# Heritability and genetic correlations

ESSGN 2026 Workshop

Espen Eilertsen & Hans Fredrik Sunde

2026-03-25

# Lecture Outline

1. **Part 0: Statistical tools**

   Variance, covariance, correlation, and the central limit theorem

2. **Part 1: Quantitative genetics and the infinitesimal model**

   Why many small Mendelian effects produce continuous variation and resemblance among relatives

3. **Part 2: Estimating genetic variance**

   Parent-offspring regression, twin designs, and mixed models (GREML)

4. **Part 3: Practical**

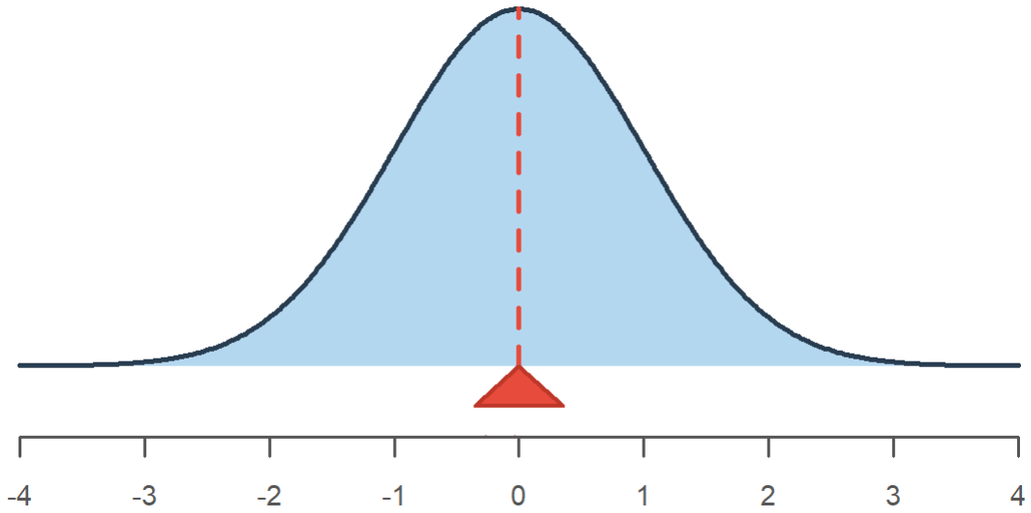
   Simulation and worked examples in R and Julia

# Part 0: Statistical Tools

# Random variables and expectation

A **random variable** $X$ is a quantity whose value is uncertain. We don't know its exact value, but we know the probability of the values it can take - it has a distribution.

The **expectation** (mean) $E[X] = \mu_x$ summarizes its center — the "balance point" of the distribution:




A linear transformation of a random variable $aX + b$ has expectation:

$$E[aX + b] = a\,E[X] + b$$

# Variance

The **variance** measures the spread of a random variable around its mean:

$$\text{Var}(X) = E\left[(X - \mu_X)^2\right]$$

For a linear transformation $aX + b$, the constant has no effect and the variance scales by $a^2$:

$$\text{Var}(aX + b) = a^2 \, \text{Var}(X)$$

> **Note**
>
> - Scaling is linear for expectation ($a$) but quadratic for variance ($a^2$).
> - Adding a constant ($b$) changes expectation but not variance.

# Covariance and correlation

The **covariance** measures how two variables move together:

$$\mathrm{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

The **correlation** is the standardized version:

$$r_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \in [-1, 1]$$

The **regression slope** of $Y$ on $X$ is:

$$b_{XY} = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}$$

**Two useful rules**

- Linear scaling: $\mathrm{Cov}(aX + c,\ bY + d) = ab\,\mathrm{Cov}(X,\ Y)$
- Covariance with itself is variance: $\mathrm{Cov}(X,\ X) = \mathrm{Var}(X)$

# Variance of a sum — an important identity

For two random variables:

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$$

**If** $X$ and $Y$ are uncorrelated ($\mathrm{Cov}(X, Y) = 0$):

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

**Why this matters**

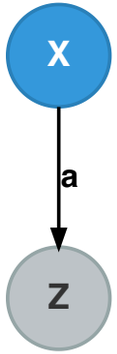
This is the **basis** for heritability. If $y = g + e$ and $\mathrm{Cov}(g, e) = 0$, then:

$$\mathrm{Var}(y) = \mathrm{Var}(g) + \mathrm{Var}(e)$$

We can **partition** the total variance into independent components.

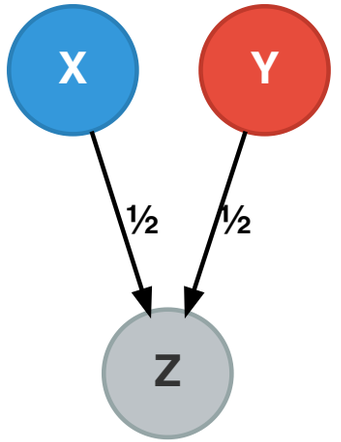
# Path diagrams — visualising the rules

A **path diagram** pictures a set of linear equations. Each circle is a variable, each arrow a scaling.




$$Z = aX \quad \Rightarrow \quad \mathrm{Var}(Z) = a^2\,\mathrm{Var}(X)$$

One arrow — **square the coefficient**.




$$Z = \tfrac{1}{2}X + \tfrac{1}{2}Y, \quad X \perp Y$$

$$\mathrm{Var}(Z) = \tfrac{1}{4}\mathrm{Var}(X) + \tfrac{1}{4}\mathrm{Var}(Y)$$

$$\mathrm{Cov}(X, Z) = \tfrac{1}{2}\,\mathrm{Var}(X)$$

# Variance of a large sum

**1.** For $M$ random variables:

$$\mathrm{Var}\left(\sum_{i=1}^{M} Z_i\right) = \sum_{i=1}^{M}\sum_{j=1}^{M} \mathrm{Cov}(Z_i, Z_j) \qquad \mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{pmatrix}$$

**2.** For $M$ uncorrelated random variables:

$$\mathrm{Var}\left(\sum_{j=1}^{M} Z_j\right) = \sum_{j=1}^{M} \mathrm{Var}(Z_j) \qquad \mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{pmatrix}$$

**3.** If each $Z_j$ also has the same variance $\sigma^2$ (i.i.d.):

$$\mathrm{Var}\left(\sum_{j=1}^{M} Z_j\right) = M\,\sigma^2 \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{pmatrix}$$

# The Central Limit Theorem (CLT)

When many small, independent effects are added together, their sum is approximately **normally distributed** — regardless of the distribution of the individual effects.

> ⓘ **Why this matters for genetics**
>
> This is why the infinitesimal model produces a Gaussian genetic value: each of the $M$ loci contributes a tiny, roughly independent effect, and their sum $g = \sum_j \beta_j x_j$ is approximately normal — even though each allele effect is discrete.

# Summary

These things are very helpful for understanding models of genetic variation - worth repeating!

We'll see these concepts in action later, but here is an example of how they come up:

| **Concept** | **Key fact** | **Where it appears** |
|---|---|---|
| Variance of a sum | $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ if uncorrelated | Heritability |
| Covariance | $\mathrm{Cov}(X,Y)$: how two variables move together | Resemblance between relatives |
| Correlation | Standardized covariance, $\in [-1,1]$ | $r_g, r_e$ |
| Regression slope | $b_{XY} = \mathrm{Cov}(X,Y)/\mathrm{Var}(X)$ | Parent-offspring regression |
| Central limit theorem | Sum of many small effects $\rightarrow$ Normal | Infinitesimal model: $M$ loci produce a normal distribution of genetic values with variance $\sigma_g^2$ |

# Part 1: Quantitative Genetics and The Infinitesimal Model

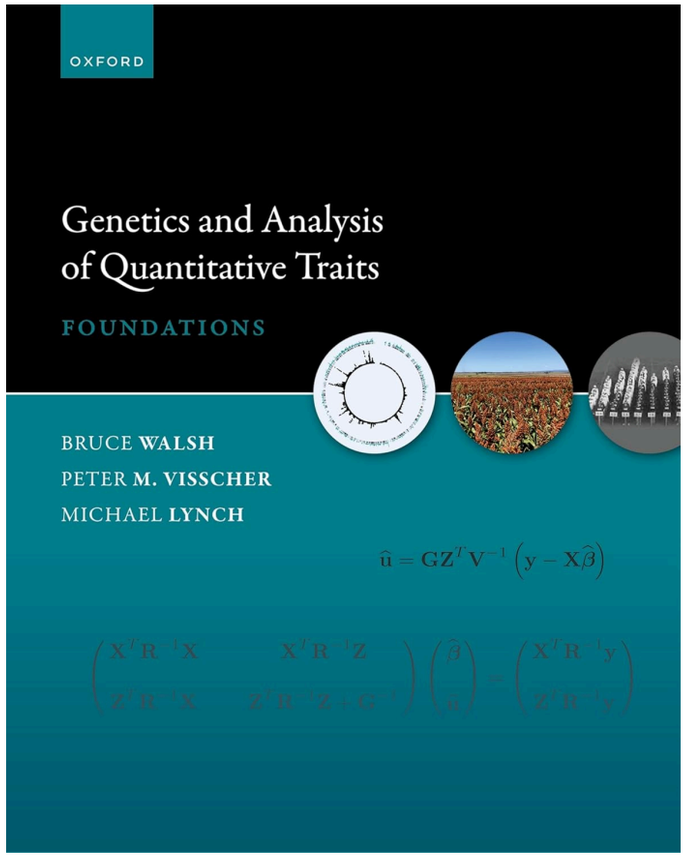
# A central question in genetics

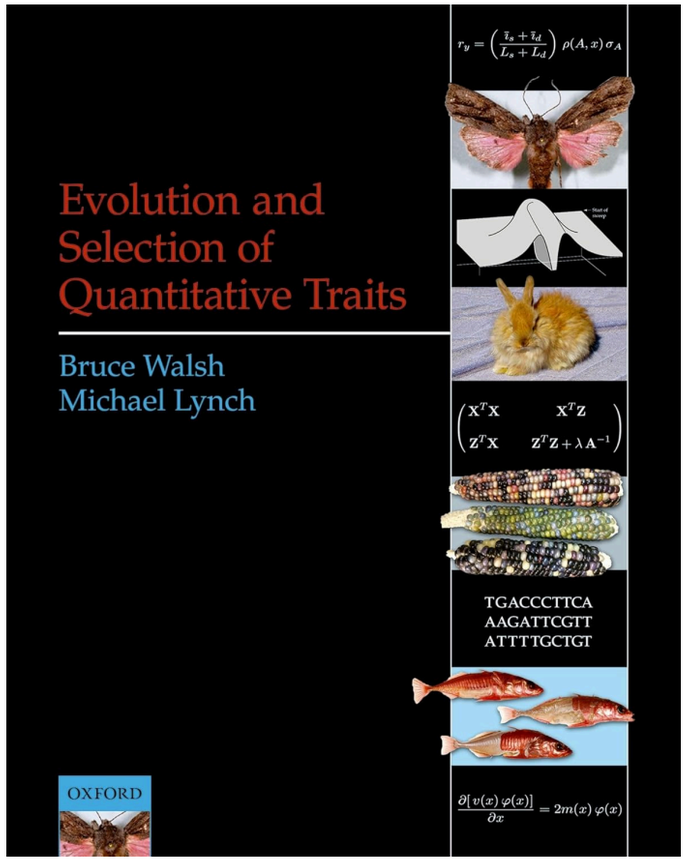
**Do genetic and environmental differences between individuals contribute to differences in outcomes?**

- In quantitative genetics this is a question about the **architecture of variation** — what are its sources?
- Usually addressed from a statistical perspective through **variance decomposition** — what proportion of variance is genetic, environmental, interactive, etc., ?

> **Important**
>
> Theory, statistics, methodology often derive from the same **theoretical model**: the infinitesimal model.

# Central textbooks




Lynch & Walsh, Vol. 1 (1998) and Vol. 2 (2018). These two books contain an extraordinary amount of interesting material. I keep returning to them.

# Fisher (1918): Unifying discrete inheritance with continuous variation

- Mendel (1866): discrete factors of inheritance give discrete traits
- Galton (1886): continuous, normally distributed traits are correlated in families
- **Fisher's resolution:** Many Mendelian loci $\Rightarrow$ approximately continuous trait

# The core idea

A **continuous** trait $y$ arises from a very large number of **discrete loci**, each of **small effect**:

$$y_i = g_i + e_i = \sum_{j=1}^{M} \beta_j \, x_{ij} + e_i$$

- $g_i$ = genetic value for individual $i$
- $e_i$ = environmental deviation for individual $i$
- $x_{ij} \in [0, 1, 2]$ = genotype at locus $j$ (the Mendelian part)
- $\beta_j$ = effect of locus $j$ (very small)
- $M$ = number of causal loci (very large)

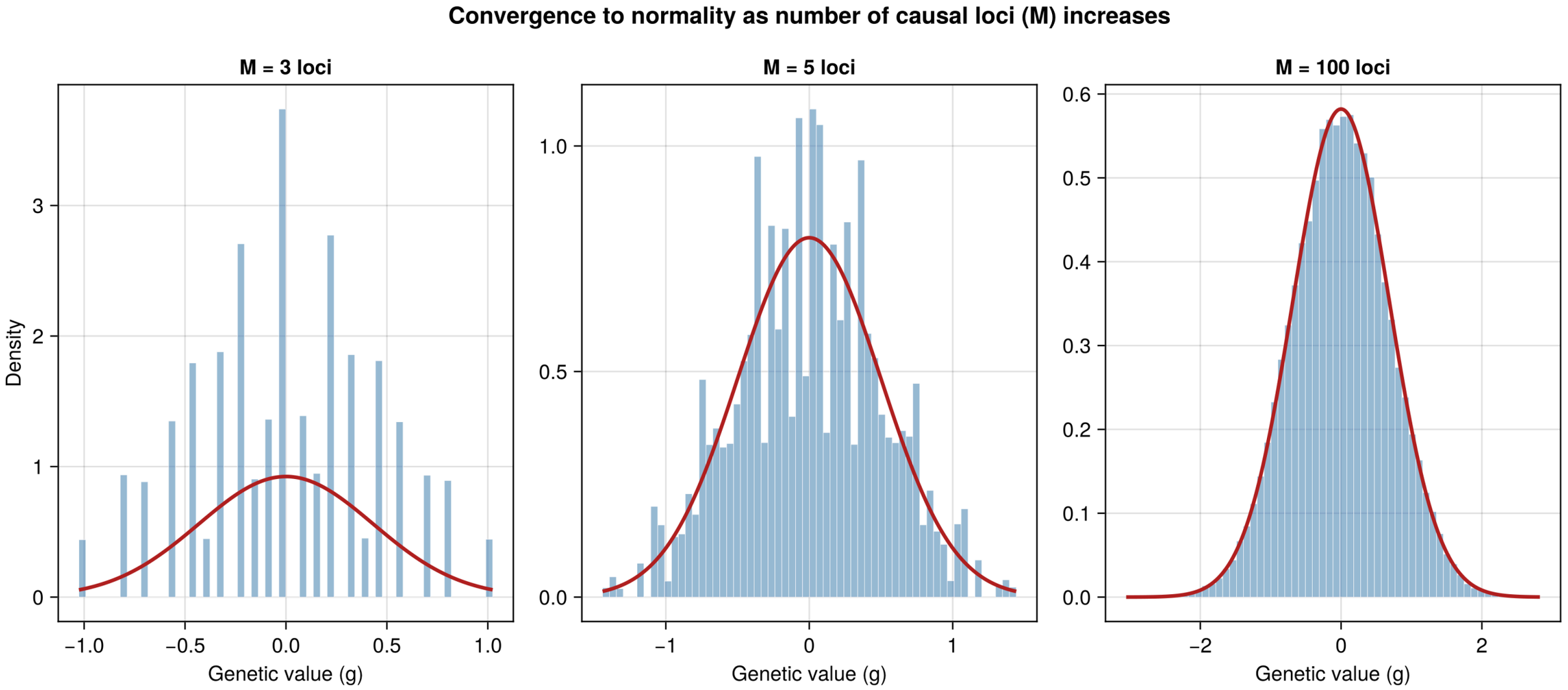
# The Central Limit Theorem does the work

As $M \to \infty$ and each $\beta_j \to 0$:

$$g_i = \sum_{j=1}^{M} \beta_j \, x_{ij} \quad \xrightarrow{d} \quad \square\left(\mu_g, \; \sigma_g^2\right)$$

**Key insight:** Even though inheritance is discrete (Mendelian), the sum over many loci converges in distribution to a **normally distributed random variable**.

# The Central Limit Theorem does the work



# From one trait to two

To talk about correlation between traits, write two outcomes for the same individual:

$$y_{1i} = g_{1i} + e_{1i} \qquad y_{2i} = g_{2i} + e_{2i}$$

with genetic values built from the same loci but trait-specific weights:

$$g_{1i} = \sum_{j=1}^{M} \beta_{1j} x_{ij} \qquad g_{2i} = \sum_{j=1}^{M} \beta_{2j} x_{ij}$$

- The same loci can contribute to both traits
- What differs across traits are the weights $\beta_{1j}$ and $\beta_{2j}$
- The environmental parts $e_1$ and $e_2$ can also be correlated

> **Note**
>
> $y_1$ and $y_2$ can be different phenotypes (for example, depression and anxiety), but they can also be the same trait measured across ages, cohorts, or populations.

# Covariance across traits

The phenotypic correlation is

$$r_y = \frac{\mathrm{Cov}(y_1, y_2)}{\sqrt{\mathrm{Var}(y_1)\mathrm{Var}(y_2)}}$$

Expanding $y_1 = g_1 + e_1$ and $y_2 = g_2 + e_2$ gives

$$\mathrm{Cov}(y_1, y_2) = \mathrm{Cov}(g_1, g_2) + \mathrm{Cov}(e_1, e_2) + \mathrm{Cov}(g_1, e_2) + \mathrm{Cov}(e_1, g_2)$$

So we can define, in exactly the same way:

$$r_g = \frac{\mathrm{Cov}(g_1, g_2)}{\sqrt{\mathrm{Var}(g_1)\mathrm{Var}(g_2)}} \qquad r_e = \frac{\mathrm{Cov}(e_1, e_2)}{\sqrt{\mathrm{Var}(e_1)\mathrm{Var}(e_2)}}$$

# A nuance about genetic correlations

Two related but distinct concepts:

- *Score correlation*, $\mathrm{Corr}(\mathbf{g}_1, \mathbf{g}_2)$: correlation in genetic values across individuals
- *Effect correlation*, $\mathrm{Corr}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$: correlation in causal effects across loci (i.e., pleiotropy)

If SNPs are i.i.d., then these are the same:

$$\mathrm{Corr}(\mathbf{g}_1, \mathbf{g}_2) = \mathrm{Corr}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$$

> **Note**
>
> This equivalence can break if:
>
> - the relevant SNPs have very different variances
> - the causal SNPs are in linkage disequilibrium (LD) with each other

# What does the model predict about resemblance among relatives?

We now have a model: $y_i = g_i + e_i$ where $g_i = \sum_j \beta_j x_{ij}$ is normally distributed with variance $\sigma_g^2$.

The model makes **clear predictions** about how much relatives should resemble each other. These follow from the fact that relatives share alleles — and therefore share parts of their genetic values.

# Genetic transmission 101

At SNP $j$, the offspring gets **one allele** from the father. Let $x_{jc}^{p} \in \{0, 1\}$ be the paternally transmitted allele.

The paternal genotype $x_{jp} \in \{0, 1, 2\}$ determines the distribution of $x_{jc}^{p}$:

| **Parental genotype** | **Transmitted allele** |
|---|---|
| $x_{jp} = 0$ | $x_{jc}^{p} = 0$ |
| $x_{jp} = 1$ | $x_{jc}^{p} \sim \text{Bernoulli}(\frac{1}{2})$ |
| $x_{jp} = 2$ | $x_{jc}^{p} = 1$ |

So the **expected** transmitted allele from the parent is:

$$E[x_{jc}^{p} \mid x_{jp}] = \frac{1}{2} x_{jp}$$

# How do relatives become correlated?

The offspring genotype at SNP $j$ is the sum of the transmitted alleles from the father and the mother:

$$x_{jc} = \underbrace{\left(\frac{1}{2}x_{jp} + \epsilon_{jp}\right)}_{x^p_{jc}} + \underbrace{\left(\frac{1}{2}x_{jm} + \epsilon_{jm}\right)}_{x^m_{jc}}$$

Here, $\epsilon_{jp}$ and $\epsilon_{jm}$ are the random segregation deviations from the parental genotypes at SNP $j$. Under random mating, $\mathrm{Cov}(x_{jp}, x_{jm}) = 0$, so:

$$\mathrm{Cov}(x_{jc}, x_{jp}) = \frac{1}{2}\mathrm{Var}(x_{jp})$$

# Covariance of genetic values among relatives

Now plug the locus-level genotype result into the genetic value model $g_i = \sum_{j=1}^{M} \beta_j x_{ij}$.

Assume cross-locus covariance is zero (linkage equilibrium). By scaling of covariance, $\mathrm{Cov}(\beta_j x_{jp},\ \beta_j x_{jc}) = \beta_j^2 \mathrm{Cov}(x_{jp}, x_{jc})$, so:

$$\mathrm{Cov}(g_p, g_c) = \sum_{j=1}^{M} \beta_j^2\ \mathrm{Cov}(x_{jp}, x_{jc}) = \frac{1}{2} \underbrace{\sum_{j=1}^{M} \beta_j^2 \mathrm{Var}(x_j)}_{\sigma_g^2} = \frac{1}{2}\sigma_g^2$$

The **genetic variance** $\sigma_g^2 = \sum_{\ j} \beta_j^2 \mathrm{Var}(x_j)$ is the total variance contributed by all loci.

**Key results**

- **For each locus** $j, \mathrm{Cov}(x_{jp}, x_{jc}) = \frac{1}{2}\mathrm{Var}(x_j)$.
- **For genetic values:** $\mathrm{Cov}(g_p, g_c) = \frac{1}{2}\sigma_g^2$

# Galton's parent–offspring regression

Galton (1886) regressed offspring height on **mid-parent height** (average of both parents) and found:

> *Tall parents have tall children — but the children are less extreme. The offspring regress toward the population mean.*

He called it **regression to mediocrity**. The infinitesimal model explains why.

**The regression slope of offspring on mid-parent** $\bar{y}_p = \frac{1}{2}(y_{\text{mother}} + y_{\text{father}})$:

$$b_{\text{mid}} = \frac{\text{Cov}(y_c, \bar{y}_p)}{\text{Var}(\bar{y}_p)} = \frac{\frac{1}{2}\sigma^2_{\text{g}}}{\frac{1}{2}\sigma^2_y} = h^2$$

Galton's slope **is** $h^2$. The regression to the mean occurs because an extreme parent is partly extreme due to environment (not transmitted) — the child inherits only the genetic part.

# The general pattern

The same logic applies to any pair of relatives — covariance of genetic values is proportional to correlation of genotypes:

| Relationship | Genotype correlation | Genetic value covariance |
|---|---|---|
| Identical twins (MZ) | $1$ | $\sigma_g^2$ |
| Parent–child | $\frac{1}{2}$ | $\frac{1}{2}\sigma_g^2$ |
| Full siblings | $\frac{1}{2}$ | $\frac{1}{2}\sigma_g^2$ |
| Half siblings | $\frac{1}{4}$ | $\frac{1}{4}\sigma_g^2$ |
| Unrelated | $0$ | $0$ |

**Note**

The structure of resemblance between relatives comes from one thing: **correlated genotypes across loci create correlated genetic values**. The $\beta_j$ are the same in both relatives — it's the $x_j$'s that are shared to different degree.

# Part 2: Estimating Genetic Variance

# Genetic values are latent variables

Only $y$ is **observed** in the model:

$$y_i = g_i + e_i$$

We never observe $g$ (or $e$) directly — it is a *latent* quantity. If we could, then we could estimate $h^2$ directly by computing $\frac{\text{Var}(g)}{\text{Var}(y)}$. *However*, we do know from theory that individuals with correlated genotypes will have correlated $g$ values.

> **Note**
>
> Although $g$ is a person-variable, it is **not** a fixed quantity. The same individual has a different genetic value for height, BMI, educational attainment, and so on.

# Family designs - traditional solution to the latent variable problem

Fit the **ACE model** to MZ and DZ twin correlations on a standardised scale ($\mathrm{Var}(y) = 1$, parameters $h^2$, $c^2$, $e^2$). Three equations:

1. Total variance: $1 = 1 \cdot h^2 + 1 \cdot c^2 + 1 \cdot e^2$
2. MZ twins share all genes and shared environments: $r_{MZ} = 1 \cdot h^2 + 1 \cdot c^2$
3. DZ twins share half their genes and shared environments: $r_{DZ} = \frac{1}{2} h^2 + 1 \cdot c^2$

We can separate the model, parameters, and data into matrices:

$$\mathbf{Ax} = \mathbf{b} \qquad \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ \frac{1}{2} & 1 & 0 \end{pmatrix} \begin{pmatrix} h^2 \\ c^2 \\ e^2 \end{pmatrix} = \begin{pmatrix} 1 \\ r_{MZ} \\ r_{DZ} \end{pmatrix}$$

# The ACE solution

Because $\mathbf{A}$ is square and invertible, the solution is exact:

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} \qquad \begin{pmatrix} h^2 \\ c^2 \\ e^2 \end{pmatrix} = \begin{pmatrix} 0 & 2 & -2 \\ 0 & -1 & 2 \\ 1 & -1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ r_{MZ} \\ r_{DZ} \end{pmatrix}$$

After rearranging these are the well-known Falconer equations:

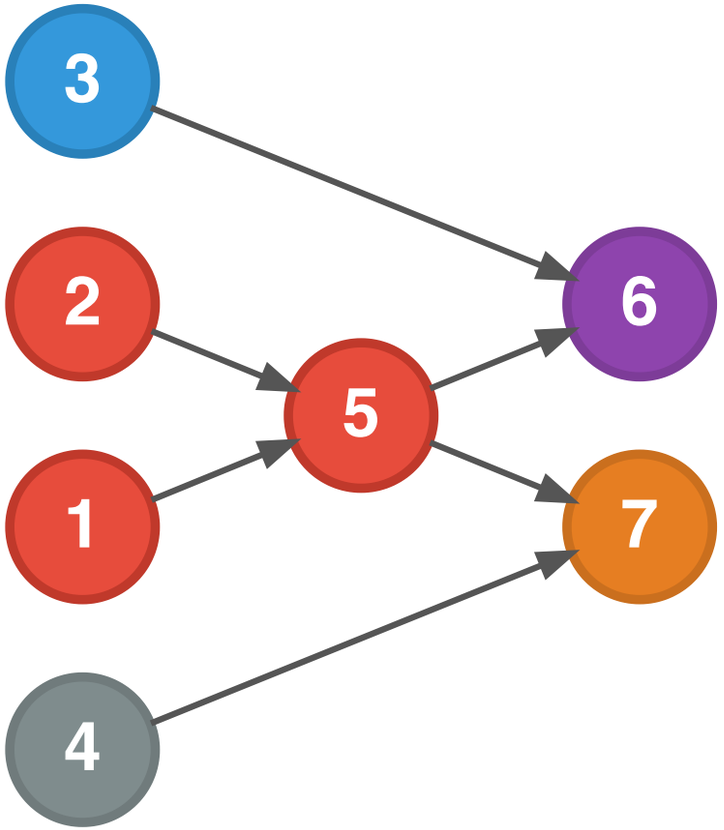
$$h^2 = 2(r_{MZ} - r_{DZ}), \qquad c^2 = 2r_{DZ} - r_{MZ}, \qquad e^2 = 1 - r_{MZ}$$

**Why are you overcomplicating this?**

Because this is generalizable to other relatives. Add more correlations to $\mathbf{b}$ and more rows to $\mathbf{A}$, and you can solve for parameters using any groups of relatives - twin-family designs, adoption studies, etc. Adding columns to $\mathbf{A}$ allows you to solve for more parameters - e.g. dominance, gene-environment interaction, etc.

# Families are complex networks

Real pedigrees create **overlapping** relatedness — individuals do not fall into exclusive groups.




**6** (purple) is simultaneously:

| Relationship | With whom |
|---|---|
| Child | 3, 5 |
| Half-sibling | 7 |
| Grandchild | 1, 2 |

We cannot partition this pedigree into exclusive groups of parent-offspring, full sibs, half sibs, etc. The relatedness structure is a **network**.

# Pedigree tables

Pedigrees are compactly stored as a three-column table with **id** for individuals, their father and mother.

| id | dad | mom |
|---|---|---|
| 1 | 0 | 0 |
| 2 | 0 | 0 |
| 3 | 0 | 0 |
| 4 | 0 | 0 |
| 5 | 1 | 2 |
| 6 | 3 | 5 |
| 7 | 4 | 5 |

- **Founders** (1–4) have no parents in the data — they *anchor* the pedigree
- Every non-founder row encodes **one transmission event**: two parents → one child
- From this table alone we can compute the **additive relatedness matrix** $\mathbf{A}$.

# The relatedness matrix and its inverse

$\mathbf{A}$ Encodes the cumulative results of gene flow — the expected genetic sharing between any two individuals:

$$\mathbf{A} = \begin{pmatrix} 1 & & & & .5 & .25 & .25 \\ & 1 & & & .5 & .25 & .25 \\ & & 1 & & & .5 & \\ & & & 1 & & & .5 \\ .5 & .5 & & & 1 & .5 & .5 \\ .25 & .25 & .5 & & .5 & 1 & .25 \\ .25 & .25 & & .5 & .5 & .25 & 1 \end{pmatrix}$$

$\mathbf{A}^{-1}$ Encodes the mechanism of inheritance — the rules of transmission between parents and offspring:

$$\mathbf{A}^{-1} = \begin{pmatrix} 1.5 & .5 & & & -1 & 0 & \\ .5 & 1.5 & & & -1 & & \\ & & 1.5 & & .5 & -1 & \\ & & & 1.5 & .5 & & -1 \\ -1 & -1 & .5 & .5 & 3 & -1 & -1 \\ 0 & & -1 & & -1 & 2 & \\ & & & -1 & -1 & & 2 \end{pmatrix}$$

Individuals 1 and 6 share genes (grandparent–grandchild), so $\mathbf{A}_{1,6} = .25$. But $\mathbf{A}^{-1}_{1,6} = 0$ because genes flow *through* 5. The green entries trace the path $1 \rightarrow 5 \rightarrow 6$.

In practice we work with $\mathbf{A}^{-1}$, not $\mathbf{A}$: it is **sparse** (nonzeros only between parents, offspring, and co-parents), it can be **built directly** from the pedigree table without ever forming $\mathbf{A}$, and it enters directly in mixed model computations.

# The linear mixed effects model

Statistical models of genetic effects with pedigrees usually rely on a mixed effects model. In the simplest case, phenotypes are modeled as a sum of genetic and environmental effects:

$$\boldsymbol{y} = \mathbf{W}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{g} + \boldsymbol{e}$$

$\mathbf{W}\boldsymbol{\beta}$ is the fixed effect part (covariates), $\mathbf{Z}$ assigns phenotypes to correct genetic effects. $\boldsymbol{g}$ and $\boldsymbol{e}$ are vectors of random effects with models

$$\boldsymbol{g} \sim \square(\mathbf{0}, \sigma_g^2 \mathbf{A}) \qquad \boldsymbol{e} \sim \square(\mathbf{0}, \sigma_e^2 \mathbf{I})$$

This is a statistical formulation of the infinitesimal model, where we treat the genetic values as *random effects*, with a covariance structure defined by the relatedness matrix $\mathbf{A}$.

**Mixed model vs. SEM/twin tradition**

The twin/SEM approach groups individuals into mutually exclusive family types (MZ pairs, DZ pairs, ...) and specifies a separate covariance model for each. The mixed model instead works directly with $\mathbf{A}$, so it handles any relationship in the pedigree simultaneously.

# Mixed model example

Suppose phenotypes are observed only for the non-founders (5, 6, 7):

$$\boldsymbol{y} = \mathbf{W}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{g} + \boldsymbol{e}$$

$$\begin{pmatrix} y_5 \\ y_6 \\ y_7 \end{pmatrix} = \mathbf{W}\beta + \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} g_1 \\ g_2 \\ g_3 \\ g_4 \\ g_5 \\ g_6 \\ g_7 \end{pmatrix} + \begin{pmatrix} e_5 \\ e_6 \\ e_7 \end{pmatrix}$$

Note that the model contains **7 genetic values** but only **3 observations**. Individuals 1–4 have no phenotypes, but they still have genetic values — it's not their fault we didn't collect data on them. Their $g$ values are informed by the pedigree and by data on their relatives.

# Henderson's mixed model equations

Henderson's equations solve simultaneously for the fixed effects and the individual genetic values:

$$\begin{pmatrix} \mathbf{W}'\mathbf{W} & \mathbf{W}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{W} & \mathbf{Z}'\mathbf{Z} + \lambda\mathbf{A}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{g}} \end{pmatrix} = \begin{pmatrix} \mathbf{W}'\boldsymbol{y} \\ \mathbf{Z}'\boldsymbol{y} \end{pmatrix}$$

This gives you **two outputs at once**:

- **Variance components** ($\sigma^2_g, \sigma^2_e$) — estimated iteratively via (RE)ML
- **Predicted genetic values** $\hat{\boldsymbol{g}}$ — individual-level predictions of the latent $g$

The ability to predict $\hat{g}$ for every individual in the pedigree — including those without phenotypes — is a distinctive feature of the mixed model framework.

> **Note**
>
> Predicting genetic values is the primary goal in animal/plant breeding — select the "best" individuals to breed. But predicted $\hat{g}$ may also be useful in social science: as control variables in regression, for studying gene–environment interplay, or for identifying individuals at genetic risk in health and educational contexts.

# Generality of the mixed effects model

The mixed-effects framework handles a wide range of genetic models — all by reshaping **how $\mathbf{Z}$ connects** phenotypes to genetic effects:

| Model | What $\mathbf{Z}$ encodes |
|---|---|
| Multi-trait | Trait indicators — each observation maps to a trait-specific genetic effect |
| G×E | Environmental covariates — genetic effects interact with measured environments |
| Growth curves | Time of measurement — genetic effects vary over age or developmental stage |
| Social genetic effects | Social adjacency matrix — genetic effects of groupmates influence focal individual |

**Kronecker structure**

When $\mathbf{Z}$ grows, the covariance model must match. A common pattern is $\mathrm{Cov}(\mathbf{g}) = \mathbf{\Sigma}_g \otimes \mathbf{A}$, separating the **within-person** covariance ($\mathbf{\Sigma}_g$) from the **between-person** covariance ($\mathbf{A}$).

# Genomic relatedness - modern solution to the latent variable problem

The mixed model uses $\mathbf{A}$ from a pedigree — **expected** relatedness based on family structure. But $\mathbf{A}$ assigns the *same* value to all full siblings ($\frac{1}{2}$), even though actual genome sharing varies.

With genome-wide SNP data, we can measure **realized** genetic similarity directly:

$$\boldsymbol{g} \sim \square(\mathbf{0},\ \sigma^2_g\,\mathbf{G})$$

The model is identical — only $\mathbf{A}$ is replaced by the **genomic relatedness matrix** $\mathbf{G}$, computed from observed genotypes.

**The key insight (Yang et al., 2010)**

Among "unrelated" individuals, $\mathbf{A}$ says relatedness is 0. But $\mathbf{G}$ reveals small variation in genetic similarity — enough to estimate $\sigma^2_g$ in population samples, without relatives.

# What does G look like?

First standardize each SNP $k$ across individuals: $\tilde{x_{ik}} = \frac{x_{ik} - 2\hat{p_k}}{\sqrt{2\hat{p_k}(1-\hat{p_k})}}$

where $\hat{p_k}$ is the sample allele frequency. After standardization, each individual $i$ has a genotype vector $\tilde{\mathbf{x}_i}$ of length $M$ — a *profile* of how they deviate from the population mean across SNPs. Then:

$$G_{ij} = \frac{1}{M}\,\tilde{\mathbf{x}_i} \cdot \tilde{\mathbf{x}_j} \qquad \Longrightarrow \qquad \mathbf{G} = \frac{1}{M}\,\tilde{\mathbf{X}}\tilde{\mathbf{X}}'$$

$G_{ij}$ is the **average co-deviation** between $i$ and $j$ across the genome — how much they tend to deviate in the same direction at a typical locus. At each locus, the contribution is **positive** when both deviate the same way, and **negative** when they deviate in opposite directions.

# Interpreting $G_{ij}$: average co-deviation

At each locus, the co-deviation between $i$ and $j$ is positive when both deviate in the same direction from the population mean, and negative when they deviate in opposite directions. Averaging over all $M$ loci:

- **Unrelated pair**: co-deviations are essentially random — positives and negatives cancel — $G_{ij} \approx 0$
- **Related pair**: shared alleles make co-deviations tend positive — $G_{ij} > 0$
- **Same person** ($i = j$): the average co-deviation with yourself is your average squared deviation from the population — $G_{ii} \approx 1$, but can exceed 1 for individuals who systematically deviate from population mean across the genome (e.g. due to inbreeding)

Since each SNP is standardized to unit variance, $G_{ij}$ is the **average genotype correlation** between two individuals across the genome. The small variation in $G_{ij}$ among unrelated individuals is the signal that identifies $\sigma^2_g$.

# Writing the GREML model from genotypes

Start from a linear model with **random SNP effects** from the standardized genotype matrix $\tilde{\mathbf{X}}$:

$$\mathbf{y} = \tilde{\mathbf{X}}\boldsymbol{\alpha} + \boldsymbol{e} \qquad \boldsymbol{\alpha} \sim \square\left(\mathbf{0},\ \frac{\sigma_g^2}{M}\,\mathbf{I}\right)$$

Each SNP has variance $\sigma_g^2/M$ so that the total genetic variance sums to $\sigma_g^2$ (recall Part 0). Defining the genetic value vector $\mathbf{g} = \tilde{\mathbf{X}}\boldsymbol{\alpha}$, its covariance is:

$$\mathrm{Cov}(\mathbf{g}) = \tilde{\mathbf{X}}\,\frac{\sigma_g^2}{M}\,\mathbf{I}\,\tilde{\mathbf{X}}' = \sigma_g^2\,\underbrace{\frac{\tilde{\mathbf{X}}\tilde{\mathbf{X}}'}{M}}_{\mathbf{G}}$$

So $\mathbf{g} \sim \square(\mathbf{0},\ \sigma_g^2\mathbf{G})$ — the GREML model falls out directly. The genomic relatedness matrix $\mathbf{G}$ is not an *assumption* — it is a **consequence** of i.i.d. random SNP effects and observed genotypes.

# Perspective: Are pedigree models still useful in the genomics era?

Social science often cares about **phenotypic architecture** — what share of variation is genetic, shared-environmental, individual-specific? The pedigree-based mixed model is the natural tool:

- **Registry data** provide millions of pedigree links with rich phenotypes
- **Only family designs** can separate $\sigma_g^2$ from $\sigma_c^2$
- **Extended family designs** identify assortative mating, cultural transmission, indirect genetic effects, …
- **Within-family analyses** are central to causal inference — and rely on pedigree structure

With today's data, pedigree models are arguably more relevant than ever.

# Part 3: Practical

# Practical 1: PO-regression and GREML

Run **sim_genomic.R** (or **.jl**) in parts 1–4:

1. Simulate genotypes for $N$ trios at $M$ SNPs
2. Simulate phenotypes with $h^2 = 0.5$
3. Estimate $h^2$ by parent–offspring regression
4. Estimate $h^2$ by GREML — computing $\mathbf{G}$ is slow; use the wait to discuss what the code does

> **Note**
>
> In practice, $\mathbf{G}$ is computed by optimized genetics libraries. You can also speed things up by pointing R to a fast BLAS (e.g., OpenBLAS, MKL).

sim_genomic.R
sim_genomic.jl

# Part 1: Simulate genotypes

R | Julia

```r
set.seed(080318)
library(gaston)

N = 3000  # Trios
M = 12000 # SNPs
p = 0.5
Xf = matrix(rbinom(N * M, 2, p), N, M)
Xm = matrix(rbinom(N * M, 2, p), N, M)
Xo = matrix(
  rbinom(N * M, 1, as.vector(Xf) / 2) +
  rbinom(N * M, 1, as.vector(Xm) / 2), N, M)
X = rbind(Xf, Xm, Xo)

idf = 1:N
idm = (N + 1):(2 * N)
ido = (2 * N + 1):(3 * N)
```

# Part 2: Simulate phenotypes

R | Julia

```r
p_est = colMeans(X) / 2
X_tilde = scale(X, 2 * p_est, sqrt(2 * p_est * (1 - p_est)))
s2_g = 0.5
s2_e = 0.5
b = rnorm(M, 0, sqrt(s2_g / M))
g = as.vector(X_tilde %*% b)
e = rnorm(3 * N, 0, sqrt(s2_e))
y = g + e

cor(cbind(g[idf], g[idm], g[ido]))
```

```
           [,1]       [,2]      [,3]
[1,]  1.0000000 -0.0235797 0.4687054
[2,] -0.0235797  1.0000000 0.4939061
[3,]  0.4687054  0.4939061 1.0000000
```

```r
cov(cbind(g[idf], g[idm], g[ido]))
```

```
            [,1]        [,2]      [,3]
[1,]  0.48251619 -0.01137793 0.2255605
[2,] -0.01137793  0.48254504 0.2376952
[3,]  0.22556055  0.23769524 0.4799704
```

# Part 3: PO-regression

R | Julia

```r
yp = 0.5 * (y[idf] + y[idm])
yo = y[ido]
m_po = lm(yo ~ yp)
summary(m_po)
```

```
Call:
lm(formula = yo ~ yp)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9114 -0.6556  0.0087  0.6526  3.4581

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.00442    0.01715  -0.258    0.797
yp           0.48668    0.02429  20.034   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9394 on 2998 degrees of freedom
Multiple R-squared:  0.1181,    Adjusted R-squared:  0.1178
F-statistic: 401.4 on 1 and 2998 DF,  p-value: < 2.2e-16
```

# Part 4: GREML

R | Julia

```
G = tcrossprod(X_tilde) / M
Goo = G[ido, ido]
m_direct = lmm.aireml(yo, K = Goo, verbose = F)
m_direct$tau
```

```
[1] 0.4651665
```