Statistical Genetics, Spring 2022

Class notes 7: Introduction to Heritability Estimation and Linear Mixed Models

Variance decomposition and heritability

Denote a generic phenotype (continuous/numeric like height for now) by P. Some of the variability in P is genetic (G, tall parents have tall children) and some is from other sources: environment (E, diet affects height), random chance(?) etc. Genetics and environment can also interact, for example if some genetic effects are more pronounced in some diets than others. We can write this generically as:

$$P = \mu + G + E + G \times E + \epsilon.$$

We usually start from a simpler model, which assumes:

$$P = G + E\ , \quad Var(P) = Var(G) + Var(E) \ \Leftrightarrow \ \sigma_p^2 = \sigma_g^2 + \sigma_e^2,$$

a model which assumes:

- The random error ϵ and possibly $G \times E$ effects are swallowed in E
- G, E are uncorrelated (so rules out existence of most $G \times E$ effects)

In this model, we define the **heritability** as:

$$H^2 = \frac{\sigma_g^2}{\sigma_p^2} = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2},$$

the percentage of variability that is due to genetic causes.

Example: Body length of flies (Robertson 1957). They created genetically identical flies and distributed them around a variety of environments (diet, temperature, etc.); in parallel they also spread "randomly mated" flies who are genetically diverse between the same set of environments. From this they measured:

$$\sigma_g^2 + \sigma_e^2 = 0.366\ , \ \ \sigma_e^2 = 0.186 \ \ \Rightarrow H^2 = \frac{0.366 - 0.186}{0.366} = 0.49.$$

We have to keep many caveats in mind:

1. If we exposed to more diverse environments, we would decrease heritability

2. If we examined more genetically diverse flies, we would increase heritability

3. If there are in fact $G \times E$ interactions, then it is unclear how to interpret the results

Conclusion: this is very useful information, but has to be interpreted with care, and inference about assumptions is important.

Estimating heritability from twin studies

If we take P_1, P_2 identical (monozygotic, MZ) twins we get:

$$P_1 = \mu + G + E_1 \ , \ \ P_2 = \mu + G + E_2 \ \ \Rightarrow \ \ P_1 - P_2 = E_1 - E_2$$

since they are genetically identical.

A naive approach now takes n pairs of identical twins and declares:

$$\mathbb{E}(P_{i1} - P_{i2})^2 = 2\sigma_e^2 \ \ \Rightarrow \ \ \hat{\sigma}_e^2 = \frac{\sum_{i=1}^n (P_{i1} - P_{i2})^2}{2n} \ , \ \ \hat{H} = \frac{\hat{\sigma}_p^2 - \hat{\sigma}_e^2}{\hat{\sigma}_p^2}.$$

Why is this approach too naive? Because E_1, E_2 are far from independent – the two twins share the environment both before and after birth!
But E_1, E_2 are not identical, since they also include random error, and important exposures might defer (like twin locations in the womb, or one twin working with radioactive material).

A less naive decomposition here would be:

$$P_1 = \mu + G + E_{Com} + E_{Sep,1} \ , \ \ P_2 = \mu + G + E_{Com} + E_{Sep,2} \ \ \Rightarrow \ \ P_1 - P_2 = E_{Sep,1} - E_{Sep,2},$$

and consequently:

$$\mathbb{E}(P_{i1} - P_{i2})^2 = 2\sigma_{Sep,e}^2 < 2\sigma_e^2 \ , \ \ Cov(P_1, P_2) = \sigma_g^2 + \sigma_{Com,e}^2.$$

To decompose the components of error, the proposal is to also take regular fraternal twins (dizygotic, DZ), who are genetically regular siblings that share 50% of their genetic material:

$$P_1 = \mu + G_{Com} + G_{Sep,1} + E_{Com} + E_{Sep,1} \ , \ \ P_2 = \mu + G_{Com} + G_{Sep,2} + E_{Com} + E_{Sep,2}.$$

Now, under the assumption that sharing half the genetic material is sharing half of G, that is $Var(G_{Com}) = Var(G_{Sep})$, we get:

$$Cov(P_{i1}, P_{i2}) = 0.5\sigma_g^2 + \sigma_{Com,e}^2.$$

Notice that we also assume that the shared environment has the same effect in MZ and DZ (also debateable).

Taking a sample of both DZ and MZ twins, we can now put all of this together, and get an unbiased estimate for the heritability:

$$\widehat{H} = 2\frac{\widehat{Cov}(MZ) - \widehat{Cov}(DZ)}{\widehat{Var}(P)} = 2\left(r_{MZ} - r_{DZ}\right).$$

Let's be explicit about the assumptions that this estimate makes:

1. Environment has an additive, independent (of G) effect, and environmental covariance is same for MZ and DZ

2. Genetic effect of 50% sharing is 0.5 of 100% sharing. This is true if G is a sum of independent effects of all alleles, so:

   - Effects of different genomic sites are additive
   - Also within site the effects are additive (rather than say dominant or recessive)

For example, assume the genetic effect is only due to one site, but there is a dominance effect rather than purely additive effect. Then we can show that $\sigma^2_{Com,g} < 0.5\sigma^2_g$, and therefore we would get $\mathbb{E}\widehat{H} > H$.

Accepted published twins heritability estimates:

- Height: 80%
- IQ: 50% - 90% (accepted number 80%)

Can GWAS explain heritability?

Now we have estimates of how much of the phenotype is heritable/genetic in nature, we should be able to find the genetic variants that underlie this in GWAS — right?

Case in point: looking for the heritability of height in GWAS. In 2008 three huge GWASs on height were published at once, total of $50K$ individuals:

Authors	# Samples	# Replication	# Sig. loci found	% variance explained
Weedon et al.	13600	16500	20	$\approx 3\%$
Lettre et al.	15800	10000	12	$\approx 2\%$
Gudbjartsson et al.	50000	-	27	$\approx 3.7\%$

Regardless of the subtle differences in population studied and other causes for the slight differences in findings — all of them together explain $< 10\%$ of the variance of height, as compared to 80% from twin estimates. Since then there have been much bigger GWASs for height (even in the millions of samples), and significant findings still explain *much* less than the 80% expected.

This has been dubbed *The mystery of missing heritability*, and has been on the statistical genetics community's mind for well over a decade. Leading theories on the explanations (obviously, the truth may be a mixture of all of them):

1. Instead of relatively small number of variants with big effect, the heritability is due to a large/huge number of variants with tiny effects
   $\Rightarrow$ Even with huge sample size, there is not enough *power* to identify these effects and declare them significant

2. Rare variants with big effect are the main cause of heritability
   $\Rightarrow$ Each "tall family" or family with cancer is has it for a different genetic reason
   Note that such variants are not even measured in traditional GWAS, which considers only *common* variants. Also, studies of unrelated individuals by definition have no power to identify such variants

3. Additive model is wrong, this can have several consequences:
    - Twin estimates may be (badly) inflated due to assuming additive model
    - GWAS which tests one site at a time might miss strong interactions
4. Epigenetics: effects which are heritable and affect the genome, but are not expressed in the sequence. This includes methylation as well as more mysterious effects.

The Linear Mixed Model (LMM) approach

To possibly deal with the first issue above (large number of small effects), we take this to extreme, and assume all variants have a tiny effect, giving the following model:

$$Y_i = \sum_{j=1}^{M} Z_{ij} b_j + \epsilon_i \ , \quad i = 1, \ldots, n,$$

where $Z_{n\times M}$ is the genotype matrix.

In the standard view, this model has M parameters (say $M = 10^6$), so it is not really that useful. But we can adopt the random effects approach, which now considers b_j as random variables, say assuming: $b_j \sim N(0, \sigma_\gamma^2)$, $\forall j$, and that the γ's are independent between them. Then, assuming $\epsilon_i \sim N(, \sigma_\epsilon^2)$, we get:

$$Y|Z \sim N\left(0, ZZ^t\sigma_b^2 + I_n\sigma_\epsilon^2\right),$$

note that by defining the coefficients as random variables, we have moved them from the mean part of Y into the covariance matrix. ZZ^t is an $n \times n$ matrix, where $(ZZ^t)_{ij} = \sum_{k=1}^{M} Z_{ik}Z_{jk}$. It is common to denote:

$$G = \frac{ZZ^t}{M} \ , \ \sigma_g^2 = M\sigma_b^2 \implies Y|Z \sim N(0, G\sigma_g^2 + I\sigma_\epsilon^2).$$

Note that this is a problem with two parameters only: $\sigma_g^2, \sigma_\epsilon^2$, and if we assume that the phenotype is standardized to have norm 1, only one parameter, σ_g^2, with $\sigma_\epsilon^2 = 1 - \sigma_g^2$, and resulting heritability $H^2 = \sigma_g^2$.

A slightly less naive view considers the following additional aspects:

1. The problem may also have fixed effects $X_{n\times p}$ like age, sex, nutrition, or known SNPs with big effects that don't fit this model. The more realistic model is then:

$$Y|Z, X \sim N(X\beta, G\sigma_g^2 + I_n\sigma_\epsilon^2), \tag{1}$$

    with $p + 2$ parameters.

2. The assumption that all coefficients are tiny and of the same magnitude may be too naive, because:
    - It may be more realistic to assume that a small fraction of SNPs have an effect, it will still be a large number and the effects can still be small. This can be expressed as a mixture distribution of zero and a normal (spike and slab)

- Properties of the SNPs, such as which genomic region they reside in, and what is the function of that region, are likely to be related to their effect size

So we may want to assume $b_j \sim F$ not normal (such as spike and slab). It is important to note, however, that as long as M is big and this mixture distribution is not too crazy, we can still use the CLT to assume

$$Y|Z \dot{\sim} N(0, G\sigma_g^2 + I\sigma_\epsilon^2).$$

3. It is possible to assume that the SNPs are divided into multiple groups, each one with its own variance parameter

The model (1) is known as the LMM as it has both fixed effects β and random effects b. The $p + 2$ parameters are:

- p fixed effects β.
- Two random effects (variance components) parameters $\sigma_g^2, \sigma_\epsilon^2$.

These are typically estimated with maximum likelihood (more accurately, restricted maximum likelihood (REML), which gives unbiased estimates of the variance components).

What do the entries in G look like? We typically assume that columns of Z are *standardized*, so approximately (why not exactly?) $\sum_k Z_{ik} \approx 0$, $\sum_k Z_{ik}^2 \approx M$. What about $G_{ij} = \sum_k Z_{ik}Z_{jk}/M$ for $j \neq k$? It is usually assumed that the individuals in the study are *unrelated*, meaning that after standardization we expect that $Z_{ik}Z_{jk}$ be centered around 0. Hence once we sum over all pairs we typically get

$$G_{ij} = \frac{1}{M}\sum_k Z_{ik}Z_{jk} \approx 0 \ , \ \ \forall i \neq j,$$

are very small. However they are not exactly 0, expressing small variations in degree of similarity between pairs of individuals. This is the key in still being able to estimate the variance components — the off-diagonal elements in G being non-zero.