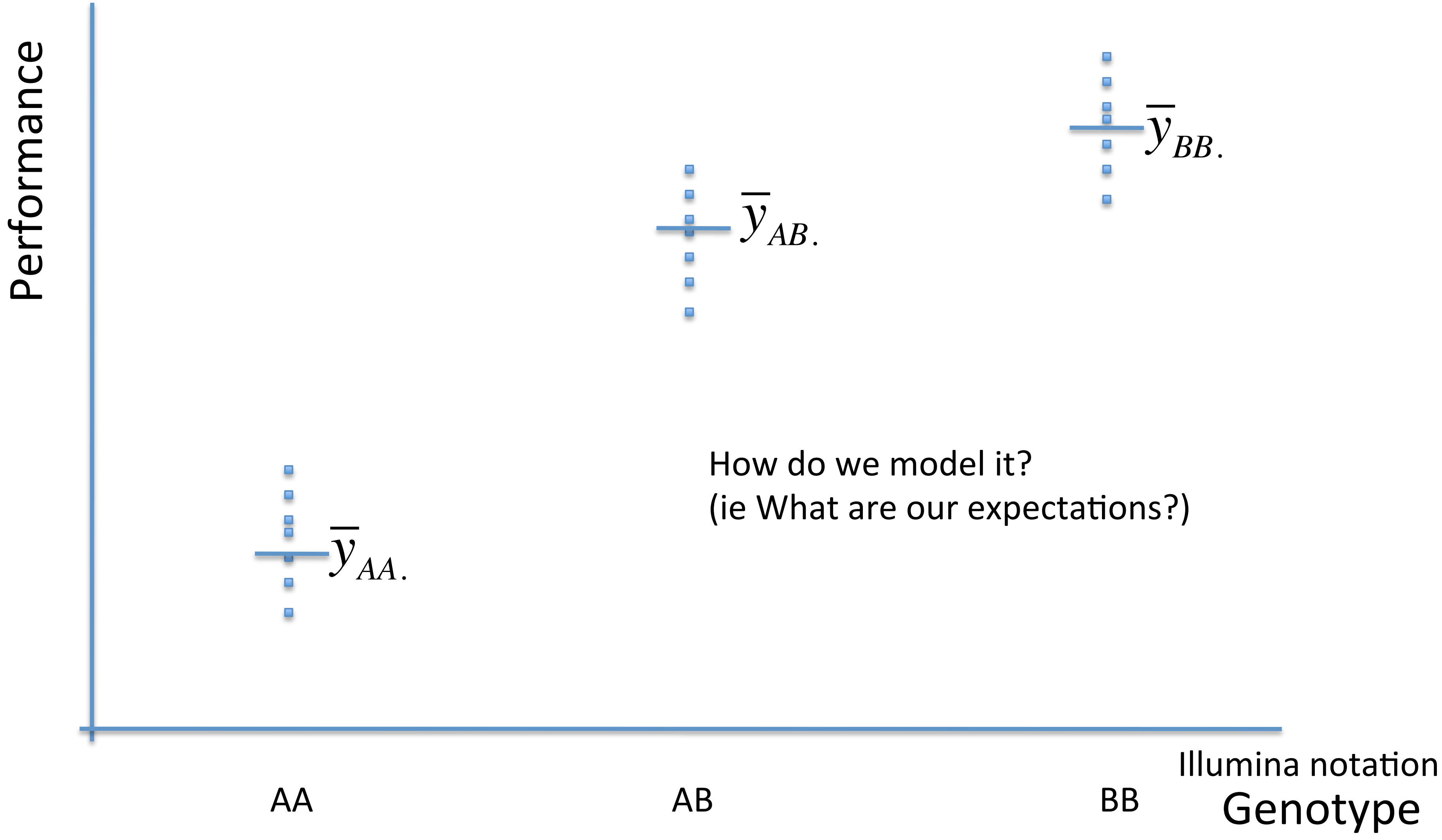
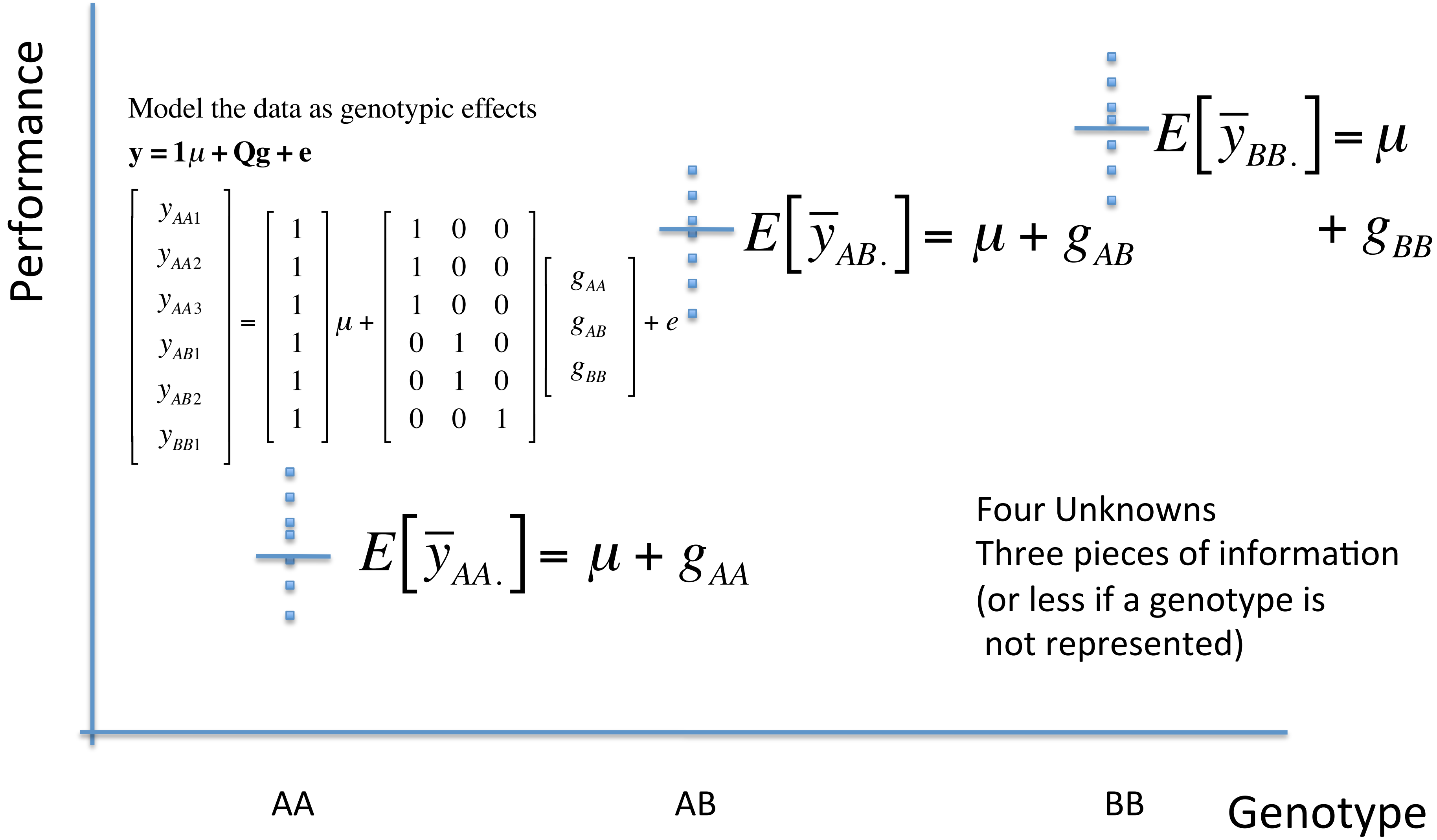
Fixed effects models to predict SNP effects

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 (independent) 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

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

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

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

Least Squares Equations

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

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

In this example
Only fixed effect is mean

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

In general 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)

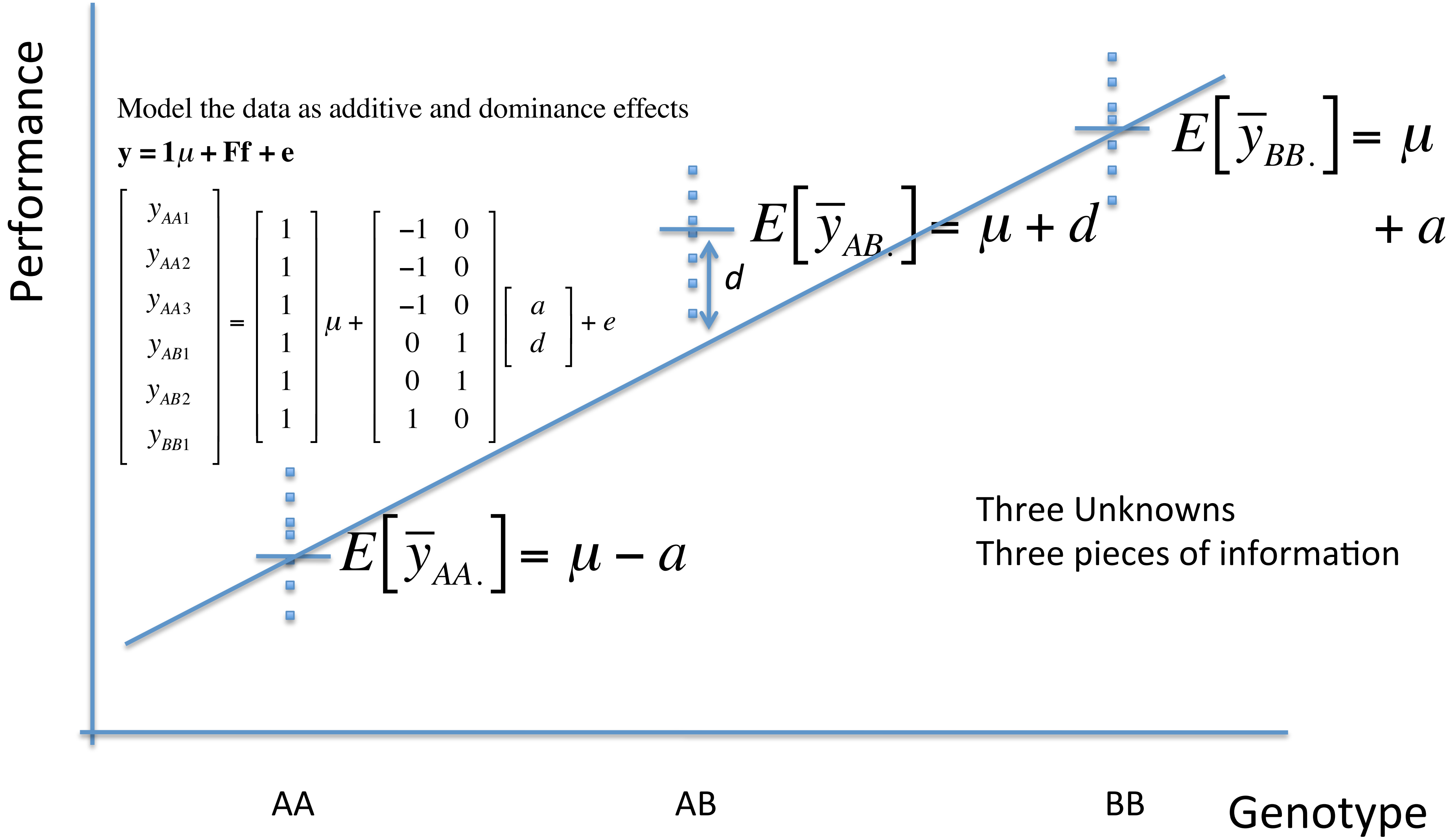
$\left[\mathbf{X'X}\right]^{-}$ is any generalized inverse of $\mathbf{X'X}$ if $\mathbf{(X'X)}\left[\mathbf{X'X}\right]^{-}\mathbf{(X'X) = X'X}$

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

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

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

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

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

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

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

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

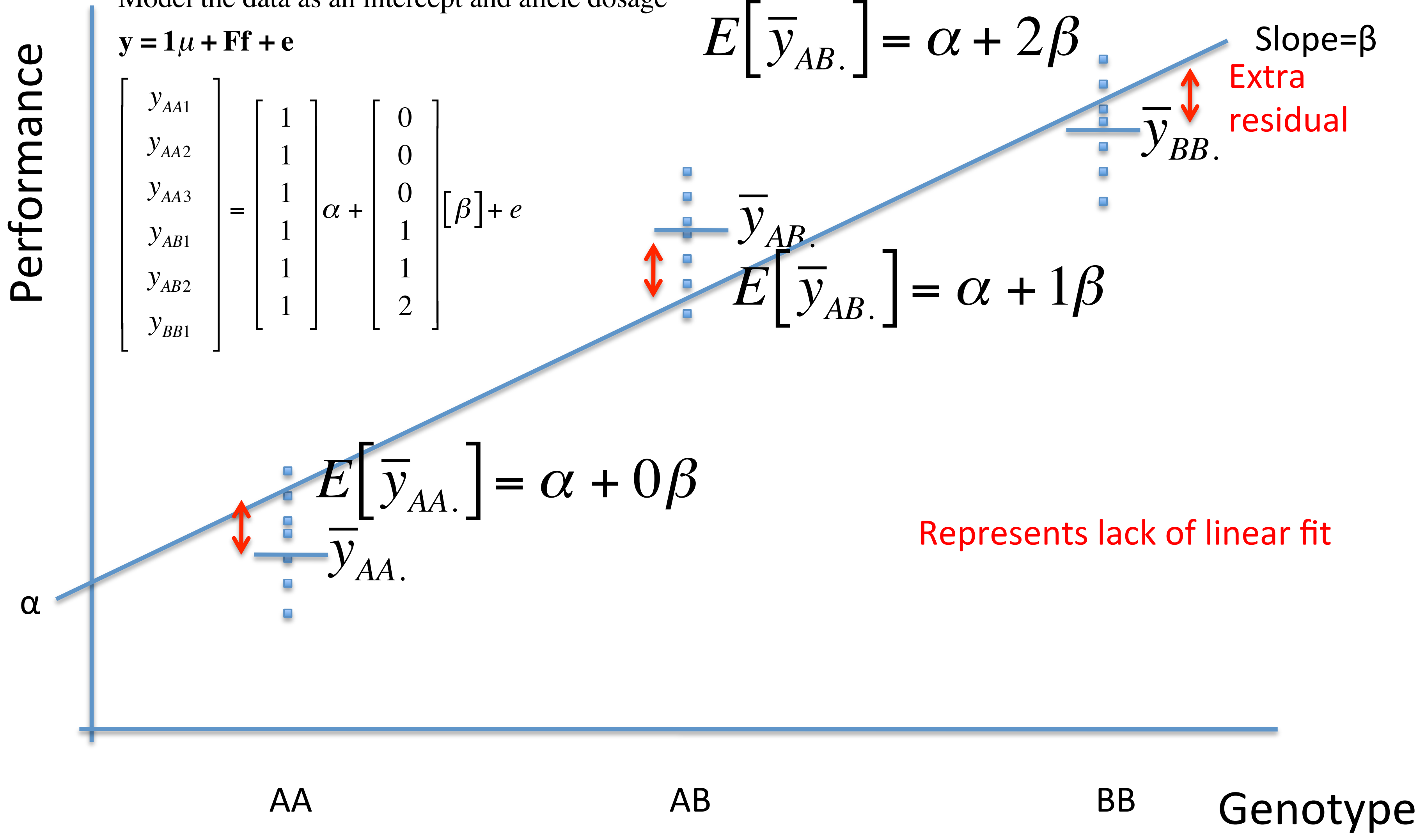
Therefore the models are equivalent
(I can fit either model and migrate from one to the other)

Suppose I ignore dominance (d=0)

Model the data as an intercept and allele dosage

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

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



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

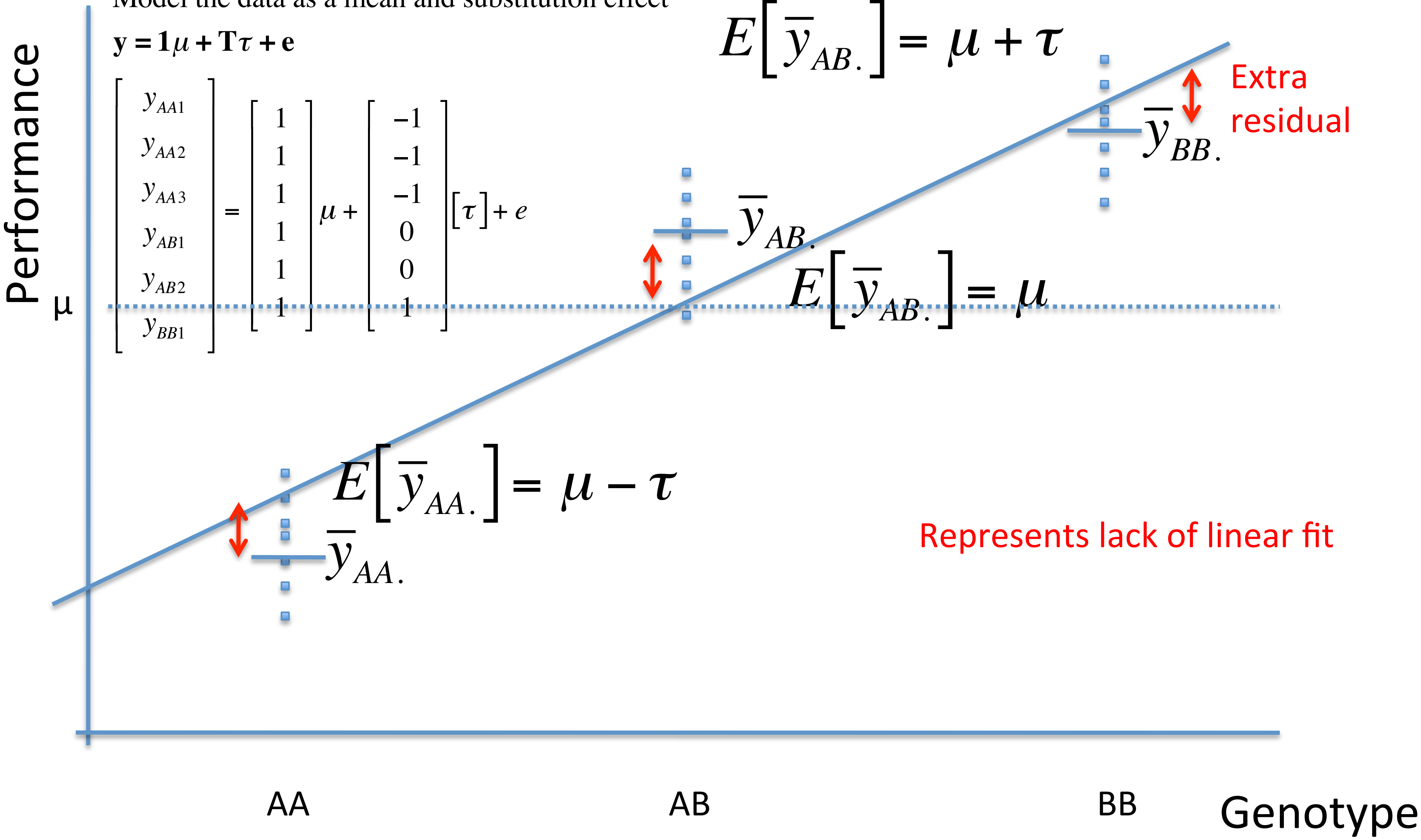
$$E\left[\bar{y}_{AB.}\right] = \alpha + 1\beta$$

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

Extra residual

Represents lack of linear fit

Suppose I ignore dominance (d=0)



Model the data as a mean and substitution effect

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

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

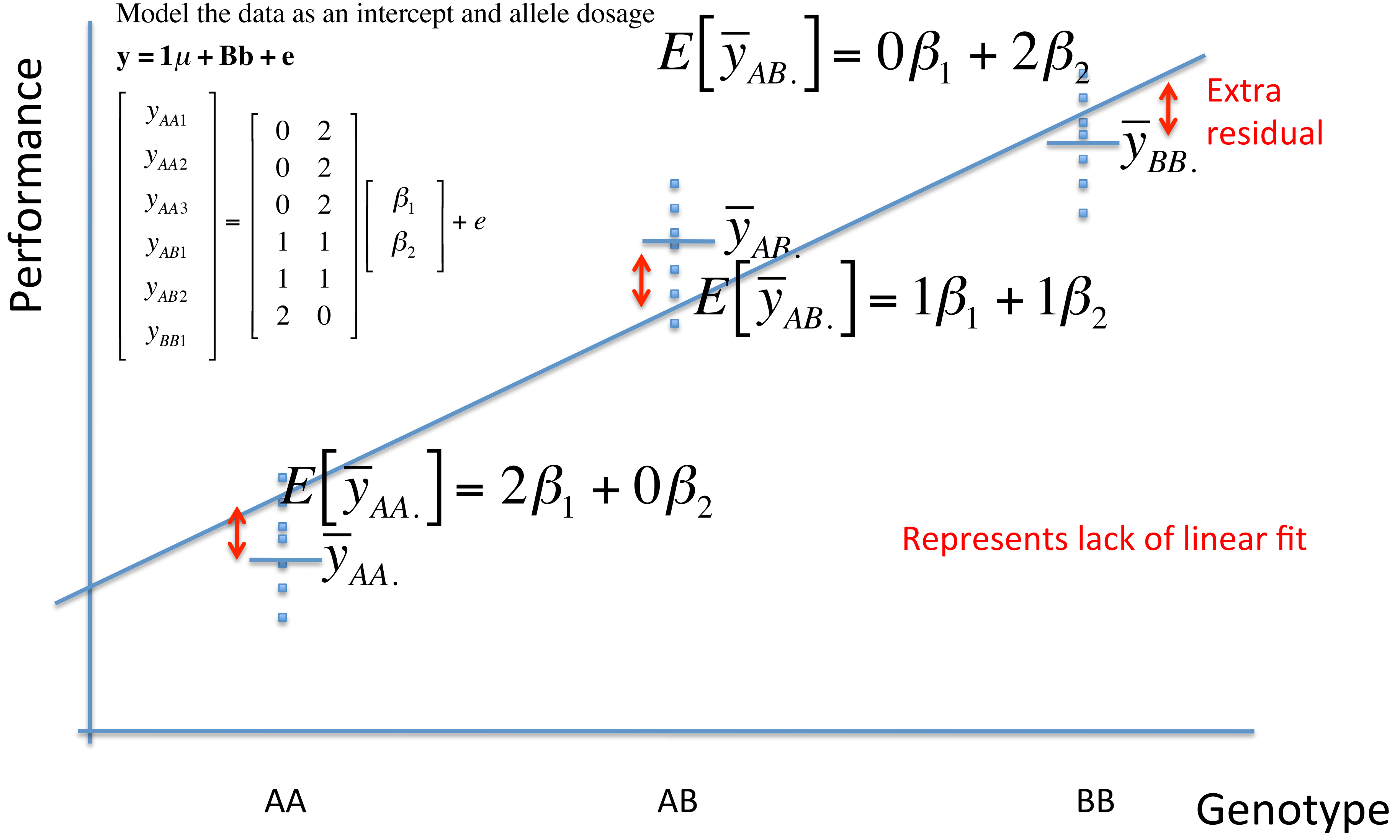
$$E\left[\bar{y}_{AB.}\right] = \mu + \tau$$

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

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

Represents lack of linear fit

Suppose I ignore dominance (d=0)



Model the data as an intercept and allele dosage

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

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

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

Extra residual

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

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

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

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

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

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

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