# Simple Derivation of BLUP and a little on ML and selection

Daniel Sorensen*

November 7, 2013

## Contents



# 1 A simple derivation of BLUP

We outline a simple derivation of BLUP that can be found in Henderson *et al.* (1959) based on maximising the joint density of $\mathbf{y}$ (vector of dimension $n \times 1$) and $\mathbf{a}$ (vector of dimension $q \times 1$) with respect to $\mathbf{b}$ (vector of dimension $p \times 1$) and $\mathbf{a}$, under the assumption of normality. When variances are known, this joint distribution takes the form

$$
\begin{aligned}
p(\mathbf{y}, \mathbf{a}) &= p(\mathbf{y}|\mathbf{a})\, p(\mathbf{a}) \\
&\propto \exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{Xb}-\mathbf{Za})' \mathbf{R}^{-1} (\mathbf{y}-\mathbf{Xb}-\mathbf{Za})\right] \exp\left[-\frac{1}{2}\mathbf{a}'\mathbf{G}^{-1}\mathbf{a}\right]. \qquad (1)
\end{aligned}
$$

It is straightforward to maximise the logarithm of (1) that is equal to

$$
\ln p(\mathbf{y}, \mathbf{a}) = \text{constant} - \frac{1}{2}(\mathbf{y}-\mathbf{Xb}-\mathbf{Za})' \mathbf{R}^{-1} (\mathbf{y}-\mathbf{Xb}-\mathbf{Za}) - \frac{1}{2}\mathbf{a}'\mathbf{G}^{-1}\mathbf{a}. \qquad (2)
$$

Using the chain rule of differentiation, the derivative with respect to $\mathbf{b}$ is

$$
\begin{aligned}
\frac{\partial \ln p(\mathbf{y}, \mathbf{a})}{\partial \mathbf{b}} &= \left(\frac{\partial (\mathbf{y}-\mathbf{Xb}-\mathbf{Za})'}{\partial \mathbf{b}}\right)_{p\times n} \left(\frac{\partial \left[-\frac{1}{2}(\mathbf{y}-\mathbf{Xb}-\mathbf{Za})' \mathbf{R}^{-1} (\mathbf{y}-\mathbf{Xb}-\mathbf{Za})\right]}{\partial (\mathbf{y}-\mathbf{Xb}-\mathbf{Za})}\right)_{n\times 1} \\
&= (-\mathbf{X}') \left[-\mathbf{R}^{-1}(\mathbf{y}-\mathbf{Xb}-\mathbf{Za})\right], \qquad (3)
\end{aligned}
$$

*Dept. of Molecular Biology and Genetics, Faculty of Science and Technology, University of Aarhus, PB 50, DK-8830 Tjele, Denmark

and with respect to $\mathbf{a}$ is

$$\begin{aligned}
\frac{\partial \ln p(\mathbf{y},\mathbf{a})}{\partial \mathbf{a}} &= \left(\frac{\partial (\mathbf{y}-\mathbf{Xb}-\mathbf{Za})'}{\partial \mathbf{a}}\right)_{q\times n}\left(\frac{\partial\left[-\frac{1}{2}(\mathbf{y}-\mathbf{Xb}-\mathbf{Za})'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{Xb}-\mathbf{Za})\right]}{\partial(\mathbf{y}-\mathbf{Xb}-\mathbf{Za})}\right)_{n\times 1} \\
&\quad + \left(\frac{\partial\left[-\frac{1}{2}\mathbf{a}'\mathbf{G}^{-1}\mathbf{a}\right]}{\partial \mathbf{a}}\right)_{q\times 1} \\
&= (-\mathbf{Z}')\left[-\mathbf{R}^{-1}(\mathbf{y}-\mathbf{Xb}-\mathbf{Za})\right] + \left(-\mathbf{G}^{-1}\mathbf{a}\right). \qquad (4)
\end{aligned}$$

Setting (3) and (4) equal to zero,

$$\begin{aligned}
\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\widehat{\mathbf{b}}+\mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\widehat{\mathbf{a}} &= \mathbf{X}'\mathbf{R}^{-1}\mathbf{y}, \\
\mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}\widehat{\mathbf{b}}+\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\widehat{\mathbf{a}}+\mathbf{G}^{-1}\widehat{\mathbf{a}} &= \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y}
\end{aligned}$$

and these equations can be rearranged in the traditional mixed model format

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{bmatrix}\begin{bmatrix} \widehat{\mathbf{b}} \\ \widehat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}. \qquad (5)$$

A transparent interpretation of the rational behind the maximisation is based on a Bayesian argument. The posterior distribution of the unobserved quantities $\mathbf{b}$ and $\mathbf{a}$ is

$$\begin{aligned}
p(\mathbf{b},\mathbf{a}|\mathbf{y}) &\propto p(\mathbf{y}|\mathbf{b},\mathbf{a})\,p(\mathbf{b},\mathbf{a}) \\
&\propto p(\mathbf{y}|\mathbf{b},\mathbf{a})\,p(\mathbf{a})
\end{aligned}$$

assuming $p(\mathbf{b},\mathbf{a}) \propto p(\mathbf{a})$. Then $\left[\widehat{\mathbf{b}},\widehat{\mathbf{a}}\right]$ is the mode (and mean, due to normality) of the posterior distribution of $[\mathbf{b},\mathbf{a}]$.

# 2 Estimation of year effects using selected data

Consider the following problem, discussed in Lush and Shrode (1950) and in Henderson *et al.* (1959). One is to obtain an estimate of the difference in milk production between years. The data on which inferences are to be based are selected: second year records are obtained only from those cows that had the highest first year records. The lowest producers of first year records are not allowed to produce in year 2 and are culled from the herd. Let $y_{i1}$ and $y_{i2}$ be age-adjusted records of the $i$th cow in years 1 and 2, respectively, with

$$y_{i1} = \mu + c_i + e_{i1}, \quad i = 1, 2, \ldots, N \qquad (6)$$

and

$$y_{i2} = \mu + c_i + \delta + e_{i2}, \quad i = 1, 2, \ldots, n, \qquad (7)$$

where $c_i$ is an effect of cow assumed $N(0, \sigma_c^2)$, $e_{i1}$ and $e_{i2} \sim N(0, \sigma_e^2)$, and $\delta$ is the difference between second and first year records that one wishes to estimate. Notice that there are $N$

cows that have a first lactation record, and $n \leq N$ that have a second lactation record. In the absence of culling, $n = N$. With culling based on first lactation information, $n < N$.

From the models,

$$\begin{aligned}
E(y_{i1}) &= \mu \\
E(y_{i2}) &= \mu + \delta \\
E(y_{i2}) - E(y_{i1}) &= \delta \\
Var(y_{i1}) &= Var(y_{i2}) = \sigma^2 = \sigma_c^2 + \sigma_e^2 \\
Cov(y_{i1}, y_{i2}) &= \sigma_c^2 = \rho\sigma^2 \\
Corr(y_{i1}, y_{i2}) &= \rho = \frac{Cov(y_{i1}, y_{i2})}{\sqrt{Var(y_{i1})}\sqrt{Var(y_{i2})}} = \frac{\sigma_c^2}{\sigma^2}.
\end{aligned}$$

Above, $\rho$ is the repeatability or correlation between $y_{i1}$ and $y_{i2}$. Assume that $y_{i1}$ and $y_{i2}$ are bivariate normally distributed in the absence of selection. Then,

$$\begin{bmatrix} y_{i1} \\ y_{i2} \end{bmatrix} \sim N \left[ \begin{bmatrix} \mu \\ \mu + \delta \end{bmatrix}, \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix} \right]. \tag{8}$$

Therefore,

$$y_{i2}|y_{i1} \sim N\left[\mu + \delta + \rho(y_{i1} - \mu), \sigma^2\left(1 - \rho^2\right)\right] \tag{9}$$

and

$$y_{i1} \sim N\left(\mu, \sigma^2\right). \tag{10}$$

If the conditional expectation of a second year record, given the first year record

$$E(y_{i2}|y_{i1}) = \mu + \delta + \rho(y_{i1} - \mu), \tag{11}$$

is averaged over all cows that have a year 1 record

$$\begin{aligned}
E_{y_{i1}}[E(y_{i2}|y_{i1})] &= E(y_{i2}) = \mu + \delta + \rho E_{y_{i1}}[(y_{i1} - \mu)] \\
&= \mu + \delta,
\end{aligned} \tag{12}$$

as expected from the model. However, only the best cows in year 1 are allowed to produce in year 2. This means that, with selection, the expected value of year 1 records is not $\mu$ but

$$E(y_{i1}|y_{i1} \in S) = \mu_s \neq \mu, \tag{13}$$

where $S$ is the set of cows with records in year 1 that are allowed to produce a second record in year 2, and $\mu_s$ is the mean of the selected records of year 1; that is, of the cows that belong in the set $S$. Then we have

$$\begin{aligned}
E_{y_{i1}}[E(y_{i2}|y_{i1} \in S)] &= \mu + \delta + \rho E_{y_{i1}}[(y_{i1} - \mu)|y_{i1} \in S] \\
&= \mu + \delta + \rho(\mu_s - \mu).
\end{aligned} \tag{14}$$

## 2.1 Estimation by least squares

Here, properties of the least squares estimator of $\delta$, the difference between milk production in year 1 and 2 are presented. A least squares estimator of this difference, including only cows that have a record on both years, is

$$\widehat{d_1} = \overline{y}_{.2} - \overline{y}_{.1}, \quad y_{i1} \in S.$$

The expected value of $\widehat{d_1}$ is

$$\begin{aligned} E\left(\widehat{d_1}\right) &= E_{y_{i1}}\left[E\left(\overline{y}_{.2} - \overline{y}_{.1} | y_{i1} \in S\right)\right] = \left(\mu + \delta + \rho\left(\mu_s - \mu\right)\right) - \mu_s \\ &= \delta - (1-\rho)\left(\mu_s - \mu\right). \end{aligned} \tag{15}$$

Since $\mu_s > \mu$, the estimator of the difference is biased downwards. With selected data, least squares estimates of year effects based on (15) give the impression of negative environmental trend with time.

Another possibility would be to estimate $\delta$ by taking the average of all year 2 records minus all year 1 records. The expected value of this second estimator is

$$\begin{aligned} E\left(\widehat{d_2}\right) &= E_{y_{i1}}\left[E\left(\overline{y}_{.2} | y_{i1} \in S\right)\right] - E\left(\overline{y}_{.1}\right) = \left(\mu + \delta + \rho\left(\mu_s - \mu\right)\right) - \mu \\ &= \delta + \rho\left(\mu_s - \mu\right), \end{aligned} \tag{16}$$

which is biased upwards unless $\rho = 0$, in which case selection would not be effective in changing milk production.

The conclusion is that with this type of selection (culling type), the least squares estimate of the difference in milk production is biased. On the other hand, BLUP yields unbiased estimates of genetic and environmental trends under certain types of selection. The paper where this is discussed at length is Henderson (1975); this is a very technical paper that belongs in a higher level course. The following subsection describes in a simplified manner estimation of year effects using maximum likelihood. An important and beautiful paper where the topic is discussed is Curnow (1961)

## 2.2 Estimation by maximum likelihood

Maximum likelihood is a very general method of estimation that has statistically attractive asymptotic properties. A historically important paper by a strong proponent of this method is Fisher (1922). Very briefly, the likelihood of parameter $\theta$, given data $\mathbf{y}$ is a function of $\theta$ proportional to the joint density of the data. A maximum likelihood estimate of $\theta$, labelled $\widehat{\theta}$, is the value of $\theta$ that maximizes the likelihood. Often it is easier to work with the log-likelihood instead, and the value of $\theta$ that maximizes the likelihood is also the value of $\theta$ that maximizes the log-likelihood.

The parameters of the model defined by (9) and (10) are

$$\theta' = \left(\mu, \delta, \sigma^2, \rho\right).$$

However, in this simplified problem, it will be assumed that $\rho$ and $\sigma^2$ are known. In view of (9) and (10), the likelihood is

$$\begin{aligned} L(\theta|\mathbf{y}) &\propto \left(2\pi\sigma^2\right)^{-\frac{N}{2}} \exp\left[-\frac{\sum_{i=1}^{N}(y_{i1}-\mu)^2}{2\sigma^2}\right] \\ &\times \left(2\pi\left(1-\rho^2\right)\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{\sum_{i=1}^{n}(y_{i2}-\mu-\delta-\rho(y_{i1}-\mu))^2}{2(1-\rho^2)\sigma^2}\right]. \end{aligned} \tag{17}$$

With $\rho$ and $\sigma^2$ assumed known, the log-likelihood, apart from an additive constant is

$$l(\mu,\delta|\mathbf{y}) = -\frac{(1-\rho^2)\sum_{i=1}^{N}(y_{i1}-\mu)^2 + \sum_{i=1}^{n}(y_{i2}-\mu-\delta-\rho(y_{i1}-\mu))^2}{2(1-\rho^2)\sigma^2}. \tag{18}$$

This log-likelihood can be written as

$$\begin{aligned} l(\mu,\delta|\mathbf{y}) &= -\frac{n(\rho^2 S_1 + S_2 - 2\rho S_{12})}{2(1-\rho^2)\sigma^2} - \frac{N(\overline{y}-\mu)^2}{2\sigma^2} \\ &\quad -\frac{n(\overline{y}_2-\mu-\delta-\rho(\overline{y}_1-\mu))^2}{2(1-\rho^2)\sigma^2}, \end{aligned} \tag{19}$$

where

$$\begin{aligned} S_1 &= \frac{\sum_{i=1}^{n}(y_{i1}-\overline{y}_1)^2}{n} \\ S_2 &= \frac{\sum_{i=1}^{n}(y_{i2}-\overline{y}_2)^2}{n} \\ S_{12} &= \frac{\sum_{i=1}^{n}(y_{i1}-\overline{y}_1)(y_{i2}-\overline{y}_2)}{n} \\ \overline{y} &= \frac{\sum_{i=1}^{N} y_{i1}}{N} \\ \overline{y}_1 &= \frac{\sum_{i=1}^{n} y_{i1}}{n} \\ \overline{y}_2 &= \frac{\sum_{i=1}^{n} y_{i2}}{n}. \end{aligned}$$

Notice that, with $\rho$ and $\sigma^2$ assumed known, the log-likelihood (19) is a function of $\mu$ and $\delta$ only. To obtain the maximum likelihood estimates of $\mu$ and $\delta$, (19) must be maximized with respect to $\mu$ and $\delta$. One way of doing this is to take partial derivatives of (19) with respect to $\mu$ and $\delta$, to set the resulting two equations equal to zero, and to solve for $\mu$ and $\delta$. The resulting maximum likelihood estimators of $\mu$ and $\delta$ are

$$\widehat{\mu} = \overline{y}, \tag{20}$$

and

$$\widehat{\delta} = \overline{y}_2 - \overline{y} - \rho(\overline{y}_1 - \overline{y}). \tag{21}$$

Notice that if all first lactation cows produce a second lactation (i.e. no selection), then $\overline{y}_1 = \overline{y}$ and the maximum likelihood estimator of $\delta$ reduces to $\overline{y}_2 - \overline{y}$, the difference between mean production in years two and one.

To study whether (20) and (21) are biased, we compute their expectations. From (10),

$$E(\widehat{\mu}) = E(\overline{y}) = \mu. \tag{22}$$

Using (13), (22) and (14), the expected value of (21) is

$$\begin{aligned} E\left(\widehat{\delta}\right) &= E_{y_{i1}}\left[E\left(\overline{y}_2 | y_{i1} \in S\right)\right] - E\left(\overline{y}\right) - \rho\left[E\left(\overline{y}_1 | y_{i1} \in S\right) - E\left(\overline{y}\right)\right] \\ &= \mu + \delta + \rho\left(\mu_s - \mu\right) - \mu - \rho\left(\mu_s - \mu\right) \\ &= \delta. \end{aligned} \tag{23}$$

Therefore, the maximum likelihood estimators of $\mu$ and $\delta$ are unbiased by selection of the records.

In general, maximum likelihood does not produce unbiased estimators in small samples, but when properly used, it produces estimators that are not affected by certain forms of selection. This point will be discussed briefly in class.

# References


Curnow, R. N. (1961). The estimation of repeatability and heritability from records subject to culling. *Biometrics* **17**, 553–566.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London Series A* **222**, 309–368.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31**, 423–447.

Henderson, C. R., Kempthorne, O., Searle, S. R. and Von Krosigk, C. N. (1959). Estimation of environmental and genetic trends from records subject to culling. *Biometrics* **15**, 192–218.

Lush, J. L. and Shrode, R. R. (1950). Changes in milk production with age and milking frequency. *Journal of Dairy Science* **33**, 338–357.