GWAS 8: Heritability and mixed models

Matti Pirinen, University of Helsinki

Latest update: April 2, 2025.

This document is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

The slide set referred to in this document is "GWAS 8".

We have seen Manhattan plots for BMI, migraine, schizophrenia etc. where the number of association regions rises to hundreds. How much these regions explain of the population variation in a trait or disease liability? How many other genomic regions, or additional variants in the current regions, might also contribute to these phenotypes? Is there something special about the genomic regions that harbor causal variants for a particular phenotype? How well can we predict these phenotypes from genetic data and how much even larger GWAS are likely to improve these predictions?

These are questions about **genetic architecture** of the phenotypes. A primary parameter of genetic architecture is **heritability**.

8.1 Heritability

Heritability measures, in a particular population, the proportion of variance of the phenotype that is due to genetic differences between individuals. A review by Visscher et al. is an excellent overview of the concept and its interpretation.

An important point is that heritability is always a property of a particular population, and its value can vary between different environmental conditions as the relative roles of genes and environment change. For example, when the level of nutrition is fairly equal between individuals, then the heritability of adult height is likely high because the observed variation is largely caused by genetics. If, instead, the population is strongly divided with respect to the level of nutrition, then the heritability may be lower because now the variation in the level of nutrition causes large variation in height.

Heritabilities have been estimated for a long time using phenotypic correlations between relatives of different degrees. (Slides 3-4) For example, the traditional twin estimates compare phenotypic similarity in monozygotic twin pairs (who share whole genome IBD = 2) to similarity in dizygotic twin pairs (who are genetically like any other pairs of full-siblings: IBD0 = 25%, IBD1 = 50%, IBD2 = 25%). Under some (strong!) assumptions, such as that the environmental contribution to the phenotypic similarity would be the same for a monozygotic as for a dizygotic twin pair, and that the genetic effects act additively over loci, it follows that the difference between the correlation estimates of these two types of twin pairs leads to an estimate of heritability.

More generally, any pedigree records can been used for estimating how the observed phenotypic correlations can be explained by the estimated genetic sharing, which, under (strong!) assumptions about covariance of the environmental effects between different relative types, again leads to estimates of heritability. However, since the close relatives also tend to share environment to some degree, it may be difficult to accurately separate the effects of the shared genetics from that of the shared environment using data on only close relatives.

Here our interest is in how much heritability we can explain with the already discovered GWAS regions, and whether we could use GWAS data, that typically have only few close relative pairs, to estimate heritability of

a trait. In particular, such estimates should be immune to the confounding factors of the shared environment that occurs between close relatives.

8.2 Heritability of GWAS loci

We consider the measure of **heritability in the narrow sense**, h^2, which is the heritability due to the **additive genetic effects**. Thus, *dominance effects* within one locus (i.e. the amount by which the heterozygotes' phenotype mean deviates from the average of the two groups of homozygotes), or interaction effects (*epistasis*) between multiple loci, are not included in this estimate. The broad-sense heritability H^2, which includes all variance due to genetics, is more difficult to estimate than h^2.

If SNP l has MAF f_l and allelic causal effect λ_l, then the phenotypic variance explained (causally) by the SNP is $2f_l(1-f_l)\lambda_l^2 = \lambda_l^{*2}$. (Here λ_l^* denotes the scaled causal effect introduced in Section 7.) If the trait variance in population is σ_Y^2, then the (additive) heritability contributed by the SNP is

$$h_l^2 = \frac{\lambda_l^{*2}}{\sigma_Y^2} = \frac{2f_l(1-f_l)\lambda_l^2}{\sigma_Y^2}.$$

In practice, the variance explained is often estimated using an estimate of the marginal effect as $(\widehat{\beta}_l^*)^2 = 2f_l(1-f_l)\widehat{\beta}_l^2$. This is the phenotypic variance that is explained by genetic variation tagged by variant l, (similar interpretation applies to heritability), but is still often interpeted as "variance explained by variant l".

Let's assume that we have standardized the phenotype ($\sigma_Y^2 = 1$) before the GWAS. According to the simplest model, assuming no LD or deviation from the additivity between the causal effects of different variants, the additive heritability over all variants is $h^2 = \sum_{l=1}^{p} \lambda_l^{*2}$. More generally, allowing LD, p variants in a region contribute to the phenotype of one individual by the quantity $\boldsymbol{x}^{*T}\boldsymbol{\lambda}^*$, and the variance contributed by this region is

$$\begin{aligned} h^2_{\text{reg}} &= \text{Var}(\boldsymbol{x}^{*T}\boldsymbol{\lambda}^*) = \boldsymbol{\lambda}^{*T}\text{Var}(\boldsymbol{x}^*)\boldsymbol{\lambda}^* = \boldsymbol{\lambda}^{*T}\boldsymbol{R}\boldsymbol{\lambda}^* \\ &= (\boldsymbol{R}^{-1}\boldsymbol{\beta}^*)^T\boldsymbol{R}(\boldsymbol{R}^{-1}\boldsymbol{\beta}^*) = \boldsymbol{\beta}^{*T}\boldsymbol{R}^{-1}\boldsymbol{\beta}^*, \end{aligned}$$

where $\boldsymbol{R}$ is the LD-matrix (Pearson correlations) of the p variants. The above formula shows how to compute the regional heritability using either the causal effect sizes or the marginal effect sizes (latter of which are directly estimated by the standard GWAS). If there is no LD ($\boldsymbol{R} = \boldsymbol{I}$) between the variants, then the heritability is the sum of the squared scaled effects (either causal or marginal effects as they are the same in this case). For example, FINEMAP (from Section 7 of course material) gives an estimate of regional heritability for each causal configuration in its .config file according to the formula above.

Example 8.1. Consider SNPs 1 and 2 whose minor allele correlation is r_{12}, MAFs are 0.24 and the marginal estimates are $\widehat{\beta}_1 = \widehat{\beta}_2 = 0.08$ in a QT GWAS with a phenotypic variance of 1. What is the estimate of heritability explained by these SNPs jointly, and how much it differs from the case where they were independent, when r_{12} is 0.99 (almost the same variant twice), 0.3 (positively correlated effect alleles), 0 (independece) or -0.2 (effect alleles are masking each other)?

```
b.est = c(0.08, 0.08)
f = c(0.24, 0.24)
b.s = b.est * sqrt(2*f*(1-f)) #scaled marginal effects
res = c()
for(r.12 in c(0.99, 0.3, 0, -0.2)){
  R = matrix(c(1, r.12, r.12, 1),2,2)
  #regional heritability b.s^T R^-1 b.s
  h2.reg = t(b.s) %*% solve(R) %*% b.s #solve(R) is the inverse of R
  h2.ind = sum(b.s^2)
  res = rbind(res,c(r.12, h2.reg, h2.ind))
```

```
}
colnames(res) = c("r.12","h2.reg","h2.ind")
res
```

```
##       r.12      h2.reg     h2.ind
## [1,]  0.99 0.002346452 0.00466944
## [2,]  0.30 0.003591877 0.00466944
## [3,]  0.00 0.004669440 0.00466944
## [4,] -0.20 0.005836800 0.00466944
```

When the effect alleles are 0.99 correlated, then they are explaining the same signal and their joint variance explained is, correctly, only half of the sum of the marginals. With a negative correlation, and a similar direction and size in the observed effects, the SNPs must been masking each other's marginal effects and their total contribution is larger than the sum of their marginal contributions.

For diseases, the heritability is often measured on the **liability scale**, which requires an estimate of the disease prevalence (see Box 5 of Visscher et al.).

A few examples of the heritability estimates summed over GWAS loci (defined by a variant having P < 5e-8), in decreasing order of heritability:

- Paraoxonase-1 level has a heritability of 70% at a single locus (unsurprisingly, harboring the gene *PON1*) Benner et al. 2018.
- Height study by Yengo et al. 2022 found that 12,111 SNPs in 7,209 loci accounted for ~40% of variation in height. Sample size was 5.4 million! The GWAS loci covered 21% of the genome and it seems that these loci cover all effects on height from common SNPs.
- LDL-cholesterol study by Graham et al. 2021 explained ~13% of the variance in HDL-C and LDL-C by up to 1750 variants in 923 loci.
- Schizophrenia study by Tubetskoy et al. 2022 explained about 2.4% of variance in liability using 277 independent GWS variants.
- Crohn's disease study by Jostins et al. 2015 explained about 14% of variance in liability from 160 loci.
- BMI study by Locke et al. 2015 found that the 97 GWS loci account for 2.7% of BMI variation.

In particular, for many complex diseases, such as schizophrenia, the variance explained by GWS loci is still very small. On the other hand, traditional ways to estimate heritability suggest a high heritability for schizophrenia. Where is that heritability, if the top GWAS loci show this little traces of it? Or are the traditional estimates grossly overestimating heritability? This gap is called the missing heritability problem (Slides 6-7).

Next, we will look at two approaches that use GWAS data to estimate the genome-wide **SNP heritability** h^2_{SNP} that considers the heritability contribution of all SNPs that are included in the study, not just those that happen to reach the genome-wide significance level. The first method, the **linear mixed model**, is based on an efficient way of correlating the variant sharing with the phenotypic similarity in the population sample. The second method, **LD-score regression (LDSC)**, is based on a link between the amount of tagging by LD and the amount of GWAS signal seen at a variant, which link is induced by a highly polygenic genetic architecture. The linear mixed model requires original phenotype-genotype data whereas LDSC works with the GWAS summary data (i.e. the marginal effect estimates and their SEs).

8.3. Linear Mixed Model

Let's write down the full linear model for additive effects across the genome for a quantitative phenotype Y:

$$y_i = \mu + \boldsymbol{z}_i^T \boldsymbol{\alpha} + \boldsymbol{x}_i^{*T} \boldsymbol{\lambda}^* + \varepsilon_i = \mu + \boldsymbol{z}_i^T \boldsymbol{\alpha} + \sum_{l=1}^{p} x_{il}^* \lambda_l^* + \varepsilon_i,$$

where $\boldsymbol{z}_i$ is the vector of covariate values for i and $\varepsilon_i \sim \mathcal{N}(0, \sigma_E^2)$ is the (environmental) error term that is assumed to be uncorrelated across individuals.

We saw in GWAS 7 that if we try to estimate the parameters λ_l^* using the ordinary least squares estimator, we run into problems because of high correlations between SNPs and, more generally, overfitting. This is because the standard linear regression model is too flexible to adapt to the data when p grows large if the effect sizes are not restricted in any way. To overcome these problems, the linear mixed model treats the effect sizes λ_l^* as **random effects** that share a common (prior) distribution, here chosen to be $\mathcal{N}(0, \tau^2)$, where τ^2 will be estimated from the data. Now the parameters λ_l^* are not allowed freely to choose their values but their magnitude is restricted by a *shared* variance parameter τ^2. The model is able to learn from the data how the whole set of values of λ^* look like when considered together, and then apply that information to keep the magnitude of λ^*s appropriate by adjusting a single variance parameter τ^2. Another way to think about the difference between this random effects model and the least squares estimation is that our focus changes from estimating each of the p values λ_l^* to estimating their *shared* distribution, as determined by the variance parameter τ^2. Hence, we reduce the number of parameters estimated from p to 1, and will avoid overfitting.

The name *mixed model* reflects that the model is a *mix* of both **fixed effects** α, whose individual values are estimated as in the standard linear model, and **random effects** λ_l^*, whose joint *distribution* is estimated, rather than the individual values of λ_l^*s.

How can we link the new parameter τ^2 to the observed values of y? The answer is to write down what the random effect assumption means in terms of the observed similarity (mathematically covariance) of the phenotypes of individuals i and j. If we follow the random effect formulation, and independently draw each $\lambda_l^* \sim \mathcal{N}(0, \tau^2)$, what is the consequence on the phenotypic covariance between i and j, induced by the terms $\eta_i = \sum_{l=1}^{p} x_{il}^* \lambda_l^*$ and $\eta_j = \sum_{l=1}^{p} x_{jl}^* \lambda_l^*$?

$$\begin{aligned}\operatorname{Cov}(\eta_i, \eta_j) &= \operatorname{Cov}\left(\sum_{l=1}^{p} x_{il}^* \lambda_l^*, \sum_{l=1}^{p} x_{jl}^* \lambda_l^*\right) = \sum_{l=1}^{p}\sum_{k=1}^{p} x_{il}^* x_{jk}^* \operatorname{Cov}(\lambda_l^*, \lambda_k^*) = \sum_{l=1}^{p} x_{il}^* x_{jl}^* \operatorname{Cov}(\lambda_l^*, \lambda_l^*) \\ &= \sum_{l=1}^{p} x_{il}^* x_{jl}^* \tau^2 = p\tau^2 \boldsymbol{G}_{ij},\end{aligned}$$

where $\boldsymbol{G}$ is the GRM-cor from chapter 5 of the course material, i.e., $\boldsymbol{G} = \frac{1}{p}\boldsymbol{X}^*\boldsymbol{X}^{*T}$ is the $n \times n$ empirical correlation matrix for individuals computed across all SNPs.

This is saying that the **additive genetic components** η_i of the trait are correlated across the individuals according to the genetic relatedness of the individuals, as measured by GRM-cor, and are scaled so that their variance is $p\tau^2$, where τ^2 is the variance of the causal effect sizes. Note that τ^2 is also the expected phenotypic variance contributed by any one causal effect as $\mathrm{E}(\lambda_l^{*2}) = \operatorname{Var}(\lambda_l^*) = \tau^2$. If we ignore LD between nearby variants, then $\sigma_G^2 = p\tau^2$ is the expected phenotypic variance contributed by all p variants together, and we would estimate the heritability as $\widehat{h}^2 = \sigma_G^2/(\sigma_G^2 + \sigma_E^2)$.

The last step is to write down the joint distribution of the phenotype vector $\boldsymbol{y}$, as defined by the variance components η and ε, from the relationship $y_i = \mu + \boldsymbol{z}_i^T\boldsymbol{\alpha} + \eta_i + \varepsilon_i$. The phenotype vector is an n-dimensional multivariate normal vector, whose mean $\boldsymbol{\mu}$ has components $\mu_i = \mu + \boldsymbol{z}_i^T\boldsymbol{\alpha}$ and whose covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\sigma^2}) = \sigma_G^2\boldsymbol{G} + \sigma_E^2\boldsymbol{I}_n$ is a function of two unknown variance parameters $\boldsymbol{\sigma^2} = (\sigma_G^2, \sigma_E^2)$. (Slide 8.)

A naive computation of such n-dimensional multivariate Normal likelihoods is expensive – $\mathcal{O}(n^3)$ operations – and in recent years many new ways to speed up the computation have been introduced. Currently, GCTA

with its **fastGWA** module remains a widely-used method and BOLT-REML is an efficient implementation of the mixed model applicable to 100,000s of samples. Recently, REGENIE has implemented an efficient version of a similar model.

In what follows, we will experiment with the mixed model by using a simple trick that can be done easily in R, but which would not generalize to multiple variance components, and therefore differs from more complex methods such as GCTA and BOLT-REML.

8.3.1 A Mixed model estimation method To simplify the setting, let's assume we first regress out the covariates from y using linear regression and consider the (quantile normalized) residuals from that regression as our covariate-adjusted phenotype y'. Our task is to maximize the multivariate Normal likelihood function of

$$\boldsymbol{y'} \sim \mathcal{N}\left(0, \boldsymbol{\Sigma}(\boldsymbol{\sigma^2})\right), \text{ with respect to } \boldsymbol{\sigma^2}.$$

The log-likelihood of the multivariate Normal is

$$L(\boldsymbol{\sigma^2}) = -\frac{1}{2}\log\det\boldsymbol{\Sigma}(\boldsymbol{\sigma^2}) - \frac{1}{2}\boldsymbol{y'}^T\boldsymbol{\Sigma}(\boldsymbol{\sigma^2})^{-1}\boldsymbol{y'}.$$

If we make an eigendecomposition of the GRM-cor matrix $\boldsymbol{G} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$, where $\boldsymbol{U}$ is an orthonormal matrix of eigenvectors and $\boldsymbol{D}$ is a diagonal matrix of eigenvalues, then the inverse and determinant of the n dimensional matrix can be transformed to those of diagonal matrices by rotating the phenotype with the eigenvectors into a new phenotype $\widetilde{\boldsymbol{y}} = \boldsymbol{U}^T\boldsymbol{y'}$:

$$L(\boldsymbol{\sigma^2}) = -\frac{1}{2}\sum_{i=1}^{n}\log(\sigma_G^2 D_i + \sigma_E^2) - \frac{1}{2}\sum_{i=1}^{n}\frac{\widetilde{y}_i^2}{\sigma_G^2 D_i + \sigma_E^2} = -\frac{1}{2}\sum_{i=1}^{n}\left(\log(\sigma_G^2 D_i + \sigma_E^2) + \frac{\widetilde{y}_i^2}{\sigma_G^2 D_i + \sigma_E^2}\right),$$

where D_i is the diagonal element i of $\boldsymbol{D}$. (The derivation of the above transformation in not explained in detail here but is given by Pirinen et al. 2013.) This version of the log-likelihood is easy to optimize in R by using `optim()`, and we can help `optim()` by giving it also the gradient of the log-likelihood. Here is a function that returns the log-likelihood which we can then maximize by `optim()`.

```
lmm.loglik <- function(sigma, y, d)
{
    sigma.sum = sigma[1] * d + sigma[2]
    res = -0.5 * sum(log(sigma.sum) + y ^ 2 / sigma.sum)
    return(res) #returns log likelihood
}

lmm.gradient <- function(sigma, y, d)
{
    sigma.sum = sigma[1] * d + sigma[2]
    tmp = y ^ 2 / sigma.sum - 1
    dsigmaG = 0.5 * sum(d / sigma.sum * tmp)
    dsigmaE = 0.5 * sum(1 / sigma.sum * tmp)
    return(c(dsigmaG, dsigmaE)) #returns gradient
}
```

Example 8.2. Let's try the mixed model with $n = 2000$ samples and $p = 10,000$ independent common variants with MAF 0.5 and simulate trait with $h^2 = 0.5$. We expect that each variant here would explain only $h^2/p = 0.00005$ of the trait variance! We don't expect to have any genome-wide significant findings with $n = 2000$ samples.

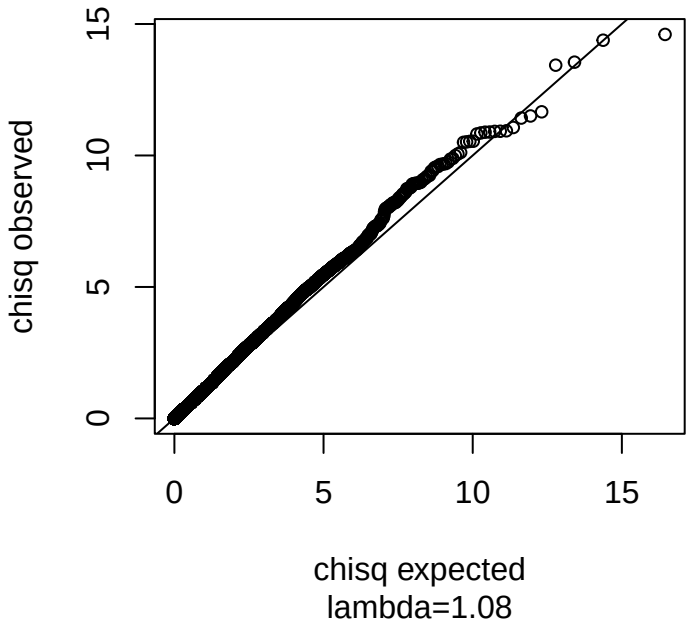
```r
p = 10000
n = 2000
f = 0.5
h2 = 0.5

X.s = scale(replicate(p, rbinom(n, size = 2, p = f) )) #use scaled genotypes

#tau^2 = var(lambda.s) = h2 / p
lambda.s = rnorm(p, 0, sqrt( h2 / p)) #scaled effects

#generate phenotype: SNP effects + random noise with var=1-h2
y = scale( X.s %*% lambda.s + rnorm(n, 0, sqrt(1-h2) )) #scaling makes mean(y)=0 as our LMM ignores int

#test individual SNPs and make a QQ-plot
pval = as.numeric(apply(X.s, 2, function(x){summary(lm( y ~ x))$coefficients[2,4]}))
expect.stats = qchisq(ppoints(p), df = 1, lower = FALSE)
obs.stats = qchisq(pval, df = 1, lower = FALSE)
lambda = median(obs.stats) / median(expect.stats) #GC lambda = ratio at medians
qqplot(expect.stats, obs.stats, xlab = "chisq expected", ylab = "chisq observed",
       sub = paste0("lambda=",signif(lambda,3)), cex = 0.8)
abline(0,1)
```




```r
sort(pval)[1:5] #show the 5 smallest P-values
```

```
## [1] 0.0001326792 0.0001490890 0.0002324405 0.0002470307 0.0006383562
```

We have no genome-wide significant SNPs and the QQ-plot doesn't look very inflated. However, the data were simulated in such a way that we expect that the SNPs explain together half of the phenotypic variance, even if we can't see any individual SNP having a clear effect.

What does the mixed model say?

```r
#G = cor(t(X.s)) #make GRM-cor matrix
G = (X.s %*% t(X.s))/p #make GRM-cor matrix (see Chapter 5.1.2.)
eig = eigen(G) #decompose G = U D t(U)
y.Ut = t(eig$vectors) %*% y #transform y to y.Ut by using U^t y
d = eig$values #eigenvalues of G

start.val = c(1, 1) / 2 #starting values for the optimization of the two variance parameters
#1st optim run with robust Nelder-Mead method without gradient
res = optim(start.val, fn = lmm.loglik, y = y.Ut, d = d,
            method = 'Nelder-Mead', control = list(fnscale = -1)) #fnscale=-1 to maximize, not minimize
#2nd optim run refining the estimate by BFGS method using gradient
res = optim(res$par, fn = lmm.loglik, gr = lmm.gradient, y = y.Ut, d = d,
            method = 'BFGS', hessian = T, control = list(fnscale = -1)) #fnscale=-1 to maximize, not mi
sigma2 = res$par #estimates of sigma.G^2 and sigma.E^2
sigma2.SE = sqrt(diag(solve(-res$hessian))) #SE of sigma_G^2 and sigma_E^2
res = cbind(sigma2, sigma2.SE) # genetic variance and environmental variance with SEs
rownames(res) = c("sigma.G^2","sigma.E^2")
res
```

```
##              sigma2  sigma2.SE
## sigma.G^2 0.4919493 0.07748576
## sigma.E^2 0.5094548 0.07062575
```

```r
h2.est = sigma2[1]/sum(sigma2)
h2.est # heritability is ~sigma.G^2 since var(y)=1. Then we also have SE(h2) = SE(sigma.G^2).
```

```
## [1] 0.4912595
```

Our estimate is close to the true value of 50%. (Note that given the SE of 0.07 we are lucky to get *this* close here.) The mixed model can indeed pick up the joint contribution of all those tiny effects! Note that we haven't derived SE for $\widehat{h}^2$ but in cases where total variance of phenotype is 1, we can use SE of $\widehat{\sigma}^2_G$ as SE of $\widehat{h}^2$.

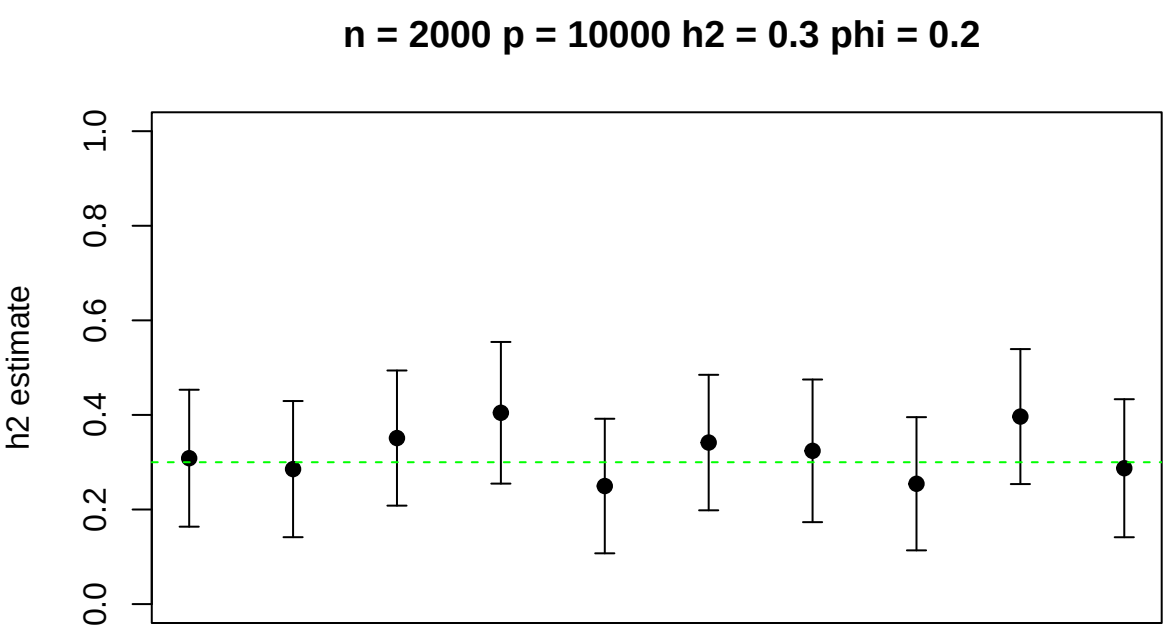
Example 8.3. LMM worked nicely above, but what happens when the true genetic architecture is not 100% polygenic, that is, when only a subset of all variants contribute to the phenotype. The mixed model was derived assuming that all effects are non-zero. What happens if we simulated non-zero effects only for, say 20% of the SNPs? Let's recycle our genotype data and G matrix and its decomposition since those take some time to make. Let's simply simulate new sets of λ^*s for a scenario where 20% of SNPs have non-zero effects and together they explain 30% of the trait variance. Let's repeat this 10 times and plot the estimates.

```r
set.seed(12)
phi = 0.2 # proportion phi of the SNPs are non-zero
h2 = 0.3
n.iter = 10 #how many simulations -- all use the same genotype data
h2.res = matrix(NA, ncol = 2, nrow = n.iter)

for(iter in 1:n.iter){
  #choose which SNPs have an effect
  c.ind = sort(sample(1:p, size = round(phi*p)))
  #var(lambda.s) = h2 / (phi*p)
  lambda.s = rnorm(length(c.ind), 0, sqrt( h2 / (phi*p)))
```

```
  #generate phenotype: SNP effects + random noise with var = 1 - h2
  y = scale( X.s[,c.ind] %*% lambda.s + rnorm(n, 0, sqrt(1-h2) ))#scale: mean(y) = 0, sd(y) = 1
  y.Ut = t(eig$vectors) %*% y #transform y to y.Ut = U^t y

  start.val = c(1, 1) / 2 #starting values for the optimization of the two variance parameters
  #1st optim run with robust Nelder-Mead method without gradient
  res = optim(start.val, fn = lmm.loglik, y = y.Ut, d = d,
              method = 'Nelder-Mead', control = list(fnscale = -1))
  #2nd optim run refining the estimate by BFGS method using gradient
  res = optim(res$par, fn = lmm.loglik, gr = lmm.gradient, y = y.Ut, d = d,
              method = 'BFGS', hessian = T, control = list(fnscale = -1))
  sigma2 = res$par #estimates of sigma.G^2 and sigma.E^2
  sigma2.SE = sqrt(diag(solve(-res$hessian))) #SE of sigma_G^2 and sigma_E^2
  h2.res[iter,] = c(sigma2[1]/sum(sigma2),sigma2.SE[1]) # h2.est with sigma.G^2 SE
}
plot(1:n.iter, h2.res[,1], pch = 19,
     main = paste0(" n = ",n," p = ",p," h2 = ",h2," phi = ",phi),
     xaxt="n", ylab="h2 estimate", ylim = c(0,1), xlab = "")
arrows(1:n.iter, h2.res[,1]-1.96*h2.res[,2], 1:n.iter, h2.res[,1]+1.96*h2.res[,2],
       code = 3, length = 0.05, angle = 90)
abline(h = h2, lty = 2, col = "green")
```



It works fine even when only 20% of the SNPs have non-zero effects. It seems that LMM is robust to a considerable proportion of zero effects among the SNPs, which is great news!

8.3.2 Mixed model heritability estimates LMMs have been used for a long time in animal breeding and pedigree analyses and this approach became popular in GWAS data in humans by the landmark publication of Yang et al. (2010) where they showed that 45% of the variation in height could be explained by about 300,000 SNPs on a genotyping chip. This was an important piece of information for the discussion surrounding the missing heritability problem, because after that publication it became more widely considered plausible that a large part of the gap between the variance explained by the GWS regions (only ~5% for height at that time) and the heritability of height as estimated by twin and sibling studies (~70-80%), may well be just because small genetic effects do not become GW-significant with the given sample sizes.

The authors have explained in their later publication Visscher et al. 2010 that "During the refereeing process (the paper was rejected by two other journals before publication in Nature Genetics) and following the publication of Yang et al. (2010) it became clear to us that the methodology we applied, the interpretation of the results and the consequences of the findings on the genetic architecture of human height and that for other traits such as complex disease are not well understood or appreciated."

Nowadays, the LMM is routinely used for SNP heritability estimation of quantitative traits. More recent applications suggest that for height and BMI there is not much of a gap anymore between the SNP heritability from very dense marker sets and the estimates of the total heritability by other means than through GWAS data Yang et al. 2015, Young 2019, Yengo et al. 2022. At the same time, the analysis has got more complex compared to the original version that only included one variance component for all the SNPs, as in our example above. It is now clear that a better model should allow SNPs with different MAFs and different amounts of LD-tagging to have different variance parameters of the effect size distribution. In practice, this means that one should compute separate GRM-cor matrices for each group of SNPs that needs to be modeled separately and then estimate their variance contribution jointly, in a single LMM, that can include tens of different GRM-cor matrices. This is possible with both GCTA and BOLT-REML.

A typical heritability analysis using LMM filters out close relatives (at least 2nd degree or closer). This is because closer relatives usually are also positively correlated in some of their environmental factors, which could create correlation in their phenotypes that LMM would falsely pick up as heritability. When the analysis is done in a homogeneous population sample of "unrelated" individuals, this concern is greatly reduced. Thus, LMM with GWAS data does something that has been impossible to imagine before: Estimating heritability using "unrelated" individuals. A downside of restricting the analysis to unrelateds is that the precision of the variance components is much smaller than if there were a larger range of possible relationships in the data. This means that large samples are needed to get useful estimates of SNP heritability from population data.

What about disease studies? The GCTA approach was also quickly applied to case-control data, but the results are much more complicated to interpret than for the quantitative phenotypes. First, there is a transformation from binary phenotype to liability model. Second, there is the case-control ascertainment which complicates statistical modeling (as we saw with the simple covariate adjustments) and makes LMM in general behave unfavorably, as reported by Golan et al. 2014. Third, typically cases and controls have been genotyped/handled/collected differently, and while a careful quality control can make GWS findings reliable and replicable, it remains a concern that the tiny effects picked up by the mixed model are not only polygenic effects but can also contain confounding effects. Hence interpreting the variance parameters of a case-control GWAS data as heritabilities of diseases makes quite a many quite strong assumptions. We will come back to this important question about whether the inflation in a QQ-plot is confoundig or polygenicity in section 8.4 below.

A perspective of using LMM to estimate heritability by Yang et al. 2017.

8.3.3 Mixed model in GWAS All the discussion above has been about estimating the variance parameters using the LMM. But mixed models have also become widely-used for running the primary GWAS analysis.

Suppose we want to test the effect of SNP s on the phenotype Y. The LMM approach does the linear regression by including in the model a random effect from **other variants except s and its LD-friends**:

$$y_i = \mu + \boldsymbol{z}_i^T \boldsymbol{\alpha} + x_{is}\beta_s + \eta_i + \varepsilon_i,$$

where $\boldsymbol{z}_i$ is the vector of covariate values for i, $\eta_i = \sum_{l=1}^{p_s} x_{il}^* \lambda_l^*$ is the additive contribution of the rest of the genome *except variant s and its LD-friends* and $\varepsilon_i \sim \mathcal{N}(0, \sigma_E^2)$ is the (environmental) error term that is assumed to be uncorrelated across individuals. The random effect assumption $\lambda_l^* \sim \mathcal{N}(0, \tau^2)$ will then lead to similar computations as above with the heritability estimation except that the GRM-cor matrix G can now be different for different variants tested, and there is an additional fixed effect in the model corresponding to the marginal effect β_s of SNP s.

In practice, when testing SNPs on, say, chromosome 1, GRM-cor can be computed for all other chromosomes and then used as the covariance structure of the random effect for all SNPs on chr 1 (e.g. BOLT-LMM and GCTA do this). Local updating of the random effect has also been studied (Listgarten et al. 2012).

There are (at least) two benefits for adding the rest of the genome as a random effect in linear regression:

1. The precision of the estimator $\widehat{\beta}_s$ will increase as the (often substantial) variation in phenotype from the rest of the genome is explained away by the model. This also leads to increased statistical power to detect new associations.
2. The rest of the genome captures confounding effects that are due to relatedness and/or population structure and hence $\widehat{\beta}_s$ from a LMM has been automatically adjusted for these confounders.

Thus, LMM is an alternative for the standard linear regression with leading PCs as covariates, and since LMM also increases power and also accounts for related individuals, it is a very useful in QT GWAS.

To be on the safe side, when the data have both close relatives and clear population structure, then these patterns may not be correctly modelled by a single joint random component, and, in general, one should generate different random effects for each major source of phenotypic covariance in the sample. Typically, we still remove one from each pair of closely related individuals because there are not many of these pairs and it is difficult to know whether their environmental correlation has been correctly modelled by LMM.

While LMMs have also been applied to GWAS of binary traits, until recently they have only been applicable to the cleanest cases of common variants and balanced case-control ratios. Recent research on efficient mixed model for binary phenotypes have been implemented by Zhou et al. 2018 into SAIGE software and by Mbatchou et al. 2020 into REGENIE.

Now, let's return to the important question that puzzled us above. We know that both confounding and true polygenic effects can cause inflation in QQ-plots and the genomic-control parameter λ. How could we tell whether the inflation is true signal from thousands of small effects, or whether it is a result of some confounding bias?

8.4. LD-score regression (LDSC)

LDSC developed by Bulik-Sullivan et al. 2015 can separate confounding from polygenicity by utilizing LD in a clever way.

Let's think about consequences of high polygenicity (a lot of non-zero causal effect sizes) on the observed marginal effect sizes when we take into account the differences in LD patterns between the variants. We know that for a region with p variants and LD-matrix $\boldsymbol{R}$, $\boldsymbol{\beta}^* = \boldsymbol{R}\boldsymbol{\lambda}^*$. If we assume a highly polygenic model, with each $\lambda_l^* \sim \mathcal{N}(0, \tau^2)$, we get that

$$\mathrm{E}(\beta_l^*) = \mathrm{E}\left(\sum_{k=1}^{p} r_{lk}\lambda_k^*\right) = \sum_{k=1}^{p} r_{lk}\mathrm{E}(\lambda_k^*) = 0,$$

$$\mathrm{E}(\beta_l^{*2}) = \mathrm{E}\left(\sum_{k=1}^{p} r_{lk}\lambda_k^*\right)^2 = \sum_{k=1}^{p} r_{lk}^2\mathrm{E}(\lambda_k^{*2}) + 2\sum_{k<m} r_{lk}r_{lm}\mathrm{E}(\lambda_k^*\lambda_m^*) = \sum_{k=1}^{p} r_{lk}^2\tau^2 + 2\cdot 0 = \tau^2 r_{l+}^2,$$

where $r_{l+}^2 = \sum_{k=1}^{p} r_{lk}^2$ is the **LD-score** of SNP l.

Thus, while the distribution of scaled causal effects was assumed independent of LD, the distribution of the marginal effects has the highest variance among the SNPs that tag heavily their neighbors, i.e., among the SNPs whose LD-scores are high. How does this property of true marginal effects translate to the observed estimates that we will obtain from a finite sample?

For scaled effects, SE of the marginal effects is approximately constant across SNPs with small marginal effects: it is $\sigma_\varepsilon/\sqrt{n}$ for quantitative traits (see below about using LD-score regression with binary traits). If

we assume that the quantitative trait variance is 1 in the population, then also $\sigma_\varepsilon \approx 1$ for small effects and SE is $1/\sqrt{n}$. Assuming no bias due to confounders, our observed marginal effect estimate is $\widehat{\beta}_l^* = \beta_l^* + \xi_l$, where error is distributed as $\xi_l \sim \mathcal{N}(0, \mathrm{SE}^2)$ and hence

$$\mathrm{E}\left(\left(\widehat{\beta}_l^*\right)^2\right) = \mathrm{E}\left(\beta_l^{*2} + \xi_l^2 + 2\beta_l^*\xi_l\right) = \mathrm{E}\left(\beta_l^{*2}\right) + \mathrm{E}(\xi_l^2) + 2\mathrm{E}\left(\beta_l^*\xi_l\right) = \tau^2 r_{l+}^2 + \mathrm{SE}^2 + 2\cdot 0 = \tau^2 r_{l+}^2 + \frac{1}{n},$$

It follows that the expected chi-square statistic observed at variant l is

$$\mathrm{E}(\chi_l^2) = \frac{\mathrm{E}\left(\left(\widehat{\beta}_l^*\right)^2\right)}{\mathrm{SE}^2} = \frac{\tau^2 r_{l+}^2 + \frac{1}{n}}{\frac{1}{n}} = n\,\tau^2\,r_{l+}^2 + 1,$$

and also increases with the LD-score of the variant. (Note that this quantity does not depend on the assumption of the trait variance being 1.)

This derivation suggests a simple and testable hypothesis about GWAS summary data. If the trait is (highly) polygenic, then we should see a linear relationship between the observed chi-square statistics and LD-score: Variants that tag more of their neighbors have higher chance of tagging causal variants and hence their marginal effects will have larger magnitude. If, instead, we see overall inflation in chi-square statistic, but the values $(\chi_l^2 - 1)$ are not proportional to the LD-scores, then the inflation is likely due to some confounding bias that is affecting the bulk of the test statistics independently of the LD-scores. We denote by such a constant bias factor in the chi-square test statistic by b.

Since under the polygenic model $h^2 = p\tau^2$, we can replace τ^2 in above formulas with h^2/p and write the LD-score regression equation as

$$\mathrm{E}(\chi_l^2) = 1 + b + \frac{nh^2}{p} r_{l+}^2.$$

If we now take our observed chi-square statistics from a GWAS and regress them on the LD-scores of the variants, we are expecting that

- the intercept is elevated from 1 if there is a confounding bias in the results (i.e. when $b > 0$),
- regardless of the possible confounding bias, the slope of the regression gives an estimate of the SNP heritability when multiplied by p/n.

Binary traits. It is a custom to apply LDSC (and other heritability estimation methods) to binary traits in two steps. First, pretend that the trait is quantitative and compute heritability h^2_{obs} on the **observed scale**, that is, when the binary trait is treated as a quantitative trait with trait values 0 and 1. Second, turn that estimate on the **liability scale** by accounting for the population prevalence of the binary trait, and by also accounting for a possible case-control ascertainment. This is explained by Lee et al. 2011.

LDSC sounds like a simple and useful tool! Let's see some results (slides 9-11).