Heritability for Unbalanced Designs

Andrey Vega Alfaro

2024-02-18

In the previous handout, we discussed plot-based heritability and entry-mean heritability. In plot-based heritability, our predictor for the true genotypic value g is the single plot phenotype (y_{ij}). In contrast, entry-mean heritability employs the mean phenotype ($\overline{y}_{i\cdot}$) across replicates as the predictor. While y_{ij} and $\overline{y}_{i\cdot}$ represent two types of predictors, other predictors are available. One notable predictor is the Best Linear Unbiased Predictor (BLUP), designed to perform well with unbalanced datasets characterized by varying degrees of replication across genotypes. Rather than delineating each predictor separately, we can adopt a general term symbolizing any predictor, denoted as $\hat{g}$. Any predictor of the true genotypic value can be defined as follows:

$$\hat{g} = g + r$$

Where:

- The term $\hat{g}$ is the predictor
- The term g is the unobserved **true genotypic value** with $g \sim N(\mu_g, \sigma^2_g)$
- and r is the prediction error

Note that we can rearrange the equation to obtain the true genotypic value:

$$g = \hat{g} - r$$

The true genotypic value is the predictor minus the prediction error r. We will leave this aside for now, but this definitions will be useful later.

BLUP definition

BLUP is defined as a linear combination of phenotypes. If we assume a completely randomized design, the formula for BLUP is:

$$BLUP(g_i) = \mu_g + r^2_i(\overline{y}_{i\cdot} - \mu_y)$$

Where:

- The term $BLUP(g_i)$ is the best linear unbiased predictor of the true genotypic value for genotype i.
- The term μ_g is the mean of the genotypic values.
- The term r^2_i is the reliability of the predictor and $r^2_i = \frac{\sigma^2_g}{\sigma^2_g + var(\hat{g}_i - g_i)}$. Recall that for a CRD $var(\hat{g}_i - g_i) = \frac{\sigma^2_r}{m_i}$ for m reps.
- The term $\overline{y}_{i\cdot}$ is the entry mean across m replicates.
- The term μ_y is the population mean.

Statistical properties of BLUP

BLUP has some interesting statistical properties which contrast with the properties of other predictors like the phenotype of a plot (y_{ij}) or the mean phenotype across several replicates ($\overline{y}_{i}.$)

- **For phenotypic predictors**, like the single plot phenotype or the mean phenotype of m replicates, it is typically assumed that the prediction errors are independent of the true genotypic value, i.e., there is no covariance between the errors and the true genotypic value, $cov(g, r) = 0$.
- **Under BLUP**, it is assumed that the errors are independent of the predictor $cov(\hat{g}, r) = 0$ but that the errors are not independent of the true genotypic value and so $cov(g, r) \neq 0$. **BLUP** adjusts the entry-mean towards the population mean, penalizing genotypes with less information or fewer replicates by shrinking them towards this mean.

Properties of correlation between the true and predicted genotypic value

As breeders, we need to predict true genotypic values. We cannot directly observe genotypic values, all we observe are phenotypes which is a combination of the true genotypic value plus an error component. To predict genotypic values we have predictors that include among others:

- the phenotype of a single plot
- the mean phenotype of a genotype across reps
- BLUP of genotype

The quantity $[corr(g, \hat{g})]^2$ is the squared correlation between the true (g) and predicted genotypic value ($\hat{g}$). It represents how **reliable** our predictor is in predicting the true genotypic value. This measure can be considered in three ways:

1. When $[corr(g, \hat{g})]^2$ is applied to the replicates of a given genotype this quantity is called **reliability**.
2. When we take the square root of the reliability $[corr(g, \hat{g})]$ this quantity is called **accuracy**
3. When this concept is applied to the entire population for your experiment this quantity is called **heritability**. A general definition of heritability in the broad-sense is the average reliability across all genotypes. For an experiment with n genotypes, the average of the reliabilities is the heritability.

$$H^2 = \frac{1}{n} \sum_{i=1}^{n} r_i^2$$

Remember that we can think of heritability as the squared correlation between the true genotypic value g and our predictor $\hat{g}$ across all genotypes in a population. Heritability measures the effectiveness of our predictor in estimating the unobserved true genotypic value. In the context of BLUP that measure is referred to as reliability. In addition, reliability is the property of a group of genotypes with a given degree of replication whereas heritability is the property for the whole population. We can use the $r_{g_i, \hat{g}_i}$ symbol to refer to the reliability of a group of replicated individuals for the i^{th} genotype.

Let's see what is the correlation between the true genotypic value g and our predictor $\hat{g}$ under the BLUP assumptions of $cov(g, r) \neq 0$ and $cov(\hat{g}, r) = 0$.

Where does reliability come from?

We can answer that question in the context of BLUP. Our starting point will be the definition of correlation between true g and predicted genotypic value $\hat{g}$:

$$r_{g,\hat{g}} = corr(g, \hat{g}) = \frac{cov(g, \hat{g})}{\sigma_g \sigma_{\hat{g}}}$$

$$r_{g,\hat{g}} = \frac{cov(g, \hat{g})}{\sigma_g \sigma_{\hat{g}}}$$

We can substitute for the value of g (see above) and we get:

$$r_{g,\hat{g}} = \frac{cov(\hat{g} - r, \hat{g})}{\sigma_g \sigma_{\hat{g}}}$$

$$r_{g,\hat{g}} = \frac{cov(\hat{g}, \hat{g}) - cov(\hat{g}, r)}{\sigma_g \sigma_{\hat{g}}}$$

The statistical property of BLUP tells us that $cov(\hat{g}, r) = 0$ and because the covariance of a variable with it self is the variance, our formula becomes:

$$r_{g,\hat{g}} = \frac{cov(\hat{g}, \hat{g})}{\sigma_g \sigma_{\hat{g}}} = \frac{var(\hat{g})}{\sigma_g \sigma_{\hat{g}}} = \frac{\sigma^2_{\hat{g}}}{\sigma_g \sigma_{\hat{g}}} = \frac{\sigma_{\hat{g}}}{\sigma_g}$$

Please note that the correlation between the true (g) and the predicted ($\hat{g}$) value is called **accuracy**. So accuracy is defined:

$$r_{g,\hat{g}} = \frac{\sigma_{\hat{g}}}{\sigma_g}$$

If we square the accuracy, we get the **reliability** under BLUP. Remember that reliability is the squared correlation between the true and predicted genotypic value.

$$r^2_{g,\hat{g}} = \frac{\sigma^2_{\hat{g}}}{\sigma^2_g}$$

Prediction Error Variance

The term σ^2_g, can be expanded it to its components:

$$\sigma^2_g = \sigma^2_{\hat{g}} - \sigma^2_r + 2cov(\hat{g}, r)$$

Because there is no covariance between the errors and the predictor,i.e. $cov(\hat{g}, r) = 0$ then:

$$\sigma^2_g = \sigma^2_{\hat{g}} - \sigma^2_r$$

We substitute σ^2_g in our formula for reliability under BLUP above and it becomes:

$$r^2_{g,\hat{g}} = \frac{\sigma^2_{\hat{g}}}{\sigma^2_{\hat{g}} - \sigma^2_r}$$

Now we need to find what σ^2_r represents. Depending on your experiment, CRD, RCBD, Multi-enviromental trial, etc, that term could increase in complexity. For now, we can simply call it prediction error variance or

PEV. If we rearrange this formula: $\hat{g} = g + r$ then we have that the prediction error is $r = \hat{g} - g$. What is the Prediction Error Variance then? Let's see: Note that $PEV = \sigma_r^2$

$$PEV = var(\hat{g} - g)$$
$$PEV = var(\hat{g}) + var(g) - 2cov(\hat{g}, g)$$

From previous results we know that $cov(\hat{g}, g)$ evaluates to $\sigma_{\hat{g}}^2$ (see above). So our formula becomes:

$$PEV = \sigma_{\hat{g}}^2 + \sigma_g - 2\sigma_{\hat{g}}^2$$
$$PEV = \sigma_g - \sigma_{\hat{g}}^2$$

The prediction error variance is simply the variance of the true genotypic values minus the variance among the predictors under BLUP.

Connection of reliability, PEV and broad-sense heritability

There is a relationship between PEV, reliability and the concept of broad-sense heritability.

For convenience, we can multiply the equation for PEV by $\frac{1}{{\sigma_g}^2}$ on both sides and simplify:

$$\frac{PEV}{\sigma_g^2} = \frac{\sigma_g - \sigma_{\hat{g}}^2}{\sigma_g^2}$$
$$\frac{PEV}{\sigma_g^2} = 1 - \frac{\sigma_{\hat{g}}^2}{\sigma_g^2}$$
$$\frac{PEV}{\sigma_g^2} = 1 - \frac{\sigma_{\hat{g}}^2}{\sigma_g^2}$$

Recall from previous results that the quantity $\frac{\sigma_{\hat{g}}^2}{\sigma_{\hat{g}}^2} = r_{g,\hat{g}}^2$ is the reliability between the predictor and the true value and so:

$$\frac{PEV}{\sigma_g^2} = 1 - r_{g,\hat{g}}^2$$

We arrange and we get:

$$r_{g_i,\hat{g_i}}^2 = 1 - \frac{PEV}{\sigma_g^2}$$

> Reliability represents the correlation between true and predicted values across replicates of a given genotype. It is thus a property of a group of individuals. Additionally, it can be interpreted as the correlation between true and predicted genotypic values for a specific genotype i across multiple replicates.

The equation above shows that minimum prediction error variance maximizes reliability. Because we are using a predictor (BLUP) which can account for differences in the replication level, the broad-sense heritability estimated as the average of the reliabilities $r_{g,\hat{g}}^2$ across n genotypes, should be more precise than entry-mean or plot-based heritability.

If heritability is the average of the reliabilities for each genotypic value identified by the index i for n genotypes then:

$$H^2 = \frac{1}{n}\sum_{i=1}^{n} r_i^2$$

And then Heritability in the broad-sense for a population of individuals is 1 minus the average prediction error variance $\overline{PEV}$ among genotypes divided by the variance among true genotypic values. The variance among true genotypic values is unknown, however, we usually can empiracally estimate it from the data.

$$H^2 = 1 - \frac{\overline{PEV}}{\sigma_g^2}$$

This formula for heritability has been published elsewhere with different notations, including Cullis et al. 2006 and Covarruvias-Pazaran.

Concluding remarks

When using BLUP, the prediction error is estimated for each genotype i over m replicates. This allows for more precision even in the presence of unbalanced data.

Reliability denotes the correlation between true and predicted values across m reps of a genotype. It is a property of a group of individuals. To relate this measure to a population we use heritability, which in this case is the average of the reliabilities for all genotypes in the experiment.

References

Selection Theory 812 Class Notes. Prof. Jeff Endelman and Prof. Natalia de Leon

Cullis BR, Smith AB, Coombes NE. On the design of early generation variety trials with correlated data. JABES. 2006;11(4):381–393. doi: 10.1198/108571106X154443

Covarruvias-Pazaran, G. Breeding Optimization, CGIAR Excellence in Breeding Platform (EiB). Heritability Manual chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://excellenceinbreeding.org/sites/default/files/manual/Heritability_v6.pdf