Training course on animal breeding data analysis and genomics prediction
Vincent Ducrocq

To cite this version:
Vincent Ducrocq. Training course on animal breeding data analysis and genomics prediction. 3rd cycle. Animal breeding Data Analysis and Genomics Prediction (Training course on Animal breeding Data Analysis and Genomics Prediction), 2017. hal-02786129

HAL Id: hal-02786129
https://hal.inrae.fr/hal-02786129
Submitted on 4 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
What is a genetic evaluation?

Vincent Ducrocq

GABI, Jouy-en-Josas
France
A farmer chooses the animals which are going to produce progeny. These offspring will replace older animals...

This farmer usually wants them to be better, i.e., they should have a phenotype different from the one obtained without selection of the parents.

The difference between the two represents a « genetic gain »

The genetic value or breeding value of an animal is (twice) what it transmits to its progeny ... on average

(twice because only half of its genes are transmitted)

So the « genetic value » is a mathematical notion it cannot be observed, it cannot be measured ... But it can be predicted / estimated using observations
Genetic values: progeny test

- The average « performance » of a large number of progeny gives a precise estimation of its genetic value …

- This is how bulls of dairy breeds have been selected during 50 years:
  - a representative sample of daughters is created we « test » each bull via the characteristics of 50-100 daughters
  - Then we keep the best bulls using the information on these characteristics but also on other related animals (mother, sisters,…)
  - these sources of information are combined into an estimated genetic value = estimated breeding value

November 20-24 2017
What is a genetic evaluation?

- **Classical** progeny test
  - Sire of sons
  - Dam of sons
  - Year n
  - Daughters with 1 lactation year n+5

- Genetic evaluation « Test » Year n+2

- Genetic evaluation « service » Selection on an global index year n+6
  - Daughters born
The traits of interest are often complex traits under the influence of **genetic** effects and **environmental** effects.

The **phenotype** (= the measure of a trait) is misleading: for example, it depends strongly on the management of the herd.

We want to avoid a confusion between the **genetic value** and the **phenotype** influenced by the environment.

Example:

**cow A** produces **3000 kg** of milk in a **herd** with a herd average of **4000 kg**

**cow B**, from the same breed, produces **2500 kg** of milk in a **herd** with a herd average of **2000 kg**

Which one has the highest genetic value for production?
**Data:**
- Phenotypes (measure of a trait, P)
- Environment (herd, season, age...)
- Pedigree

**Models:**
- Description of the data
  \[ P = a + M + e \]
- Genetic model of « transmission » of
  \[ a_i = -\frac{a_s}{2} + -\frac{a_d}{2} + \phi_i \]
### What is a genetic evaluation?

A practical example: 5 recorded cows

<table>
<thead>
<tr>
<th>cow</th>
<th>sire</th>
<th>dam</th>
<th>Sire of sire</th>
<th>Dam of sire</th>
<th>Sire of dam</th>
<th>Dam of dam</th>
<th>305d milk yield</th>
<th>Calving date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bandi</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>5000</td>
<td>14/10/2010</td>
</tr>
<tr>
<td>Honde</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4500</td>
<td>23/11/2010</td>
</tr>
<tr>
<td>Kapila</td>
<td>Hiren</td>
<td>Bandi</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>5500</td>
<td>6/7/2013</td>
</tr>
<tr>
<td>Batali</td>
<td>Akot</td>
<td>Honde</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>5000</td>
<td>15/9/2013</td>
</tr>
<tr>
<td>Mangli</td>
<td>Gary</td>
<td>Kapila</td>
<td>Hiren</td>
<td>Bandi</td>
<td>Hiren</td>
<td>Bandi</td>
<td>6000</td>
<td>25/04/2016</td>
</tr>
</tbody>
</table>
Choose a model:

\[
milk\ yield \ of\ cow\ i = \text{effect of year } j + \text{genetic effect of } i + \text{unexplained “residual”}
\]
What is a genetic evaluation?

The chosen model:

\[ y_i = \beta_j + a_i + e_i \]

We would like to estimate:

- the year effects (\( \beta_1, \beta_2, \beta_3 \))
- the genetic values (a1 to a8)

**Data file**: animal n°, year, production

<table>
<thead>
<tr>
<th>Number</th>
<th>Year</th>
<th>Production</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>5000</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>4500</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>5500</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>5000</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>6000</td>
</tr>
</tbody>
</table>

**Pedigree file**: animal n°, sire n°, dam n°

<table>
<thead>
<tr>
<th>Number</th>
<th>Sire n°</th>
<th>Dam n°</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>
What is a genetic evaluation?

- "BLUP" theory is a system of equations, in matrix form:

\[
\begin{bmatrix}
X'R^{-1}X & X'R^{-1}Z \\
Z'R^{-1}X & Z'R^{-1}Z + G^{-1}
\end{bmatrix}
\begin{bmatrix}
\beta \\
a
\end{bmatrix}
= \begin{bmatrix}
X'R^{-1}y \\
Z'R^{-1}y
\end{bmatrix}
\]

- Mixed model equations

- In red: contribution of pedigree
- In black: contribution for records
- \( \hat{a}_i \): genetic value

\[
\begin{pmatrix}
2 & 0 & 0 \\
0 & 2 & 0 \\
0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
0 & 1 & 0 \\
0 & 0 & 1 \\
0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
0 & 1 & 0 \\
0 & 0 & 1 \\
0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0
\end{pmatrix}
\begin{pmatrix}
\beta_1 \\
\beta_2 \\
\beta_3
\end{pmatrix}
= \begin{pmatrix}
5000 + 4500 \\
5500 + 5000 \\
6000
\end{pmatrix}
\begin{pmatrix}
a_1 \\
a_2 \\
a_3 \\
a_4 \\
a_5 \\
a_6 \\
a_7 \\
a_8
\end{pmatrix}
\]

- \( \begin{pmatrix}
0 \\
5000 \\
0 \\
4500 \\
0 \\
5500 \\
5000 \\
6000
\end{pmatrix}
\]
What is a genetic evaluation?

- characteristics: a lot of 0’s, simple numbers easy to interpret and easy to “construct” just by reading two files (data + pedigree)
- the elements in red representing the links between related animals are easy to construct from a pedigree file

\[ \alpha = \text{a simple function of the heritability } h^2 \text{ of the trait} \]
- Also, the equations can be interpreted intuitively ....

\[
\begin{bmatrix}
2 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 2 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 2.5 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1.5 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0
\end{bmatrix} \begin{bmatrix}
\beta_1 \\
\beta_2 \\
\beta_3 \\
\beta_4 \\
\beta_5 \\
\beta_6 \\
\beta_7 \\
\beta_8 \\
\end{bmatrix} = \begin{bmatrix}
0 \\
0 \\
0 \\
0 \\
5000 + 4500 \\
5500 + 5000 \\
6000 \\
\end{bmatrix}
\]
developing the system of equations:

\[
\begin{align*}
2 \beta_1 + 0 \beta_2 + 0 \beta_3 + 0 a_1 + 1a_2 + 0 a_3 + 1a_4 + 0 a_5 + 0 a_6 + 0 a_7 + 0 a_8 &= 5000 + 4500 \\
0 \beta_1 + 2 \beta_2 + 0 \beta_3 + 0 a_1 + 0 a_2 + 0 a_3 + 0 a_4 + 0 a_5 + 1a_6 + 1a_7 + 0 a_8 &= 5500 + 5000 \\
0 \beta_1 + 0 \beta_2 + 1 \beta_3 + 0 a_1 + 0 a_2 + 0 a_3 + 0 a_4 + 0 a_5 + 0 a_6 + 0 a_7 + 1a_8 &= 6000 \\
0 \beta_1 + 0 \beta_2 + 0 \beta_3 + 2\alpha a_1 + \alpha a_2 + 0 a_3 + 0 a_4 - \alpha a_5 - \alpha a_6 + 0 a_7 + 0 a_8 &= 0 \\
1\beta_1 + 0 \beta_2 + 0 \beta_3 + \alpha a_1 + (1 + 2 \alpha) a_2 + 0 a_3 + 0 a_4 - \alpha a_5 - \alpha a_6 + 0 a_7 + 0 a_8 &= 5000 \\
0 \beta_1 + 0 \beta_2 + 0 \beta_3 + 0 a_1 + 0 a_2 + 1.5 \alpha a_3 + 0.5 \alpha a_4 + 0 a_5 + 0 a_6 - \alpha a_7 + 0 a_8 &= 0 \\
1\beta_1 + 0 \beta_2 + 0 \beta_3 + 0 a_1 + 0 a_2 + 0.5 \alpha a_3 + (1 + 1.5 \alpha) a_4 + 0 a_5 + 0 a_6 - \alpha a_7 + 0 a_8 &= 4500 \\
0 \beta_1 + 0 \beta_2 + 0 \beta_3 - \alpha a_1 - \alpha a_2 + 0 a_3 + 0 a_4 + 2.5 \alpha a_5 + 0.5 \alpha a_6 + 0 a_7 - \alpha a_8 &= 0 \\
0 \beta_1 + 1 \beta_2 + 0 \beta_3 - \alpha a_1 - \alpha a_2 + 0 a_3 + 0 a_4 + 0.5 \alpha a_5 + (1 + 2.5 \alpha) a_6 + 0 a_7 - \alpha a_8 &= 5500 \\
0 \beta_1 + 1 \beta_2 + 0 \beta_3 + 0 a_1 + 0 a_2 - \alpha a_3 - \alpha a_4 + 0 a_5 + 0 a_6 + 2 \alpha a_7 + 0 a_8 &= 5000 \\
0 \beta_1 + 0 \beta_2 + 1 \beta_3 + 0 a_1 + 0 a_2 + 0 a_3 + 0 a_4 - \alpha a_5 - \alpha a_6 + 0 a_7 + (1 + 2 \alpha) a_8 &= 6000 
\end{align*}
\]
We obtain this:

\[ 2\beta_1 + 1a_2 + 1a_4 = y_2 + y_4 \]
\[ 2\beta_2 + 1a_6 + 1a_7 = y_6 + y_7 \]
\[ 1\beta_3 + 1a_8 = y_8 \]

2\alpha a_1 + \alpha a_2 - \alpha a_5 - \alpha a_6 = 0
\[ 1\beta_1 + \alpha a_1 + (1 + 2\alpha) a_2 - \alpha a_5 - \alpha a_6 = y_2 \]
1.5\alpha a_3 + 0.5\alpha a_4 - \alpha a_7 = 0
\[ 1\beta_1 + 0.5\alpha a_3 + (1 + 1.5\alpha) a_4 - \alpha a_7 = y_4 \]
\[ -\alpha a_1 - \alpha a_2 + 2.5\alpha a_5 + 0.5\alpha a_6 - \alpha a_8 = 0 \]
\[ 1\beta_2 - \alpha a_1 - \alpha a_2 + 0.5\alpha a_5 + (1 + 2.5\alpha) a_6 - \alpha a_8 = y_6 \]
\[ 1\beta_2 - \alpha a_3 - \alpha a_4 + 2\alpha a_7 = y_7 \]
\[ 1\beta_3 - \alpha a_5 - \alpha a_6 + (1 + 2\alpha) a_8 = y_8 \]
What is a genetic evaluation?

2\beta_1 + 1a_2 + 1a_4 = y_2 + y_4 
\Rightarrow \hat{\beta}_1 = \frac{1}{2}[(y_2 - \hat{a}_2) + (y_4 - \hat{a}_4)]

2\beta_2 + 1a_6 + 1a_7 = y_6 + y_7 
\Rightarrow \hat{\beta}_2 = \frac{1}{2}[(y_6 - \hat{a}_6) + (y_7 - \hat{a}_7)]

1\beta_3 + 1a_8 = y_8 
\Rightarrow \hat{\beta}_3 = (y_8 - \hat{a}_8)

= average of performances corrected for the genetic effect
What is a genetic evaluation?

Information from parents corrected own performance information on progeny = weighted average and accounting for the average genetic values of parents and of progeny

\[2\alpha a_1 + \alpha a_2 - \alpha a_5 - \alpha a_6 = 0\]
\[1\beta_3 - \alpha a_5 - \alpha a_6 + (1 + 2\alpha) a_8 = y_8\]
\[1\beta_2 - \alpha a_1 - \alpha a_2 + 0.5\alpha a_5 + (1 + 2.5\alpha) a_6 - \alpha a_8 = y_6\]
What is a genetic evaluation?

Sum for «base animals » = 0

average = 18.6
average = 70

Genetic trend / generation

$y_8 = 6000 = 5930 + 70$  □ connexion through pedigree
What is a genetic evaluation?

The need for a mathematical tool: Matrix Algebra

(BLUP) Mixed Model Equations: univariate case

\[
\begin{bmatrix}
X'R^{-1}X & X'R^{-1}Z \\
Z'R^{-1}X & Z'R^{-1}Z + G^{-1}
\end{bmatrix}
\begin{bmatrix}
\beta \\
a
\end{bmatrix}
= 
\begin{bmatrix}
X'R^{-1}y \\
Z'R^{-1}y
\end{bmatrix}
\]

(BLUP) Mixed Model Equations: multivariate case

\[
\begin{bmatrix}
X'R^{11}X_1 & X'R^{12}X_2 \\
X'R^{21}X_1 & X'R^{22}X_2 \\
\end{bmatrix}
\begin{bmatrix}
X'R^{11}Z_1 & X'R^{12}Z_2 \\
X'R^{21}Z_1 & X'R^{22}Z_2 \\
\end{bmatrix}
\begin{bmatrix}
\beta_1 \\
\beta_2 \\
a_1 \\
a_2
\end{bmatrix}
= 
\begin{bmatrix}
X'R^{11}y_1 + X'R^{12}y_2 \\
X'R^{21}y_1 + X'R^{22}y_2 \\
Z'R^{11}y_1 + Z'R^{12}y_2 \\
Z'R^{21}y_1 + Z'R^{22}y_2
\end{bmatrix}
\]
P = G + E

P: Phenotype
- may be standardized, corrected, combined ...

G: Genetic component
- depends on the objective of the analysis
- may be incomplete (e.g., just « sire » or « additive »)

\[ P = G + E \]

\[ \phi_i = \text{mendelian sampling} \]

\[ a_i = \frac{1}{2} a_s + \frac{1}{2} a_d + \phi_i \]
What is a genetic evaluation?

- $P = G + E$
- $P$: Phenotype
  - may be standardized, corrected, combined ...
- $G$: Genotype
  - depends on the objective of the analysis
  - may be incomplete (e.g., just « sire » or « additive » component)
- $E$: Environment
  - « = P – E »
  - may contain a genetic component (which was ignored in G)
  - includes what is not explained (residual)
Environment

- $E: \ll P - G \gg$
  - may contain a genetic component
  - includes what is not explained (residual)

- $E = \text{an overall mean}$
  - + effects that are identified and recorded
  - + effects that are identified and not recorded
  - + effects that are not identified
  - + inaccuracies, errors

- By definition: $\text{residual} = e = y - E(y)$
How to represent an environmental effect?

Covariates: example: age, stage of lactation
- define a relation between the observation \((y_i)\) and age \((x_i)\)
  - parametric: linear \((A)\), quadratic \((B)\) polynomial
  - non parametric: define classes \((C)\)
  - semi-parametric: splines (linear \((D)\), quadratic \((E)\), cubic)
How to represent an environmental effect? (2)

Factors:

example: herd, year

- use a 0/1 covariable

\[ y_i = \ldots + 0^* h_1 + 0^* h_2 + \ldots + 1^* h_i + \ldots 0^* h_n + \ldots \]

\[ y_i = \ldots + h_i + \ldots \]  

\( n \) levels
If we ignore important effects ...

- Biased estimation (e.g., season effect)
- Accuracy overestimated
If we include useless effects ...

- example: season effect
  - semester
  - trimester
  - month
  - week
  - day ...

- no bias
- less precise estimation
- risk of disconnection
What is a genetic evaluation?

- Disconnection
  - = absence of information for a fair estimation
    - absence of information for a fair estimation

<table>
<thead>
<tr>
<th></th>
<th>herd 1</th>
<th>herd 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>bull A</td>
<td>2000</td>
<td></td>
</tr>
<tr>
<td>bull B</td>
<td>3000</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>herd 1</th>
<th>herd 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>bull A</td>
<td>2000</td>
<td>3500</td>
</tr>
<tr>
<td>bull B</td>
<td></td>
<td>3000</td>
</tr>
</tbody>
</table>

- Interactions
  - example: herd x year x season x lactation number
    - balance between risk of bias, disconnection, loss of accuracy ...

 disconnected
c connected
Description of the genetic part

- **(individual) animal model** \[ y_i = \ldots + a_i + e_i \]
- many effects to estimate (often more animals than observations)
- but: \[ a_i = \frac{1}{2} a_s + \frac{1}{2} a_d + \phi_i \]
- = sire-dam model
- = sire model
- interest: fewer parameters to estimate (e.g., only BV of males)
- but stricter assumptions !!! (eg: \( E(ud) = 0 \) )
If G x E interaction...

- 1 trait per environment (1 G for each E group
e.g., production in each state = 1 trait)
  - multivariate approach
- Included in the model
  (scale factors, heterogeneous variances, random regression models
  « reaction norms »)
Milk Yield varies with DIM

\[ a(t) = \lambda_0 + \lambda_1 t + \lambda_2 t^2 + \lambda_3 t^3 + \lambda_4 t^4 \]

Mean production

Persistency: explain 96.5-99.6% of total genetic variance
Fixed and random effects

Assumption: **Polygenic model:**
many (infinite?) genes, each one with a small effect
sum of effects = $a_i = \text{additive genetic effect}$

Law of large numbers $\Rightarrow a_i \sim \mathcal{N}(\mu_a, \sigma_a^2)$ (assume $\mu_a = 0$)
$a_i$ is a *random* effect

Similarly, residual = sum of many small effects = *random effect*
$e_i \sim \mathcal{N}(0, \sigma_e^2)$

but « Year » or « age at calving » ? No distribution $\Rightarrow$ *fixed* effect

Quite arbitrary distinction (eg, herd effect?)
... unless you are a Bayesian…
Bayesian perspective (cf D. Gianola)

Bayes (1763), Laplace (1774)

\[ p(A, B) = p(A|B) \cdot p(B) = p(B|A) \cdot p(A) \]

\[ p(y, \theta) = p(y|\theta) \cdot p(\theta) = p(\theta|y) \cdot p(y) \]

**Prior distribution**: what we know about \( \theta \) before looking at the sample.

**Likelihood function**: Information coming from the sample.

**Marginal density** of the data: No information about \( \theta \).

**Posterior density**: a summary of sample and non sample information about \( \theta \).

**Remark**: the formulas are complex but only the « philosophy » is important here: see next example.

vincent.ducrocq@inra.fr

What is a genetic evaluation?

November 20-24 2017
What is a genetic evaluation?

A and B have very different priors but closer posteriors.
What is a genetic evaluation?

Data « dominate » priors

An example

100 observations

- prior A
- prior B

- likelihood

posterior A
posterior B
What we are trying to do in a Bayesian analysis is to describe our « knowledge » through a distribution (posterior density).

- For a Bayesian, all effects are random.

In many instances, there exists a Bayesian «point estimate» similar with what we get with a frequentist approach (e.g., BLUP).

To mimic the frequentist estimate of a fixed effect: assume no prior.

My practical definition of fixed and random effects:

- **fixed effect**: no prior information (or we don’t use it)
- **random effect**: we use prior information
What is a genetic evaluation?