THE ESTIMATION OF ENVIRONMENTAL AND GENETIC TRENDS FROM RECORDS SUBJECT TO CULLING*

C. R. HENDERSON  
Animal Husbandry Department, Cornell University, Ithaca, N. Y., U.S.A.

Oscar KEMPTHORNE  
Statistical Laboratory, Iowa State College, Ames, Iowa, U.S.A.

S. R. SEARLE  
Animal Husbandry Department, Cornell University, Ithaca, N. Y., U.S.A.

AND  

C. M. von KROSIGK  
Animal Husbandry Department, Iowa State College, Ames, Iowa, U.S.A.

INTRODUCTION

A very common problem which arises in animal or plant breeding research is that of assessing the gain which has resulted from a selection program carried on over a number of years. To be specific, let us suppose that we have a closed dairy herd which has been maintained over a number of years with selection being practiced. The records available for assessing any genetic improvement consist of production records of cows in the various years and can be represented by a two-way classification, cow by year. At first it might be thought that such a two-way classification could be analyzed by the method of fitting constants [Yates, 1934]. Applications of this technique have, however, led to the apparent conclusion that the environment gradually deteriorated over the period of years, as indicated by the fact that the constants fitted for years tend to decrease year by year.

Henderson [1949] pointed out that a least squares procedure in which the cow effects are regarded as fixed leads to biased estimates. Lush and Shrode [1950] gave a simple explanation of the biases arising in the estimation of age correction factors; similar considerations apply to the estimation of year effects. The present paper is the combined

result of work in this problem carried out over the past few years independently at Iowa and Cornell.

During the summer of 1957 the two of us at Iowa (Kempthorne and von Krosigk), after preparing a paper for this journal, had correspondence and personal discussion with the two at Cornell (Henderson and Searle), from which it was felt that a combined paper would best suit our present state of knowledge of the problem. A method for estimating environmental trend when repeatability is assumed known was outlined by Henderson in mimeographed material that has been circulating for some years; a method not requiring an assumed value for repeatability and now being proposed by two of us (K. and K.) appeared at first to differ, but when one of us (S.) showed that for given repeatability these two methods are equivalent, it seemed more suitable to present all the work in one publication rather than two separate ones. This paper therefore presents both methods and the manner in which they are equivalent, attributing the various sections to their appropriate authors.

ORIGIN OF THE BIAS IN ROUTINE LEAST SQUARES
(KEMPThORNE AND VON KROSIGK)

Lush and Shrode [1950] have shown how bias enters into the estimation of age correction factors due to culling. Similar arguments apply when estimating year effects. These admit of easy explanation in the simple situation in which a cow's first-year record \( x_{i1} \), and her second-year record \( x_{i2} \), conform to a bivariate normal distribution, with a strictly additive difference \( \gamma \) between first and second year records, based on the model

\[
x_{i1} = \mu + c_i + e_{i1} , \quad x_{i2} = \mu + \gamma + c_i + e_{i2} .
\]

The \( e \)'s are errors arising from differing environments among the same animal's records from year to year, and from inaccuracies of measurement, and \( c_i \) is common to all records of cow \( i \). The \( c \)'s and \( e \)'s are random variables with zero means and variances \( \sigma_c^2 \) and \( \sigma_e^2 \) respectively, and all covariances are zero. The expectation with regard to the errors is over a hypothetical set of repetitions in the particular year with any particular cow. The expectation with regard to the cow effects is over the population of cows which could have entered the records, it being hypothesized that the particular set entering the records is a random sample of the possibilities that could have arisen. The fact that cows in a herd will probably be somewhat related would vitiate the assumption that the covariances between any two \( c \)'s is zero. We shall, however, in this paper not extend the argument to take care of this situation.
It should be emphasized that we are supposing no age or lactation-
number effect, on the assumption that certain correction factors obtained
apart from the data to be analyzed are appropriate. The validity of
this assumption will be discussed at the end of the paper.

With this model the herd average for the first year (mean value of \(x\))
is \(\mu\) and that for the second year (mean value of \(y\)) is \(\mu + \gamma\). \(x\) and \(y\)
both have variances \(\sigma_x^2\) and \(\sigma_y^2\) and the covariance between them is \(\sigma_{xy}\),
so that based on a bivariate normal distribution the mean value of \(y\)
given \(x\) is

\[
\mu + \gamma + r(x - \mu)
\]

where \(r = \sigma_y^2 / (\sigma_x^2 + \sigma_y^2)\), a ratio known in animal breeding work as
repeatability [Lush, 1949]. If we take the expectation of this con-
ditional mean over all the cows who had a first record we obtain of
course \(\mu + \gamma\), the population mean for \(y\). But the culling has the effect
that the expectation of \(x\) over the cows retained in the herd after the
first record is equal to a number say \(\mu'\), which would not in general be
equal to \(\mu\), unless the culling were either based only on some attributes
statistically independent of \(x\), or were a peculiar type of balanced culling
which did not affect the mean. The usual thing is probably some sort
of truncation selection, though probably based on an index of which \(x\)
is one component rather than on \(x\) alone. It therefore follows that the
mean of the second-year records of cows which are retained is

\[
\mu + \gamma + r(\mu' - \mu).
\]

In this simple case the method of least squares gives an estimate of
the year difference \(\gamma\), as the average of second-year records minus the
average of first-year records of those cows that had a second-year
record, and this estimate has expected value,

\[
\mu + \gamma + r(\mu' - \mu) - \mu' = \gamma - (1-r)(\mu' - \mu).
\]

This is the paired comparison method of estimation (method \(B\) of Lush
and Shrode); with the usual type of selection \(\mu'\) will be greater than \(\mu\),
so this estimate of the year difference is biased downward to the extent

\[
(1-r)(\mu' - \mu).
\]

With the gross comparison method (method \(A\) of Lush and Shrode) the
year difference is estimated by the difference between the mean of all
second-year records and the mean of all first-year records; such an
estimate has expectation

\[
\mu + \gamma + r(\mu' - \mu) - \mu,
\]

and hence is biased upward by an amount \(r(\mu' - \mu)\).
Suppose the linear model for a cow’s record is

\[ y_{ikt} = \mu + d_k + g_t + c_{it} + e_{ikt} \]

where \( y_{ikt} \) is the record in the \( k \)th year made by the \( i \)th cow of the \( t \)th group of cows in a herd. These groups might for example represent daughters of a bull, sets of cows born within a specified period, or groups of cows that enter the herd together. \( \mu \) is the population average, \( d_k \) is the environmental effect of the \( k \)th year, \( g_t \) is the mean real producing ability of the \( t \)th group of cows, \( c_{it} \) is the real producing ability of the \( i \)th cow of the \( t \)th group, and \( e_{ikt} \) is a random environmental effect peculiar to the individual record. We will assume that the \( c_{it} \) are normally and independently distributed with zero means and variance \( \sigma_c^2 \), that the \( e_{ikt} \) are normally and independently distributed with zero means and variance \( \sigma_e^2 \), and that the \( c \)'s and \( e \)'s are uncorrelated. These assumptions imply that the cows of a particular group are randomly drawn from a normal population with mean \( \mu + g_t \) and variance \( \sigma_c^2 \). Furthermore, temporary environment is not correlated with real producing ability. The problem is to estimate differences among the \( d \)'s and \( g \)'s assuming that repeatability is known.

The following method for maximum likelihood estimation of fixed elements of mixed linear models has been derived by one of us (H.).

Let the mixed linear model be

\[ y = X\beta + Zu + e \]

where \( \beta \) is a vector of fixed effects, while \( u \) and \( e \) are independent vectors of variables that are normally distributed with zero means and variance-covariance matrices \( D\sigma^2 \) and \( R\sigma^2 \) respectively. Then \( y \) has a multivariate normal distribution with means \( X\beta \) and variance-covariance matrix \( (R + ZDZ')\sigma^2 \), and the m.l. estimator of \( \beta \), say \( \hat{\beta} \), is the solution to

\[
X'(R + ZDZ')^{-1}X\hat{\beta} = X'(R + ZDZ')^{-1}y \tag{1}
\]

assuming that the coefficient matrix is non-singular and that \( \beta \) is estimable. The difficulty in applying this method is that \( R + ZDZ' \) is, in practice, often large and non-diagonal.

Now the same estimator can be obtained by maximizing for variation in \( \beta \) and \( u \) the joint density function of \( y \) and \( u \). This function is

\[
f(y, u) = g(y|u)h(u) = \text{Const.} \exp \left[ -\frac{1}{2\sigma^2} (y - X\beta - Zu)'R^{-1}(y - X\beta - Zu) \right. \\
\left. - \frac{1}{2\sigma^2} u'D^{-1}u \right].
\]
Differentiating this with respect to \( \beta \) and \( u \) and equating to zero gives the following equations:

\[
X'R^{-1}X\beta + X'R^{-1}Z\bar{u} = X'R^{-1}y
\]

\[
Z'R^{-1}X\beta + (Z'R^{-1}Z + D^{-1})\bar{u} = Z'R^{-1}y.
\]

These equations except for \( D^{-1} \) are identical to those for the m.l. estimation of \( \beta \) and \( u \) regarding \( u \) as fixed. In many problems the equations are easy to write since \( R \) and \( D \) are diagonal.

To prove that \( \beta \) and \( \hat{\beta} \) are identical we eliminate \( \bar{u} \) obtaining

\[
X'WX\beta = X'Wy
\]

where

\[
W = R^{-1} - R^{-1}Z(Z'R^{-1}Z + D^{-1})^{-1}Z'R^{-1}.
\]

Now, if it can be shown that \( W = (R + ZDZ')^{-1} \) these equations are identical to those given in (1) and then \( \beta = \hat{\beta} \). We show this by proving that \( (R + ZDZ')W = I \).

\[
(R + ZDZ')W = (R + ZDZ')[R^{-1} - R^{-1}Z(Z'R^{-1}Z + D^{-1})^{-1}Z'R^{-1}]
\]

\[
= I + ZDZ'R^{-1} - Z(Z'R^{-1}Z + D^{-1})^{-1}Z'R^{-1}
\]

\[
- ZDZ'R^{-1}Z(Z'R^{-1}Z + D^{-1})^{-1}Z'R^{-1}
\]

\[
= I + ZDZ'R^{-1} - Z(I + DZ'R^{-1}Z)(Z'R^{-1}Z + D^{-1})^{-1}Z'R^{-1}
\]

\[
= I + ZDZ'R^{-1} - ZD(D^{-1} + Z'R^{-1}Z)(Z'R^{-1}Z + D^{-1})^{-1}Z'R^{-1}
\]

\[
= I + ZDZ'R^{-1} - ZDZ'R^{-1}
\]

\[
= I, \text{ thus completing the proof.}
\]

In applying this method to the present problem \( \mu, \ d' \text{s, and } g' \text{s correspond to } \beta, \text{ and } c' \text{s correspond to } u \). The joint distribution of the \( y_{ikt} \) and \( c_{it} \), is

\[
\prod_{ikt} \left\{ \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp -\frac{1}{2\sigma_e^2} (y_{ikt} - \mu - d - g_i - c_{it})^2 \right\}
\]

\[
\times \left\{ \prod_{it} \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp -\frac{c_{it}^2}{2\sigma_e^2} \right\}.
\]
The maximizing values for variations in \( \mu, g's, d's, \) and \( c's \) are the solutions to the following set of equations:

\[
\begin{align*}
n_{..}r + \sum_k n_{..}c_k - d_k + \sum_i n_{..}c_i = y.., \\
n_{..}r + \sum_k n_{..}c_k - d_k + \sum_i n_{..}c_i = y.., \\
\end{align*}
\]

and analogous equations for other years \( (k) \);

\[
\begin{align*}
n_{..}r + \sum_k n_{..}c_k - d_k + \sum_i n_{..}c_i = y.., \\
n_{..}r + \sum_k n_{..}c_k - d_k + \sum_i n_{..}c_i = y.., \\
\end{align*}
\]

and analogous equations for other groups \( (t) \);

\[
\begin{align*}
n_{..}r + \sum_k n_{..}c_k - d_k + \sum_i n_{..}c_i = y.., \\
n_{..}r + \sum_k n_{..}c_k - d_k + \sum_i n_{..}c_i = y.., \\
\end{align*}
\]

and analogous equations for other cows \( (i, t) \).

\( n_{..i} = 1 \) if the \( it \)th cow has a record in the \( k \)th year, and is zero otherwise. A dot in the subscript denotes summation over that particular subscript.

It will be noted that the term \( (1 - r)/r \) appears in the coefficient of the \( c' \)'s in equations (5). This arises from the ratio \( \sigma^2/\sigma^2_e \) and the definition of repeatability, namely \( \sigma^2_e/(\sigma^2 + \sigma^2_c) \). As used here repeatability is not intra-herd as Lush [1949] uses it, but rather intra-group, (within \( g \) groups) and therefore depends on the choice of groups.

Solving equations (5) gives

\[
\hat{c}_{it} = \frac{n_{..i}^r}{1 + (n_{..i} - 1)r} (\bar{y}_{..i} - \bar{\mu} - \bar{g}_i + \bar{D}).
\]

\( \bar{d} \) is the mean of \( d_k \) associated with that particular cow. Using these expressions we can eliminate \( \hat{c}_{it} \) from equations (2), (3), and (4). These reduced equations are of the form

\[
\sum_k m_{k} \hat{d}_k = \sum_i m_i (\bar{\mu} + \bar{g}_i) = z_i.
\]

and analogous equations for other years \( (k') \);

\[
\sum_k m_{k} \hat{d}_k = \sum_i m_i (\bar{\mu} + \bar{g}_i) = z_i
\]

and analogous equations for other groups \( (t) \).

The \( m's \) of these equations are functions of the \( n's \), and the \( z's \) are

\*Animal breeders will be interested in the fact that \( \hat{c}_{it} \) is Lush's (1949) "most probable producing ability" modified by correcting the records for years and then expressing the result as a deviation from the group mean.
functions of \( n's \) and \( y's \) of equations (2), (3), (4), (5). As a check on the computations it should be noted that

\[
\sum_k m_{k',k} = \sum_t m_{k',t} \quad \text{for each year (} k' \text{)}
\]

\[
\sum_k m_{kt} = m_{t} \quad \text{for each group (} t \text{)}
\]

\[
\sum_k Z_{k'} = \sum_t Z_{t} .
\]

It is shown later that these equations are equivalent to the m.l. equations of Method II when \( r \) is assumed known.

Now from (7)

\[
\hat{\mu} + \hat{g}_t = \frac{1}{m_{t}} (z_{t} - \sum_k m_{kt} d_{k}).
\]

Substituting this in (6) we get equations in \( \hat{d}_k \) as follows

\[
\sum_k w_{kk'} \hat{d}_k = v_{k'} , \tag{8}
\]

and analogous equations for other years (\( k' \)). The fact that

\[
\sum_k w_{kk'} = 0 \quad \text{for each year (} k' \text{)}
\]

and

\[
\sum_k v_{k'} = 0
\]

can be used as a computational check. Because of this linear relationship among the coefficients (8) does not have a unique solution. Consequently, we impose one constraint on the estimators, a convenient one being \( \hat{d}_f = 0 \), where \( f \) refers to the final year. To solve (8) with this constraint we delete the last equation and the last unknown of the remaining equations. Then \( \hat{d}_k \) is the m.l. estimator of \( d_k - d_f \), and \( \hat{\mu} + \hat{g}_f \) is the m.l. estimator of \( \mu + g_t + d_f \). Therefore \( \hat{d}_k - \hat{d}_k' \) is the m.l. estimator of \( d_k - d_k' \), and \( (\hat{\mu} + \hat{g}_f) - (\hat{\mu} + \hat{g}_f') \) is the m.l. estimator of \( g_t - g_t' \).

Use of an incorrect value of repeatability biases the estimates of environmental and genetic trends. For example, if too large a value of \( r \) is used and if cows that were culled had lower records than their contemporaries, the estimate of environmental trend is biased downward.

If \( r \) is known, the sampling variance-covariance matrix of \( \hat{\mu} + \hat{g}_t \) and \( \hat{d}_k \), when \( \hat{d}_f = 0 \), is \( \sigma^2 \) times the inverse of the coefficients of (6) and (7) with the \( f \)th row and column deleted. The upper left submatrix
of this inverse can be obtained by inverting the coefficients of (8) with the last row and last column deleted.

The following methods for testing hypotheses concerning the $d$'s and $g$'s are appropriate if $r$ is assumed known.

The denominator sum of squares in the $F$ tests is

$$
\sum_i \sum_k \sum_t y_{ikt}^2 - \sum_i \sum_t \left[ \frac{r}{1 + (n_{i.t} - 1)r} \right] y_{i.t}^2 
- \sum_i \frac{z_{i.t}^2}{m_{i.t}} \text{ [from (7)]} - \sum_k \hat{d}_k \nu_k \text{ [from (8)].}
$$

This has degrees of freedom = number of records - number of years - number of groups + 1. The numerator sum of squares for testing the hypothesis that all $d_k$ are equal is

$$\sum_k \hat{d}_k \nu_k \text{ from (8).}$$

The numerator sum of squares for testing the hypothesis that all $g_t$ are equal is

$$\sum_t \frac{z_{i.t}^2}{m_{i.t}} \text{ [from (7)]} + \sum_k \hat{d}_k \nu_k \text{ [from (8)]} - \sum_k \tilde{d}_k z_k.$$

where $\tilde{d}_k$ are solutions to equations (6) after first deleting ($\bar{d} + \hat{g}_i$).

**A NUMERICAL EXAMPLE OF METHOD I (HENDERSON)**

Suppose we have the following records, classified according to group, cow in group, and year of freshening.

<table>
<thead>
<tr>
<th>Group</th>
<th>Cow</th>
<th>Year Fresh</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>450</td>
<td>420</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>500</td>
<td>470</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>410</td>
<td>430</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>390</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>400</td>
<td>430</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>430</td>
<td>380</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>380</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1750</td>
<td>2530</td>
</tr>
</tbody>
</table>
We shall take a value of 1/2 for \( r \); then as in (5), \( (1 - r)/r = 1 \) is added to \( n_{i,i} \) to form the diagonal coefficients of these equations, (5). Then the complete set of equations to be solved, namely those in (2) through (5) is

\[
\begin{bmatrix}
16 & 4 & 6 & 6 & 9 & 5 & 2 & 3 & 3 & 2 & 1 & 2 & 2 & 1 & 1 & 1 \\
4 & 4 & 0 & 0 & 4 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
6 & 0 & 6 & 0 & 3 & 3 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 \\
6 & 0 & 0 & 6 & 2 & 2 & 2 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\
9 & 4 & 3 & 2 & 9 & 0 & 0 & 3 & 3 & 2 & 1 & 0 & 0 & 0 & 0 \\
5 & 0 & 3 & 2 & 0 & 5 & 0 & 0 & 0 & 0 & 2 & 2 & 1 & 0 & 0 \\
2 & 0 & 0 & 2 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
3 & 1 & 1 & 1 & 3 & 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
3 & 1 & 1 & 1 & 3 & 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
2 & 1 & 1 & 0 & 2 & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\
2 & 0 & 1 & 1 & 0 & 2 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 \\
2 & 0 & 1 & 1 & 0 & 2 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
\end{bmatrix}
\begin{bmatrix}
\mu \\
\delta_1 \\
\delta_2 \\
\delta_3 \\
\delta_4 \\
\delta_5 \\
\delta_6 \\
\delta_7 \\
\delta_8 \\
\delta_9 \\
\end{bmatrix}
= \begin{bmatrix}
6960 \\
1750 \\
2530 \\
2680 \\
3980 \\
2020 \\
960 \\
1280 \\
1470 \\
840 \\
390 \\
830 \\
810 \\
380 \\
500 \\
460 \\
\end{bmatrix}
\]

The \( \delta \)'s are eliminated to obtain (6) and (7) as illustrated by the coefficient of \( \mu + \delta \), in the second equation of (6), which is

\[
3 - \frac{1(3)}{4} - \frac{1(3)}{4} - \frac{1(2)}{3} = \frac{5}{6}.
\]

The right hand member of the first equation reduces to

\[
1750 - \frac{1(1280)}{4} - \frac{1(1470)}{4} - \cdots - \frac{0(460)}{2} = \frac{3525}{6}.
\]

Multiplying all equations by 6 to eliminate fractions gives the following equations (6) and (7):
Note that the sum of the first three equations equals the sum of the last three.

Now we eliminate $\hat{\mu} + \hat{\beta}_i$ to obtain (8). For example, the coefficient of $d_2$ in the first equation is

$$-5 - \frac{8(5)}{16} - \frac{0(7)}{11} - \frac{0(0)}{6} = -\frac{15}{2}$$

and the first right hand member is

$$3525 - \frac{8(6975)}{16} - \frac{0(4420)}{11} - \frac{0(2880)}{6} = \frac{75}{2}.$$

Then multiplying each equation by 528 we get

$$\begin{bmatrix} 6336 & -3960 & -2376 & \hat{d}_1 \\ -3960 & 9495 & -5535 & \hat{d}_2 \\ -2376 & -5535 & 7911 & \hat{d}_3 \end{bmatrix} = \begin{bmatrix} 19,800 \\ -19,755 \\ -45 \end{bmatrix}.$$

These are equations (8). Note that their sum is zero.

Imposing the restriction $\hat{d}_3 = 0$, the solution is

$$\hat{d}_1 = 2.468, \quad \hat{d}_2 = -1.051$$

and substituting in the second set of equations and solving for $\hat{\mu} + \hat{\beta}$ we get

$$\hat{\mu} + \hat{\beta}_1 = 435.032$$

$$\hat{\mu} + \hat{\beta}_2 = 402.487$$

$$\hat{\mu} + \hat{\beta}_3 = 480.000.$$

**METHOD II (KEMPTHORNE AND VON KROSIGK)**

The discussion of the origin of the bias in the method of fitting
constants leads to an obvious estimation procedure for the case of two years. We use the following notation:

\[ \bar{x} = \text{average of all first-year records,} \]
\[ \bar{x}_1 = \text{average of first-year records of cows which are retained in the herd,} \]
\[ \bar{y} = \text{mean of second-year records (of cows retained, of course).} \]

The regression line of the joint distribution of the \( y \)'s and \( x \)'s is estimated unbiasedly even if selection of any sort is made with regard to the \( x \)'s, provided only that the true regression, or relationship to \( x \) of conditional mean of \( y \) for given \( x \), is linear throughout the entire range.* This happens automatically if the joint distribution of \( y \) and \( x \) in the unselected data is bivariate normal. It follows that the line

\[ y - \bar{y} = \hat{r}(x - \bar{x}_1) \]

is an unbiased estimate of the line

\[ y - (\mu + \gamma) = r(x - \mu), \]

where

\[ \hat{r} = \frac{\sum (y - \bar{y})(x - \bar{x})}{\sum (x - \bar{x}_1)^2}, \]

summations being over cows with both first- and second-year records. It follows then that

\[ \bar{y} + \hat{r}(\bar{x} - \bar{x}_1) \]

is an unbiased estimate of

\[ \mu + \gamma, \]

since \( \bar{x} \) is an unbiased estimate of \( \mu \) with errors independent of the errors of estimation of \( r \). Hence an unbiased estimate of \( \gamma \) is

\[ \bar{y} + \hat{r}(\bar{x} - \bar{x}_1) - \bar{x} = (\bar{y} - \bar{x}) - \hat{r}(\bar{x}_1 - \bar{x}). \]

We shall find the variance conditionally on the \( x \)'s for which second records exist. We have

\[ \hat{r} = \bar{y} + \hat{r}(\bar{x} - \bar{x}_1) - \bar{x} = (\bar{y} - \hat{r}\bar{x}_1) - (1 - \hat{r})\bar{x}, \]
\[ V(\hat{r}) = V(\bar{y} - \hat{r}\bar{x}_1) + V[(1 - \hat{r})\bar{x}] - 2 \text{Cov} [\bar{y} - \hat{r}\bar{x}_1, (1 - \hat{r})\bar{x}] \]
\[ = \sigma_{\hat{r}x}^2 \left( \frac{1}{n_2} + \frac{\bar{x}_1^2}{S} \right) + (1 - \hat{r})^2 \sigma^2_x \frac{\sigma^2_x}{n_1} + [\mu^2 + V(\bar{x})] \frac{\sigma^2_x}{S} \]
\[ + 2\mu \text{Cov} (\hat{r}, \bar{y} - \hat{r}\bar{x}_1), \]

*A bias would occur if culling were based also on some attribute correlated with both \( x \) and \( y \). This would occur if selection were based on dam’s performance for example. The bias can be seen to be of the order of \( r_1^2(1 - r) \) where \( r_1 \) is the correlation of the attribute with \( x \) and \( y \). This will be small in many cases, and for the particular case discussed here.
with \( S = \sum (x - \bar{x}_1)^2 \) and \( n_1, n_2 \) equal to the total number of first and second records respectively.

\[
\therefore V(\gamma) = \sigma_{v|x}^2 \left( \frac{1}{n_2} + \frac{\bar{x}_1^2}{S} \right) + (1 - r)^2 \frac{s^2}{n_1} + [\mu^2 + V(\bar{x})] \frac{\sigma_{y|x}^2}{S} - 2\mu \bar{x}_1 \frac{\sigma_{y|x}^2}{S}.
\]

An unbiased estimate of this is given by

\[
V(\gamma) = \sigma_{v|x}^2 \left( \frac{1}{n_2} + \frac{\bar{x}_1^2}{S} \right) + (1 - r)^2 \frac{s^2}{n_1} + \frac{s_{v|x}^2 s^2}{n_1 S} + \bar{x}^2 \frac{\sigma_{y|x}^2}{S} - 2\bar{x}_1 \frac{\sigma_{y|x}^2}{S}.
\]

with \( s^2 = \) mean square among first records
and \( s_{v|x}^2 = \) mean square about regression of second on first records
and is an unbiased estimate of \( \sigma_{v|x}^2 \).

The unconditional variance would depend also on Cov \( (\bar{x}, \bar{x}_1) \) which would be based on the culling procedure.

The procedure given above leads to an unbiased estimate of the year effect but the estimate is not efficient in the sense of exhausting the information even in the data from two years with successive records, because we have estimated \( r \) only by the regression of second-year records on first-year records. Under the assumptions the mean square of totality of first-year records is an estimate of \( \sigma^2 \), while the mean square of the second-year records about the fitted line is an estimate of \( \sigma^2 (1 - r^2) \). So we could obtain an estimate of \( r \), apart from sign, purely from these two mean squares. One problem of fitting is therefore to get the estimate of \( r \) which exhausts the data. Also data will extend over more than 2 years for many cows.

For these reasons and because a solution can be obtained by the method of maximum likelihood, we shall now use this method to attack the general case. The material given above is to be regarded as a first approximation to an exhaustive solution and indicates the sort of thing that can happen when one deals with data culled on the basis of part of the actual data. The approximate solution given above will also be of use in obtaining the maximum likelihood solution, because this solution will be obtainable only by iterative solution of non-linear equations.

As a framework for analysis of the general situation, suppose we have \( t \) records of a sample of the population of individuals and that there is no systematic environmental effect and no culling. Then it

\*This is independent of the regression estimate because the sum of squares of deviations of \( y \) about its regression on \( x \) is independent of that regression and of the \( x \)'s.
is reasonable to regard the $j$th record of the $i$th individual as being given by the equation

$$y_{ij} = \mu + c_i + e_{ij}$$

in which the portion $c_i$ is common to all records of the $i$th cow. Also we suppose that the $c_i$'s are normally and independently distributed with mean zero and variance $\sigma^2_c$, that the $e_{ij}$ are normally and independently distributed with mean zero and variance $\sigma^2_e$, and that the $e_{ij}$ and $c_i$ are independent. Then we have, with $V$ denoting variance, and Cov denoting covariance,

$$V(y_{ij}) = \sigma^2_c + \sigma^2_e = \sigma^2$$

$$\text{Cov}(y_{ij}, y_{ij'}) = \sigma^2_e = \tau \sigma^2,$$

with

$$\tau = \frac{\sigma^2_c}{\sigma^2_e + \sigma^2_e}.$$

The variance-covariance matrix of the $t$ records of an individual is therefore

$$
\begin{bmatrix}
1 & & & \\
& \ddots & \ddots & \ddots \\
& & \ddots & \ddots \\
& & & 1
\end{bmatrix}
$$

i.e. $\sigma^2$ times a matrix with unity on the diagonal and $r$ in every off-diagonal position. Let us now consider the regression of each record on the preceding ones. We have the following relationships, in which $z_{ij}$ equals $y_{ij} - \mu$:

$$z_{i2} = rz_{i1} + e_{i2.1}, \quad V(e_{i2.1}) = \sigma^2(1 - r^2)$$

$$z_{i3} = \frac{r}{1 + r} (z_{i1} + z_{i2}) + e_{i3.12}, \quad V(e_{i3.12}) = \sigma^2 \left(1 - \frac{2r^2}{1 + r}\right)$$

$$z_{i4} = \frac{r}{1 + 2r} (z_{i1} + z_{i2} + z_{i3}) + e_{i4.123}, \quad V(e_{i4.123}) = \sigma^2 \left(1 - \frac{3r^2}{1 + 2r}\right)$$

$$z_{i5} = \frac{r}{1 + 3r} (z_{i1} + z_{i2} + z_{i3} + z_{i4}) + e_{i5.1234}, \quad V(e_{i5.1234}) = \sigma^2 \left(1 - \frac{4r^2}{1 + 3r}\right)$$
and so on. These regression equations may be obtained by the use of the method of least squares. As an example, to find the regression of \( z_{i4} \) on \( z_{i1}, z_{i2}, \) and \( z_{i3} \) we have to fit the equation

\[
z_{i4} = \beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 z_{i3}
\]

which gives the normal equations

\[
\begin{align*}
\beta_1 \sigma^2 + \beta_2 r \sigma^2 + \beta_3 r \sigma^2 &= r \sigma^2 \\
\beta_1 r \sigma^2 + \beta_2 r \sigma^2 + \beta_3 r \sigma^2 &= r \sigma^2 \\
\beta_1 r \sigma^2 + \beta_2 r \sigma^2 + \beta_3 r \sigma^2 &= r \sigma^2
\end{align*}
\]

to which the solution is

\[
\beta_1 = \beta_2 = \beta_3 = \frac{r}{1 + 2r}.
\]

The residual variance is the total variance of \( z_{i4} \) minus the sum of products of regression coefficients and right-hand sides of the normal equations. In addition, from the general regression theorem that the residual from a regression is uncorrelated with the "independent" variables that are included in the regression we know that

- \( e_{i2.1} \) is independent of \( z_{i1} \),
- \( e_{i3.12} \) is independent of \( z_{i1} \) and \( z_{i2} \) and hence of \( e_{i2.1} \)
- \( e_{i4.123} \) is independent of \( z_{i1}, z_{i3} \) and hence of \( e_{i3.12} \) and so on.

The joint distribution of the \( y_{ij}, j = 1, 2, \cdots, t \) with \( i \) fixed is therefore expressible as

\[
\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{z_{i1}^2}{2\sigma^2} \right] \times \frac{1}{\sqrt{2\pi\sigma^2}} \sqrt{1 - r^2} \exp \left[ -\frac{1}{2} \frac{e_{i2.1}^2}{\sigma^2(1 - r^2)} \right] \\
\times \frac{1}{\sqrt{2\pi\sigma^2}} \sqrt{1 - \frac{2r^2}{1 + r}} \exp \left[ -\frac{e_{i3.12}^2}{2\sigma^2(1 - \frac{2r^2}{1 + r})} \right] \times \cdots
\]

We now apply these elementary distributional results, which are probably widely known even if not stated explicitly somewhere in the literature, to the problem at hand.

We suppose that the first records of an individual are not subject to culling* at all and that individuals are culled after the first record on any basis. The existence of a second record usually depends in some way on the value of the first record, of the third record on the value of

---

*Any such culling is more properly to be regarded purely as selection of incoming animals and the effects of such selection will be included in genetic parameters to be estimated.
the first two records. We suppose that the group of individuals contributing first records in year $t$ deviates from the population mean systematically by an amount $g_t$, and that records made in the $k$th year deviate from the population mean systematically by an amount $d_k$. We can therefore write down the joint distribution of the totality of records as follows. Let $y_{ikt}$ denote the $j$th record of individual $i$ which was made in year $k$, this individual having entered the herd in year $t$. Then the distribution of the $y_{ikt}$ is

$$F_1 \times F_{2|1} \times F_{3|12} \times \cdots,$$

where

$$F_1 = \prod_{(1)} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(y_{1kt} - \mu - g_t - d_k)^2}{2\sigma^2} \right],$$

$\prod_{(1)}$ denoting the product over all first records, and

$$F_{2|1} = \prod_{(2)} \frac{1}{\sqrt{2\pi\sigma^2}} \sqrt{1 - r^2} \cdot \exp \left[ -\frac{1}{2\sigma^2(1 - r^2)} \left\{ (y_{1kt} - \mu - g_t - d_k) - r(y_{1kt'} - \mu - g_t - d_{k'}) \right\}^2 \right],$$

$\prod_{(2)}$ denoting the product over all second records, $k'$

denoting the year of the first record corresponding to the second record $y_{ikt}$,

and

$$F_{3|12} = \prod_{(3)} \frac{1}{\sqrt{2\pi\sigma^2}} \sqrt{1 - \frac{2r^2}{1 + r}} \cdot \exp \left[ -\frac{1}{1 + r} \left( (y_{3kt} - \mu - g_t - d_k) - \frac{r}{1 + r} (y_{1kt'} - \mu - g_t - d_{k'}) \right) \right.

\left. - \frac{r}{1 + r} (\mu_{1k,t'} - \mu - g_t - d_{k'}) \right]^2 \bigg/ 2\sigma^2 \left( 1 - \frac{2r^2}{1 + r} \right),$$

$\prod_{(3)}$ denoting the product over all third records, and $k'$, $k''$ denoting the years in which the first and second records corresponding to the third record $y_{13kt}$ were made, and so on.

This joint distribution takes account of the fact that culling on prior records to an unknown extent was made. It incorporates, superficially at least, all the ingredients which one would like, namely
\[ \mu, \] the mean in an arbitrary base population;
\[ g_t, \] the amount by which the group of individuals entering the herd in year \( t \) deviate from the base population;
\[ d_k, \] the amount by which the records made in year \( k \) deviate from those made in an arbitrary base year;
\[ r, \] the repeatability;
\[ \sigma^2, \] the variance of records of a population of unculled individuals.

The only regrettable feature of the joint distribution is the assumption that \( \sigma^2 \) does not change, though if repeatability is not high and culling and selection not intense, one would not expect the genetic part of \( \sigma^2 \) to change appreciably over a period of many years. We have no easy alternative to assuming the environmental component remains constant.* Without this assumption the notion of repeatability would, of course, break down, at least partially.

The logarithm of the likelihood is equal, apart from a constant, to

\[
\log L = \sum_{(1)} \left[ -\frac{1}{2} \log \sigma^2 - \frac{(y_{i1k} - \mu - g_t - d_k)^2}{2\sigma^2} \right] \\
+ \sum_{(2)} \left[ -\frac{1}{2} \log \sigma^2 - \frac{1}{2} \log (1 - r^2) - \frac{(y_{i2k} - \mu - g_t - d_k - r(y_{i1k'} - \mu - g_t - d_{k'}))^2}{2\sigma^2(1 - r^2)} \right] \\
+ \sum_{(3)} \left[ -\frac{1}{2} \log \sigma^2 - \frac{1}{2} \log \left(1 - \frac{2r^2}{1 + r}\right) - \frac{r}{1 + r} (y_{i3k} - \mu - g_t - d_k - \frac{r}{1 + r} (y_{i2k'} - \mu - g_t - d_{k'}))^2 \right] / 2\sigma^2 \left(1 - \frac{2r^2}{1 + r}\right) \\
+ \text{etc.,}
\]

where \( \sum_{(1)} \) denotes summation over all first records, \( \sum_{(2)} \) summation over all second records, and so on. An alternative expression for \( \log L \) is more suitable for visualizing the likelihood and the problem of maximizing this likelihood. Let \( N \) denote the total number of records available, \( n_1 \) the total number of first records, \( n_2 \) the number of second records, and so on. Also let \( \sum_{(1)} \) denote summation over individuals.

*The extent to which this assumption is false could be examined by fitting a more general model.
with only a first record, $\sum_{(2)}$ summation over individuals with only first and second records, and so on. Then

$$\log L = -\frac{N}{2} \log \sigma^2 - \frac{n_2}{2} \log (1 - r^2) - \frac{n_3}{2} \log \left(1 - \frac{2r^2}{1 + r}\right)$$

$$- \frac{n_4}{2} \log \left(1 - \frac{3r^2}{1 + 2r}\right) - \ldots - \sum_{(1)} \frac{(y_{i1kt} - \mu - g_i - d_k)^2}{2\sigma^2}$$

$$- \sum_{(2)} \left[ \frac{1}{2\sigma^2(1 - r^2)} \left\{ \right. \right.$$

$$- \sum_{(3)} \left[ \frac{1}{2\sigma^2(1 - r^2)} \left. \right\} \right. \right.$$

$$- \frac{r}{1 + r} \left( y_{i2k't} - \mu - g_i - d_k \right)^2 \right. \right.$$

$$+ \left. \left. \frac{(y_{i3kt} - \mu - g_i - d_k)^2}{2\sigma^2(1 - r^2)} \right. \right.$$

$$+ \frac{(y_{i1k't} - \mu - g_i - d_k)^2}{2\sigma^2(1 - r^2)} \right. \right.$$

$$+ \ldots \text{ etc.}$$

With records covering $s$ years we shall have $(2s + 3)$ parameters to fit, of which one $g$ and one $d$ may be chosen arbitrarily. We may as well carry on the mathematical manipulations symmetrically in all the $g$'s and $d$'s, and impose necessary conditions at as late a stage as possible. We find

$$\frac{\partial \log L}{\partial g_p} = \sum_{(1)p} \frac{(y_{i1kp} - \mu - g_p - d_k)}{\sigma^2}$$

$$+ \sum_{(2)p} \frac{(y_{i2kp} - \mu - g_p - d_k - r(y_{i1k',p} - \mu - g_p - d_k'))}{\sigma^2(1 + r)}$$

$$+ \sum_{(3)p} \frac{(y_{i3kp} - \mu - g_p - d_k - r(y_{i2k',p} - \mu - g_p - d_k'))}{\sigma^2(1 + 2r)}$$

$$- \frac{r}{1 + r} \left( y_{i1k',p} - \mu - g_p - d_k' \right) \right. \right.$$

$$\left. \right. \left. \right\} \right. \right.$$

$$\left. \right. \left. \right. \right. \right.$$

$$\left. \right. \left. \right. \right. \right.$$

$$\left. \right. \left. \right. \right. \right.$$

$$\left. \right. \left. \right. \right. \right.$$

$$\left. \right. \left. \right. \right. \right.$$

$$\left. \right. \left. \right. \right. \right.$$

$$\left. \right. \left. \right. \right. \right.$$

$$\left. \right. \left. \right. \right. \right.$$

$$\left. \right. \left. \right. \right. \right.$$

$$\left. \right. \left. \right. \right. \right.$$

$$\left. \right. \left. \right. \right. \right.$$

$$\left. \right. \left. \right. \right. \right.$$

$$\left. \right. \left. \right. \right. \right.$$

$$\left. \right. \left. \right. \right. \right.$$

$$\left. \right. \left. \right. \right. \right.$$

$$\left. \right. \left. \right. \right. \right.$$

$$\left. \right. \left. \right. \right. \right.$$

$$\left. \right. \left. \right. \right. \right.$$

$$\left. \right. \left. \right. \right. \right.$$

$$\left. \right. \left. \right. \right. \right.$$

$$\left. \right. \left. \right. \right. \right.$$

$$\left. \right. \left. \right. \right. \right.$$

$$\left. \right. \left. \right. \right. \right.$$
\[
\frac{\partial L}{\partial d_q} = \sum_{(q,q')}' (y_{i1q'} - \mu - g_i - d_q) / \sigma^2 \\
+ \sum_{(qq')}'(y_{i2q} - \mu - g_i - d_q - r(y_{i1q} - \mu - g_i - d_q')) / \sigma^2 (1 - r^2) \\
- \sum_{(q,q')}' (y_{i2q'} - \mu - g_i - d_q') - r(y_{i1q} - \mu - g_i - d_q)) / \sigma^2 (1 - r^2) \\
+ \sum_{(qq'q''q''')}'(y_{i3q} - \mu - g_i - d_q) - \frac{r}{1 + r} (y_{i2q} - \mu - g_i - d_q) \\
- \frac{r}{1 + r} (y_{i1q'} - \mu - g_i - d_q') \bigg/ \sigma^2 \left(1 - \frac{2r^2}{1 + r}\right) \\
- \sum_{(qq'q''q''')}'(y_{i3q'} - \mu - g_i - d_q') - \frac{r}{1 + r} (y_{i2q'} - \mu - g_i - d_q') \\
- \frac{r}{1 + r} (y_{i1q} - \mu - g_i - d_q) \bigg/ \sigma^2 \left(1 - \frac{2r^2}{1 + r}\right) \\
+ \sum_{(qq'q''q''')}'(y_{i4q} - \mu - g_i - d_q) - \frac{r}{1 + 2r} (y_{i3q'} - \mu - g_i - d_q') \\
- \frac{r}{1 + 2r} (y_{i2q'} - \mu - g_i - d_q') \bigg/ \sigma^2 \left(1 - \frac{3r^2}{1 + 2r}\right) \\
- \sum_{(qq'q''q''')}'(y_{i4q'} - \mu - g_i - d_q') - \frac{r}{1 + 2r} (y_{i3q} - \mu - g_i - d_q) \\
- \frac{r}{1 + 2r} (y_{i2q} - \mu - g_i - d_q) \bigg/ \sigma^2 \left(1 - \frac{3r^2}{1 + 2r}\right) \\
- \sum_{(qq'q''q''')}'(y_{i4q'} - \mu - g_i - d_q') - \frac{r}{1 + 2r} (y_{i3q'} - \mu - g_i - d_q') \\
- \frac{r}{1 + 2r} (y_{i2q} - \mu - g_i - d_q)
\]
\[- \frac{r}{1 + 2r} \left( y_{i1q'''}t - \mu - g_t - d_{q'''} \right) \frac{r}{1 + 2r} / \sigma^2 \left( 1 - \frac{3r^2}{1 + 2r} \right) \]

\[- \sum' \left\{ (y_{i1q'}t - \mu - g_t - d_{q'}) - \frac{r}{1 + 2r} (y_{i2q''}t - \mu - g_t - d_{q''}) \right. \]

\[- \frac{r}{1 + 2r} (y_{i2q'''}t - \mu - g_t - d_{q'''}) \]

\[- \frac{r}{1 + 2r} (y_{i1et} - \mu - g_t - d_{e}) \frac{1}{1 + 2r} / \sigma^2 \left( 1 - \frac{3r^2}{1 + 2r} \right) + \text{etc.} \]

where, as an example of the summation notation, \( \sum'_{(q''q''',q''')q} \) means summation over the function in the brackets for those individuals whose third record is in year \( q \), and \( q', q''', q'''' \) denote the years in which the first, second, and fourth record respectively of each such individual were made. These derivatives are indeed formidable but the method of obtaining the maximum likelihood estimates will be iterative and one will make rough estimates of the parameters which will be used to compute the first derivatives numerically. Adjustments to the rough estimates will be obtained by solution of linear equations. Alternatively if one can determine the best estimate of \( r \), the estimates of \( \mu, g_t, d_k \) are obtained as the solution of linear equations.

The derivative of log \( L \) with respect to \( \mu \) is the sum of the derivatives with respect to the \( g \)'s or the \( d \)'s.

In addition we have

\[ \frac{\partial \log L}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \sum_{(1)} (y_{i1kt} - \mu - g_t - d_k)^2 \cdot 1/2\sigma^4 \]

\[ + \sum_{(2)} \left\{ (y_{i2kt} - \mu - g_t - d_k) - r(y_{i2k't} - \mu - g_t - d_{k'})^2 \cdot 1/2\sigma^4(1 - r^2) \right. \]

\[ + \sum_{(3)} \left\{ (y_{i3kt} - \mu - g_t - d_k) - \frac{r}{1 + r} (y_{i3k't} - \mu - g_t - d_{k'}) \right. \]

\[ - \frac{r}{1 + r} (y_{i3k''t} - \mu - g_t - d_{k''}) \left\} \cdot 1/2\sigma^4 \left( 1 - \frac{3r^2}{1 + 2r} \right) + \text{etc.} \]

It is easy to evaluate \( \sigma^2 \) by equating this derivative to zero if one has values for \( \mu \), the \( g \)'s, the \( d \)'s, and \( r \). The derivative with respect to \( r \) is not as simple, for we have
\[
\frac{\partial \log L}{\partial r} = \frac{n_0 r}{(1 - r^2)} + \frac{n_3 r(2 + r)}{(1 - r)(1 + r)(1 + 2r)} \\
+ \frac{n_3 r(2 + 2r)}{2(1 - r)(1 + 2r)(1 + 3r)} + \frac{n_3 r(2 + 3r)}{2(1 - r)(1 + 3r)(1 + 4r)} + \cdots \\
- \sum_{(2)} \frac{(e_{i'k'} - r e_{i'k''})^2}{2\sigma^2} \frac{2r}{(1 - r^2)^2} \\
- \sum_{(3)} \left( \frac{e_{ik'} - \frac{r}{1 + r} e_{i'k'} - \frac{r}{1 + r} e_{i'k''}}{2\sigma^2} \right)^2 \frac{2r(2 + r)}{(1 + 2r)^2(1 - r)^2} \\
- \sum_{(4)} \left( \frac{e_{i'k'} - \frac{r}{1 + 2r} e_{i'k''} - \frac{r}{1 + 2r} e_{i'k'''} - \frac{r}{1 + 2r} e_{i'k''''}}{2\sigma^2} \right)^2 \\
\cdot \frac{3r(2 + 2r)}{(1 + 3r)^2(1 - r)^2} - \text{etc.}
\]

where \(e_{i'k'} = y_{i'k'} - \mu - g_1 - d_k\).

The maximum likelihood estimates for \(\mu, g_1, d_k, r, \sigma^2\) are the values for these quantities which make the derivatives equal to zero. The equations obtained by setting the derivatives equal to zero are non-linear in the parameters so the method of solution followed will be iterative. In a situation with estimation of parameters \(\theta_1, \theta_2, \ldots, \theta_r\), a general procedure is to expand the quantities \(\partial \log L / \partial \theta_i\) by a Taylor series around guessed values, ignoring terms of higher order than quadratic, giving for \(\theta_1\), for instance,

\[
\frac{\partial \log L}{\partial \theta_1} \bigg|_{\hat{\theta}} = \frac{\partial \log L}{\partial \theta_1} \bigg|_{\theta_0} + (\hat{\theta}_1 - \theta_0) \frac{\partial^2 \log L}{\partial \theta_1^2} \bigg|_{\theta_0} \\
+ (\hat{\theta}_2 - \theta_0) \frac{\partial^2 \log L}{\partial \theta_1 \partial \theta_2} \bigg|_{\theta_0} + \cdots (\hat{\theta}_r - \theta_0) \frac{\partial^2 \log L}{\partial \theta_1 \partial \theta_r} \bigg|_{\theta_0}.
\]
In this equation \( \hat{\theta} \) is one arbitrary point \((\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_r)\) in \(r\)-dimensional space and \(\theta_0\) or \((\theta_{10}, \theta_{20}, \cdots, \theta_{r0})\) is another arbitrary point, and all derivatives on the right-hand side are evaluated at \((\theta_{10}, \theta_{20}, \cdots, \theta_{r0})\). The maximum likelihood estimates are those for which the left-hand sides are zero, so by putting these left-hand sides equal to zero and solving for \(\hat{\theta}_1 - \theta_{10}, \hat{\theta}_2 - \theta_{20}, \cdots\), we get an approximation to the maximum likelihood estimates. The closer the original \(\theta_{10}, \theta_{20}, \cdots, \theta_{r0}\) are to the maximum likelihood values, the smaller will be the steps from \(\theta_{10}\) to \(\hat{\theta}_1\) and when these steps are negligible we take \(\hat{\theta}_i\) or \(\theta_{10}\) to be the solution.

A modification of the purely mathematical procedure for finding the maximum of a function is to use in a statistical situation the expected values of the second derivatives rather than the actual derivatives. It was in fact in these terms that the iterative method of maximum likelihood was first presented. As a general rule the authors prefer to use actual derivatives but in the present case it can be seen there are considerable advantages in using the expected values of the second derivatives.

The problem is strictly a computational one from this point on. It does not seem appropriate to pursue the matter further herein, particularly as we have had no experience with a set of data of the dimensions that ordinarily arise. We shall therefore close this section with some remarks on the situation, leaving the computational details to a later publication.

It is clear that if we possess a good estimate of \(r\) (see below, however) the problem is strictly one of weighted least squares with known relative weights. This suggests that one should first estimate \(r\) in as good a way as possible, and then use weighted least squares on linear functions of the records. One would then presumably go through at least one cycle of the iteration which would result in estimates and estimated variances and covariances of the estimates.

The extent and direction of bias resulting from estimating \(r\) without taking genetic trend into account is an important and unsolved problem. Another question lies in the propriety of applying an \(r\) value for one herd to another herd when there is no reason to suppose that \(r\) is constant.

A PARTIAL EXAMINATION OF A SET OF DATA
(KEMPThORNE AND VON KROSIGK)

The basic ideas of the estimation procedure of Method II have been explored with the data of the Iowa Board of Control herd at Woodward for the period 1940 to 1954. The aspects we have examined are the
consistency of the regression equations of a record on preceding records and the estimation of $r$.

Considering only first and second records made in successive years, we examined the homogeneity of the regressions between starting groups, which are defined as the groups of cows entering the records in each year. The analysis of variance is given in Table 1.

**TABLE 1**

**Test for Homogeneity of Within Starting Year Regressions**

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>S.S.</th>
<th>M.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Among groups</td>
<td>14</td>
<td>38,210,685</td>
<td></td>
</tr>
<tr>
<td>Common regression</td>
<td>1</td>
<td>257,323</td>
<td></td>
</tr>
<tr>
<td>Additional due to separate</td>
<td>13</td>
<td>52,514</td>
<td>4,040</td>
</tr>
<tr>
<td>regressions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>205</td>
<td>1,067,441</td>
<td>5,207</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>233</td>
<td>39,587,963</td>
<td></td>
</tr>
</tbody>
</table>

There is no evidence at all for different regressions so we take, as the estimate from this portion of the data, $r$ equal to 0.501 with a standard error of 0.071.

For the portion of the data with first, second, and third records in three successive years, we found that the within-group partial regression of third record on first record was $0.278 \pm 0.115$ and of the second record on first record was $0.501 \pm 0.133$. The difference of these two coefficients leads to a $t$ value equal to 1.1, which in no way contradicts the hypothesis of equality of the two coefficients. The estimate forcing the two partial regression coefficients to be equal was $0.378 \pm 0.071$.

For the portion of the data with four records in successive years the multiple regression of fourth record on first, second, and third gave partial regression coefficients as follows:

- on first: $0.145 \pm 0.218$
- on second: $0.268 \pm 0.184$
- on third: $0.228 \pm 0.140$

The $F$ value for differences of these was 0.07, which does not contradict the hypothesis of equality. The estimate forcing the three partial coefficients to be the same was $0.218 \pm 0.066$.

These three estimates are independent. They can best be represented in tabular form. Table 2 shows the estimates, the resulting estimates
of $r$, and the weights. The best estimate of $r$ from the regression coefficients is $0.503 \pm 0.063$. This estimate has been used to obtain estimates of yearly environmental trends from first and second records only. The computations are summarized in Table 3. The sum of the estimates of the yearly environmental trends gives an estimate of $31 \pm 98$ pounds of butterfat for the total change in environment.

**TABLE 3**

**Estimated Yearly Environmental Effects**

<table>
<thead>
<tr>
<th>Year</th>
<th>$n_1$</th>
<th>$n_2$</th>
<th>$\bar{x}$</th>
<th>$\bar{z}_1$</th>
<th>$\bar{y}$</th>
<th>$(\bar{y} - \bar{x}) - \hat{\rho}(\bar{z}_1 - \bar{x})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1940</td>
<td>26</td>
<td>20</td>
<td>360.8</td>
<td>363.7</td>
<td>422.0</td>
<td>$59.7 \pm 18.0$</td>
</tr>
<tr>
<td>1941</td>
<td>14</td>
<td>11</td>
<td>382.7</td>
<td>403.6</td>
<td>431.6</td>
<td>$38.4 \pm 24.3$</td>
</tr>
<tr>
<td>1942</td>
<td>22</td>
<td>22</td>
<td>449.3</td>
<td>449.3</td>
<td>428.1</td>
<td>$-21.2 \pm 17.6$</td>
</tr>
<tr>
<td>1943</td>
<td>13</td>
<td>9</td>
<td>373.1</td>
<td>414.6</td>
<td>414.1</td>
<td>$20.1 \pm 26.7$</td>
</tr>
<tr>
<td>1944</td>
<td>6</td>
<td>3</td>
<td>423.4</td>
<td>413.3</td>
<td>355.0</td>
<td>$-63.3 \pm 44.8$</td>
</tr>
<tr>
<td>1945</td>
<td>13</td>
<td>10</td>
<td>352.2</td>
<td>344.2</td>
<td>335.8</td>
<td>$-12.4 \pm 25.4$</td>
</tr>
<tr>
<td>1946</td>
<td>5</td>
<td>2</td>
<td>363.4</td>
<td>327.0</td>
<td>363.0</td>
<td>$17.9 \pm 54.2$</td>
</tr>
<tr>
<td>1947</td>
<td>37</td>
<td>31</td>
<td>354.5</td>
<td>357.7</td>
<td>345.2</td>
<td>$-10.9 \pm 14.6$</td>
</tr>
<tr>
<td>1948</td>
<td>21</td>
<td>16</td>
<td>385.1</td>
<td>411.7</td>
<td>388.3</td>
<td>$-10.2 \pm 20.0$</td>
</tr>
<tr>
<td>1949</td>
<td>31</td>
<td>23</td>
<td>388.3</td>
<td>415.7</td>
<td>441.0</td>
<td>$38.9 \pm 16.8$</td>
</tr>
<tr>
<td>1950</td>
<td>31</td>
<td>22</td>
<td>411.6</td>
<td>447.3</td>
<td>372.3</td>
<td>$-57.3 \pm 17.2$</td>
</tr>
<tr>
<td>1951</td>
<td>28</td>
<td>18</td>
<td>361.4</td>
<td>389.8</td>
<td>375.3</td>
<td>$-0.4 \pm 18.7$</td>
</tr>
<tr>
<td>1952</td>
<td>30</td>
<td>19</td>
<td>376.2</td>
<td>401.0</td>
<td>444.2</td>
<td>$55.5 \pm 17.8$</td>
</tr>
<tr>
<td>1953</td>
<td>40</td>
<td>27</td>
<td>464.0</td>
<td>473.1</td>
<td>444.6</td>
<td>$-24.0 \pm 15.3$</td>
</tr>
</tbody>
</table>

To illustrate the importance of the assumption that age effects have been eliminated suppose all of the cows in the above example freshened for the first time at two years and two months. Then if these first records were multiplied by, say, 1.25 instead of 1.28 the estimate of the total environmental change would be increased by approximately 120 pounds. Of course, if records at all ages were used any effect of age being incorrectly discounted would be damped down. However, first and second records will always comprise the major share of the data.
If a term were added to the model for ages so that the effect of ages could be estimated simultaneously with the other factors, one would have more confidence in the estimates. However, it can be seen, as pointed out by Rendel and Robertson [1950], that if ages and starting dates were both classified with the same accuracy they would be perfectly confounded and there would not be a unique solution to the equations. It does not seem logically defensible to classify the records by years for start of lactation and by, say, months for ages in order to break down the perfect correlation between the two.

EQUIVALENCE OF THE METHODS (SEARLE)

Method II uses four subscripts on \( y, y_{ikt} \) being the record made in year \( k \) by the \( i \)th cow of the group of cows whose first records were made in year \( t \), its being the cow's \( j \)th record. Thus \( j \) is no more than an ordinal indicator of which record of the cow \( y_{ikt} \) is and in terms of the elements of the model it plays no part. In the case where a cow's second record is always made in the year immediately following that of her first record, and her third record always comes in the year following that of her second, and so on, \( j = k + 1 - t \). The presence of the \( j \) as a subscript to \( y \) merely emphasizes that records may not occur in concurrent years—when they do, as in many situations, the \( j \) is redundant because of the above relationship. In either case the model of Method I is applicable:

\[
y_{ikt} = \mu + d_k + g_i + c_{ij} + e_{ikt}.
\]

In this model the random variables are the \( c \)'s and \( e \)'s. For the cow \((i, t)\) having \( n_{i,t} \) records it is convenient to define the following column vectors: (i) \( y_{it} \), her \( n_{i,t} \) records, (ii) \( y_{it}^t \), a vector of \( n_{i,t} \) 1's, and (iii) \( d_{it} \) and \( e_{it} \) as the \( d \)'s and \( e \)'s appropriate to her \( n_{i,t} \) records. Then the likelihood of the sample of \( y \)'s for all cows is

\[
L = \prod_{i,t} \left\{ \frac{1}{(\sqrt{2\pi})^{n_{i,t}} | A_{it}|^{1/2}} \exp \left[ -1/2(c_{it}y_{it}^t + e_{it})A_{it}^{-1}(c_{it}y_{it} + e_{it}) \right] \right\}
\]

where \( A_{it} \) is a square matrix of order \( n_{i,t} \), all diagonal terms being \( \sigma_s^2 = \sigma_c^2 + \sigma_e^2 \), and all non-diagonal terms being \( \tau \sigma \), where \( \tau \) is repeatability. This is the variance-covariance matrix of the \( n_{i,t} \) records of the cow \((i, t)\). Due to its special form the likelihood can be expanded and be shown to be equal to

\[
L = \frac{1}{(2\pi \sigma_s^2)^{n_{i,t}/2}} \exp \left[ -\frac{\sum c^2_{it}}{2\sigma_s^2} \right] \prod_{i,t} \left[ 1 + (n_{i,t} - 1)\tau \right]^{1/2} \exp \left[ -\frac{\sum c^2_{it}}{2\tau} \right] \prod_{i,t} \left( 2\pi \sigma_c^2(1 - \tau)[1 + (n_{i,t} - 1)\tau] \right)^{-1/2} \exp \left[ -\frac{(1 - \tau)c_{it} - \tau \sum c_{it}}{2\tau_c^2(1 - \tau)[1 + (n_{i,t} - 1)\tau]} \right]
\]
where $M$ is the number of cows. This expression is the likelihood of the sample of $y$'s. It is therefore the likelihood used in Method II, although it was there written as the product over all cows of the conditional likelihoods

$$F_{1}F_{2|1}F_{3|2,1}\ldots$$

equal to

$$L_{1}(y_{i1t})L_{2}(y_{i2t} | y_{i1t})L_{3}(y_{i3t} | y_{i2t}, y_{i1t})\ldots.$$

With the substitution

$$c_{ikt} = y_{ikt} - \mu - d_{k} - g_{i} - c_{it}$$

the numerator of $L$ as given above is the joint distribution function which has been maximized in Method I with respect to $\mu$, the $d$'s, $g$'s, and the $c$'s. In Method II, $L$ has been maximized with respect to $\mu$, the $d$'s, $g$'s, and $r$. Therefore the Method I equations with the following three amendments will yield the Method II equations:

(i) Include the equations arising from maximizing $L$ with respect to $r$.

(ii) Subtract from the equations arising from the differentiation with respect to $\mu$, the $d$'s, and $g$'s, those terms arising from differentiating the logarithm of the denominator of $L$ with respect to these same parameters.

(iii) Delete the equations arising from maximizing $L$ with respect to the $c_{i,i}'s$.

In the situation where $r$ is assumed known and is not to be estimated the equations of (i) will not occur.

As an example of the terms in (ii), consider the equation arising from differentiation with respect to $\mu$. In Method I this will come from equating to zero the expression

$$\frac{\partial}{\partial \mu} \left\{ \sum_{i} \sum_{k} \sum_{l} \frac{(y_{ikt} - \mu - d_{k} - g_{i} - c_{it})^{2}}{2\sigma^{2}} \right\}$$

and this is

$$\frac{\sum_{i} \sum_{k} \sum_{l} (y_{ikt} - \mu - d_{k} - g_{i} - c_{it})(-1)}{(1 - r)\sigma^{2}}.$$
In Method II there will be added to this expression the following

$$\frac{\partial}{\partial \mu} \left\{ \sum_{i,t} \left[ \frac{(1 - r)c_{i,t} - r \sum_k (y_{ikt} - \mu - d_k - g_t - c_{i,t})^2}{2\sigma^2(1 - r)[1 + (n_{i,t} - 1)r]} \right] \right\}$$

which equals

$$\sum_{i,t} \frac{n_{i,t}r[1 + (n_{i,t} - 1)]\hat{c}_{i,t} - r[y_{i,t} - n_{i,t}(\bar{y} + \bar{g}_t) - \sum_k n_{ikt} \hat{d}_k]}{r(1 - r)\sigma^2[1 + (n_{i,t} - 1)r]}.$$

In the difference the terms in \(\hat{c}_{i,t}\) are

$$\sum_{i,t} \frac{\hat{c}_{i,t}}{(1 - r)^2} (-n_{i,t} + n_{i,t}) = 0,$$

i.e. the \(\hat{c}'s\) are eliminated from the equation; and this will be found to be true for the equations corresponding to the differentiations with respect to the \(d's\) and the \(g's\). Thus together with (iii) we see that the equations of Method II are those of Method I after the \(c's\) have been eliminated, namely equations (6) and (7) of Method I.

Hence we have shown that when the repeatability \(r\) is assumed known, the two methods are equivalent. That is, maximizing the likelihood with respect to \(\mu\), the \(d_k's\), and the \(g_t's\) (as in Method II), gives the same equations as result from eliminating the \(\hat{c}_{i,t}'s\) from the equations obtained by maximizing the joint distribution function of the \(y's\) and \(c's\), with respect to \(\mu\), the \(d_k's\), the \(g_t's\), and the \(c_{i,t}'s\) (as in Method I). Thus when \(r\) is assumed known, the equations of Method II are simply those of Method I with the \(\hat{c}'s\) eliminated; and hence the two methods give the same estimates of estimable linear functions of the fixed effects \(\mu, d_k, g_t\). Furthermore, since it has been proved in Method II that these estimates are unbiased by selection, such unbiasedness also applies to Method I.

**SUMMARY**

Two methods for the separation of genetic and environmental time trends are presented and compared. The reason the classical least squares approach to this problem yields biased estimates is described in standard statistical terminology. The first method is illustrated with a small numerical example and a partial examination of a set of data is used to illustrate some of the assumptions and the estimation of parameters in Method II.

The methodology is presented and discussed in terms of dairy records. However, the same technique would apply to many other cases, often without the complication of age-correction factors.
REFERENCES


