Motivation
The problem of predicting genetic merit

What's wrong with what we do now?

The Prediction Problem

Model Equation

$$\mathbf{y = Xb + Zu + e}$$

Other aspects of the model

First moments $\mathbf{E[u] = 0, E[e] = 0}$, therefore $\mathbf{E[y] = Xb}$

Second moments $\mathbf{var[u] = G, var[e] = R}$, $\mathbf{cov[u,e'] = 0}$

Distributional Assumptions e.g. $\mathbf{u, e} \sim \mathrm{MVN}$

Want to predict $\mathbf{u}$ or linear functions like $\mathbf{k'u}$

Original Solution

Generalized Least Squares (GLS)

For estimable $\mathbf{q'b}$, $\mathbf{q'\hat{b}^0}$ is BLUE (Best Linear Unbiased Estimator)

where $\mathbf{\hat{b}^0} = \left(\mathbf{X'V^{-1}X}\right)^{-}\mathbf{X'V^{-1}y}$ for $\mathbf{V} = \mathbf{ZGZ'} + \mathbf{R}$

then $\mathbf{\hat{u}} = \mathbf{GZ'V^{-1}}\left(\mathbf{y} - \mathbf{X\hat{b}^0}\right)$, is BLUP (BLU Predictor)

(same as Selection Index/BLP except $\left(\mathbf{y} - \mathbf{X\hat{b}^0}\right)$ in place of $\left(\mathbf{y} - \mathbf{Xb}\right)$

obtained by exploiting (genetic) covariances between animals

In traditional animal breeding practice

- $\mathbf{G}$ is large and dense and determined by $\mathbf{A}$ the numerator relp matrix
- $\mathbf{V}$ is too big to compute $\mathbf{X'V^{-1}}$

BLP vs GLS BLUP

$\mathbf{y} = \mathbf{X}\beta + \mathbf{Zu} + \mathbf{e}$

$\mathbf{y} - \mathbf{X}\beta = \mathbf{Zu} + \mathbf{e}$, a fully random model

Selection Index Equations $\mathbf{Pb} = \mathbf{Gv}$

$\mathbf{b} = \mathbf{P}^{-1}\mathbf{Gv}$, defines the best linear function to predict $\mathbf{u}$

the "weights" are the same for every animal with the same sources of information (ie same traits observed)

BLP $\hat{\mathbf{u}} = \mathbf{b}'\left(\mathbf{y} - \mathbf{X}\beta\right) = \mathbf{vGP}^{-1}\left(\mathbf{y} - \mathbf{X}\beta\right)$

cf GLS BLUP $\hat{\mathbf{u}} = \mathbf{GZ}'\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\beta}^{0}\right)$

Henderson’s Contributions One

Developed methods to compute **G** and **R** from field data

Henderson’s Method I (not his!), II and III

Including circumstances that involved selection

Henderson's Contributions Two

Invented the Mixed Model Equations

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z+G^{-1}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}}^0 \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{Z'R^{-1}y} \end{bmatrix}, \textit{for full rank } \mathbf{G}$$

and jointly showed $\mathbf{k'\hat{b}^0}$ and $\hat{\mathbf{u}}$ were BLUE and BLUP

Computationally tractable if **G** and **R** assumed diagonal or block-diagonal

(eg sire model with relationships ignored)

(Order 40 matrix takes weeks to invert by hand)

MME typically sparse in national animal evaluation

Henderson's Contributions Three

Invented an algorithm to directly form $\mathbf{A}^{-1}$ from a pedigree list

Then $\mathbf{G}^{-1}$ can be formed as a scalar product or kronecker product

define d to be "mendelian" sampling variance

d=(1, 3/4, 1/2) for 0, 1 or 2 parents known

define $\mathbf{s}'$=(-1/2, -1/2, 1) to represent sire (if known), dam (if known) and individual equations

accumulate $\mathbf{s}d^{-1}\mathbf{s}'$ in the sire, dam and individual rows/columns for every trio of animals in the pedigree list

Consequence of $\mathbf{A}^{-1}$ structure

Accumulate for each animal

$$\begin{array}{c|ccc|}
 & \textit{sire} & \textit{dam} & i \\
\hline
\textit{sire} & 0.25 & 0.25 & -0.5 \\
\textit{dam} & 0.25 & 0.25 & -0.5 \\
i & -0.5 & -0.5 & 1
\end{array} d^{-1}$$

When both parents are known

Nonparents (ie terminal offspring)

Own equation (ie row) has 2 on diagonal, -1 in sire column -1 in dam column

Parent with one offspring

Own equation has 2+1/2 on diagonal, -1 in sire and dam columns

in addition to -1/2 in the column of its mate, -1 in column of offspring

Parent with many offspring to different mates

accumulates a large diagonal element, many small negative offdiagonals

Consider rearranging the MME

In general,

$$\begin{bmatrix} \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z + G^{-1}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}}^{0} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{Z'R^{-1}y} \end{bmatrix}$$

$$\textit{or equivalently} \left[\mathbf{Z'R^{-1}Z + G^{-1}}\right]\left[\hat{\mathbf{u}}\right] = \left[\mathbf{Z'R^{-1}}\left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}^{\mathbf{0}}\right)\right]$$

Single trait animal model $\mathbf{R = I}\sigma_e^2,\qquad \mathbf{G = A}\sigma_g^2,\ \ \mathbf{G}^{-1} = \mathbf{A}^{-1}\sigma_g^{-2}$

$$\textit{or multiplying } \sigma_e^2, \left[\mathbf{Z'Z} + \lambda\mathbf{A^{-1}}\right]\left[\hat{\mathbf{u}}\right] = \left[\mathbf{Z'}\left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}^{\mathbf{0}}\right)\right],\ \textit{with } \lambda = {\sigma_e^2}/{\sigma_g^2}$$

Consider the MME for a nonparent

$$\left[\mathbf{Z'Z} + \lambda\mathbf{A^{-1}}\right]\left[\hat{\mathbf{u}}\right] = \left[\mathbf{Z'}\left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}^{\mathbf{0}}\right)\right]$$

Nonparent animal with one record

$$\left(1+2\lambda\right)\hat{u}_{animal} - \lambda\hat{u}_{sire} - \lambda\hat{u}_{dam} = adjusted\_y$$

$$\hat{u}_{animal} = \frac{2\lambda\left(\hat{u}_{sire} + \hat{u}_{dam}\right)}{\left(1+2\lambda\right)2} + \frac{\left(adjusted\_y\right)}{\left(1+2\lambda\right)}$$

$$= \left(1-w\right)PA + w\left(adjusted\_y\right) \;\; for \;\;\; w = \frac{1}{\left(1+2\lambda\right)}$$

Consider the MME for a nonparent

$$\hat{u}_{animal} = (1-w)PA + w\left(adjusted\_y\right) \;\; for \;\; w = \frac{1}{\left(1+2\lambda\right)}$$

$$\lambda = \frac{1-h^2}{h^2} so \;\; for \;\; h^2 = 1, \;\; \lambda = 0, w = 1, \;\; (no \;\; shrinkage)$$

$$for\, h^2 = low, \;\; \lambda = big, \;\; w = small, \;\; (shrink \; the \; deviation)$$

Two sources of BV information are pooled

The parent average PA

The individual prediction (shrunk deviation)

with heritability influencing shrinkage

Consider the MME for a nonparent

$$\left[\mathbf{Z'Z} + \lambda\mathbf{A}^{\mathbf{-1}}\right][\hat{\mathbf{u}}] = \left[\mathbf{Z'}\left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}^{\mathbf{0}}\right)\right]$$

Nonparent animal with one record

$$\hat{u}_{animal} \quad = (1-w)PA + w\left(adjusted\_y\right)$$

Nonparent animal with no record

$$2\lambda\hat{u}_{animal} - \lambda\hat{u}_{sire} - \lambda\hat{u}_{dam} = 0$$

$$\hat{u}_{animal} = \frac{\lambda\left(\hat{u}_{sire} + \hat{u}_{dam}\right)}{\lambda 2} = \frac{\left(\hat{u}_{sire} + \hat{u}_{dam}\right)}{2} = PA$$

Reliability of nonparents

Property of BLP/BLUP is $\mathrm{cov}(u, \hat{u}) = \mathrm{var}(\hat{u})$ *so* $\mathrm{r}^2 = \dfrac{\mathrm{var}(\hat{u})}{\mathrm{var}(u)}$

but $\hat{u}_{nonparent} = \dfrac{\hat{u}_{sire}}{2} + \dfrac{\hat{u}_{dam}}{2}$, *for nonparent without a record*

so $\mathrm{r}^2_{nonparent} = \dfrac{r^2_{sire}}{4} + \dfrac{r^2_{dam}}{4} \leq \dfrac{1}{2}$

Finally $\Delta G = \dfrac{ir_{nonparent}\sigma_g}{L}$, limiting selection response

when candidates at puberty lack phenotypic information

An option to do better

Solution

- We need a different representation of the covariance between relatives, that allows relatives other than parents to directly contribute to the prediction of nonparents without records
- The NRM or **A**-matrix is an expectation of relationships in the context of repeated sampling of the pedigree (conditional on pedigree)

A-matrix

- Relationship with self is 1+F (noninbred F=0)
- (Additive) relationship of ½ between non-inbred full-sibs and between parents and non-inbred offspring
- Relationship of ¼ between non-inbred half-sibs and between grandparents and offspring
- But particular individuals can have greater or lesser values
  - If we know their genotype we can compute relationships conditional on the chromosome regions they inherited

Relationship matrix

A matrix

$$\begin{bmatrix} 1 & 0 & .5 & .5 & .5 & .5 \\ 0 & 1 & .5 & .5 & .5 & .5 \\ .5 & .5 & 1 & .5 & .5 & .5 \\ .5 & .5 & .5 & 1 & .5 & .5 \\ .5 & .5 & .5 & .5 & 1 & .5 \\ .5 & .5 & .5 & .5 & .5 & 1 \end{bmatrix}$$

A-inverse matrix

$$\begin{bmatrix} 3 & 2 & -1 & -1 & -1 & -1 \\ 2 & 3 & -1 & -1 & -1 & -1 \\ -1 & -1 & 2 & 0 & 0 & 0 \\ -1 & -1 & 0 & 2 & 0 & 0 \\ -1 & -1 & 0 & 0 & 2 & 0 \\ -1 & -1 & 0 & 0 & 0 & 2 \end{bmatrix}$$

Consider a sire, dam and 4 full sibs

Relationship matrix

A matrix

$$\begin{bmatrix} 1 & 0 & .5 & .5 & .5 & .5 \\ 0 & 1 & .5 & .5 & .5 & .5 \\ .5 & .5 & 1 & .5 & .5 & .5 \\ .5 & .5 & .5 & 1 & .5 & .5 \\ .5 & .5 & .5 & .5 & 1 & .5 \\ .5 & .5 & .5 & .5 & .5 & 1 \end{bmatrix}$$

G matrix

$$\begin{bmatrix} 1 & 0 & .5 & .5 & .5 & .5 \\ 0 & 1 & .5 & .5 & .5 & .5 \\ .5 & .5 & 1 & .6 & .4 & .4 \\ .5 & .5 & .6 & 1 & .4 & .4 \\ .5 & .5 & .4 & .4 & 1 & .6 \\ .5 & .5 & .4 & .4 & .6 & 1 \end{bmatrix}$$

A-inverse matrix

$$\begin{bmatrix} 3 & 2 & -1 & -1 & -1 & -1 \\ 2 & 3 & -1 & -1 & -1 & -1 \\ -1 & -1 & 2 & 0 & 0 & 0 \\ -1 & -1 & 0 & 2 & 0 & 0 \\ -1 & -1 & 0 & 0 & 2 & 0 \\ -1 & -1 & 0 & 0 & 0 & 2 \end{bmatrix}$$

G-inverse matrix

$$\begin{bmatrix} 3.5 & 2.5 & -1.25 & -1.25 & -1.25 & -1.25 \\ 2.5 & 3.5 & -1.25 & -1.25 & -1.25 & -1.25 \\ -1.25 & -1.25 & 2.1875 & -0.3125 & 0.3125 & 0.3125 \\ -1.25 & -1.25 & -0.3125 & 2.1875 & 0.3125 & 0.3125 \\ -1.25 & -1.25 & 0.3125 & 0.3125 & 2.1875 & -0.3125 \\ -1.25 & -1.25 & 0.3125 & 0.3125 & -0.3125 & 2.1875 \end{bmatrix}$$

Predict the last animal with no data

$$\begin{bmatrix} -1.25\hat{u}_{sire} & -1.25\hat{u}_{dam} & .3125\hat{u}_{sib1} & .3125\hat{u}_{sib2} & -.3125\hat{u}_{sib3} & 2.1875\hat{u}_{candidate} \end{bmatrix} = [0]$$

$$\hat{u}_{candidate} = \frac{1.25\left(\hat{u}_{sire} + \hat{u}_{dam}\right) - 0.3125\left(\hat{u}_{sib1} + \hat{u}_{sib2}\right) + 0.3125\hat{u}_{sib3}}{2.1875}$$

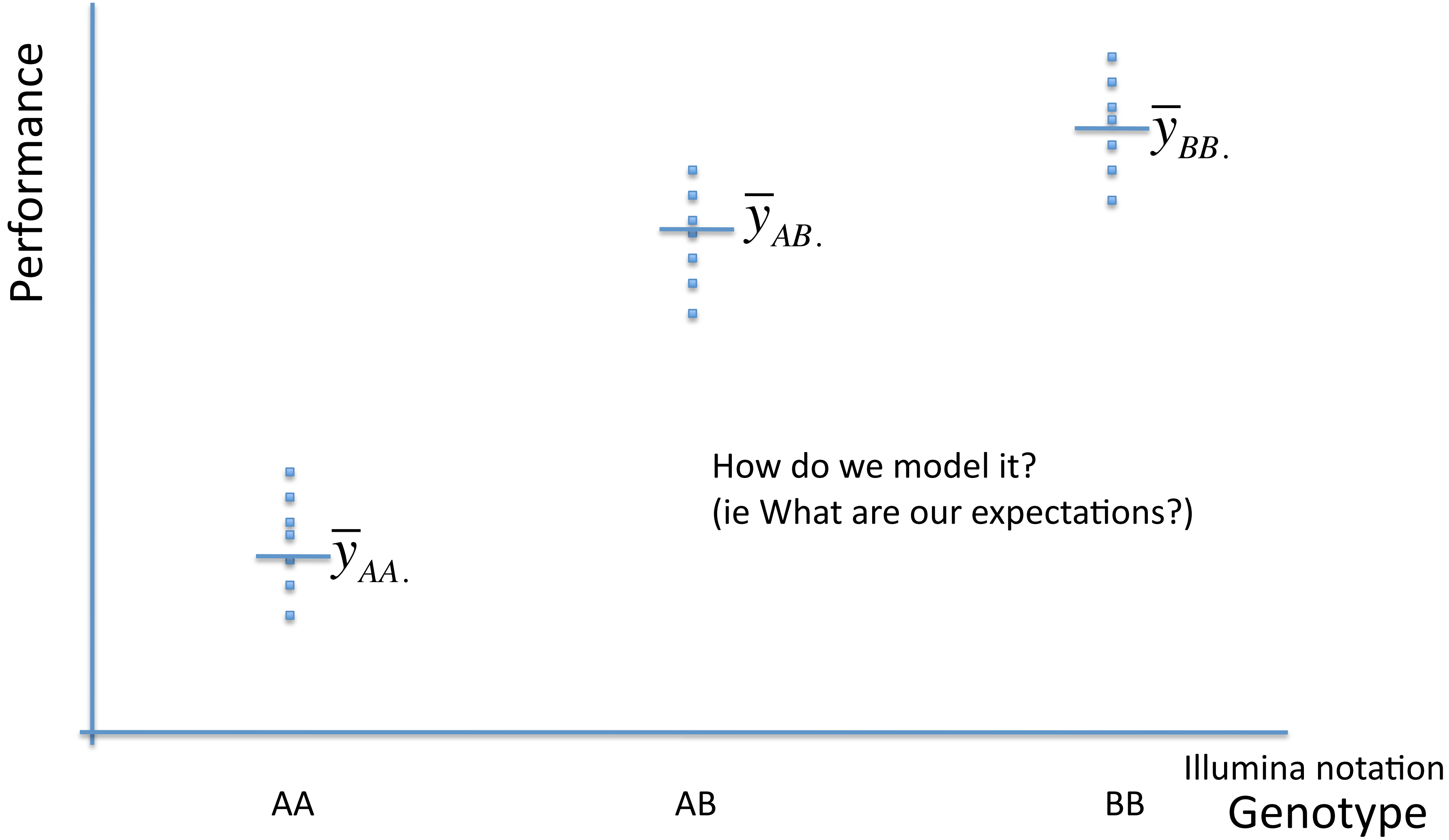
But to form **G**, we needed to know which loci/QTL contribute to variation in performance

Fixed effects models to predict SNP effects

Genomic Prediction

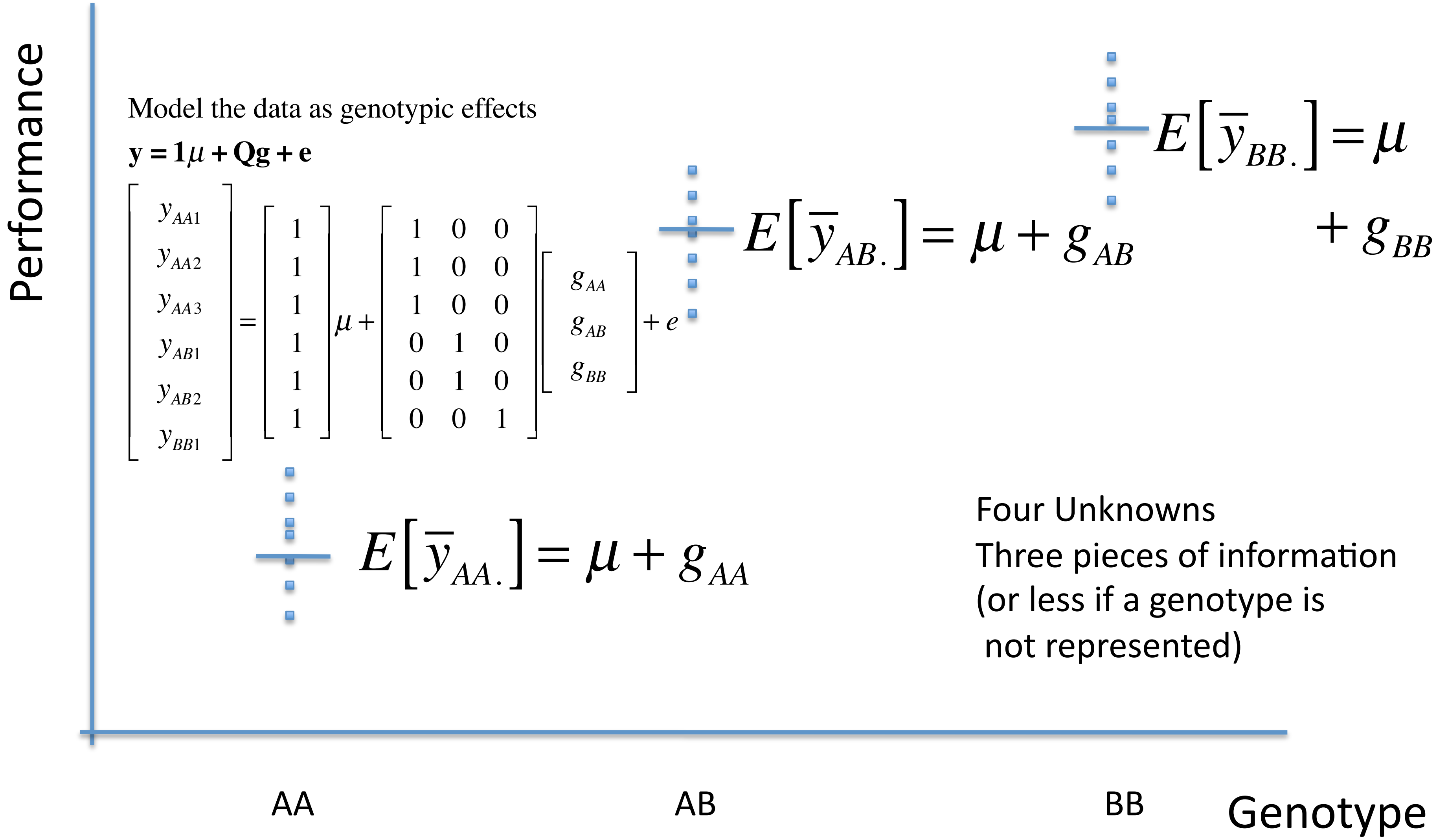
- Two-step process
  - Training population
    - Predict the breeding value of (every) (small) genomic region (to find the informative regions ie QTL)
  - Target population
    - Predict the breeding value of the selection candidates by summing up the breeding values of all the genomic regions they inherited

Data on some locus



How do we model it?
(ie What are our expectations?)

Data on some locus



Model the data as genotypic effects

$\mathbf{y} = \mathbf{1}\mu + \mathbf{Qg} + \mathbf{e}$

$$\begin{bmatrix} y_{AA1} \\ y_{AA2} \\ y_{AA3} \\ y_{AB1} \\ y_{AB2} \\ y_{BB1} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} g_{AA} \\ g_{AB} \\ g_{BB} \end{bmatrix} + e$$

$$E\left[\bar{y}_{BB.}\right] = \mu$$
$$+ g_{BB}$$

$$E\left[\bar{y}_{AB.}\right] = \mu + g_{AB}$$

$$E\left[\bar{y}_{AA.}\right] = \mu + g_{AA}$$

Four Unknowns
Three pieces of information
(or less if a genotype is
not represented)

Parameters and Information Content

- The information content (in fixed effects model) is partly reflected in the degrees of freedom
  - Some degrees of freedom are available to estimate functions of fitted parameters
  - The remainder, if any, contribute to the error sum of squares
- Overparameterized models have more parameters than estimable functions

Fixed Effects Model for Genotypes

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Wq} + \mathbf{e}$$

$\mathbf{b}$ *contains the usual fixed effects*

$$\mathbf{q} = \begin{bmatrix} q_{AA} \\ q_{AB} \\ q_{BB} \end{bmatrix}, \textit{ defines a class effect}$$

$\mathbf{W}$ *is the incidence matrix for* AA, AB, BB *genotypes and has* 3 *columns – one for each genotype class and* N *rows – one for each animal with exactly one* 1 *in each row according to the genotype of the animal*

Fixed Effects Model for Genotypes

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Wq} + \mathbf{e}$$

$$\mathbf{E[y]} = \mathbf{Xb} + \mathbf{Wq}$$

$$\mathbf{var[y]} = \mathbf{var[e]} = \mathbf{I}\sigma_{\mathbf{e}}^{2}$$

Least Squares Equations

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'W} \\ \mathbf{W'X} & \mathbf{W'W} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{q}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{W'y} \end{bmatrix}$$

$$For\ [\mathbf{b}] = [\mu],\ \mathbf{X} = \mathbf{1}$$

$$LHS = \begin{bmatrix} N & n_{AA} & n_{AB} & n_{BB} \\ n_{AA} & n_{AA} & 0 & 0 \\ n_{AB} & 0 & n_{AB} & 0 \\ n_{BB} & 0 & 0 & n_{BB} \end{bmatrix} \quad RHS = \begin{bmatrix} y_{..} \\ y_{AA}. \\ y_{AB}. \\ y_{BB}. \end{bmatrix}$$

Equations have order equal to number of fixed effects plus genotypes

No unique solution

$$LHS = \begin{bmatrix} \cancel{N} & \cancel{n_{AA}} & \cancel{n_{AB}} & \cancel{n_{BB}} \\ \cancel{n_{AA}} & n_{AA} & 0 & 0 \\ \cancel{n_{AB}} & 0 & n_{AB} & 0 \\ \cancel{n_{BB}} & 0 & 0 & n_{BB} \end{bmatrix} \quad RHS = \begin{bmatrix} \cancel{y_{..}} \\ y_{AA\cdot} \\ y_{AB\cdot} \\ y_{BB\cdot} \end{bmatrix}$$

$$\hat{\mathbf{b}} = \begin{bmatrix} 0 \\ \widehat{\mu + q_{AA}} \\ \widehat{\mu + q_{AB}} \\ \widehat{\mu + q_{BB}} \end{bmatrix}, \quad \textit{is one possible solution}$$

No unique solution

$$\hat{\mathbf{b}} = \begin{bmatrix} \widehat{\mu + q_{BB}} \\ \widehat{q_{AA} - q_{BB}} \\ \widehat{q_{AB} - q_{BB}} \\ 0 \end{bmatrix}, \textit{is another possible solution}$$

$$LHS = \begin{bmatrix} N & n_{AA} & n_{AB} & n_{BB} \\ n_{AA} & n_{AA} & 0 & 0 \\ n_{AB} & 0 & n_{AB} & 0 \\ n_{BB} & 0 & 0 & n_{BB} \end{bmatrix} \quad RHS = \begin{bmatrix} y_{..} \\ y_{AA}. \\ y_{AB}. \\ y_{BB}. \end{bmatrix}$$

Different Solutions have same Estimable Functions

$$\hat{\mathbf{b}}_1 = \begin{bmatrix} \widehat{\mu + q_{BB}} \\ \widehat{q_{AA} - q_{BB}} \\ \widehat{q_{AB} - q_{BB}} \\ 0 \end{bmatrix} \qquad \hat{\mathbf{b}}_2 = \begin{bmatrix} 0 \\ \widehat{\mu + q_{AA}} \\ \widehat{\mu + q_{AB}} \\ \widehat{\mu + q_{BB}} \end{bmatrix}$$

Interesting contrasts

$$\mathbf{k}' = \begin{bmatrix} 1 & 1 & 0 & 0 \end{bmatrix} \; then \; \mathbf{k}'\hat{\mathbf{b}}_1 = \mathbf{k}'\hat{\mathbf{b}}_2 = \widehat{\mu + q_{AA}}$$

$$\mathbf{k}' = \begin{bmatrix} 0 & 1 & -1 & 0 \end{bmatrix} \; then \; \mathbf{k}'\hat{\mathbf{b}}_1 = \mathbf{k}'\hat{\mathbf{b}}_2 = \widehat{q_{AA} - q_{AB}}$$

Estimable Functions

- In fixed effects models, many model parameters or functions of model parameters are not estimable, even though a numeric value can be obtained by solving the least squares equations (eg by generalized inverse)

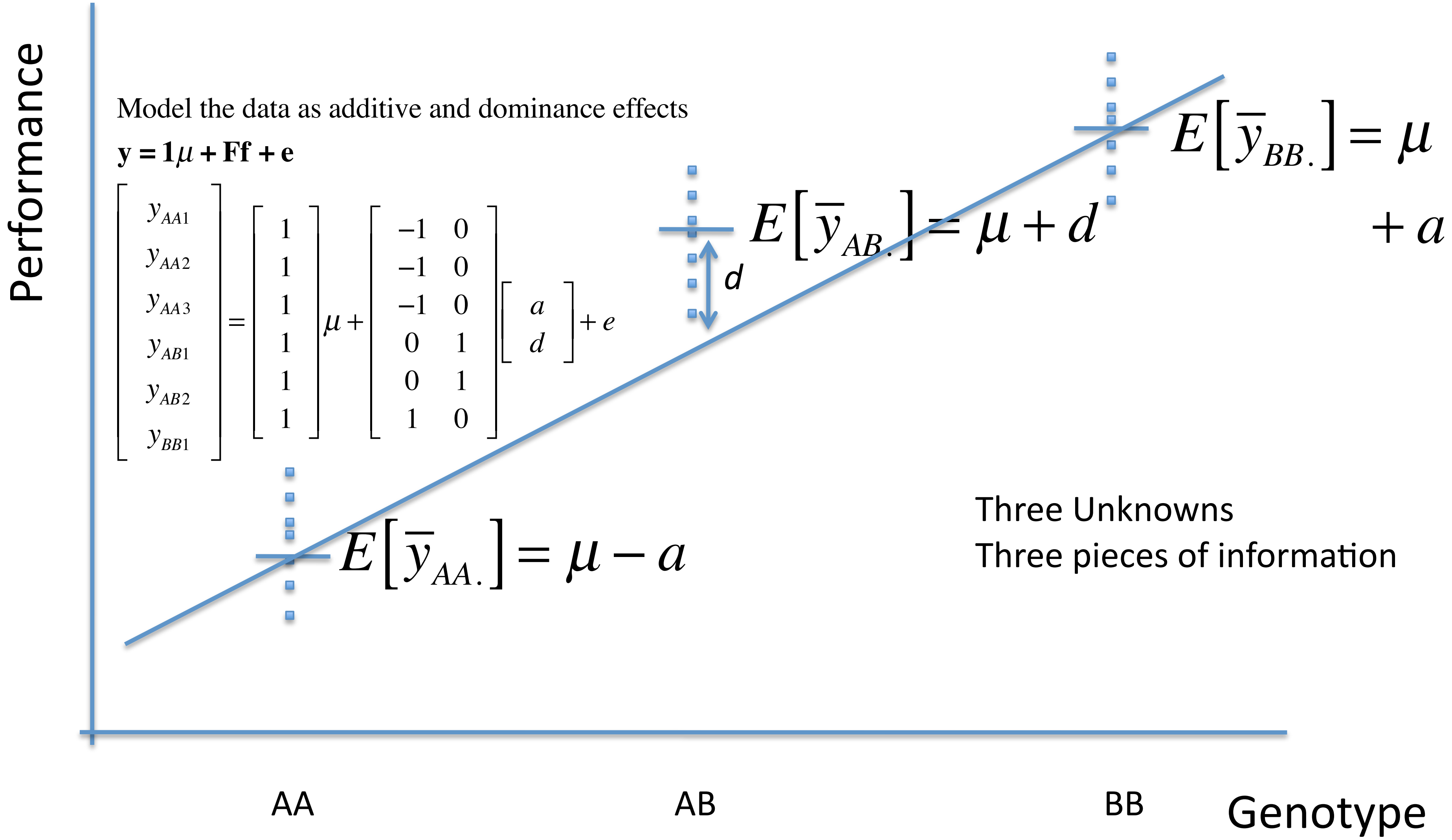
$\left[\mathbf{X'X}\right]^{-}$ is any generalized inverse of $\mathbf{X'X}$ if $\mathbf{(X'X)}\left[\mathbf{X'X}\right]^{-}\mathbf{(X'X) = X'X}$

Define $\mathbf{H = }\left[\mathbf{X'X}\right]^{-}\mathbf{(X'X)}$

A linear function $\mathbf{k'b}^{0}$ is estimable if $\mathbf{k'H = k'}$

$\mathrm{var}(\mathbf{k'b^0}) = \mathbf{k'}\left[\mathbf{X'X}\right]^{-}\mathbf{k}\ \left\{ \mathit{or}\,\mathbf{k'}\left[\mathbf{X'X}\right]^{-}\mathbf{k}\,\sigma^{2}\ \text{(if } \mathbf{R} \text{ was not explicitly fitted)}\right\}$

Data on some locus



Model the data as additive and dominance effects

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Ff} + \mathbf{e}$$

$$\begin{bmatrix} y_{AA1} \\ y_{AA2} \\ y_{AA3} \\ y_{AB1} \\ y_{AB2} \\ y_{BB1} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} -1 & 0 \\ -1 & 0 \\ -1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} a \\ d \end{bmatrix} + e$$

$$E\left[\bar{y}_{BB.}\right] = \mu$$
$$+ a$$

$$E\left[\bar{y}_{AB.}\right] = \mu + d$$

$$E\left[\bar{y}_{AA.}\right] = \mu - a$$

Three Unknowns
Three pieces of information

Genotypic vs genetic effects

$$\mathbf{g} = \begin{bmatrix} g_{AA} \\ g_{AB} \\ g_{BB} \end{bmatrix}, \ \textit{genotypic class effects} \qquad \mathbf{a} = \begin{bmatrix} -a \\ d \\ a \end{bmatrix}, \ \textit{additive and dominance effects}$$

$$a = \frac{g_{BB} - g_{AA}}{2}, \textit{and } d = g_{AB} - \frac{g_{AA} + g_{BB}}{2}$$

$$\mathbf{K}' = \begin{bmatrix} \mathbf{k}_1^{'} \\ \mathbf{k}_2^{'} \end{bmatrix} = \begin{bmatrix} \frac{-1}{2} & 0 & \frac{1}{2} \\ \frac{-1}{2} & 1 & \frac{-1}{2} \end{bmatrix}, \mathbf{K}'\mathbf{q} = \mathbf{a}, \ \textit{columns of } \mathbf{K} \textit{ are othogonal } \ \mathbf{k}_1^{'}\mathbf{k}_2 = 0$$

but note $\mathbf{g}$ *itself is not estimable, but functions like* $g_{BB} - g_{AA}$ *are*

Equivalent Models

	Genotypic	E[ ]	Falconer	E[ ]
AA	$\mu+g_{AA}$	10	μ-a	10=13-3
AB	$\mu+g_{AB}$	14	μ+d	14=13+1
BB	$\mu+g_{BB}$	16	μ+a	16=13+3

$\mu=0$	$\mu=10$	$\mu=16$	$\mu=13$
$g_{AA}= 10$	$g_{AA}= 0$	$g_{AA}= -6$	a= 3
$g_{AB}= 14$	$g_{AB}= 4$	$g_{AB}= -2$	d= 1
$g_{BB}= 16$	$g_{BB}= 6$	$g_{BB}= 0$	

Both models have the same expectation
Both models have the same variance

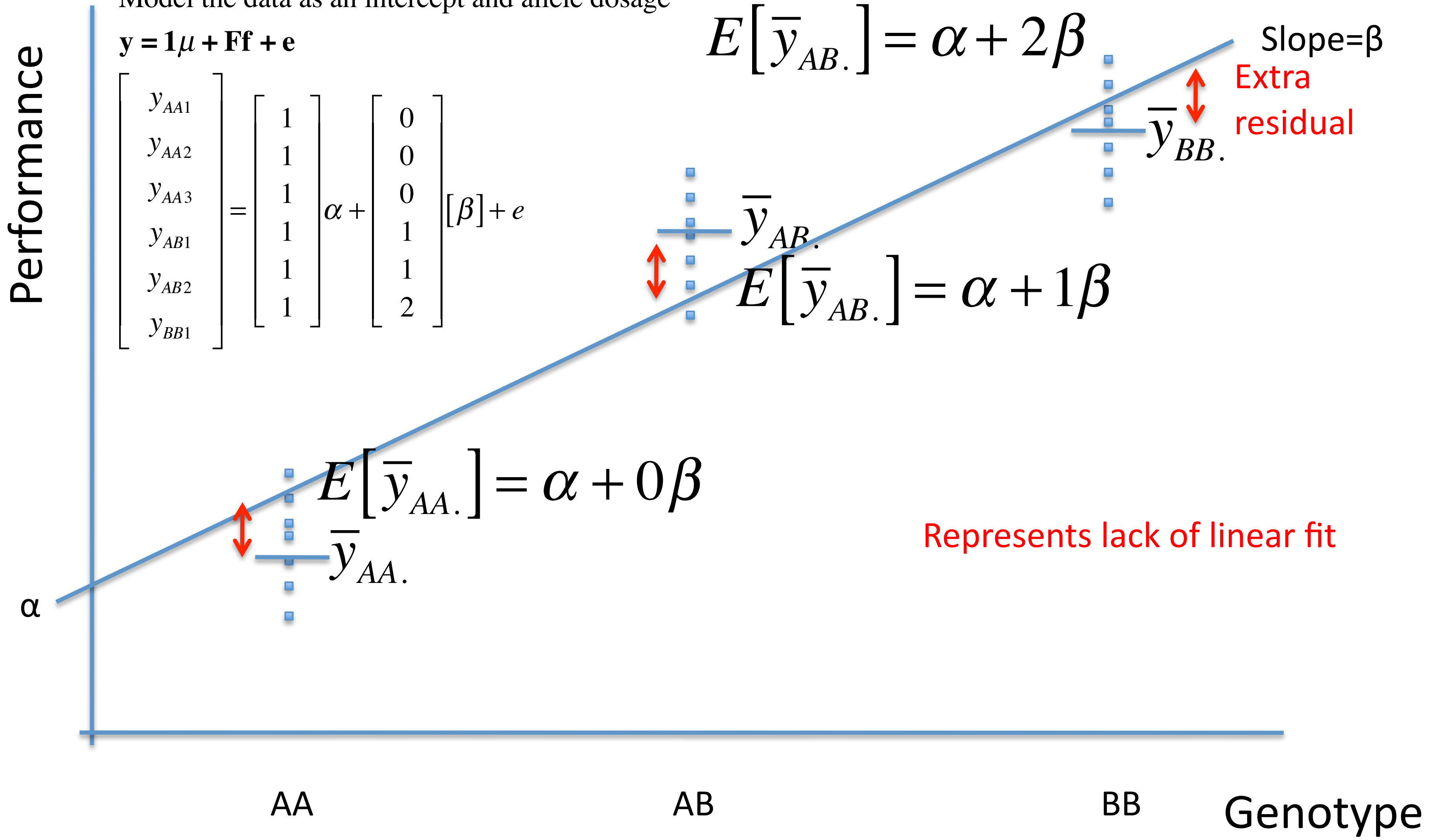
Therefore the models are equivalent
(I can fit either model and migrate from one to the other)

Suppose I ignore dominance (d=0)

Model the data as an intercept and allele dosage

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Ff} + \mathbf{e}$$

$$\begin{bmatrix} y_{AA1} \\ y_{AA2} \\ y_{AA3} \\ y_{AB1} \\ y_{AB2} \\ y_{BB1} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \alpha + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 2 \end{bmatrix} [\beta] + e$$



$$E\left[\bar{y}_{AB.}\right] = \alpha + 2\beta$$

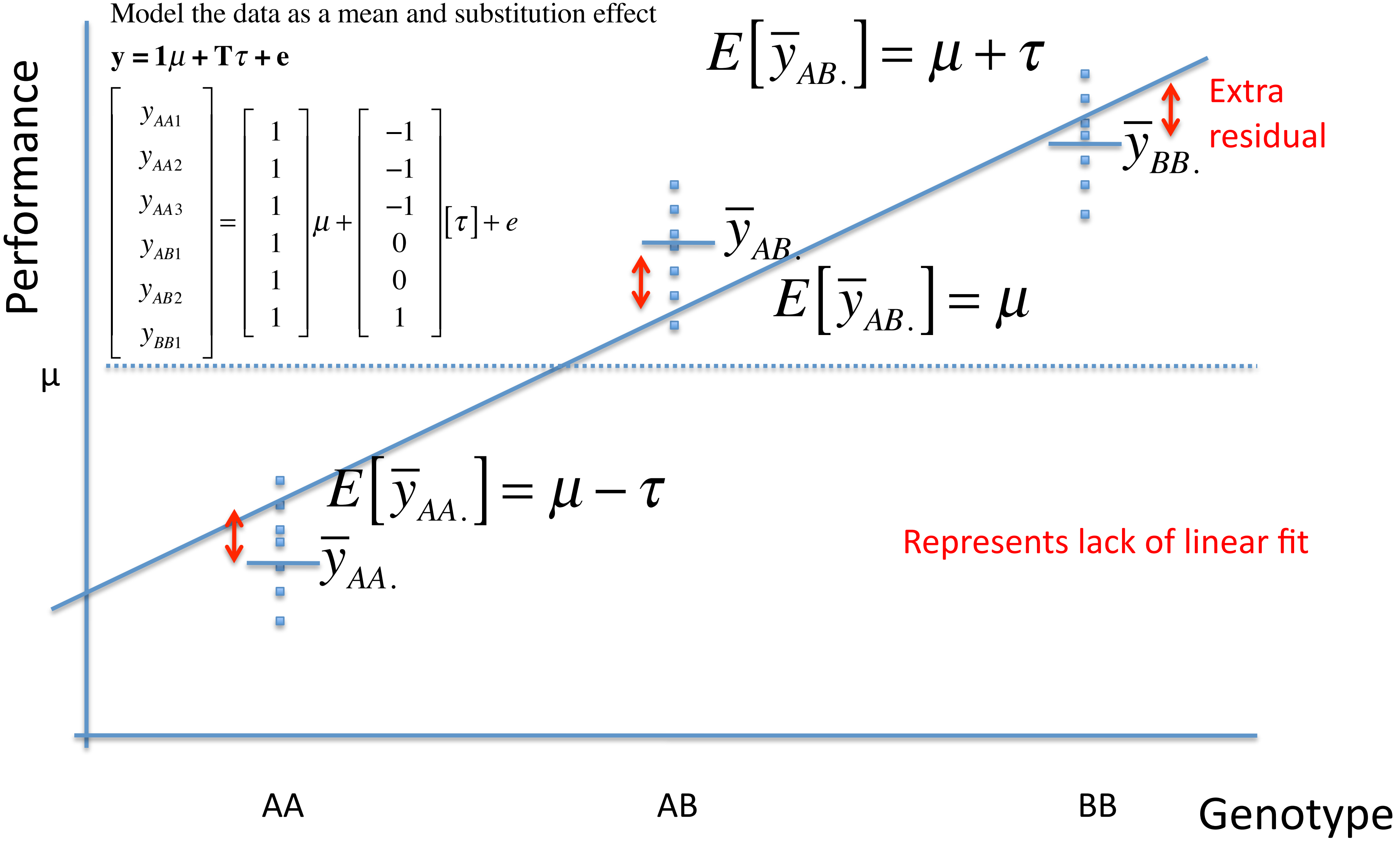
$$E\left[\bar{y}_{AB.}\right] = \alpha + 1\beta$$

$$E\left[\bar{y}_{AA.}\right] = \alpha + 0\beta$$

Extra residual

Represents lack of linear fit

Suppose I ignore dominance (d=0)



Model the data as a mean and substitution effect

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{T}\tau + \mathbf{e}$$

$$\begin{bmatrix} y_{AA1} \\ y_{AA2} \\ y_{AA3} \\ y_{AB1} \\ y_{AB2} \\ y_{BB1} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} -1 \\ -1 \\ -1 \\ 0 \\ 0 \\ 1 \end{bmatrix} [\tau] + e$$

$$E\left[\bar{y}_{AB.}\right] = \mu + \tau$$

$$E\left[\bar{y}_{AB.}\right] = \mu$$

$$E\left[\bar{y}_{AA.}\right] = \mu - \tau$$

$\bar{y}_{BB.}$

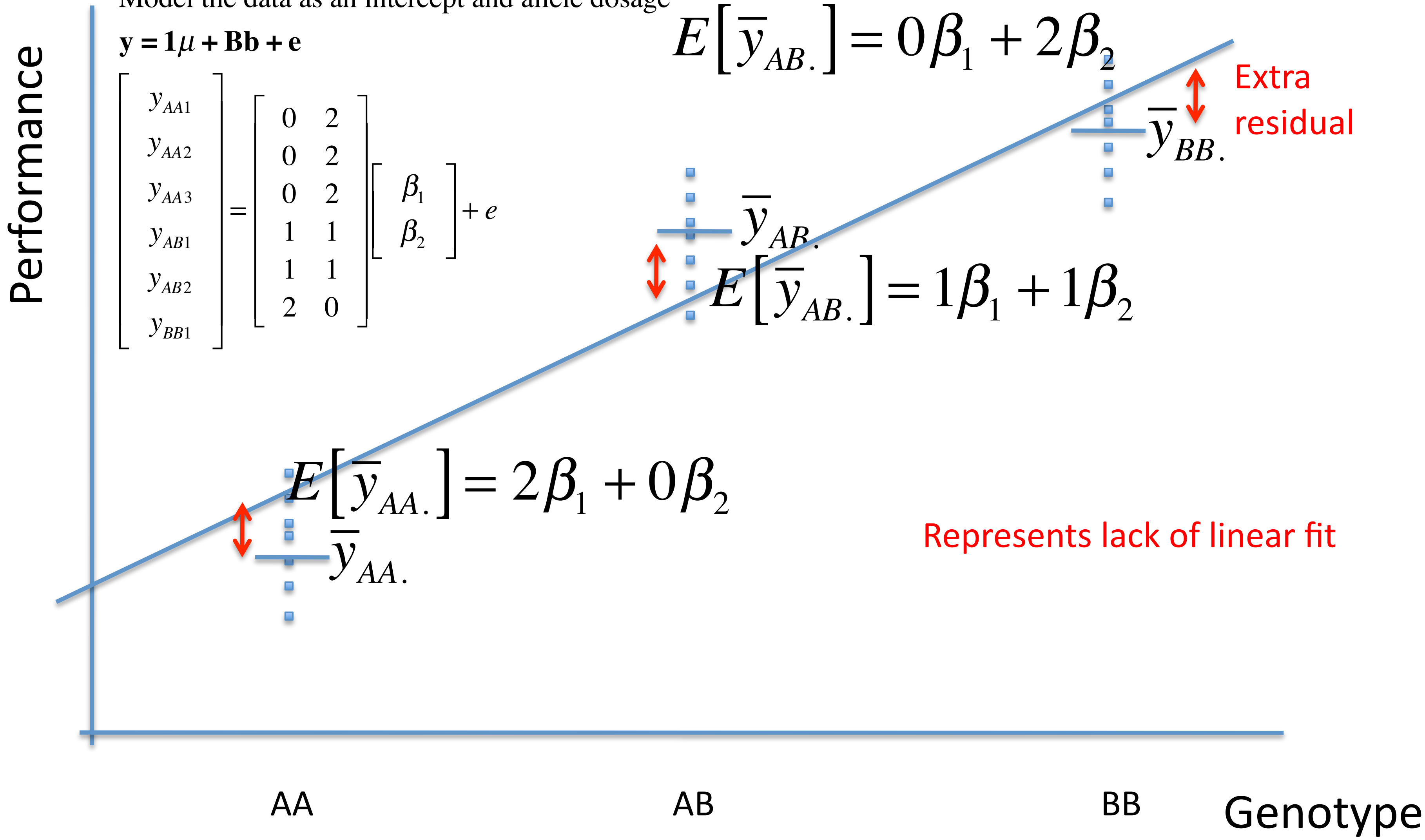
$\bar{y}_{AB.}$

$\bar{y}_{AA.}$

Extra residual

Represents lack of linear fit

Suppose I ignore dominance (d=0)



Model the data as an intercept and allele dosage

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Bb} + \mathbf{e}$$

$$\begin{bmatrix} y_{AA1} \\ y_{AA2} \\ y_{AA3} \\ y_{AB1} \\ y_{AB2} \\ y_{BB1} \end{bmatrix} = \begin{bmatrix} 0 & 2 \\ 0 & 2 \\ 0 & 2 \\ 1 & 1 \\ 1 & 1 \\ 2 & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + e$$

$$E\left[\overline{y}_{AB.}\right] = 0\beta_1 + 2\beta_2$$

$$E\left[\overline{y}_{AB.}\right] = 1\beta_1 + 1\beta_2$$

$$E\left[\overline{y}_{AA.}\right] = 2\beta_1 + 0\beta_2$$

Extra residual

Represents lack of linear fit

Equivalent Models

	Slope & intercept	E[ ]	Mean & Substitution	E[]	Two allelic effects	E[ ]
AA	$\alpha+0\beta$	10	$\mu-\tau$	10	$2\beta_1+0\beta_2$	10=2x5
AB	$\alpha+1\beta$	13	μ	13	$1\beta_1+1\beta_2$	13=5+8
BB	$\alpha+2\beta$	16	$\mu+\tau$	16	$0\beta_1+2\beta_2$	16=2x8

$\alpha=10$
$\beta=3$

$\mu=13$
$\tau=3$

$\beta_1=5$
$\beta_2=8$
NB $\beta_2-\beta_1=3$

All models have the same expectation
All models have the same variance

Therefore the models are equivalent
(I can fit any of the models and migrate from one to the other)

Summary Fixed Effects Models

	Fixed Effects		Random Effects		
	dominance	d=0	dominance	d=0	d=0
Model df	3	2			
Genotypic	yes ↑	no			
All alleles	yes	yes ↑			
Substitution	yes ↓	yes ↓			
Animals	n/a	n/a			

↕ Equivalent models

Summary Fixed Effects Models

	Fixed Effects		Random Effects		
	dominance	d=0	dominance	d=0	d=0
Model df	3	2			
Genotypic	yes	no			
All alleles	yes	yes			
Substitution	yes	yes			
Animals	n/a	n/a			

↕ Equivalent models

↔ Non equivalent models

Fitting SNPs as random effects

Fixed or Random

- Reasonable to consider animal effects as random in the usual context
  - Variation in alleles (ie genotype) between animals that contributes to the genetic variance
    - Not variation in allelic value at a particular locus
- Not so clear that an individual locus (or every loci) should be treated as random
  - Especially when the genotypes are observed and treated as known in the incidence matrix

Suppose we have many loci

The obvious solution is to fit the *a* effects jointly for every locus

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Ma} + \mathbf{e}$$

$$= \mathbf{Xb} + \sum_{i=1}^{i=\mathrm{nmarkers}} \mathbf{m}_i a_i + \mathbf{e}$$

a_i is the substitution effect for the ith locus

Singular Coefficient Matrix

- The incidence matrix of genotypes, **M**, has *n* rows (= number of genotyped animals) and *p* columns (= number of loci/markers/haplotypes)
- Typically using Illumina livestock chips (cattle, horses, pigs, sheep, chickens, dogs) $n < 10{,}000$ and $p > 40{,}000$
- If no 2 animals have the same *p* genotypes, then **M** has full row rank
- The **M'M** component of the coefficient matrix cannot be full rank (rank **M'M** is $n << p$)
  - Rank(AB) is at most the lesser of rank(A) and rank(B)

Practical Consequence

- It is not possible using ordinary least squares to simultaneously estimate more than *n* effects of loci plus other fixed effects
  - Can use stepwise approaches to successively add loci and determine a subset of markers that are informative in the training data
    - But least squares tend to produce upwards biased estimates of effects (especially when power is limiting)
  - Cannot use all markers to predict genomic merit

Alternative Approaches

- Modifications to Least Squares
  - Ridge Regression, Partial Least Squares etc
- Treat *a* effects as random rather than fixed
  - We routinely fit single and multi-trait animal models with many more effects than observations
  - Provides opportunities for many mixed model procedures, such as BLUP, REML, Bayesian analyses
  - These methods will also "shrink" estimates

Summary Fixed Effects Models

Natural (but incorrect) progression to fitting loci as random
Simply augment the coefficient matrix with a variance ratio

	Fixed Effects		Random Effects		
	dominance	d=0	dominance	d=0	d=0
Model df	3	2			
Genotypic	yes	no			
All alleles	yes	yes			
Substitution	yes	yes			
Animals	n/a	n/a			

Everything random is estimable

The random models for substitution effects are NOT equivalent to the other random models unless you are very careful

Random locus effects

- Following the treatment of locus effects as fixed, we could consider the following possible models for random locus effects
  - A) fitting every genotype at a locus
    - This would require us to describe the variance-covariance matrix between the alternative genotypes
    - That matrix is singular in the absence of dominance
  - B) fitting every allele at a locus
  - C) fitting substitution effect at each locus

Mixed Model Theory

- Prediction and estimation follow logically once we define relevant variance-covariance matrices
  - All effects are estimable (unlike least squares)

$$\mathrm{var}(\mathbf{g}) = \mathbf{G} \quad \mathrm{var}(\hat{\mathbf{g}}) = \mathbf{G} - \mathbf{C}^{\mathbf{22}} \quad \mathrm{var}(\hat{\mathbf{g}} - \mathbf{g}) = \mathbf{C}^{\mathbf{22}} \quad r^2_{g\hat{g}} = {}^{\mathrm{var}(\hat{g})}\!\diagup_{\mathrm{var}(g)}$$

$$\mathrm{var}(\mathbf{k'g}) = \mathbf{k'Gk} \quad \mathrm{var}(\mathbf{k'}\hat{\mathbf{g}}) = \mathbf{k'}\left(\mathbf{G} - \mathbf{C}^{\mathbf{22}}\right)\mathbf{k}$$

- The analogous terms in routinely applied animal models are the numerator relationship matrix, genetic and residual variances
  - Random effects might be interpreted in the context of resampling in repeat experiments

Summary of Model Alternatives

	Fixed Effects		Random Effects		
	dominance	d=0	dominance	d=0	d=0
Model df	3	2	Not Relevant		
Genotypic	yes	no	Not considered in this course		
All alleles	yes	yes		This model follows	
Substitution	yes	yes		R≠D	R=D
Animals	n/a	n/a			

Fit all allelic effects as random

- Assuming no dominance we could fit effects of two (or more) individual alleles

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Ma} + \mathbf{e}$$

- ***M*** is a matrix of covariates, one column for each allele (or haplotype), that counts the number of copies – each row sums to two

$$rows\ of\ \mathbf{M}\ are\ one\ of \begin{bmatrix} 0 & 2 \\ 1 & 1 \\ 2 & 0 \end{bmatrix},\ \mathbf{a} = \begin{bmatrix} a_A \\ a_B \end{bmatrix},\ \ for \begin{bmatrix} y_{AA} \\ y_{AB} \\ y_{BB} \end{bmatrix}$$

Estimable Functions in Fixed Models

- Class variables of fixed effects are not estimable
  - Differences between levels in the same class are estimable
  - The sum of any one level and the mean are estimable (in a 1-way model)
  - Fitting a fixed class variable is typically done by
    - deleting the row and column of the coefficient matrix for any one level of the class
    - Introducing a lagrange multiplier to fit a sigma constraint

Sum to Zero in Random Models

- Class variables of random effects (e.g. sire or animal) are all estimable
  - Typically all levels are fitted, even though interest may be focused on differences between levels
    (eg one sire compared to another)
- A feature of BLUP(u) is that certain sums of the elements are zero
  - A biallelic factor fitting say a_1 and a_2 will have solutions that sum to zero (ie a-hat$_1$ = - a-hat$_2$)
  - In a model fitting many biallelic loci as random effects, the number of equations can be halved

Var(**a**) (ie allelic effects)

$$\mathrm{var}(\mathbf{a}) = \mathbf{A} = \mathrm{var}\begin{bmatrix} a_A \\ a_B \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\sigma_A^2 = \mathbf{I}\sigma_A^2$$

For the 3 possible biallelic genotypes

$$\mathrm{var}(\mathbf{MA}) = \mathbf{MAM'} = \begin{bmatrix} 0 & 2 \\ 1 & 1 \\ 2 & 0 \end{bmatrix}\mathbf{A}\begin{bmatrix} 0 & 1 & 2 \\ 2 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 4 & 2 & 0 \\ 2 & 2 & 2 \\ 0 & 2 & 4 \end{bmatrix}\sigma_A^2$$

Note this **A** is the variance-covariance matrix of allelic effects, not the NRM

Peculiar Feature of this Model

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{m_1}a_1 + \mathbf{m_2}a_2 + \mathbf{e} \quad but \;\; \mathbf{m_2} = 2\mathbf{1} - \mathbf{m_1}$$

$$= \mathbf{1}\mu + \mathbf{m_1}a_1 + \left(2\,\mathbf{1} - \mathbf{m_1}\right)a_2 + \mathbf{e}$$

$$= \mathbf{1}\mu + \mathbf{m_1}a_1 - \mathbf{m_1}a_2 + 2\mathbf{1}a_2 + \mathbf{e}$$

$$but \quad 2a_2 = k_2 = \text{constant}$$

$$= \mathbf{1}\left(\mu + k_2\right) + \mathbf{m_1}a_1 - \mathbf{m_1}a_2 + \mathbf{e}$$

Peculiar Feature (cont)

$$\mathbf{y} = \mathbf{1}\mu^* + \mathbf{m_1}a_1 - \mathbf{m_1}a_2 + \mathbf{e} \quad (\textit{last slide})$$

$$\begin{bmatrix} \mathrm{N} & \mathbf{1'm_1} & -\mathbf{1'm_1} \\ \mathbf{m_1'1} & \mathbf{m_1'm_1} + \lambda & -\mathbf{m_1'm_1} \\ -\mathbf{m_1'1} & -\mathbf{m_1'm_1} & \mathbf{m_1'm_1} + \lambda \end{bmatrix} \begin{bmatrix} \hat{\mu}^* \\ \hat{\mathrm{a}}_{\mathbf{1}} \\ \hat{\mathrm{a}}_{\mathbf{2}} \end{bmatrix} = \begin{bmatrix} \mathbf{1'y} \\ \mathbf{m_1'y} \\ -\mathbf{m_1'y} \end{bmatrix}$$

$\textit{Now add equations } 2 \textit{ and } 3$

$$\lambda\hat{\mathrm{a}}_{\mathbf{1}} + \lambda\hat{\mathrm{a}}_{\mathbf{2}} = 0$$

$$\lambda\left(\hat{\mathrm{a}}_{\mathbf{1}} + \hat{\mathrm{a}}_{\mathbf{2}}\right) = 0$$

$$\hat{\mathrm{a}}_{\mathbf{1}} = -\hat{\mathrm{a}}_{\mathbf{2}} \quad \textit{and therefore} \;\; \hat{\mathrm{a}}_{\mathbf{1}} - \hat{\mathrm{a}}_{\mathbf{2}} = 2\hat{\mathrm{a}}_{\mathbf{1}} = -2\hat{\mathrm{a}}_{\mathbf{2}}$$

This “sum to zero” feature is common to all mixed models with factors

Extension to multiple loci

Allelic effects

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Ma} + \mathbf{e} \quad (1\ locus)$$

$$\mathbf{y} = \mathbf{1}\mu + \sum_{i=1}^{i=ploci} \mathbf{M}_i\mathbf{a}_i + \mathbf{e} \quad (p\,loci)$$

MME for two uncorrelated loci (order is 1+ 2 x 2 = 4 allelic effects)

$$\begin{bmatrix} \mathrm{N} & \mathbf{1'M_1} & \mathbf{1'M_2} \\ \mathbf{M_1'1} & \mathbf{M_1'M_1} + \lambda_1 & \mathbf{M_1'M_2} \\ \mathbf{M_2'1} & \mathbf{M_2'M_1} & \mathbf{M_2'M_2} + \lambda_2 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{a}}_1 \\ \hat{\mathbf{a}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{1'y} \\ \mathbf{M_1'y} \\ \mathbf{M_2'y} \end{bmatrix}$$

$$\lambda_i = \frac{\sigma_e^2}{\sigma_{ai}^2}$$

Order of MME is number of fixed effects plus twice number loci (if biallelic)
Consider the implications for 100-1,000 animals with 50,000 loci

Summary of Model Alternatives

	Fixed Effects		Random Effects		
	dominance	d=0	dominance	d=0	d=0
Model df	3	2	Not Relevant		
Genotypic	yes	no	Not considered in this course		
All alleles	yes	yes			
Substitution	yes	yes		R≠D	R=D
Animals	n/a	n/a		This model follows	

An equivalent (animal) model for genomic prediction

More loci than animals

Allellic effects – but for selection we are more interested in animal (not allelic) merit

$$\mathbf{y} = \mathbf{1}\mu + \sum_{i=1}^{i=ploci} \mathbf{M}_i \mathbf{a}_i + \mathbf{e}$$

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{I}\left\{ \sum_{i=1}^{i=ploci} \mathbf{M}_i \mathbf{a}_i \right\} + \mathbf{e}$$

$$\mathbf{y} = \mathbf{1}\mu + "\mathbf{Z}""\mathbf{u}" + \mathbf{e}$$

Order of MME is number of fixed effects plus number of animals
Consider the implications for 100-1,000 animals with 50,000 loci

Mixed Model Equations

$$\mathbf{y} = \mathbf{1}'\mu + \mathbf{Zu} + \mathbf{e}$$

$$\begin{bmatrix} \mathrm{N} & \mathbf{1'Z} \\ \mathbf{Z'1} & \mathbf{Z'Z} + \sigma_e^2 \mathbf{G}^{\mathbf{-1}} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{1'y} \\ \mathbf{Z'y} \end{bmatrix}, \textit{ for full rank } \mathbf{G} = \mathrm{var}(\mathbf{u})$$

$$\mathbf{y} = \mathbf{1}'\mu + \mathbf{I}\sum \mathbf{M}_i \mathbf{a}_i + \mathbf{e}$$

$$\begin{bmatrix} \mathrm{N} & \mathbf{1'} \\ \mathbf{1} & \mathbf{I} + \sigma_e^2 \left[\mathrm{var}\left(\sum \mathbf{M}_i \mathbf{a}_i \right) \right]^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \widehat{\sum \mathbf{M}_i \mathbf{a}_i} \end{bmatrix} = \begin{bmatrix} \mathbf{1'y} \\ \mathbf{y} \end{bmatrix}$$

Order of MME is number of fixed effects plus number of animals
Consider the implications for 100-1,000 animals with 50,000 loci

Mixed Model Equations

$$\mathbf{y} = \mathbf{1'}\mu + \mathbf{I}\sum \mathbf{M}_i \mathbf{a}_i + \mathbf{e}$$

$$\begin{bmatrix} \mathrm{N} & \mathbf{1'} \\ \mathbf{1} & \mathbf{I} + \sigma_e^2 \left[\mathrm{var}\left(\sum \mathbf{M}_i \mathbf{a}_i \right) \right]^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \widehat{\sum \mathbf{M}_i \mathbf{a}_i} \end{bmatrix} = \begin{bmatrix} \mathbf{1'y} \\ \mathbf{y} \end{bmatrix}$$

$$\mathrm{var}\left(\sum \mathbf{M}_i \mathbf{a}_i \right) = \sum \mathrm{var}\left\{ \mathbf{M}_i \mathbf{a}_i \right\} = \sum \mathbf{M_i A_i M_i'} = \sum \mathbf{M_i M_i'} \sigma_{a\mathbf{i}}^2 = like\ \mathbf{A}\sigma_g^2$$

numerator relationship matrix=**A**

$$\begin{bmatrix} \mathrm{N} & \mathbf{1'} \\ \mathbf{1} & \mathbf{I} + \sigma_e^2 \left[\sum \mathbf{M_i M_i'} \sigma_{a\mathbf{i}}^2 \right]^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \widehat{\sum \mathbf{M}_i \mathbf{a}_i} \end{bmatrix} = \begin{bmatrix} \mathbf{1'y} \\ \mathbf{y} \end{bmatrix}$$

An Equivalent Animal Model

$$\mathbf{M}_i\mathbf{M}_i{}'\sigma^2_{ai} \;\; \textit{contains elements like} \begin{bmatrix} 2 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 2 \end{bmatrix} 2\sigma^2_{ai}$$

$\mathbf{M}_i\mathbf{M}_i{}'$ has order equal to number of animals (N)

$\sum \mathbf{M}_i\mathbf{M}_i{}'$ is summed over p loci

A diagonal element for a totally heterozygous animal is $1\times 2\sum\sigma^2_{ai}$

Therefore σ^2_u in a typical animal model is (at least) $2\sum\sigma^2_{ai}$

A diagonal element for a totally homozygous animals is (1+F)=$2\times 2\sum\sigma^2_{ai}$

A typical offdiagonal element is a weighted function of 0, 1 or 2

The number of 0's is the number of loci that the 2 animals are alternate homozygotes

The number of 2's is the number of loci that the 2 animals are the same homozygote

The number of 1's is N minus the number of 0's and 2's

Non-inbred animal

- In the usual context, a non-inbred animal is IBS but not IBD (with a_{ii}=1)
- The fraction of homozygosity across loci is expected to be the sum over all loci of p^2+q^2 in the absence of inbreeding
- Such an animal would have an average diagonal of the genomic matrix >> 1

Summary of Model Alternatives

	Fixed Effects		Random Effects		
	dominance	d=0	dominance	d=0	d=0
Model df	3	2	Not Relevant		
Genotypic	yes	no	Not considered in this course		
All alleles	yes	yes			
Substitution	yes	yes		R≠D	R=D This model follows
Animals	n/a	n/a			

Some alternative computing strategies that are not equivalent models

Reconsider a single locus

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Ma} + \mathbf{e} \qquad \textit{or} \qquad \mathbf{y} = \mathbf{1}\mu + \mathbf{m_1}a_1 + \mathbf{m_2}a_2 + \mathbf{e}$$

$$\begin{bmatrix} \mathrm{N} & \mathbf{1'm_1} & \mathbf{1'm_2} \\ \mathbf{m_1^{'}1} & \mathbf{m_1^{'}m_1} + \lambda & \mathbf{m_1^{'}m_2} \\ \mathbf{m_2^{'}1} & \mathbf{m_2^{'}m_1} & \mathbf{m_2^{'}m_2} + \lambda \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\mathrm{a}}_1 \\ \hat{\mathrm{a}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{1'y} \\ \mathbf{m_1^{'}y} \\ \mathbf{m_2^{'}y} \end{bmatrix}$$

For $\lambda = \dfrac{\sigma_e^2}{\sigma_a^2}$, *these MME have the same solution for* $\hat{a}_1 - \hat{a}_2$ $(\textit{but not } \hat{\mu})$ *as*

$$\begin{bmatrix} N & \mathbf{1'm_1} \\ \mathbf{m_1^{'}1} & \mathbf{m_1^{'}m_1} + \lambda/2 \end{bmatrix} \begin{bmatrix} \hat{\mu}\,* \\ \widehat{a_1 - a_2} \end{bmatrix} = \begin{bmatrix} \mathbf{1'y} \\ \mathbf{m_1^{'}y} \end{bmatrix}$$

As if we fitted $\mathbf{y} = \mathbf{1}\mu + \mathbf{m_1}a_1 + \mathbf{e}$ *with different* λ

Proof of Identical Solutions

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Ma} + \mathbf{e} \text{ (Model I)}, \ \textit{with} \ \ \mathbf{M'1} = 2\mathbf{1}$$

$$E[y] = \mu, \ \text{var}[\mathbf{y}] = \mathbf{MM'}\sigma_a^2 + \mathbf{I}\sigma_e^2 \quad \lambda_I = \sigma_e^2 / \sigma_a^2$$

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{m_1}a_1 + \mathbf{m_2}a_2 + \mathbf{e} \quad \textit{but} \ \ \mathbf{m_2} = 2\mathbf{1} - \mathbf{m_1}$$

$$= \mathbf{1}\mu + \mathbf{m_1}a_1 + \left(2\,\mathbf{1} - \mathbf{m_1}\right)a_2 + \mathbf{e} \quad \textit{but} \ \ 2a_2 = k_2 = \text{constant}$$

$$= \mathbf{1}\left(\mu + k_2\right) + \mathbf{m_1}a_1 - \mathbf{m_1}a_2 + \mathbf{e}$$

$$= \mathbf{1}\left(\mu + k_2\right) + \mathbf{m_1}\left(a_1 - a_2\right) + \mathbf{e} \quad \text{(Model II)}$$

$$E[y] = \left(\mu + k_2\right), \ \text{var}[\mathbf{y}] = \mathbf{m_1}\mathbf{m_1}' 2\sigma_a^2 + \mathbf{I}\sigma_e^2 \quad \lambda_{II} = \sigma_e^2 / 2\sigma_a^2 = \lambda_I / 2$$

Clearly the first and second moments are different in models I and II

More Alternatives

Previously $\quad \mathbf{y} = \mathbf{1}\left(\mu + k_2\right) + \mathbf{m_1}\left(a_1 - a_2\right) + \mathbf{e}$

Note $\mathbf{m_1}$ $\left(\text{and } \mathbf{m_2}\right)$ contain covariate values of 0, 1 or 2

another model with $\mathrm{k}_{12} = \left(a_1 - a_2\right)$ is

$$\mathbf{y} = \mathbf{1}\left(\mu + k_2 + k_{12}\right) + \mathbf{m_1}\left(a_1 - a_2\right) \text{-} \mathbf{1}\left(a_1 - a_2\right) + \mathbf{e}$$

$$\mathbf{y} = \mathbf{1}\left(\mu + k_2 + k_{12}\right) + \left(\mathbf{m_1} \text{ -} \mathbf{1}\right)\left(a_1 - a_2\right) + \mathbf{e}$$

whereby the covariate values are now -1, 0 and 1

Computational Alternatives

	covariates
$\mathbf{y} = \mathbf{1}\mu + \mathbf{Ma} + \mathbf{e}$	$0, 1, 2$ and $2, 1, 0$
$\mathbf{y} = \mathbf{1}(\mu + k_2) + \mathbf{m_1}(a_1 - a_2) + \mathbf{e}$	$0, 1, 2$
$\mathbf{y} = \mathbf{1}(\mu + k_2 + k_{12}) + (\mathbf{m_1} \mathbf{-1})(a_1 - a_2) + \mathbf{e}$	$-1, 0, 1$
$\mathbf{y} = \mathbf{1}(\mu + k_1) + \mathbf{m_2}(a_2 - a_1) + \mathbf{e}$	$2, 1, 0$
$\mathbf{y} = \mathbf{1}(\mu + k_1 + k_{21}) + (\mathbf{m_2} \mathbf{-1})(a_2 - a_1) + \mathbf{e}$	$1, 0, -1$

All these models have different E[y]

All these models have identical predictions of random effects

Only the first model has the correct PEV for the random effect if **e** assumed diagonal

Consider the genetic part of var[y]

$$\mathbf{M} = \begin{bmatrix} \mathbf{m}_1 & \mathbf{m}_2 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 1 & 1 \\ 0 & 2 \end{bmatrix}$$

covariate

M

$\mathbf{m}_1$

$\mathbf{m}_1$ **-1**

$\mathbf{m}_2$

$\mathbf{m}_2$ **-1**

genetic variance (**ZGZ'**)

$$\text{var}\left[\mathbf{Ma}\right] = \mathbf{MAM'} = \begin{bmatrix} 4 & 2 & 0 \\ 2 & 2 & 2 \\ 0 & 2 & 4 \end{bmatrix} \sigma_a^2$$

$$\text{var}\left[\mathbf{m_1}\left(a_1 - a_2\right)\right] = 2\sigma_a^2 \mathbf{m_1 m}_1^{'} = \begin{bmatrix} 4 & 2 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} 2\sigma_a^2$$

$$\text{var}\left[\left(\mathbf{m_1} - 1\right)\left(a_1 - a_2\right)\right] = 2\sigma_a^2 \left(\mathbf{m_1} - 1\right)\left(\mathbf{m_1} - 1\right)' = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix} 2\sigma_a^2$$

$$\text{var}\left[\mathbf{m}_2\left(a_2 - a_1\right)\right] = 2\sigma_a^2 \mathbf{m_2 m}_2^{'} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 2 & 4 \end{bmatrix} 2\sigma_a^2$$

$$\text{var}\left[\left(\mathbf{m}_2 - 1\right)\left(a_2 - a_1\right)\right] = 2\sigma_a^2 \left(\mathbf{m_2} - 1\right)\left(\mathbf{m_2} - 1\right)' = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix} 2\sigma_a^2$$

These are typically singular, unless there are more loci than animals

Animal Model Counterpart

Any full rank inverse of the following can be used in place of $\mathbf{A}^{\mathbf{-1}}\boldsymbol{\sigma}_a^2$ *in MME to predict animal merit*

$$\sum \mathbf{M}_i\mathbf{M}_i{}'\sigma_{ai}^2 = \sum\left(m_{1i}m_{1i}^{'} + m_{2i}m_{2i}^{'}\right)\sigma_{ai}^2$$

$$\sum m_{1i}m_{1i}^{'}2\sigma_{ai}^2$$

$$\sum m_{2i}m_{2i}^{'}2\sigma_{ai}^2$$

$$\sum\left(m_{1i}-1\right)\left(m_{1i}-1\right)'2\sigma_{ai}^2$$

$$\sum\left(m_{2i}-1\right)\left(m_{2i}-1\right)'2\sigma_{ai}^2$$

Only the first can be used for PEV or r^2

Summary of Model Alternatives

	Fixed Effects		Random Effects		
	dominance	d=0	dominance	d=0	d=0
Model df	3	2	Not Relevant		
Genotypic	yes	no	Not considered in this course		
All alleles	yes	yes			
Substitution	yes	yes		R≠D This model follows	R=D
Animals	n/a	n/a			

Correct handling of the model

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Ma} + \mathbf{e} \quad \textit{with} \quad \mathbf{M'1} = 2\mathbf{1}$$

$$E[y] = \mu,\ \mathrm{var}[\mathbf{y}] = \mathbf{MM'}\sigma_a^2 + \mathbf{I}\sigma_e^2 \qquad \lambda_I = \sigma_e^2 / \sigma_a^2$$

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{m_1}a_1 + \mathbf{m_2}a_2 + \mathbf{e} \quad \textit{but} \quad \mathbf{m_2} = 2\mathbf{1} - \mathbf{m_1}$$

$$= \mathbf{1}\mu + \mathbf{m_1}a_1 + \left(2\,\mathbf{1} - \mathbf{m_1}\right)a_2 + \mathbf{e}$$

$$= \mathbf{1}\mu + \mathbf{m_1}a_1 - \mathbf{m_1}a_2 + \left(\mathbf{1}2a_2 + \mathbf{e}\right)$$

$$= \mathbf{1}\mu + \mathbf{m_1}\left(a_1 - a_2\right) + \mathbf{e}^*$$

$$\textit{with}\ \mathrm{var}(\mathbf{e}^*) = \mathrm{var}\left(\mathbf{1}2a_2 + \mathbf{e}\right) = 4\mathbf{11'}\sigma_a^2 + \mathbf{I}\sigma_e^2$$

$$\textit{but}\ \mathrm{cov}\left[\left(a_1 - a_2\right), \mathbf{e}^{*}{}'\right] = -2\mathbf{1}'\mathrm{var}\,a_2 \neq 0 \Rightarrow \textit{no MME},\ \textit{GLS OK}$$

Summary of Model Alternatives

	Fixed Effects		Random Effects		
	dominance	d=0	dominance	d=0	d=0
Model df	3	2	Not Relevant		
Genotypic	yes	no	Not considered in this course		
All alleles	yes	yes		(1)	
Substitution	yes	yes		R≠D Not MME	R=D (2)
Animals	n/a	n/a		(1)	(2)

Models (1) are equivalent

Models (2) are equivalent (if both use 1st allele, or 2nd allele, or -1,0,1 etc)

Models (1) and (2) give the same BLUP solutions, but not PEV or r^2

Equivalent “Animal” Model

- Any of these models with equivalent computations for loci effects, can be formulated to solve for animal effects rather than locus effects
  - Give identical estimates for every animal
  - Will not all give the same PEV for animal (or locus) effects
    - This has implications in quantifying accuracy/reliability

Two practical problems in high-density genomic prediction

The Genomic Selection Problem

- Estimating SNP effects using BLUP and predicting the merit of new animals is straightforward
  - Given the correct model equation
    - That is, knowledge of informative/uninformative loci
    - Given SNPs in perfect LD with QTL
  - Given the second moments
    - That is known variance of informative SNP loci (assuming they really are random effects) and known residual variance
- Real life neither of these requisites are known

SNPs not in perfect LD with QTL

- Ability of the SNP to act as a surrogate (or marker) for the QTL will erode as you use the SNP to compute covariances between relatives separated by a greater number of meioses
  - Hope that genomic training will identify the SNPs in highest LD with (and closest) to the QTL

Simplest genomic selection model

- Partition genetic variance "equally" among all loci
- MHG (sadly) referred to this method as BLUP
  - Sometimes called GBLUP or RR-BLUP (for random regression or ridge regression)

More complex model

- Partition variance unequally among every locus (Bayes A)
  - Practical impact of this will depend upon shrinkage
- Partition variance unequally among a subset of the loci (Bayes B)
  - But which subset?
  - And how do you assume the size of the subset, a parameter they referred to as π

The variance component problem

- We need to jointly estimate the residual and genetic variances for perhaps tens of thousands of loci, simultaneously considering model selection criteria to discard models with low levels of support
  - 50k 1-locus additive models
  - About $50k^2$ 2-locus models and so on
  - Little knowledge of how many loci might be needed but it could be hundreds

Fitted Model

- We will use the model that fits a substitution effect for each locus, recognizing that we cannot use the equations for estimating reliabilities
  - Equations are too big anyway
  - Bayesian posteriors can be used for reliability of SNP effects