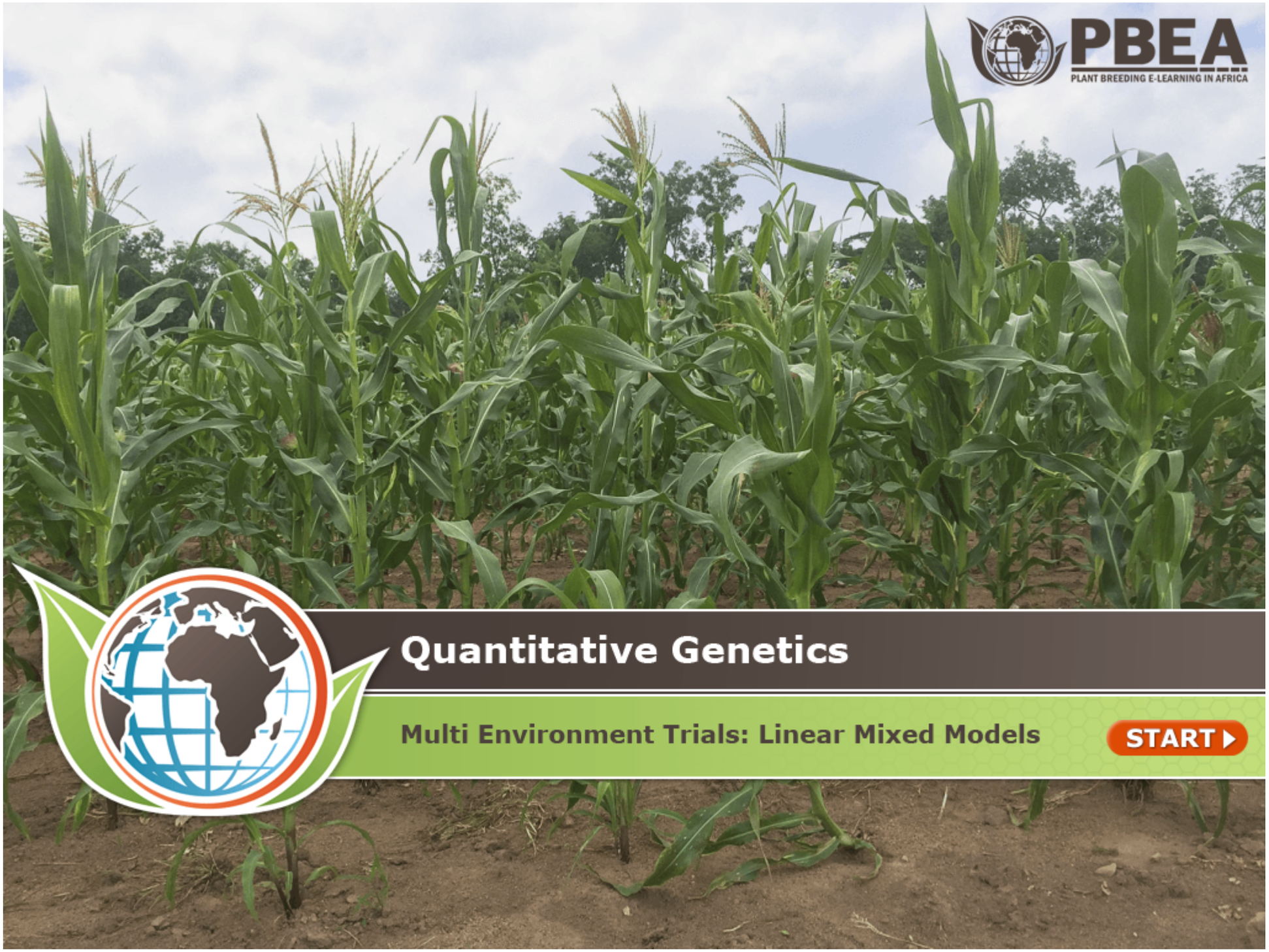
PBEA
PLANT BREEDING E-LEARNING IN AFRICA
Quantitative Genetics
Multi Environment Trials: Linear Mixed Models
START

Objectives

- Conceptual basis of mixed linear models
- Review matrix algebra
- The meaning of BLUE and BLUP

Two Linear Models

SCALAR NOTATION

Throughout this course we utilize two types of models to analyze data:

$$Y_i = \beta_0 + \beta_1 G_1 + \varepsilon_{ij}$$

Equation 1

$$Y_{ij} = \mu + g_i + r_j + \varepsilon_{ij}$$

Equation 2

The parameters of Equation 1 represent the intercept and slope of a line that can be fit to data consisting of pairs of genotypic values, G_i, and Phenotypic responses, where the genotypic values are continuous and known (i.e., measured without error) while the phenotypic data are measured with error in plots (experimental units). The parameters of Equation 2 represent a population mean, genotypic units, g_i, r_j replicates of the genotypic units and the phenotypic, Y_{ij} responses. The genotypes are usually categorical designators of distinct segregating lines and cultivars while the phenotypic data are measured with error on these genotypes in replicated plots (experimental units).

Scalar Notation

We typically estimate the parameters of Equation 1 using least squares regression methods. These methods are based on the idea of minimizing the squared differences between the model parameters and the measured phenotypic value:

$$\min\left(Y_i - [\beta_0 + \beta_1 G_i]\right)^2$$

Equation 3

Taking the partial derivatives of Equation 3 with respect to β_0 and β_1 and setting the resulting two equations = 0, we find that

$$\beta_1 = [V(G_i)]^{-1}[\mathrm{Cov}(G_i Y_i)] \text{ and } \beta_0 = \bar{Y} - \beta_1 \bar{X}$$

Equation 4

The result is a prediction equation:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Equation 5

Note that the predicted values are placed on the fitted line. Such values are sometimes referred to as 'shrunken' estimates because relative to the observed values they show much less variability.

If it were possible to obtain the true genotypic values, G_i, then we could routinely use [LMM.1] to predict phenotypic performance of individual i. Instead, plant breeders have used [LMM.2] and its expanded versions to evaluate segregating lines and cultivars.

Two Linear Models

MATRIX NOTATION

Equation 1 also can be represented as

$$y = X\beta + \varepsilon$$

Equation 6

with $\beta = (X'X)^{-1}(X'\bar{Y})$

Equation 7

and Equation 2 could be represented as

$$y = Xr + Zg + \varepsilon$$

although Equation 2 is usually represented as

$$y = Xb + Zu + \varepsilon$$

which the beginning student often mis-interprets as the matrix form of Equation 1 with an added set of parameters **Z**. The matrix form of Equation 2 is actually a mixed linear model equation and not a simple expansion of the matrix form of Equation 1.

Henderson's Concept

C.R. Henderson recognized the challenge of prediction using models such as Equation 2 and addressed it using the concept of shrinkage estimators for the genotypic units in the model. Note that the fitted regression line provides predictions that are 'shrunken' to the line rather than scattered around the line. Henderson's idea, first published in 1963, was framed in the context of the matrix form of Equation 2, but can be explained using scalar algebra.

First, let's obtain phenotypic averages for each genotypic unit. Next minimize the difference of $\mathrm{E}(w_i[\bar{\mathrm{Y}}_{i.} - \mu] - g_i)^2$, where E represents the expectation,represents the average for the genotypic unit, m is the population mean and g_i is the genotypic value from the scalar version of model Equation 2. In this case we need to find a value of w_i that will assure that the sum of the squared differences is minimal. As with Equation 3, a little knowledge of how to obtain partial derivatives provides the answer:

$$w_i = \frac{\sigma_g^2}{\left(\sigma_g^2 + \sigma_e^2\right)/r}$$

Equation 6

This is known as the intra-class correlation coefficient. It is also known as the broad sense heritability, but for now we will refer to it as a shrinkage factor. When w_i is multiplied by (Y_i - m) it will provide the Best Linear Unbiased Predictor of g_i. Notice that the predictions of genotypic values are scaled towards the mean BV, which by definition is zero.

Example Prediction 1

If the overall mean is the only fixed effect (one environment), all lines are unrelated to each other, and the data are balanced, then:

$$\hat{u}_j = w\,(Y_j - \hat{Y})$$

Shrinkage factor

If w is equal to zero, $\hat{u}_j$ would be zero.

If w is equal to one, $\hat{u}_j$ equals the phenotypic values.

Let's demonstrate this with a simple data set in which four unrelated lines (A,B,C,D) were evaluated (t/ha) in hybrid combination with a single tester (Z) in single rep tests at N environments. For this simple example we are only interested in the impact of number of environments (replicates) on w_i and its subsequent impact on the predicted value for each g_i. Also, assume that the residual variance, $\sigma_e^2 = 40$.

Summary Data

Hybrid	$\bar{Y}_i$	$\bar{Y}_i$	$\bar{Y}_i$-$\bar{Y}_i$	N_1	$w_i(Y_{i.}-\bar{Y}..)$	N2	$w_i(Y_{i.}-\bar{Y}..)$
AxZ	7	10	-3	10	-2.5	2	-1.5
BxZ	9	10	-1	10	-0.83	2	-0.5
CxZ	11	10	1	10	0.83	2	0.5
DxZ	13	10	3	10	2.5	2	1.5

Table 1. Summary data of four inbreds evaluated in hybrid combination with one tester (Z) in single rep tests at 10 environments.

Prove for yourself that the estimated $\sigma_e^2 = 20$.

Some things to notice from the table:

- The data are from balanced trials, i.e., all genotypic units are evaluated in the same number of environments (either 2 or 10).
- With a large N, the observed differences will be equal to the predicted values.
- For balanced trials, shrinkage does not change the relative ranking.

In essence the shrinkage predictor provides us with a value that not only includes the difference relative to the mean, but also weights it by our confidence in the magnitudes of the differences from the overall mean.

We need to consider how to obtain predictions for genotypic units in the more likely situations where not all genotypic units (lines, cultivars, hybrids) will be evaluated equally in all environments. Indeed, we now find it possible with marker technologies to predict the values of the genotypes before they have been grown.

Matrix Algebra

DEFINITIONS AND NOTATION

A matrix is a collection of numerical values arranged in rows and columns. Herein, the elements of a matrix are enclosed in brackets. For example,

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

is a matrix with 4 elements arranged in 2 rows and two columns. Matrices with more than two or more rows and columns are denoted with upper case bold letters. Vectors are a special type of matrix with only one row or one column.

For example, $\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ \end{bmatrix}$ or $\boldsymbol{y} = \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix}$.

Special Kinds of Matrices

Vector matrices are denoted with lower case bold italicized letters. A matrix consisting of only one row and one column is referred to as a **scalar**. A **square matrix** has the same number of rows and columns. A **diagonal matrix** is a square matrix with off-diagonal elements equal to 0. An **identity matrix** is a diagonal matrix with diagonal elements = 1. The identity matrix is almost always denoted **I**.

Matrix Algebra

OPERATIONS

Matrices must be conformable, i.e., matrix operations have requirements on the numbers of rows and columns.

It is possible to add or subtract two matrices, but only if they have the same numbers of rows and columns. For example,

$$C = A - B = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix} = \begin{bmatrix} a_{11} - b_{11} & a_{12} - b_{12} & a_{13} - b_{13} \\ a_{21} - b_{21} & a_{22} - b_{22} & a_{23} - b_{23} \\ a_{31} - b_{31} & a_{32} - b_{32} & a_{33} - b_{33} \end{bmatrix}$$

It is possible to multiply a matrix by a scalar by a matrix by simply multiplying all elements of the matrix by the scalar value, v.

Thus $\boldsymbol{D} = v\boldsymbol{A} = \boldsymbol{A}v = \boldsymbol{D} = \begin{bmatrix} va_{11} & va_{12} & va_{13} \\ va_{21} & va_{22} & va_{23} \\ va_{31} & va_{32} & va_{33} \end{bmatrix}$

Multiplying Vectors

It is possible to multiply two vectors, but only if 1) one of the vectors is a row vector, 2) the second is a column vector, 3) the row vector has as many elements as the column vector. For example,

$$\begin{bmatrix}1 & 3 & 5\end{bmatrix}\begin{bmatrix}2\\4\\6\end{bmatrix}$$ is a legal operation, whereas $$\begin{bmatrix}1\\3\end{bmatrix}\begin{bmatrix}2 & 4 & 6\end{bmatrix}$$ is not.

The operation of vector multiplication in the first instance indicates that we have a 1x3 matrix multiplied by a 3x1 matrix. The way we carry out the vector multiplication is to multiply the elements from each matrix in a pairwise manner, then sum the results of all 3 pairs:

$$\begin{bmatrix}1 & 3 & 5\end{bmatrix}\begin{bmatrix}2\\4\\6\end{bmatrix} = 1\times 2 + 3\times 4 + 5\times 6 = 44\,.$$

Multiplying Vectors

We could also apply the rule of multiplying and summing pairs of elements to the reverse arrangement of these two vectors:

$$\begin{bmatrix}2\\4\\6\end{bmatrix}\begin{bmatrix}1 & 3 & 5\end{bmatrix}=\begin{bmatrix}2 & 6 & 10\\4 & 12 & 20\\6 & 18 & 30\end{bmatrix}$$

Notice that the order of arrangement of vectors matters. Likewise, the arrangement of matrices that are to be multiplied matters. Virtually all types of matrix multiplication involve the multiplication of a row vector by a column vector. In essence we partition each matrix into a set of row and column vectors, then apply the rules of vector multiplication.

Matrix Multiplication

Let's consider **C=AB**. c_{ij} = $\mathbf{a}_{i}.\mathbf{b}._{j}$, where $\mathbf{a}_{i}.$ is the i^{th} row vector of **A** and $\mathbf{b}._{j}$ is the j^{th} column vector of **B**. For example,

$$\boldsymbol{A}=\begin{bmatrix}2 & 8 & -1\\ 3 & 6 & 4\end{bmatrix}, \boldsymbol{B}=\begin{bmatrix}1 & 7\\ 9 & -2\\ 6 & 3\end{bmatrix}$$

then $$c_{11}=a_1.b._1=\begin{bmatrix}2 & 8 & -1\end{bmatrix}\begin{bmatrix}1\\ 9\\ 6\end{bmatrix}=2\times 1+8\times 9-1\times 6=68$$

and $$c_{12}=a_1.b._2=1, c_{21}=a_2.b._1=81, c_{21}=a_2.b._1=21$$

and $$\boldsymbol{AB}=\begin{bmatrix}68 & 1\\ 81 & 21\end{bmatrix}=\boldsymbol{C}$$

Notice that matrix multiplication requires that the first matrix must have as many columns as the second matrix has rows. Thus, **AB** is usually not equal to **BA**. Indeed, while **AB** may be possible, **BA** may not. Lastly verify for yourself that **IA**, **IB** and **Ix** = **A**, **B** and **x** respectively.

Additional Important Operations

The transpose of a matrix, denoted as **A'** (or $\mathbf{A}^t$ or $\mathbf{A}^T$) is a useful operation in which the first row of a matrix becomes the first column of its transpose, while the second, third, … etc rows become the second, third, … etc columns of its transpose. For example,

$$\boldsymbol{A} = \begin{bmatrix} 2 & 8 & -1 \\ 3 & 6 & 4 \end{bmatrix}, \boldsymbol{A}' = \begin{bmatrix} 2 & 3 \\ 8 & 6 \\ -1 & 4 \end{bmatrix}$$

The inverse of a matrix is best understood by recalling that in scalar algebra the inverse of a number multiplied by the number will be = 1. Thus the inverse of x is x-1. In matrix algebra the inverse of a matrix is a matrix when multiplied by the original matrix is **I**. That is $\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. Only square matrices will have an inverse, although not all square matrices will have an inverse. Bernardo describes how to calculate the inverse of a simple 2x2 matrix and it is possible to calculate inverse matrices consisting of 3x3 elements, but calculations of inverses of larger matrices are better left to software.

Best Linear Unbiased Prediction

Henderson's shrinkage predictor can now be considered in the context of the matrix form of the mixed model equation:

$$y = Xb + Zu + \varepsilon$$

Equation 7

y = Vector of observations (phenotypes)
X = Design matrix for fixed effects
b = Vector of unknown fixed effects (to be estimated)
Z = Design matrix of random effects
u = a vector of random effects (genotypic values to be estimated)
ε = a vector of residual errors (random effects to be estimated)

The random effects are assumed to be distributed as u ~ MVN (0,A) and ε ~ MVN (0,R)

Just as estimates for ß in the matrix form of Equation 1 can be found using the normal equations Equation 4, the normal equations for Equation 2 can be used to find

$$\begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}X + A^{-1}(V_r/V_A) \end{bmatrix} - \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}$$

Equation 8

BLUEs and BLUPs

The values for $\hat{\mathbf{b}}$ represent the Best Linear Unbiased Estimators (BLUE) of the fixed effects, while the values for $\hat{\mathbf{u}}$ represent the Best Linear Unbiased Predictors (BLUP) of the random effects. It is important to remember that BLUE's and BLUP's are **not** methods, they are statistical properties of methods (there are many) that are capable of producing such values. These statistical properties include

- **Best**: the sampling variance of what is being estimated or predicted is minimized.
- **Linear**: estimates or predictions are linear functions of the observations.
- **Unbiased**: in BLUE indicates that the expected values of the estimates are equal to their true values. In BLUP, indicates that the sum of the predictions have an expectation of zero.
- **Estimators and Predictors**: refer to algorithms that generate the estimated or predicted values.

For BLUE's the effects are considered fixed. Examples include the overall mean, effects of different soil types, fertilizer treatments, etc. From a practical perspective, fixed effects do not have a covariance structure. Due to the practical perspective, we often consider environments as fixed effects.

Effects of BLUPs

The effects of BLUPs are considered random and it is possible to define covariance structures associated with these effects. Examples include breeding values, dominance effects, tester effects, etc. The challenge for application of methods that provide BLUPs is that Equation 8 assumes covariances and variances are known. The truth is that the variances of genetic and non-genetic random effects are not known. Rather in practice we estimate these values. Thus, all implementations of methods that provide BLUPs from mixed linear model equations provide only approximations of the unknown vector values.

Nonetheless, BLUP values are useful in practical plant breeding trials where designs are unbalanced. Indeed, a method that produces a BLUP value enables the estimation of genetic variances without having to resort to mating designs to obtain estimates of heritability. A typical trial will have different numbers of genotypic units from different families evaluated in different sets of environments, some replicated some not. BLUPs utilize covariance structures (covariances among genotypic units grown in the same sets of environments and covariances among relatives) to maximize information on the traits of interest. Thus, the true purpose of a plant breeding trial (to compare genotypes for purposes of selection), is enabled with the best possible values for comparison because BLUPs maximize the correlation between the true genotypic values and predicted values.

Example

While Equation 8 initially appears to be daunting, with the little bit of matrix algebra, introduced above, you have the skill to do these analyses using EXCEL. For example, consider the simple data set in the following table (adapted from Chapter 11 of Bernardo, 2010):

Environments	n Env	Line	Yield
Low yield	18	1	4.45
Low yield	18	2	4.61
Low yield	18	3	5.27
High yield	9	2	5.00
High yield	9	4	5.82
High yield	9	3	5.79

We desire to translate this into the following model.

$$Y_{ijk} = \mu + G_i + E_j + GE_{ij} + \varepsilon_{(ij)k}$$

$$i = 1, ..., g;\ j = 1, ..., e;\ k = 1, ..., n$$

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{Z}v + \varepsilon$$

In matrix notation the data are represented in the model as:

$$\begin{bmatrix} 4.45 \\ 4.61 \\ 5.27 \\ 5.00 \\ 5.82 \\ 5.79 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \end{bmatrix}$$

Linear Mixed Model Solution

The LMM solution is represented as:

$$\begin{bmatrix}\hat{\beta}\\ \hat{u}\end{bmatrix}=\begin{bmatrix}\boldsymbol{X'R^{-1}X} & \boldsymbol{X'R^{-1}Z}\\ \boldsymbol{Z'R^{-1}X} & \boldsymbol{Z'R^{-1}Z}+\boldsymbol{A}^{-1}\left(V_R/V_A\right)\end{bmatrix}-\begin{bmatrix}\boldsymbol{X'R^{-1}y}\\ \boldsymbol{Z'R^{-1}y}\end{bmatrix}$$

where

$$\begin{bmatrix}\hat{\beta}\\ \hat{u}\end{bmatrix}=\begin{bmatrix}\hat{b}_1\\ \hat{b}_2\\ \hat{u}_1\\ \hat{u}_2\\ \hat{u}_3\\ \hat{u}_4\end{bmatrix},\ \boldsymbol{X}=\begin{bmatrix}1&0\\1&0\\1&0\\0&1\\0&1\\0&1\end{bmatrix},\ \boldsymbol{y}=\begin{bmatrix}4.45\\4.61\\5.27\\5.00\\5.82\\5.79\end{bmatrix},\ \boldsymbol{Z}=\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix},\ \boldsymbol{R}=\begin{bmatrix}1/18&0&0&0&0&0\\0&1/18&0&0&0&0\\0&0&1/18&0&0&0\\0&0&0&1/9&0&0\\0&0&0&0&1/9&0\\0&0&0&0&0&1/9\end{bmatrix}$$

Thus, R represent a matrix that weights the calculations by the number of observations that contribute to the estimated mean values of each cultivar in each type of environment.

Estimated Residual Variance

Assuming that the lines are unrelated to each other, $\boldsymbol{A} = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}$ and V_R/V_A is the ratio of the estimated residual variance (sometimes incorrectly referred to as the estimate of the experimental error) to the estimated additive genetic variance. For purposes of illustration let's consider this estimated ratio to be 5, i.e., the estimated additive genetic variance is 20% as large as the residual variability.

Calculations for the example have been implemented in an EXCEL spreadsheet "BLUEs and BLUPs of 4 barley varieties.xlsx".

As an exercise to conduct on your own, consider implementing the LMM.7 for the example on estimation of means using **lsmeans** discussed in the review module "Review of EDA and Estimation".

Quantitative Genetics

Multi Environment Trials: Linear Mixed Models

This module was developed as part of the Bill & Melinda Gates Foundation Contract No. 24576 for Plant Breeding E-Learning in Africa.

Funding provided by:

BILL & MELINDA GATES foundation

Other collaborating organizations:

AGRA Growing Africa's Agriculture

CGIAR

Partnering universities:

IOWA STATE UNIVERSITY OF SCIENCE AND TECHNOLOGY

Quantitative Genetics **Module 3 Author:**
William D. Beavis *(ISU)*

Multimedia Developers:
Gretchen Anderson, Todd Hartnell, and Andy Rohrback (*ISU*)

Quantitative Genetics **Course Team:**
William D. Beavis *(ISU)*; Richard Akromah, Joseph Sarkodie-Addo, Maxwell Asante *(KNUST)*; Richard Edema, Paul Gibson, Settumba Mukasa, Patrick Ongom *(MAK)*; John Derera, Pangirayi Tongoona, Hussein Shimelis *(UKZN)*