Statistical Genetics, Fall 2025-6

Class notes 10: Heritability estimation and LMMs in case-control studies

Basic binormal identities

Assume

$$\left(\begin{array}{c} X \\ Y \end{array}\right) \sim N\left(0, \left(\begin{array}{cc} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{array}\right)\right).$$

Then:

$$Y|X = x \sim N\left(\rho\frac{\sigma_Y}{\sigma_X}x, (1-\rho)^2\sigma_Y^2\right).$$

Heritability estimations and LMMs for disease studies

Assume we have a disease with prevalence K in the population ($\mathbb{P}(Y=1) = K$). Going back to basic definitions, we can adopt the *liability threshold* model, assuming an underlying continuous pheontype:

$$L = G + E \ , \ Y = \mathbb{I}\{L > t\}.$$

Since the *liability* L is unobserved, we can safely assume $\sigma_L^2 = \sigma_g^2 + \sigma_e^2 = 1$, and even $L \sim N(0,1) \Rightarrow t = Z_{1-K}$

For GWAS data, it is typically assumed as in continuous phenotypes that:

$$L \in \mathbb{R}^n \sim N\left(0, G\sigma_g^2 + I(1-\sigma^2 g)\right),$$

with the actual phenotype $Y = \mathbb{I}\{L > t\}$, applied coordinate-wise.

Now we can investigate the connections between the observed Y and unobserved L:

$$\begin{aligned}
\mathbb{E}(L|Y=1) &= \mathbb{E}(L|L>t) = \frac{\int_t^\infty u\phi(u)du}{1-\Phi(t)} = \frac{1}{K}\int_t^\infty (2\pi)^{-1/2}\exp\{-0.5\cdot u^2\}udu = \\
&= \frac{1}{K} - (2\pi)^{-1/2}\exp\{-0.5\cdot u^2\}|_{u=t}^{\infty} = \frac{\phi(t)}{1-\Phi(t)} \\
\mathbb{E}(L^2|Y=1) &= \mathbb{E}(L^2|L>t) = \frac{\int_t^\infty u^2\phi(u)du}{1-\Phi(t)} = \frac{-\phi(u)u|_{u=t}^{\infty} + \int_t^\infty \phi(u)du}{1-\Phi(t)} = \\
&= \frac{t\phi(t)+(1-\Phi(t))}{1-\Phi(t)} = 1 + \frac{t\phi(t)}{1-\Phi(t)}. \\
Cov(L,Y) &= \mathbb{E}(YL) - \mathbb{E}(Y)\mathbb{E}(L) = \mathbb{P}(Y=1)\mathbb{E}(L|Y=1) = K\frac{\phi(t)}{K} = \phi(t). \\
Cor(L,Y) &= \frac{\phi(Z_{1-K})}{\sqrt{K(1-K)}}.
\end{aligned}$$

In twins studies, we can still use MZ and DZ twins to estimate the *observed-level heritability* using the same formula:

$$\hat{H}^2_{ad,obs} = 2(r(MZ) - r(DZ)).$$

The question is, what is the connection between $H^2_{ad,obs}$ and $\sigma^2_g = H^2_{ad,liability}$?
This turns out to be a difficult problem, which we will get back to later. For now, let's deal with an easier problem.

Assume we observe G, and therefore can estimate $Cor(Y,G)$ — can we use this to estimate $\sigma^2_g = Cov(L,G)$? The famous result which addresses that is by Dempster and Lerner (1950), for which we need to derive some additional formulas:

$$\mathbb{E}(G|L>t) = \sigma_g^2\frac{\phi(t)}{K} \quad \Rightarrow \quad Cor(Y,G) = \sigma_g\frac{\phi(t)K}{K\sqrt{K(1-K)}} = \frac{\sigma_g\phi(t)}{\sqrt{K(1-K)}}.$$

Therefore, we can conclude:

$$H^2_l = \sigma^2_g = Cor^2(Y,G)\cdot\frac{K(1-K)}{\phi(t)^2}.$$

For rare diseases (K small), we have $\phi(Z_{1-K}) \approx K$ (same order of magnitude), and therefore $H^2_l >> Cor^2(Y,G) = H^2_{obs}$, in other words: the heritability is much bigger on the unobserved liability scale than on the observed scale.

Thus, a methodology which arises;

1. Somehow estimate "heritability"$= Cor^2(Y,G)$ on the observed scale

2. Perform the transformation to transform it to the liability scale

This approach was adopted by Lee et al. (2011), who used the LMM approach, meaning they assumed (for absolutely no good reason) the standard LMM model for the 0/1 phenotype:

$$\frac{Y}{\sqrt{K(1-K)}}|Z,X \dot{\sim} N(X\beta, G\sigma^2_g + I_n(1-\sigma^2_g)),$$

then in that model, estimated $\widehat{H^2_{obs}} = \hat{\sigma}^2_g$, and applied the Depmster-Lerner correction:

$$\widehat{H^2_l} = \widehat{H^2_{obs}} \cdot \frac{K(1-K)}{\phi(t)^2}.$$

Surprisingly(?), this approach generally gives pretty good (unbiased) estimates of H^2_l in simulations.

An equally unsubstantiated approach uses twins to estimate H^2_{obs} and applies the same correction.

These are both fundamentally unsubstantiated because there is no reason to assume additivity between G and E on the observed scale, which is fundamental to the formulas we derived. A second (major?) concern for the Lee et al. method is that normality is impossible here by definition, since $Y \in \{0, 1\}$.

Finding a legitimate model

(This section follows Golan, Lander, Rosset (2014))

We want to see whether we can find a legitimate way of estimating H^2_l from disease data under the liability threshold model? The approach we choose generalizes *Haseman-Elston* regression.

Aside: Haesman-Elston for continuous phenoytpe. Assume the usual model

$$Y \sim N(0, G\sigma^2_g + I_n(1 - \sigma^2_g)), \tag{1}$$

then it is easy to see:

$$E(Y_iY_j) = G_{ij}\sigma^2_g.$$

Therefore, if we regress the (observed) pairs $Y_i \times Y_j$ on the (observed) entries G_{ij}, the slope is an unbiased estimate of $H^2 = \sigma^2_g$.

Now we want to apply the same thinking to our binary model. Let $G_{ij} = \rho$ for brevity. It is easy to see that:

$$\begin{pmatrix} l_i \\ l_j \end{pmatrix} \sim N\left(0, \begin{pmatrix} 1 & \rho\sigma^2_g \\ \rho\sigma^2_g & 1 \end{pmatrix}\right).$$

Therefore:

$$\mathbb{E}(Y_iY_j) = P(l_i > t, l_j > t) = \int_t^\infty \phi(u)\left(1 - \Phi\left(\frac{t - \sigma^2_g\rho u}{\sqrt{1 - \rho^2\sigma^4_g}}\right)\right)du.$$

Clearly, there is no simple Haesman-Elston formula. Luckily, we can take advantage of the fact that $\rho \approx 0$ for GWAS (off-diagonal elements in G are small), and apply a first-order approximation:

$$\mathbb{E}(Y_iY_j) \approx (1-\Phi(t))\int_t^\infty \phi(u)du + \int_t^\infty \phi(u)\rho\phi(t)\sigma^2_g u du = K^2 + \rho\sigma^2_g\phi(t)\int_t^\infty \phi(u)udu = K^2 + \rho\sigma^2_g\phi(t)^2.$$

Denote $Z_i = \frac{Y_i - K}{\sqrt{K(1-K)}}$ the standardized phenotypes, and get:

$$\mathbb{E}(Z_i Z_j) = \mathbb{E}\left(\frac{(Y_i - K)(Y_j - K)}{K(1-K)}\right) \approx \frac{K^2 + \rho\sigma_g^2\phi(t)^2}{K(1-K)} - \frac{K^2}{K(1-K)} = \frac{\sigma_g^2\phi(t)^2}{K(1-K)}\rho.$$

Conclusion: If we regress $Z_i Z_j$ on $\rho = G_{ij}$ the resulting slope $\hat{b}$ is unbiased estimate of the quantity above, and therefore an unbiased estimate of heritability is:

$$\widehat{H_l^2} = \widehat{\sigma_g^2} = \hat{b}\frac{K(1-K)}{\phi(t)^2},$$

which turns out to be exactly equal to the Dempster-Lerner correction — is that surprising? Perhaps yes, since this is actually an approximation in this case!

Dealing with case-control sampling

Assume now that in addition, we perform case-control sampling. That is now, we assume:

- $L = G + E$ as before
- Y is not taken randomly from the population (where $\mathbb{P}(Y = 1) = K$), but from a "rebalanced" population, where $\mathbb{P}(Y = 1) = P >> K$.

To address this setting, we can "imagine" that we now have a new population, where case prevalence is P instead of K, by inflating the tail of the normal liability distribution. The advantage of this approach is that we can now imagine that we do random sampling in this new population, rather than dealing with non-random sampling in the original population.

In this setting, many interesting and surprising changes occur. For the rebalanced population, the distribution of liability is no longer normal. Furthermore, assuming that G and E are independent in the original population (as in the standard liability threshold model), they are no longer independent but highly correlated after the case control sampling, due to the fact that high values of $L = G + E$ are preferably selected, meaning when G is high in the case control population, E tends to be high as well.

The Lee et al. (2010) approach no longer gives reasonable results in this setting (and no one should expect it to give).

However we can extend the first-order moment-based approach to this setting. A slightly more complicated calculation than above (but similar in nature) now gives that if we standardize Y in the rebalanced population: $Z_i = \frac{Y_i - P}{P(1-P)}$, and regress $Z_i Z_j$ on the matrix G, with a similar first-order approximation, we get:

$$\mathbb{E}(Z_i Z_j) \approx \frac{P(1-P)}{(K(1-K))^2}\sigma_g^2\phi(t)^2\rho.$$

Now, the correction for the slope of the regression becomes:

$$\widehat{H_l^2} = \hat{b}\frac{(K(1-K))^2}{P(1-P)\phi(t)^2}.$$