r2redux

Moksedul Momin and Hong Lee

01/08/2022

# r2redux

The ‘r2redux’ package can be used to derive test statistics for R^2 values from polygenic risk score (PGS) models (variance and covariance of R^2 values, p-value and 95% confidence intervals (CI)). For example, it can test if two sets of R^2 values from two different PGS models are significantly different to each other whether the two sets of PGS are independent or dependent. Because R^2 value is often regarded as the predictive ability of PGS, r2redux package can be useful to assess the performances of PGS methods or multiple sets of PGS based on different information sources. Furthermore, the package can derive the information matrix of β ̂\_1^2and β ̂\_2^2 from a multiple regression (see olkin\_beta1\_2 or olkin\_beta\_info function in the manual), which is a basis of a novel PGS-based genomic partitioning method (see r2\_enrich or r2\_enrich\_beta function in the manual). It is recommended that the target sample size in the PGS study should be more than 2,000 for quantitative traits (Figure S27) and more than 5,000 for binary responses or case-control studies (Figures S28 and S29). The P-value generated from the r2redux package provides two types of p-values (for one- and two-tailed test) unless the comparison is for nested models (e.g. y=〖PGS〗\_1+〖PGS〗\_2+e vs. y=〖PGS〗\_2+e) where the R^2 of the full model is expected to be always higher than the reduced model. When there are multiple covariates (e.g. age, sex and other demographic variables), the phenotypes can be adjusted for the covariates, and pre-adjusted phenotypes (residuals) should be used in the r2redux.

# INSTALLATION

To use r2redux: - install.packages(“r2redux”) - library(r2redux) or - install.packages(“devtools”) - library(devtools) - devtools::install\_github(“mommy003/r2redux”) - library(r2redux)

# QUICK START

We illustrate the usage of r2redux using multiple sets of PGS estimated based on GWAS summary statistics from UK Biobank or Biobank Japan (reference datasets). In a target dataset, the phenotypes of target samples (y) can be predicted with PGS (a PGS model, e.g. y=PRS+e, where y and PGS are column-standardised 1. Note that the target individuals should be independent from reference individuals. We can test the significant differences of the predictive ability (R^2) between a pair of PGS (see r2\_diff function and example in the manual).

# DATA PREPARATION

**a. Statistical testing of significant difference between R2 values for p-value thresholds:** r2redux requires only phenotype and estimated PGS (from PLINK or any other software). Note that any missing value in the phenotypes and PGS tested in the model should be removed. If we want to test the significant difference of R^2 values for p-value thresholds, r2\_diff function can be used with an input file that includes the following fields (also see test\_ukbb\_thresholds\_scaled in the example directory form github (<https://github.com/mommy003/r2redux>) or read dat1 file embedded within the package and r2\_diff function in the manual).

* Phenotype (y)
* PGS for p value 1 (x1)
* PGS for p value 0.5 (x2)
* PGS for p value 0.4 (x3)
* PGS for p value 0.3 (x4)
* PGS for p value 0.2 (x5)
* PGS for p value 0.1 (x6)
* PGS for p value 0.05 (x7)
* PGS for p value 0.01 (x8)
* PGS for p value 0.001 (x9)
* PGS for p value 0.0001 (x10)

To get the test statistics for the difference between R2(y=x[,v1]) and R2(yx[,v2]). (here we define R\_1^2= R^2(y=x[,v1])) and R\_22=R2(y=x[,v2]))) - dat=read.table(“test\_ukbb\_thresholds\_scaled”) (see example files) or - dat=dat1 (this example embedded within the package) - nv=length(dat$V1) - v1=c(1) - v2=c(2) - output=r2\_diff(dat,v1,v2,nv)

* r2redux output
* output$var1 (variance of R\_1^2)
* 0.0001437583
* output$var2 (variance of R\_2^2)
* 0.0001452828
* output$var\_diff (variance of difference between R\_1^2and R\_2^2)
* 5.678517e-07
* output$r2\_based\_p (p-value for significant difference between R\_1^2 and R\_2^2)
* 0.5514562
* output$mean\_diff (differences between R\_1^2 and R\_2^2)
* -0.0004488044
* output$upper\_diff (upper limit of 95% CI for the difference)
* 0.001028172
* output$lower\_diff (lower limit of 95% CI for the difference)
* -0.001925781

**b. PGS-based genomic enrichment analysis:** If we want to perform some enrichment analysis (e.g., regulatory vs non\_regulatory) in the PGS context to test significantly different from the expectation (p\_exp= # SNPs in the regulatory / total # SNPs = 4%). We simultaneously fit two sets of PGS from regulatory and non-regulatory to get β ̂\_regu^2 and β ̂\_(non-regu)^2, using a multiple regression, and assess if the ratio, (β ̂\_1^2)/(r\_(y,〖(x〗\_1,x\_2))^2 ) are significantly different from the expectation, p\_exp. To test this, we need to prepare input file for r2redux that includes the following fields (e.g. test\_ukbb\_enrichment\_choles in example directory or read dat2 file embedded within the package and r2\_enrich\_beta function in the manual).

* Phenotype (y)
* PGS for regulatory region (x1)
* PGS for non-regulatory region (x2)

To get the test statistic for the ratio which is significantly different from the expectation. var(β ̂\_12/r\_(y,(x\_1,x\_2))2), where β ̂\_1^2 is the squared regression coefficient of x\_1 from a multiple regression model, i.e. y=x\_1 β\_1+ x\_2 β\_2+e, and r\_(y,(x\_1,x\_2))^2 is the coefficient of determination of the model. It is noted that y, x\_1 and x\_2 are column standardised (mean 0 and variance 1).

* dat=read.table(“test\_ukbb\_enrichment\_choles”) (see example file) or
* dat=dat2 (this example embedded within the package)
* nv=length(dat$V1)
* v1=c(1)
* v2=c(2)
* dat=dat2
* nv=length(dat$V1)
* v1=c(1)
* v2=c(2)
* output=r2\_beta\_var(dat,v1,v2,nv)
* r2redux output
* output$beta1\_sq (beta1^2)
* 0.01118301
* output$beta2\_sq (beta2^2)
* 0.004980285
* output$var1 (variance of beta1^2)
* 7.072931e-05
* output$var2 (variance of beta2^2)
* 3.161929e-05
* output$var1\_2 (variance of difference between beta1^2 and beta2^2)
* 0.000162113
* output$cov (covariance between beta1^2 and beta2^2)
* -2.988221e-05
* output$upper\_beta1\_sq (upper limit of 95% CI for beta1^2)
* 0.03037793
* output$lower\_beta1\_sq (lower limit of 95% CI for beta1^2)
* -0.00123582
* output$upper\_beta2\_sq (upper limit of 95% CI for beta2^2)
* 0.02490076
* output$lower\_beta2\_sq (lower limit of 95% CI for beta2^2)
* -0.005127546
* dat=dat2 (this example embedded within the package)
* nv=length(dat$V1)
* v1=c(1)
* v2=c(2)
* expected\_ratio=0.04
* output=r2\_enrich\_beta(dat,v1,v2,nv,expected\_ratio)
* r2redux output
* output$beta1\_sq (beta1^2)
* 0.01118301
* output$beta2\_sq (beta2^2)
* 0.004980285
* output$ratio1 (beta12/R2)
* 0.4392572
* output$ratio2 (beta22/R2)
* 0.1956205
* output$ratio\_var1 (variance of ratio 1)
* 0.08042288
* output$ratio\_var2 (variance of ratio 2)
* 0.0431134
* output$upper\_ratio1 (upper limit of 95% CI for ratio 1)
* 0.9950922
* output$lower\_ratio1 (lower limit of 95% CI for ratio 1)
* -0.1165778
* output$upper\_ratio2 upper limit of 95% CI for ratio 2)
* 0.6025904
* output$lower\_ratio2 (lower limit of 95% CI for ratio 2)
* -0.2113493
* output$enrich\_p1 (two tailed P-value for beta12/R2 is significantly different from exp1)
* 0.1591692
* output$enrich\_p1\_one\_tail (one tailed P-value for beta12/R2 is significantly different from exp1)
* 0.07958459
* output$enrich\_p2 (P-value for beta2^2/R2 is significantly different from (1-exp1))
* 0.000232035
* output$enrich\_p2\_one\_tail (one tailed P-value for beta2^2/R2 is significantly different from (1-exp1))
* 0.0001160175

# References

1. Olkin, I. and Finn, J.D. Correlations redux. Psychological Bulletin, 1995. 118(1): p. 155.
2. Momin, M.M., Lee, S., Wray, N.R. and Lee S.h. 2022. Significance tests for R2 of out-of-sample prediction using polygenic scores. bioRxiv.

# Contact information

Please contact Hong Lee ([hong.lee@unisa.edu.au](mailto:hong.lee@unisa.edu.au)) or Moksedul Momin ([cvasu.momin@gmail.com](mailto:cvasu.momin@gmail.com)) if you have any queries.