error error message