Linear Regression

Nuno Carvalho

April 16, 2025

1 Mixed linear model associations

These derivations are based on Supplementary Note 1 of [Yang et al., 2014]. They describe the model behind GCTA's implementation of mixed linear model (MLM) associations for determining SNP effects in a dataset containing cryptic relatedness.

1.1 GCTA implementation of MLM

1.1.1 Phenotype model

A phenotype is modeled as:

$$\mathbf{y} = \mathbf{Kc} + \mathbf{g} + \boldsymbol{\epsilon}$$

Where:

- $\mathbf{y}$: outcomes. $N \times 1$ vector.
  - N: number of samples.
- $\mathbf{K}$: design matrix of covariates and intercept placeholder term. $N \times (K+1)$ matrix.
  - K: number of covariates.
- $\mathbf{c}$: covariate coefficients, including intercept term. $(K+1) \times 1$ vector.
- $\mathbf{g}$: genetic effects. $N \times 1$ vector.
- $\boldsymbol{\epsilon}$: non-genetic effects (i.e. error or noise term). $N \times 1$ vector.

The $\mathbf{Kc}$ term consists of fixed effects, while the $\mathbf{g}$ and $\boldsymbol{\epsilon}$ terms are random effects. The genetic effect, $\mathbf{g}$, is drawn from a multivariate normal distribution with a covariance structure determined by the genetic relatedness matrix (GRM, $\mathbf{A}$) scaled by the genetic variance component $varg$. That is:

$$\mathbf{g} \sim \mathcal{N}(0, \mathbf{A}\sigma_g^2)$$

Where the $\mathbf{A}$ is a $N \times N$ matrix with elements denoting the genetic relatedness between individuals j and k as:

$$\mathbf{A}_{jk} = \frac{1}{M} \sum_{i=1}^{M} \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1-p_i)}$$

Where M is the number of markers, x_{ij} is the genotype (0, 1, or 2) at SNP i for individual j, and p_i is the allele frequency of SNP i: $p_i = \frac{1}{2N}\sum_{j=1}^{N} x_{ij}$. If the genotype matrix, $\mathbf{X}$, has been standardized such that each row containing genotypes for a given variant, $\mathbf{X}_i$, is $\mathrm{E}[\mathbf{X}_i] = 0$ and $\mathrm{Var}[\mathbf{X}_i] = 1$, then the formula for the GRM is simply:

$$\mathbf{A} = \frac{\mathbf{X}^\intercal \mathbf{X}}{M}$$

The non-genetic effect, $\boldsymbol{\epsilon}$, is drawn from a multivariate normal distribution with no covariance structure and scaled by the environmental variance component, σ^2_ϵ. That is:

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma^2_\epsilon)$$

We can therefore assume that $\mathbf{g}$ and $\boldsymbol{\epsilon}$ are independent from each other. We can then define the SNP heritability as the proportion of non-fixed-effects variance that is explained by the genotyped markers, which is equivalent to:

$$\begin{aligned} h^2_g &= \frac{\mathrm{Var}[\mathbf{g}]}{\mathrm{Var}[\mathbf{g} + \boldsymbol{\epsilon}]} \\ &= \frac{\mathrm{Var}[\mathbf{g}]}{\mathrm{Var}[\mathbf{g}] + \mathrm{Var}[\boldsymbol{\epsilon}]} \\ &= \frac{\sigma^2_g}{\sigma^2_g + \sigma^2_\epsilon} \end{aligned}$$

The SNP heritability and the variance components, σ^2_g and $\mathbf{I}\sigma^2_\epsilon$, can be estimated by maximum likelihood methods (not covered here). We can combine the covariance structure of both random effects into a single covariance matrix:

$$\mathbf{V} = \mathbf{A}\sigma^2_g + \mathbf{I}\sigma^2_\epsilon$$

1.1.2 Testing for the effect of a SNP

The GCTA method of MLMi, where the i stands for inclusion, assumes that any SNP's effect on the phenotype is small enough to not make up a large amount of the variance of the phenotype explained by $\mathbf{A}$. Therefore, this allows the SNP of interest to be pulled out as a fixed term to estimate the effect of while also keeping it in the GRM we calculated earlier. In contrast, MLMe methods construct a GRM that excludes the SNP being tested at any given time, as well as other SNPs in correlation with it. GCTA implements this through the Leave-One-Chromosome-Out (LOCO) method, where a GRM is constructed out of all the genotyped SNPs except those belonging to the same chromosome as the SNP being tested. This results in 22 GRMs (assuming autosomes only). We modify our above model:

$$\mathbf{y} = \mathbf{K}\mathbf{c} + \mathbf{w}_i b_i + \mathbf{g} + \boldsymbol{\epsilon}$$

Where $\mathbf{w}_i$ is a $N \times 1$ vector of mean-adjusted genotypes, $w_{ij} = x_{ij} - 2p_i$, and b_i is its fixed effect we are trying to estimate. We can combine our fixed effects as follows:

$$\mathbf{q} = \begin{bmatrix} \mathbf{c} \\ b_i \end{bmatrix} \quad \text{and} \quad \mathbf{Q} = \begin{bmatrix} \mathbf{K} & \mathbf{w}_i \end{bmatrix}$$

Our phenotype is now modeled as:

$$\mathbf{y} = \mathbf{Qq} + (\mathbf{g} + \boldsymbol{\epsilon})$$

Where the first term are fixed effects and the second term (in parantheses) are random effects with covariance structure $\mathbf{V}$. Thus, we can think of our of phenotype as being drawn from:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{Qq}, \mathbf{V})$$

Since we have a covariance structure for the non-fixed term, estimating $\mathbf{q}$ is therefore a matter of performing **Generalized Least Squares (GLS)**, where $\boldsymbol{\beta} = \mathbf{q}$, $\mathbf{X} = \mathbf{Q}$, and $\boldsymbol{\Omega} = \mathbf{V}$. From GLS, we know:

$$\hat{\mathbf{q}} = (\mathbf{Q}^\intercal \mathbf{V}^{-1} \mathbf{Q})^{-1} \mathbf{Q}^\intercal \mathbf{V}^{-1} \mathbf{y} \quad \text{and} \quad \text{Var}[\hat{\mathbf{q}}] = (\mathbf{Q}^\intercal \mathbf{V}^{-1} \mathbf{Q})^{-1}$$

We can pull the SNP fixed effect, $\hat{b}_i$ from the last element of $\hat{\mathbf{q}}$ and its variance, $\text{Var}[\hat{b}_i]$, from the last diagonal element of $\text{Var}[\hat{\mathbf{q}}]$.

The χ^2 test statistic is appropriate here because $\hat{b}_i$ is normally distributed. A χ^2 distribution with 1 degree of freedom is equivalent to:

$$\chi^2_{df=1} = Z^2$$

Where Z is a standard normal variable with $\text{E}[Z] = 0$ and $\text{Var}[Z] = 0$. $\hat{b}_i$ already has mean 0 but non-standard variance. We can divide $\hat{b}_i$ by $\sqrt{\text{Var}[\hat{b}_i]}$ to obtain a standard normal variable, and then square it to obtain the χ^2:

$$\chi^2 = (\frac{\hat{b}_i}{\sqrt{\text{Var}[\hat{b}_i]}})^2 = \frac{\hat{b}_i^2}{\text{Var}[\hat{b}_i]}$$

1.1.3 Case where there are no covariates

In the case where there are no covariates and the phenotype y has been centered by the intercept term:

$$\mathbf{y}^* = \mathbf{y} - \mathbf{1}\hat{\mathbf{c}} \quad \text{where} \quad \hat{\mathbf{c}} = (\mathbf{1}^\intercal \mathbf{V}^{-1} 1)^{-1} 1^\intercal \mathbf{y}$$

then the SNP effect estimate can be simplified to:

$$\hat{b}_i = \frac{\mathbf{w}_i^\intercal \mathbf{V}^{-1} \mathbf{y}^*}{\mathbf{w}_i^\intercal \mathbf{V}^{-1} \mathbf{w}_i} \quad \text{and} \quad \text{Var}[\hat{b}_i] = \frac{1}{\mathbf{w}_i^\intercal \mathbf{V}^{-1} \mathbf{w}_i}$$

The GCTA authors recommend adjusting for the covariates jointly with the SNP, rather than pre-adjusting the phenotype. This is because if the SNP is correlated with the covariates, pre-adjusting the phenotype can reduce power.