mvtests Software Documentation

April 1, 2019

Version 0.2

Date April 1, 2019

Title Multivariate Tests of Association for Multiple Phenotypes

Correspondence Debashree Ray, Ph.D. `<dray@jhu.edu>`

Description Currently, mvtests package implements a single-variant genetic association test of multiple phenotypes using a proportional odds regression model of genotype on phenotypes.

Depends MASS, lmtest, R (>= 3.0.1)

`pom` *Genetic association test of multiple phenotypes using proportional odds model*

Description

POM uses a proportional odds model for testing association between a genetic marker and multiple traits using individual-level genotype phenotype data. The traits may be continuous and/or binary, correlated and/or independent. One may use likelihood ratio test (LRT) or the Wald test. The R function `pom` implements this association test.

Usage

```
pom(Y, X, COV=NULL, test.method="LRT",

     msg.mute=FALSE, no.format.check=FALSE, ...)
```

Arguments

`Y` The $n \times K$ phenotype matrix, where n is the number of individuals and K is the number of phenotypes. The joint association of all K phenotypes with the single marker will be tested. `Y` needs to be in R data frame format.

`X` The $n \times 1$ column matrix for the single genetic marker, where n is the number of individuals. `X` needs to be in R data frame format.

`COV` The $n \times q$ matrix of covariates that need to be adjusted in the model. q is the number of such covariates. `COV` needs to be in R data frame format. The default value is `NULL` assuming there is no covariate in the model.

`test.method` Either the likelihood ratio test (`LRT`) or the Wald test (`Wald`) is used for testing genetic association of the traits. Default value is `LRT`.

`msg.mute` Default value is `FALSE`, which allows messages to print when the analysis is in progress.

`no.format.check` Default value is `FALSE`, which allows the code to check if the input parameters conform to the required format.

`...` Additional arguments to be passed to `polr()` (used when proportional odds model framework is used) or `glm()` (used when logistic model framework is used).

Details

For testing joint association of multiple phenotypes $\boldsymbol{Y}$ ($n \times K$ matrix) and a single genetic marker $\boldsymbol{X}$ ($n \times 1$ matrix), one can consider a reverse regression approach: regressing $\boldsymbol{X}$ on $\boldsymbol{Y}$ assuming a proportional odds model (POM). For a given individual with genotype X (taking values $0, 1$ or 2) and phenotypes $\boldsymbol{y}$ (binary and/or continuous) and no covariate (for simplicity), the assumed model is

$$\text{logit}\left(\mathbb{P}(X \leq j|\boldsymbol{y})\right) = \alpha_j + \boldsymbol{\beta}'\boldsymbol{y}, \;\; j = 0, 1$$

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_K)'$ are the genetic effects of interest. This model can also be written as

$$\mathbb{P}\left(X = 0|\boldsymbol{y}\right) \;\; = \;\; \frac{1}{1 + e^{-\alpha_0 - \sum_{k=1}^{K} \beta_k y_k}}$$

$$\begin{aligned}\mathbb{P}\left(X=1|\boldsymbol{y}\right) &= \frac{1}{1+e^{-\alpha_1-\sum_{k=1}^{K}\beta_k y_k}}-\frac{1}{1+e^{-\alpha_0-\sum_{k=1}^{K}\beta_k y_k}}\\ \mathbb{P}\left(X=2|\boldsymbol{y}\right) &= \frac{1}{1+e^{\alpha_1+\sum_{k=1}^{K}\beta_k y_k}}\end{aligned}$$

For a genetic variant with low allele frequency, it is possible for X to take only 2 values (instead of 3). In that case, a logistic regression is used.

The null hypothesis of no association of the genetic marker (say, single nucleotide polymorphism or SNP) with the phenotypes is $H_0 : \beta_1 = ... = \beta_K = 0$. The `test.method="LRT"` (default) implements likelihood ratio test to test H_0 against the alternative hypothesis H_a : $\beta_k \neq 0$ for at least one k. Under the null H_0, the LRT statistic has an asymptotic chi-squared distribution with K degrees of freedom.

When `test.method="Wald"`, the test statistic is $\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\Sigma}}_\beta\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimate (MLE) of $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\Sigma}}_\beta$ is the estimated variance-covariance matrix of estimates $\hat{\boldsymbol{\beta}}$. Under the null H_0, the Wald test statistic, too, has an asymptotic chi-squared distribution with K degrees of freedom. Additionally, a version of the Wald test using F distribution is implemented using the `lmtest` R package, which requires fitting the full model as well as the null model (model under H_0).

We request that the reference for Ray and Chatterjee (2019+) be cited if this software is used in any publication. In the Ray and Chatterjee (2019+) article, we explored advantages and pitfalls of some of the currently used single-variant cross-phenotype methods in genome-wide analysis of rare, low-frequency and common variants when the basic assumption of multivariate normality is not satisfied. Based on our findings, we recommend extra caution when applying cross-phenotype association tests in GWAS with low-frequency or rare variants due to possible violation of multivariate normality assumption. However, we found that robust association testing is still possible for variants with minor allele count (MAC) > 30 by applying the POM-LRT approach (`pom()` with `test.method="LRT"`) when individual-level data are available. Our recommendation is based on an MAC threshold (instead of an minor allele frequency - MAF - threshold as is commonly used) because we found consistent type I error calibration of methods when the MAC is kept constant.

Value

`coef` The estimated coefficients in the model (e.g., α_0, α_1, β_1,...,β_K). If `COV` is not `NULL`, estimated coefficients for covariates are also included.

`SE.coef` The estimated standard errors of the coefficient estimates.

`stat.lrt` (If `test.method="LRT"`) The LRT statistic for testing the null hypothesis H_0.

`df.lrt` (If `test.method="LRT"`) The degrees of freedom of the asymptotic null distribution of the LRT statistic.

`pvalue.lrt` (If `test.method="LRT"`) The p-value from testing H_0 using LRT.

`stat.wald.chisq` (If `test.method="Wald"`) The Wald test statistic for testing the null hypothesis H_0.

`df.wald.chisq` (If `test.method="Wald"`) The degrees of freedom of the asymptotic chi-squared distribution of the Wald test statistic under H_0.

`pvalue.wald.chisq` (If `test.method="Wald"`) The p-value from testing H_0 using Wald test assuming chi-squared distribution.

`stat.wald.F` (If `test.method="Wald"`) The Wald test statistic for testing the null hypothesis H_0 using the F distribution approximation.

`pvalue.wald.F` (If `test.method="Wald"`) The p-value from testing H_0 using Wald test assuming F distribution.

`n.obs` Number of individuals used for testing association. Individuals with missing observations in `Y`, `X` or `COV` are removed.

`error.msg` Saves any error message that arises during the analysis. If no error is encountered, message `"OK"` is returned.

Reference

Ray, D., Chatterjee, N. Effect of Non-Normality and Low Count Variants on Cross-Phenotype Association Tests in GWAS. *In revision.* 2019.
(Contact `dray@jhu.edu` for updated citation)

Example

```
source("mvtests_v0.2.R")
set.seed(1)
# simulate 2 phenotypes on 1000 individuals
Y<-mvrnorm(n=1000, mu=c(0,0), Sigma=matrix(c(1,0.2,0.2,1),2,2))
# simulate a single marker for 1000 individuals
X<-matrix(rbinom(n=1000, size=2, prob=0.2), ncol=1) # additive model
# required data-frame formats
Y<-as.data.frame(Y)
X<-as.data.frame(X)
# unique column names for the data-frames
colnames(Y)<-paste("Y",1:2,sep="")
colnames(X)<-"X"
## apply POM to test association of X with Y1 and Y2
out1<-pom(Y=Y, X=X, COV=NULL, test.method="LRT")
out1
out2<-pom(Y=Y, X=X, COV=NULL, test.method="Wald")
out2
## optimization parameters may be changed
## (e.g., when convergence issues come up)
out3<-pom(X=X, Y=Y, test.method="Wald", control=list(maxit=1000))
out3
```