POM-PS Software Documentation

February 28, 2017

Version 1.15

Date February 28, 2017

Title Proportional Odds Model adjusted for Propensity Score

Correspondence Debashree Ray, Ph.D. <debashr@umich.edu>;
Saonli Basu, Ph.D. <saonli@umn.edu>

Description POM-PS uses a likelihood ratio test in a proportional odds model framework to test association of secondary phenotypes (from a case-control study) with a single marker.

Depends MASS, survey, R (>= 2.14.0)

`pomps` *Proportional Odds Model adjusted for Propensity Score*

Description

POM-PS uses a proportional odds model to test genetic association of one or more secondary phenotypes from a case-control study. It adjusts propensity score (estimated probability of being a case) as a covariate and uses a likelihood ratio test (LRT) for testing the null hypothesis of no association. It is designed for unrelated individuals. The R function `pomps` implements this association test.

Usage

```
pomps(Y, X, D=NULL, COV=NULL, PS=NULL, method="POM-PS",
     add.D.as.COV=FALSE, 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 secondary 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.
D	The $n \times 1$ column matrix for the case-control status (primary phenotype), where n is the number of individuals. D 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, i.e., it is assumed there is no covariate in the model. The propensity score, although adjusted as a covariate, is input separately and not included in COV.
PS	The $n \times 1$ column matrix for the propensity score (estimated probability of being a case), where n is the number of individuals. It can be calculated using the getPS() function. PS needs to be in R data frame format.
method	The method to be used for testing genetic association of traits. Default value is POM-PS. The other possible value is POM, which tests the genetic association of traits in a proportional odds framework without any adjustment of the propensity score. POM may be used when the traits are from a random sample.
add.D.as.COV	Default value is FALSE. If TRUE, the case-control status D is adjusted as a covariate.
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), the assumed model is

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

This model can also be written as

$$\begin{aligned}
\mathbb{P}\left(X = 0|\boldsymbol{y}\right) &= \frac{1}{1 + e^{-\alpha_0 - \sum_{k=1}^{K} \beta_k y_k}} \\
\mathbb{P}\left(X = 1|\boldsymbol{y}\right) &= \frac{1}{1 + e^{-\alpha_1 - \sum_{k=1}^{K} \beta_k y_k}} - \frac{1}{1 + e^{-\alpha_0 - \sum_{k=1}^{K} \beta_k y_k}} \\
\mathbb{P}\left(X = 2|\boldsymbol{y}\right) &= \frac{1}{1 + e^{\alpha_1 + \sum_{k=1}^{K} \beta_k y_k}}
\end{aligned}$$

The null hypothesis of no association of the genetic marker (say, single nucleotide polymorphism or SNP) with the phenotypes is $H_0 : \beta_1 = ... = \beta_K = 0$. LRT is used to test H_0 and is implemented by `method="POM"`. Under the null, the LRT statistic has an asymptotic chi-squared distribution with K degrees of freedom.

When the phenotypes are collected from a case-control sample (an ascertained sample as opposed to a random sample), one needs to be mindful of the ascertainment bias that may arise from the over-representation of cases in the sample. When testing H_0, this may result in numerous spurious association signals (inflated type I error). To control for type I error, Ray and Basu (2017) proposed adjusting the estimated propensity score in the proportional odds model for genotype on the secondary phenotype(s). If S be the propensity score for an individual and $\boldsymbol{S}$ be the $n \times 1$ vector of propensity scores for the sample, the assumed model for POM-PS is

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

LRT is used to test H_0 and is implemented by `method="POM-PS"`. The LRT statistic has an asymptotic $\chi^2_{(K)}$ distribution under H_0. Note that the propensity scores $\boldsymbol{S}$ can be calculated using the function `getPS()`. Since the estimation of propensity score depends only on the secondary phenotypes, the case-control status and other relevant covariates, it

needs to be calculated only once for a given dataset.

The framework for POM-PS can accommodate one or more secondary phenotypes (binary and/or continuous) from a case-control genome-wide association study (GWAS). Since the phenotypes are treated as predictors in the model, POM-PS does not assume multivariate normality of the phenotypes. Unlike other existing approaches, it does not require knowledge of disease prevalence. The fast & simple implementation of POM-PS makes it applicable to large case-control GWASs. It took 52 minutes to test $10,530$ single marker associations on a real dataset with $n = 2,762$ individuals measured on $K = 4$ phenotypes using a single core of an Intel(R) Xeon(R) CPU X5660 @2.80GHz processor.

For details, please refer Ray and Basu (2017). We request that the reference for Ray and Basu (2017) be cited if this software is used in any publication.

Value

`coef`	The coefficients in the model (e.g., α_0, α_1, β_1,...,β_K,δ).
`stat`	The LRT statistic for testing the null hypothesis H_0.
`df`	The degrees of freedom of the asymptotic null distribution.
`pvalue`	The p-value from testing H_0.
`n.obs`	Number of individuals used for testing association. Individuals with missing observations in `Y`, `X`, `D`, `PS` or `COV` are removed.
`error.msg`	Saves any error message that arises during the analysis. If no error is encountered, message `"OK"` is returned.

Note

The genotype X takes values 0, 1 or 2. For a given sample, if X has only two possible values, a logistic regression (`glm`) is used instead of proportional odds regression (`polr`).

The current version can not take user-specified value for the parameter `start` (starting values of regression coefficients for optimization) for function `optim()` used within `polr()` or `glm()`. This is because LRT is used for testing association within `pomps()` that involves two different models: full model and the null (reduced) model.

Reference

Ray, D. and Basu, S. (2017). A Novel Association Test for Multiple Secondary Phenotypes from a Case-Control GWAS. *Genetic Epidemiology*, 41, DOI:10.1002/gepi.22045.

Example

```
source("pomps_v1.15.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 3 covariates for disease status
Sig<-matrix(c(1,0.1,-0.45,0.1,1,0.6,-0.45,0.6,1),3,3)
Z<-mvrnorm(n=1000, mu=c(0,0,0), Sigma=Sig)
# simulate error for disease status
E<-rt(n=1000, df=1,ncp=0)
# simulate disease status
D<-rep(0,1000)
D[which(rowSums(cbind(Y,Z,E))>=0)]<-1
# 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)
D<-as.data.frame(D)
Z<-as.data.frame(Z)
X<-as.data.frame(X)
# unique column names for the data-frames
colnames(Y)<-paste("Y",1:2,sep="")
colnames(Z)<-paste("Z",1:3,sep="")
colnames(X)<-"X"
colnames(D)<-"D"
## apply POM to test association
out1<-pomps(Y=Y, X=X, COV=NULL, method="POM")
# POM test statistic and p-value
t1<-out1$stat
p1<-out1$pvalue
## apply POM-PS to test association
S=getPS(Y=Y, D=D, covars=Z)
out2<-pomps(Y=Y, X=X, COV=NULL, PS=S)
# POM-PS test statistic and p-value
t2<-out2$stat
p2<-out2$pvalue
```

`getPS` *Function for calculating propensity score*

Description

The R function `getPS` calculates propensity score using secondary phenotypes, case-control status and other covariates relevant for the disease status. The propensity score, in this context, is defined as the estimated conditional probability of being diseased. This needs to be calculated only once for a given GWAS.

Usage

```
getPS(Y, D, covars=NULL, 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 secondary phenotypes. The joint association of all K phenotypes with the single marker will be tested. `Y` needs to be in R data frame format.
`D`	The $n \times 1$ column matrix for the case-control status (primary phenotype), where n is the number of individuals. `D` needs to be in R data frame format.
`covars`	The $n \times q$ matrix of covariates that are relevant for predicting disease status. q is the number of such covariates. `covars` needs to be in R data frame format. The default value is `NULL`, i.e., it is assumed there is no covariate in the model.
`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 `glm()` for estimating the probability of being diseased.

Value

An $n \times 1$ R data frame of propensity scores for all individuals in the sample.

Note

`getPS` works for a binary primary phenotype (such as case-control status). If the primary phenotype has three or more categories, one may use a proportional odds regression for

calculating the propensity score[1].

`pom.ipw` *Inverse probability weighted proportional odds model*

Description

This R function implements a inverse probability weighted proportional odds regression model for testing genetic association of one or more phenotypes measured on unrelated individuals.

Usage

```
pom.ipw(Y, X, D=NULL, COV=NULL, PS=NULL, weights=NULL,
     method="POM-IPSW", 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 secondary 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.
`D`	The $n \times 1$ column matrix for the case-control status (primary phenotype), where n is the number of individuals. `D` 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`, i.e., it is assumed there is no covariate in the model. The propensity score is not adjusted as a covariate and should not be included in `COV`.

[1]Joffe, M. M. and Rosenbaum, P. R. (1999). Invited commentary: propensity scores. *American Journal of Epidemiology*, 150:327-333.

PS	The $n \times 1$ column matrix for the propensity score (estimated probability of being a case), where n is the number of individuals. It can be calculated using the getPS() function. PS needs to be in R data frame format.
weights	A vector of size n of inverse probability weights (or sampling weights). It will be used with method="POM-IPW" only. Default value is NULL. PS and weights will not be simultaneously used; method determines which will be used.
method	The method to be used for modeling the genetic association of secondary phenotypes. Default value is POM-IPSW, where inverse propensity scores are used as weights (i.e., each subject's weight is equal to the inverse of his/her estimated propensity score). The other possible value is POM-IPW, which uses user-specified probability weights. In both cases, proportional odds model framework is used when there are > 2 categories for the genotype.
test.method	The method for testing association: LRT (default) or Wald. Note that Wald chi-squared test is much faster than LRT; however, LRT is more accurate. Also, test.method="LRT" will not work if starting values of regression coefficients (i.e., parameter start of optim() used within svyolr/svyglm) are supplied.
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 svyolr() (used when proportional odds model framework is used) or svyglm() (used when logistic model framework is used).

Details

Inverse probability weighting using propensity scores is a form of model-based direct standardization[2] that is traditionally used to estimate average treatment effect by comparing

[2]Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, 82:387-394.

treated and untreated groups in a non-randomized study. In a similar fashion, one may consider such a weighting scheme within the framework of a proportional odds model, which is implemented by `pom.ipw()` with `method="POM-IPSW"`. Ray and Basu (2017) compared the performances of POM-IPSW and POM-PS on simulated datasets. It was found that POM-IPSW can not control for type I error arising out of over-representation of cases in a case-control (ascertained) sample. Hence, POM-IPSW is not recommended for testing genetic association of secondary phenotype(s).

One may use other pertinent weights for each subject in the sample (such as sampling weights). This is implemented by `pom.ipw()` with `method="POM-IPW"`. Ray and Basu (2017) did not extensively study the performance of such an inverse-probability-weighted proportional odds model.

Value

`coef`	The coefficients in the weighted model.
`stat`	The LRT statistic for testing the null hypothesis H_0.
`df`	The degrees of freedom of the asymptotic null distribution.
`pvalue`	The p-value from testing H_0.
`test.method`	The test method used.
`n.obs`	Number of individuals (with complete observations) used for testing association. Individuals with missing observations in `Y`, `X`, `D`, `PS` or `COV` are removed.
`error.msg`	Saves any error message that arises during the analysis. If no error is encountered, message `"OK"` is returned.

Note

`pom.ipw()` uses function `svyolr()` or `svyglm()` from R package `survey`[3,4] to perform weighted ordinal regression. Please be aware that these functions often encounter convergence problems with starting values of regression coefficients for optimization, and errors in saddlepoint approximation of LRT. `pom.ipw()` can be technically used to implement POM-PS. However, one may encounter the afore-mentioned convergence issues with `pom.ipw()`, and hence not recommended.

[3]Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9:1-19.
[4]Lumley, T. (2016). survey: analysis of complex survey samples. R package version 3.31.

Example

```
# continuation of last example
## apply POM-IPSW to test association
out3<-pom.ipw(Y=Y, X=X, D=D, PS=S, COV=NULL)
## implement POM-PS using pom.ipw()
out4<-pom.ipw(Y=Y, X=X, COV=S, weights=rep(1,1000), method="POM-IPW")
```