On the impact of relatedness on SNP association analysis

Arnd Gross, Markus Scholz

Institute for Medical Informatics, Statistics and Epidemiology (IMISE)

Background

When testing for SNP (single nucleotide polymorphism) associations in related individuals, observations are not independent and simple linear regression analysis results in an increased type I error and the power of the test is also affected in a more complicate manner. We investigate how heritability and strength of relatedness contribute to variance inflation of the effect estimate and study the consequences of variance inflation on hypothesis testing.

Modelling a SNP-Phenotype association

True model

Phenotypes follow the mixed model

$$\mathbf{y} = b_1 + b_2\mathbf{s} + \mathbf{g} + \mathbf{e}$$

with

- phenotypes $\mathbf{y}$ for n samples
- intercept b_1, genetic effect b_2
- SNP genotypes $\mathbf{s}$, $s_i \in \{0, 1, 2\}$
- polygenic random effects $\mathbf{g} \sim \mathrm{N}_n(0, \sigma^2_\mathrm{g}\mathbf{G})$, variance σ^2_g, relatedness matrix $\mathbf{G}$
- uncorrelated residuals $\mathbf{e} \sim \mathrm{N}_n(0, \sigma^2_\mathrm{e}\mathbf{I})$, variance σ^2_e, identity matrix $\mathbf{I}$.

Simplified model

Phenotypes are analysed with the model

$$\mathbf{y} = \beta_1 + \beta_2\mathbf{s} + \boldsymbol{\epsilon}$$

assuming uncorrelated residuals $\boldsymbol{\epsilon}$ only. It can be shown that $\mathrm{E}(\hat{\beta}_2) = b_2$ and

$$\mathrm{V}(\hat{\beta}_2) = \frac{\lambda}{1 - R^2_\mathrm{h}} V_\beta$$

with

- inflation factor λ
- heritability R^2_h
- variance V_β of $\hat{\beta}_2$ without heritability.

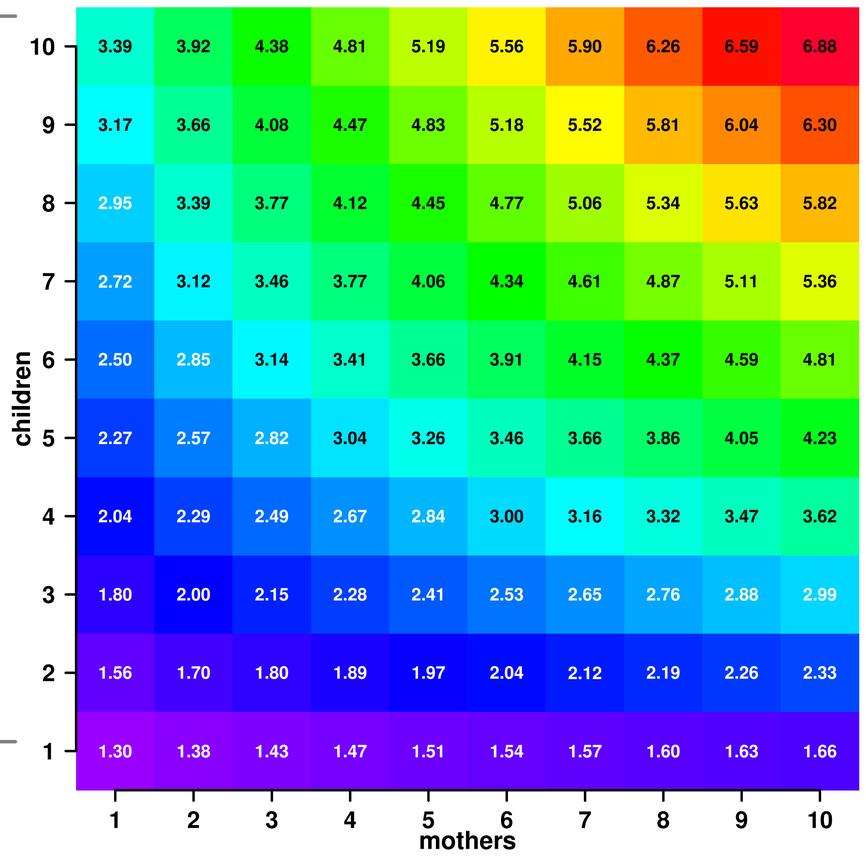
Inflation factor

Expected variance inflation is

$$\lambda = 1 + R^2_\mathrm{h} \frac{\sum_i \sum_{j \neq i} G^2_{ij} - \frac{2}{n} \sum_i \left(\sum_{j \neq i} G_{ij} \right)^2}{n - 1}$$

Properties are

- stronger relatedness increases inflation
- higher heritability increases inflation
- inflation is independent from allele frequency.



The figure presents the expected variance inflation for 90% heritability and family studies with varying numbers of children per mother and mothers per family/father. The number of families is constrained by *n*=1000 individuals.

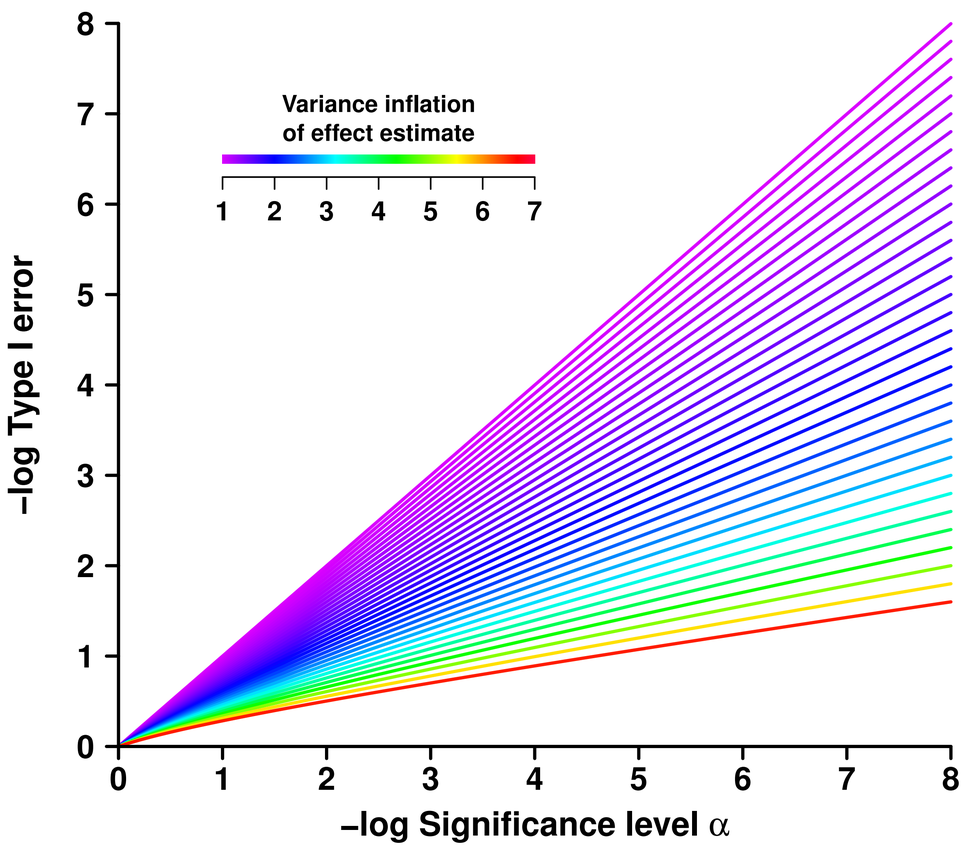
Hypothesis testing

Type I error

Under the null hypothesis $b_2 = 0$, it approximately holds that

$$T \sim \mathrm{N}(0, \lambda)$$

for test statistic $T = \hat{\beta}_2 / S_\beta$ with empirical variance S^2_β of $\hat{\beta}_2$.



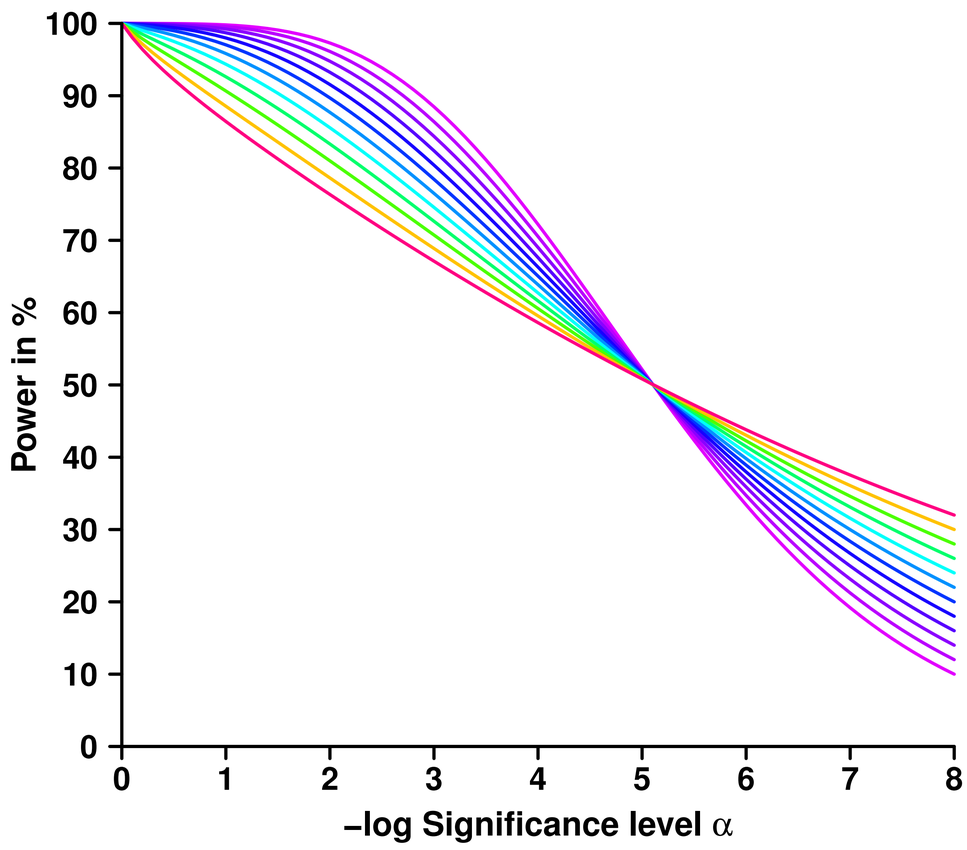
Comparison of type I errors with respect to different degrees of variance inflation. The negative common logarithm is presented for significance level α as well as the type I error.

Power of test

Under the alternative hypothesis $b_2 \neq 0$, it approximately holds that

$$T \sim \mathrm{N}(\sqrt{(n-1)R^2_\mathrm{s}}, \lambda)$$

with sample size n and explained variance by the SNP R^2_s.



Comparison of power with respect to different degrees of variance inflation assuming *n*=1000 and 2% explained variance by the SNP. The negative common logarithm is presented for significance level α.

Conclusions

We provide a simple formula for estimating variance inflation given the relatedness structure and the heritability of a phenotype. Stronger relatedness as well as higher heritability result in increased variance of the effect estimate of simple linear regression. While type I error rates are generally inflated, the behaviour of power is more complex since power can be increased or reduced in dependence on the significance level. For additional information, have a look at https://bmcgenet.biomedcentral.com/articles/10.1186/s12863-017-0571-x or use the QR Code: