Genet. Sel. Evol. (1989) 21, 399-414

Original article

Likelihood inferences in animal breeding under selection: a missing-data theory view point

S. Im[1], R.L. Fernando[2] and D. Gianola[2]

[1] *Institut National de la Recherche Agronomique, laboratoire de biométrie, BP 27, 31326 Castanet-Tolosan, France;*
[2] *University of Illinois at Urbana-Champaign, 126 Animal Sciences Laboratory, 1207 West Gregory Drive, Urbana, Illinois 61801, USA*

(received 28 October 1988; accepted 20 June 1989)

The Editorial Board here introduces a new kind of scientific report in the Journal, whereby a current field of research and debate is given emphasis, being the subject of an open discussion within these columns.

As a first essay, we propose a discussion about a difficult and somehow trouble some question in applied animal genetics: how to take proper account of the observed data being *selected* data? Several attempts have been carried out in the past 15 years, without any clear and unanimous solution. In the following, Im, Fernando and Gianola propose a general approach that should make it possible to deal with every problem. In addition to the interest of an original article, we hope that their own discussion and response to the comments given by Henderson and Thompson will provide the reader with a sound insight into this complex topic.

This paper is dedicated to the memory of Professor Henderson, who gave us here one of his latest contributions.

The Editorial Board

Summary – Data available in animal breeding are often subject to selection. Such data can be viewed as data with missing values. In this paper, inferences based on likelihoods derived from statistical models for missing data are applied to production records subject to selection. Conditions for ignoring the selection process are discussed.

animal genetics – selected data – missing data – likelihood inference

Résumé – Les méthodes d'inférence fondées sur la vraisemblance en génétique animale: prise en compte de données issues de la sélection au moyen de la théorie des données manquantes. *Les données disponibles en génétique animale sont souvent issues d'un processus préalable de sélection. On peut donc considérer comme manquants les attributs (non observés) associés aux individus éliminés, et analyser les données recueillies comme provenant d'un échantillon avec données manquantes. Dans cet article,*

on développe les méthodes d'inférence fondées sur les vraisemblances, en explicitant dans leur calcul le processus, dû à la sélection, qui induit les données manquantes. On discute les conditions dans lesquelles on peut ignorer la sélection, et donc considérer seulement la vraisemblance des données effectivement recueillies.

génétique animale – sélection – données manquantes – vraisemblance

INTRODUCTION

Data available in animal breeding often come from populations undergoing selection. Several authors have considered methods for the proper treatment of data subject to selection in animal breeding. Examples are Henderson *et al.* (1959), Curnow (1961), Thompson (1973), Henderson (1975), Rothshild *et al.* (1979), Goffinet (1983), Meyer and Thompson (1984), Fernando and Gianola (1989), and Schaeffer (1987).

Data subject to selection can be viewed as data with missing values, selection being the process that causes missing data. The statistical literature discusses missing data that arise intentionally. Rubin (1976) has given a mathematically precise treatment which encompasses frequentist approaches that are not based on likelihoods as well as inferences from likelihoods (including maximum likelihood and Bayesien approaches). Whether it is appropriate to ignore the process that causes the missing data depends on the method of inference and on the process that causes the missing values. Rubin (1976) suggested that in many practical problems, inferences based on likelihoods are less sensitive than sampling distribution inferences to the process that causes data. Goffinet (1987) gave alternative conditions to those of Rubin (1976) for ignoring the process that causes missing-data when making sampling distribution inferences, with an application to animal breeding.

The objective of this paper is to consider inferences based on likelihoods derived from statistical models for the data and the missing-data process, in analysis of data from populations undergoing selection. As in Little and Rubin (1987), we consider inferences based on likelihoods, in the sense described above, because of their flexibility and avoidance of *ad-hoc* methods. Assumptions underlying the resulting methods can be displayed and evaluated, and large sample estimates of variances based on second derivatives of the log-likelihood taking into account the missing data process, can be obtained.

MODELING THE MISSING-DATA PROCESS

Ideas described by Little and Rubin (1987) are employed in subsequent developments. Let $\boldsymbol{y}$, the realized value of a random vector $\boldsymbol{Y}$, denote the data that would occur in the absence of missing values, or complete data. The vector $\boldsymbol{y}$ is partitioned into observed values, $\boldsymbol{y}_{\text{obs}}$, and missing values, $\boldsymbol{y}_{\text{mis}}$. Let

$$f(\boldsymbol{y}|\boldsymbol{\theta}) \equiv f(\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}|\boldsymbol{\theta}) \tag{1}$$

be the probability density function of the joint distribution of $\boldsymbol{Y} = (\boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{mis}})$, and $\boldsymbol{\theta}$ be an unknown parameter vector. We define for each component of $\boldsymbol{Y}$ an indicator variable, R_i (with realized value r_i), taking the value 1 if the component is observed and 0 if it is missing. In order to illustrate the notation, 3 types of

missing data are described in table I. Consider 2 correlated traits measured on n unrelated individuals; for example, first and second lactation yields of n cows. The 'complete' data are $\boldsymbol{y} = (y_{ij})$, where y_{ij} is the realized value of trait j in individual i ($j = 1,2$; $i = 1 \ldots n$). Suppose that selection acts on the first trait (case (a) in Table I). As a result, a subset of $\boldsymbol{y}$, $\boldsymbol{y}_{\text{obs}}$, becomes available for analysis. The pattern of the available data is a random variable. For example, if the better of two cows ($n = 2$) is selected to have a second lactation, the complete data would be

$$\boldsymbol{y} = (y_{11}, y_{21}, y_{12}, y_{22})$$

Then when $y_{11} > y_{21}$:

$$\boldsymbol{y}_{\text{obs}} = y_{11}, y_{21}, y_{12}), \ \boldsymbol{y}_{\text{mis}} = (y_{22}), \ \boldsymbol{r} = (1,1,1,0)$$

and when $y_{11} \leq y_{21}$:

$$\boldsymbol{y}_{\text{obs}} = (y_{11}, y_{21}, y_{22}), \ \boldsymbol{y}_{\text{mis}} = (y_{12}), \ \boldsymbol{r} = (1,1,0,1)$$

Thus, in analysis of selected data, the pattern of records available for analysis, characterized by the value of $\boldsymbol{r}$, should be considered as part of the data. If this is not done, there will be a loss of information.

To treat $R = (R_i)$ as a random variable, we need to specify the conditional probability that $R = \boldsymbol{r}$, $f(\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{\psi})$, given the 'complete' data $\boldsymbol{Y} = \boldsymbol{y}$; the vector $\boldsymbol{\psi}$

Table I. An example of 3 types of missing-data

(a)		(b)		(c)	
y_{11}	y_{12}	y_{11}	y_{12}	*	y_{12}
y_{21}	y_{22}	y_{21}	y_{22}	*	y_{22}
⋮	⋮	⋮	⋮	⋮	⋮
y_{m1}	y_{m2}	y_{m1}	y_{m2}	*	y_{m2}
$y_{(m+1)1}$	*	*	*	*	*
⋮	⋮	⋮	⋮	⋮	⋮
y_{n1}	*	*	*	*	*

* Stands for missing value

(a) $\boldsymbol{y}_{\text{obs}} = (y_{11}, y_{21} \cdots y_{n1}, y_{12} \cdots y_{m2})$
$\boldsymbol{y}_{\text{mis}} = (y_{(m+1)2} \cdots y_{n2})$
$r_{i1} = 1$ for $i = 1 \ldots n$
$r_{i2} = 1$ for $i = 1 \ldots m$ and 0 for $i = m+1 \ldots n$.

(b) $\boldsymbol{y}_{\text{obs}} = (y_{11} \cdots y_{m1}, y_{12} \cdots y_{m2})$
$\boldsymbol{y}_{\text{mis}} = (y_{(m+1)1} \cdots y_{n1}, y_{(m+1)2} \cdots y_{n2})$
$r_{ij} = 1$ for $i = 1 \ldots m$; $j = 1,2$
$r_{ij} = 0$ for $i = m+1 \ldots n$; $j = 1,2$.

(c) $\boldsymbol{y}_{\text{obs}} = (y_{12} \cdots y_{m2})$
$\boldsymbol{y}_{\text{mis}} = (y_{11} \cdots y_{n1}, y_{(m+1)2} \cdots y_{n2})$
$r_{ij} = 1$ for $i = 1 \ldots m$ and $j = 2$
$= 0$ otherwise.

is a parameter of this conditional distribution. The density of the joint distribution of $\boldsymbol{Y}$ and $\boldsymbol{R}$ is

$$f(\boldsymbol{y}, \boldsymbol{r}|\boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}|\boldsymbol{\theta}) \cdot f(\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{\psi}) \tag{2}$$

The likelihood ignoring the missing-data process, or marginal density of $\boldsymbol{y}_{\text{obs}}$ in the absence of selection, is obtained by integrating out the missing data $\boldsymbol{y}_{\text{mis}}$ from (equ.(1))

$$f(\boldsymbol{y}_{\text{obs}}|\boldsymbol{\theta}) = \int f(\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}|\boldsymbol{\theta}) \mathrm{d}\boldsymbol{y}_{\text{mis}} \tag{3}$$

The problem with using $f(\boldsymbol{y}_{\text{obs}}|\boldsymbol{\theta})$ as a basis for inferences is that it does not take into account the selection process. The information about $\boldsymbol{R}$, a random variable whose value $\boldsymbol{r}$ is also observed, is ignored. The actual likelihood is

$$f(\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}|\boldsymbol{\theta}, \boldsymbol{\psi}) = \int f(\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}|\boldsymbol{\theta}) \cdot f(\boldsymbol{r}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}, \boldsymbol{\psi}) \mathrm{d}\boldsymbol{y}_{\text{mis}} \tag{4}$$

The question now arises as to when inferences on $\boldsymbol{\theta}$ should be based on the joint likelihood (equ.(4)), and when can it based on equ.(3), which ignores the missing data process. Rubin (1976) has studied conditions under which inferences from equ.(3) are equivalent to those obtained from equ.(4). If these hold, one can say that the missing data process can be ignored. The conditions given by Rubin (1976) are: 1) the missing data are missing at random, *ie*, $f(\boldsymbol{r}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}, \boldsymbol{\psi}) = f(\boldsymbol{r}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{\psi})$ for all $\boldsymbol{\psi}$ and $\boldsymbol{y}_{\text{mis}}$ evaluated at the observed values $\boldsymbol{r}$ and $\boldsymbol{y}_{\text{obs}}$; and 2) the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are distinct, in the sense that the joint parameter space of $(\boldsymbol{\theta}, \boldsymbol{\psi})$ is the product of the parameter space of $\boldsymbol{\theta}$ and the parameter space of $\boldsymbol{\psi}$. Within the contexte of Bayesian inference, the missing data process is ignorable when 1) the missing data are missing at random, and 2) the prior density of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ is the product of the marginal prior density of $\boldsymbol{\theta}$ and the marginal prior density of $\boldsymbol{\psi}$.

IGNORABLE OR NON-IGNORABLE SELECTION

Without loss of generality, we examine ignorability of selection when making likelihood inferences about $\boldsymbol{\theta}$ for each of the three examples given in Table I. Suppose individuals $1, 2 \ldots m$ $(< n)$ are selected.

Cases (a)

Selection based on observations on the first trait, which are a part of the observed data and all the data used to make selection decisions are available. The likelihood for the observed data, ignoring selection, is

$$f(\boldsymbol{y}_{\text{obs}}|\boldsymbol{\theta}) = \Big[\prod_{i=1}^{m} f(y_{i1}, y_{i2}|\boldsymbol{\theta})\Big]\Big[\prod_{i=m+1}^{n} f(y_{i1}|\boldsymbol{\theta})\Big] \tag{5}$$

$$= \Big[\prod_{i=1}^{n} f(y_{i1}|\boldsymbol{\theta})\Big]\Big[\prod_{i=1}^{m} f(y_{i2}|y_{i1}, \boldsymbol{\theta})\Big] \tag{6}$$

Because selection is based on the observed data only, the conditional probability $f(\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{\psi}) = f(\boldsymbol{r}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{\psi})$ because it does not depend on the missing data. Applying this condition in equ.(4) one obtains as likelihood function

$$f(\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}|\boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{r}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{\psi}) \cdot f(\boldsymbol{y}_{\text{obs}}|\boldsymbol{\theta}) \quad (7)$$

It follows that maximization of equ.(7) with respect to $\boldsymbol{\theta}$ will give the same estimates of this parameter as maximization of equ.(6). Thus, knowledge of the selection process is not required, *i.e.*, selection is ignorable. Note that with or without normality, $f(\boldsymbol{y}_{\text{obs}}|\boldsymbol{\theta})$ can always be written as equ.(5) or (6). Under normality of the joint distribution of Y_{i1} and Y_{i2}, Kempthorne and Von Krosigk (Henderson *et al.*, 1959) and Curnow (1961) expressed the likelihood as equ.(6). These authors, however, did not justify clearly why the missing data process could be ignored.

In order to illustrate the meaning of the parameter $\boldsymbol{\psi}$ of the conditional probability of $\boldsymbol{R} = \boldsymbol{r}$ given $\boldsymbol{Y} = \boldsymbol{y}$, we consider a 'stochastic' form of selection: individual i is selected with probability $g(\psi_0 + \psi_1 y_{i1})$, so $\boldsymbol{\psi} = (\psi_0, \psi_1)$. This type of selection can be regarded as selection based on survival, which depends on the first trait *via* the function $g(\psi_0 + \psi_1 y_{i1})$. We have for the data in Table I

$$Pr(R_{i1} = 1|\boldsymbol{y}) = 1 \quad \text{for } i = 1 \ldots n$$

$$Pr(R_{i2} = 1|\boldsymbol{y}\boldsymbol{\psi}) = g(\psi_0 + \psi_1 y_{i1}) \quad \text{for } i = 1 \ldots n$$

The actual likelihood for the observed data $\boldsymbol{y}_{\text{obs}}$ and $\boldsymbol{r}$ is

$$f(\boldsymbol{y}_{\text{obs}}, r|\boldsymbol{\theta}, \boldsymbol{\psi}) = \prod_{i=1}^{m} [f(y_{i1}, y_{i2}|\boldsymbol{\theta}) g(\psi_0 + \psi_1 y_{i1})] \cdot \prod_{i=m+1}^{n} \left\{ \int f(y_{i1}, y_{i2}|\boldsymbol{\theta})[1 - g(\psi_0 + \psi_1 y_{i1})] \mathrm{d}y_{i2} \right\}$$

$$= \left\{ \prod_{i=1}^{m} g(\psi_0 + \psi_1 y_{i1}) \right\} \left\{ \prod_{i=m+1}^{n} [1 - g(\psi_0 + \psi_1 y_{i1})] \right\} f(\boldsymbol{Y}_{\text{obs}}|\boldsymbol{\theta}) \quad (8)$$

It follows that when $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$ are distinct, inference about $\boldsymbol{\theta}$ based on the actual likelihood, $f(\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}|\boldsymbol{\theta}, \boldsymbol{\psi})$, will be equivalent to that based on the likelihood ignoring selection, $f(\boldsymbol{y}_{\text{obs}}|\boldsymbol{\theta})$. As shown in equ.(8), the two likelihoods differ by a multiplicative constant which does not depend on $\boldsymbol{\theta}$.

It should be noted that in general, although the conditional distribution of R_{i2} given $\boldsymbol{y}$ does not depend on $\boldsymbol{\theta}$, this is not with the marginal distribution. For example, when Y_{i1} is normal with mean μ_1 and variance σ_1^2, and g is the standard normal function (Φ) we have

$$Pr(R_{i2} = 1|\boldsymbol{\theta}, \boldsymbol{\psi}) = \Phi[(\psi_0 + \psi_1 \mu_1)/(1 + \psi_1^2 \sigma_1^2)^{1/2}]$$

The condition (b) in Goffinet (1987) for ignoring the process that causes missing data is not satisfied in this situation.

Cases (b)

Data are available only in selected individuals because observations are missing in the unselected ones. In what follows, we will consider truncation selection: individual i is selected when $y_{i1} > t$, where t is a known threshold.

The likelihood of the observed data ($\boldsymbol{y}_{\text{obs}}$) ignoring selection is

$$f(\boldsymbol{y}_{\text{obs}}|\boldsymbol{\theta}) = \prod_{i=1}^{m} f(y_{i1}, y_{i2}|\boldsymbol{\theta}) \tag{9}$$

The conditional probability that $\boldsymbol{R} = \boldsymbol{r}$ given $\boldsymbol{Y} = \boldsymbol{y}$ depends on the observed and on the missing data. We have

$$Pr = (R_{ij} = 1|\boldsymbol{y}) = 1_{(t,\infty)}(y_{i1}) \text{ for } j = 1,2;\ i = 1 \ldots n$$

where $1_{(t,\infty)}(y_{i1}) = 1$ if $y_{i1} > t$, and 0 if $y_{i1} \leq t$.

The actual likelihood, accounting for selection, is

$$f(\boldsymbol{y}_{\text{obs}}, r|\boldsymbol{\theta}) = \Big[\prod_{i=1}^{m} f(y_{i1}, y_{i2}|\boldsymbol{\theta})\Big]\Big[\prod_{i=m+1}^{n} \int\int f(y_{i1}, y_{i2}|\boldsymbol{\theta}) 1_{(-\infty,t)}(y_{i1}) \mathrm{d}y_{i1} \mathrm{d}y_{i2}\Big]$$

$$= f(\boldsymbol{y}_{\text{obs}}|\boldsymbol{\theta})\Big[\prod_{i=m+1}^{n} Pr(Y_{i1} < t|\boldsymbol{\theta})\Big] \tag{10}$$

where $1_{(-\infty,t)}(y_{i1}) = 1$ if $y_{i1} \leq t$ and 0 if $y_{i1} > t$.

Comparison of equs.(9) and (10) indicates that one should make inferences about $\boldsymbol{\theta}$ using equ.(10), which takes selection into account. If equ.(9), is used, the information about $\boldsymbol{\theta}$ contained in the second term in equ.(10) would be neglected. Clearly selection is not ignorable in this situation.

Cases (c)

Often selection is based on an unknown trait correlated with the trait for which data are available (Thompson, 1979). As in case (c) in Table I, suppose the data are available for the second trait on selected individuals only, following selection, *e.g.* by truncation, on the first trait. The likelihood ignoring selection is

$$f(\boldsymbol{y}_{\text{obs}}|\boldsymbol{\theta}) = \prod_{i=1}^{m} f(y_{i2}|\boldsymbol{\theta}) \tag{11}$$

We have

$$Pr(R_{i1} = 0|\boldsymbol{y}) = 1 \text{ for } i = 1 \ldots n$$

$$Pr(R_{i2} = 1|\boldsymbol{y}) = 1_{(t,\infty)}(y_{i1}) \text{ for } i = 1 \ldots n$$

The likelihood of the observed data, $\boldsymbol{y}_{\text{obs}}$ and $\boldsymbol{r}$ is

$$f(\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}|\boldsymbol{\theta}) = \Big[\prod_{i=1}^{m} \int f(y_{11}, y_{i2}|\boldsymbol{\theta}) 1_{(t,\infty)}(y_{i1}) \mathrm{d}y_{i1}\Big]$$

$$\cdot \prod_{i=m+1}^{n} \int\int f(y_{11}, y_{i2}|\boldsymbol{\theta}) 1_{(-\infty,t)}(y_{i1}) \mathrm{d}y_{i1} \mathrm{d}y_{i2}$$

$$= \Big[\prod_{i=1}^{m} f(y_{i2}|\boldsymbol{\theta})\Big]\Big[\prod_{i=1}^{m} Pr(Y_{i1} > t|y_{i2}, \boldsymbol{\theta})\Big]\Big[\prod_{i=m+1}^{n} Pr(Y_{i1} < t|\boldsymbol{\theta})\Big] \tag{12}$$

Inferences based on the likelihood (equ.(11)) would be affected by a loss of information represented by the second and the third terms in equ.(12).

Under certain conditions one could use $f(\boldsymbol{y}_{obs}|\boldsymbol{\theta})$ to make inferences about parameters of the marginal distribution of the second trait after selection. Suppose the marginal distribution of the second trait depends only on parameters $\boldsymbol{\theta}_2$, and that the marginal and conditional (given the second trait) distributions of the first trait do not depend on $\boldsymbol{\theta}_2$. In this case, likelihood inferences on $\boldsymbol{\theta}_2$ from equs.(11) and (12) will be the same.

In summary, the results obtained for the 3 cases discussed indicate that when selection is based only on the observed data it is ignorable, and knowledge of the selection process is not required for making correct inferences about parameters of the data. When the selection process depends on observed and also on missing data, selection is generally not ignorable. Here, making correct inferences about parameters of the data requires knowledge of the selection process to appropriately construct the likelihood.

A GENERAL TYPE OF SELECTION

Selection based on data

In this section, we consider the more general type of selection described by Goffinet (1983) and Fernando and Gianola (1987). The data $\boldsymbol{y}_0$ are observed in a 'base population' and used to make selection decisions which lead to observe a set of data, $\boldsymbol{y}_{1obs}$, among n_1 possible sets of values $\boldsymbol{y}_{11}, \boldsymbol{y}_{12} \cdots \boldsymbol{y}_{1n_1}$. Each $\boldsymbol{y}_{1k}(k = 1 \ldots n_1)$ is a vector of measurements corresponding to a selection decision. The observed data at the first stage, $\boldsymbol{y}_{1obs}$, are themselves used (jointly with $\boldsymbol{y}_0$) to make selection decisions at a second stage, and so forth. At stage j $(j = 1 \ldots J)$, let $\boldsymbol{y}_j$ be the vector of all elements from $\boldsymbol{y}_{j1} \ldots \boldsymbol{y}_{jn_j}$, without duplication. The vector $\boldsymbol{y}_j$ can be partitioned as

$$\boldsymbol{y}_j = (\boldsymbol{y}_{jobs}, \boldsymbol{y}_{jmis})$$

where $\boldsymbol{y}_{jobs}$ and $\boldsymbol{y}_{jmis}$ are the observed and the missing data, respectively. For the J stages, the data

$$\boldsymbol{y} = (\boldsymbol{y}_0, \boldsymbol{y}_1, \boldsymbol{y}_2 \cdots \boldsymbol{y}_j \cdots \boldsymbol{y}_J)$$

can be partitioned as $\boldsymbol{y} = (\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis})$, where

$$\boldsymbol{y}_{obs} = (\boldsymbol{y}_0, \boldsymbol{y}_{1obs} \cdots \boldsymbol{y}_{Jobs})$$

and

$$\boldsymbol{y}_{mis} = (\boldsymbol{y}_{1mis}, \cdots \boldsymbol{y}_{Jmis})$$

are the observed and missing parts, respectively, of the complete data set. The complete data set $\boldsymbol{y}$ is a realized value of a random variable $\boldsymbol{Y}$.

When the selection process is based only on the observed data, $\boldsymbol{y}_{obs}$, the observed missing data pattern, $\boldsymbol{r}$ is entirely determined by $\boldsymbol{y}_{obs}$. Thus,

$$f(\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{\psi}) = f(\boldsymbol{r}|\boldsymbol{y}_{obs}, \boldsymbol{\psi})$$

and the actual likelihood can be written as in equ.(7). In this case, the selection process is ignorable and inferences about $\boldsymbol{\theta}$ can be based on the likelihood of the observed data, $f(\boldsymbol{y}_{obs}|\boldsymbol{\theta})$. This agrees Gianola and Fernando (1986) and Fernando and Gianola (1989).

Selection based on data plus 'externalities'

Suppose that external variables, represented by a random vector $\boldsymbol{E}$, and the observed data $\boldsymbol{y}_{obs}$ are jointly used to make selection decisions. Let $f(\boldsymbol{y}, \boldsymbol{e}|\boldsymbol{\theta}, \boldsymbol{\xi})$ be the joint density of the complete data $\boldsymbol{Y}$ and $\boldsymbol{E}$, with an additional parameter $\boldsymbol{\xi}$ such that $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ are distinct. The actual likelihood, density of the joint distribution of $\boldsymbol{Y}_{obs}$ and $\boldsymbol{R}$, is

$$f(\boldsymbol{y}_{obs}, \boldsymbol{r}|\boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\psi}) = \int\int f(\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}, \boldsymbol{e}, \boldsymbol{\xi}) \cdot f(\boldsymbol{r}|\boldsymbol{y}_{obs}, \boldsymbol{e}, \boldsymbol{\psi}) \mathrm{d}\boldsymbol{y}_{mis} \mathrm{d}\boldsymbol{e} \tag{13}$$

where $f(\boldsymbol{r}|\boldsymbol{y}_{obs}, \boldsymbol{e}, \boldsymbol{\psi})$ is the distribution of the missing data process (selection process).

In general, inferences about $\boldsymbol{\theta}$ based on $f(\boldsymbol{y}_{obs}, \boldsymbol{r}|\boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\psi})$ are not equivalent to those based on $f(\boldsymbol{y}_{obs}|\boldsymbol{\theta})$. However, if for the observed data, $\boldsymbol{y}_{obs}$

$$f(\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}, \boldsymbol{e}|\boldsymbol{\theta}, \boldsymbol{\xi}) = f(\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}|\boldsymbol{\theta}) \cdot f(\boldsymbol{e}|\boldsymbol{\xi})$$

for all $\boldsymbol{y}_{mis}$ and $\boldsymbol{e}$, then equ.(13) can be written as

$$\begin{aligned} f(\boldsymbol{y}_{obs}, \boldsymbol{r}|\boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\psi}) &= \int\int f(\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}|\boldsymbol{\theta}) \cdot f(\boldsymbol{e}|\boldsymbol{\xi}) \cdot f(\boldsymbol{r}|\boldsymbol{y}_{obs}, \boldsymbol{e}, \boldsymbol{\psi}) \mathrm{d}\boldsymbol{y}_{mis} \mathrm{d}\boldsymbol{e} \\ &= \int f(\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}|\boldsymbol{\theta}) \mathrm{d}\boldsymbol{y}_{mis} \cdot \int f(\boldsymbol{e}|\boldsymbol{\xi}) \cdot f(\boldsymbol{r}|\boldsymbol{y}_{obs}, \boldsymbol{e}, \boldsymbol{\psi}) \mathrm{d}\boldsymbol{e} \\ &= f(\boldsymbol{y}_{obs}|\boldsymbol{\theta}) \cdot \int f(\boldsymbol{e}|\boldsymbol{\xi}) \cdot f(\boldsymbol{r}|\boldsymbol{y}_{obs}, \boldsymbol{e}, \boldsymbol{\psi}) \mathrm{d}\boldsymbol{e} \end{aligned} \tag{14}$$

Thus, under the above condition, which is satisfied when $\boldsymbol{Y}$ and $\boldsymbol{E}$ are independent, inferences about $\boldsymbol{\theta}$ based on the actual likelihood $f(\boldsymbol{y}_{obs}, \boldsymbol{r}|\boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\psi})$ and those based on $f(\boldsymbol{y}_{obs}|\boldsymbol{\theta})$ are equivalent. Consequently, the selection process is ignorable. Note that the condition

$$f(\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}, \boldsymbol{e}|\boldsymbol{\theta}, \boldsymbol{\xi}) = f(\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}|\boldsymbol{\theta}) \cdot f(\boldsymbol{e}|\boldsymbol{\xi})$$

for all $\boldsymbol{y}_{mis}$ and $\boldsymbol{e}$ does not require independence between $\boldsymbol{Y}$ and $\boldsymbol{E}$ because it holds only for the observed data $\boldsymbol{y}_{obs}$ and not for all values of the random variable $\boldsymbol{Y}_{obs}$.

The results can be summarized as follows: 1) the selection process is ignorable when it is only on the observed data, or on observed data and independent externalities; 2) the selection process is not ignorable when it is based on the observed data plus dependent externalities. In the latter case, knowledge of the selection process is required for making correct inferences.

DISCUSSION

Maximum likelihood (ML) is a widely used estimation procedure in animal breeding applications and has been suggested as the method of choice (Thompson, 1973) when selection occurs. Simulation studies (Rothschild *et al.*, 1979, Meyer and Thompson, 1984) have indicated that there is essentially no bias in ML estimates of variance and covariance components under forms of selection, *e.g.*, data-based selection.

Rubin's (1976) results for analysis of missing data provide a powerful tool for making inferences about parameters when data are subject to selection. We have considered ignorability of the selection process when making inferences based on likelihood and given conditions for ignoring it. The conditions differ from those given by Henderson (1975) for estimation of fixed effects and prediction of breeding value under selection in a multivariate normal model. For example, Henderson (1975)

requires that selection be carried out on a linear, translation invariant function. This requirement does not appear in our treatment because we argue from a likelihood viewpoint.

In this paper, the likelihood was defined as the density of the joint distribution of the observed data pattern. In Henderson's (1975) treatment of prediction, the pattern of missing data is fixed, rather than random, and this results in a loss of information about parameters (Cox and Hinkley, 1974). It is possible to use the conditional distribution of the observed data given the missing data pattern. Gianola *et al.* (submitted) studied this problem from a conditional likelihood viewpoint and found conditions for ignorability of selection even more restrictive that those of Henderson (1975). Schaeffer (1987) arrived to similar conclusions, but this author worked with quadratic forms, rather than with likelihood. The fact that these quadratic forms appear in an algorithm to maximize likelihood is not sufficient to guarantee that the conditions apply to the method *per se*.

If the conditions for ignorability of selection discussed in this study are met, the consequence is that the likelihood to be maximized is that of the observed data, *i.e.*, the missing data process can be completely ignored. Further, if selection is ignorable $f(\mathbf{y}_{obs}, \mathbf{r}, |\boldsymbol{\theta}) \propto f(\mathbf{y}_{obs}|\boldsymbol{\theta})$, so

$$\frac{\partial^2 \log f(\mathbf{y}_{obs}, \mathbf{r}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'} = \frac{\partial^2 \log f(\mathbf{y}_{obs}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}$$

Efron and Hinkley (1978) suggested using observed rather than expected information to obtain the asymptotic variance-covariance matrix of the maximum likelihood estimates. Because the observed data are generally not independent or identically distributed, simple results that imply asymptotic normality of the maximum likelihood estimates do not immediately apply. For further discussion see Rubin (1976).

We have emphasized likelihoods and little has been said on Bayesian inference. It is worth noticing that likelihoods constitute the 'main' part of posterior distributions, which are the basis of Bayesian inference. The results also hold for Bayesian inference provided the parameters are distinct, *i.e.*, their prior distributions are independent. For data-based selection, our results agree with those of Gianola and Fernando (1986) and Fernando and Gianola (1989) who used Bayesian arguments. In general, inferences based on likelihoods or posterior distributions have been found more attractive by animal breeders working with data subject to selection than those based on other methods. This choice is confirmed and strengthened by application of Rubin's (1976) results to this type of problem.

REFERENCES

Cox D.R. & Hinkley D.V. (1974) *Theoretical Statistics*. Chapman and Hall, London

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

Efron B. & Hinkley D.V. (1978) Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. *Biometrika* 65, 457-482

Fernando R.L. & Gianola D. (1989) Statistical inferences in populations undergoing selection and non-random mating. *In: Advances in Statistical Methods for Genetic Improvement of Livestock* Springer-Verlag, in press

Gianola D. & Fernando R.L. (1986) Bayesian methods in animal breeding theory. *J. Anim. Sci.* 63, 217-244

Gianola D., Im S. Fernando R.L. & Foulley J.L. (1989) Maximum likelihood estimation of genetic parameters under a "Pearsonian" selection model. *J. Dairy Sci.* (submitted)

Goffinet B. (1983) Selection on selected records. *Genet. Sel. Evol.* 15, 91-98

Goffinet B. (1987) Alternative conditions for ignoring the process that causes missing data. *Biometrika* 71, 437-439

Henderson C.R. (1975) Best linear unbiased estimation and prediction under a selection model. *Biometrics* 31, 423-439

Henderson C.R., Kempthorne O., Searle S.R. & Von Krosigk C.M. (1959) The estimation of environmental and genetic trends from records subject to culling. *Biometrics* 15, 192-218

Little R.J.A. & Rubin D.B. (1987) *Statistical Analysis with Missing Data* Wiley, New York

Meyer K. & Thompson R. (1984) Bias in variance and covariance component estimators due to selection on a correlated trait. *Z. Tierz. Zuchtungsbiol.* 101, 33-50

Rothschild M.F., Henderson C.R. & Quaas R.L. (1979) Effects of selection on variances and covariances of simulated first and second lactations. *J. Dairy Sci.* 62, 996-1002

Rubin D.B. (1976) Inference and missing data. *Biometrika* 63, 581-592

Schaeffer L.R. (1987) Estimation of variance components under a selection model. *J. Dairy Sci.* 70, 661-671

Thompson R. (1973) The estimation of variance and covariance components when records are subject to culling. *Biometrics* 29, 527-550

Thompson R. (1979) Sire evaluation. *Biometrics* 35, 339-353

COMMENT

C.R. Henderson †*

The paper by Im, Fernando and Gianola provides an interesting and invaluable contribution to estimation and prediction in an almost universal situation in animal breeding. Very few data are available for parameter estimation or prediction of breeding values that have not arisen from either selection experiments or from field data in herds that have undergone selection.

For several years after the adoption of BLUP, a mixed linear model was assumed, and the usual description of the model was that $E(Y)$ and $E(e)$ are both null, and in an additive genetic model $\text{Var}(U) = A\sigma_a^2$. The assumption of $E(U) = 0$ is clearly untenable, because if selection has been effective, the expectations of a subvectors for successive generations are increasing.

A serious attempt to model for selection was made in my 1975 Biometrics paper cited by Im *et al.* It must be emphasized, as has been done in the paper under review, that my model is different from the model of the present paper. Consequently, the solution to prediction and estimation differs. I do not disagree with the authors' conclusions from their model, and I think, based on long discussions with Gianola and Fernando, that they do not disagree with my conclusions based on my model. The really critical question is, "What is the best model for describing selection?" I have no intention of addressing this issue because I neither have strong convictions about my model nor about any others.

Our models differ in that mine is considerably more restrictive, requiring as it does, a fixed incidence matrix with conceptual repeated sampling. This of course is the traditional approach taken by classical statisticians. The problem is more difficult, however, with selection problems, as compared to nicely designed experimental situations. No attempt was made in the 1975 paper to solve the problem of estimation of variances and covariances. Rather, I solved the problem of BLUE of estimable functions of β and BLUP of random variables, given multivariate normality and with variances and covariances known to proportionality. I pointed out that, in contrast to no selection models, the estimators and predictors are biased if incorrect ratios are employed. Thus, it is critical to obtain the best possible of these parameters. Im *et al.* address this problem.

Several workers have speculated that REML applied to a selection model estimates the variances and covariances that existed prior to selection and which may have been altered by selection. In contrast to most of these speculations, I suggested that when selection is on observed records, the linear selection functions should be translation invariant. I think this is true under my selection model but may well not be true for other selection models.

Im *et al.* strongly emphasize the desirability of likelihood methods. I agree with them, and in many meetings and papers have recommended these methods over some of my own, such as Method 3. I doubt the accuracy of the last sentence of the paper under review which states that animal breeders find likelihood methods more attractive. A study of animal breeding literature of the past 5 years would probably disclose that animal breeders have used Method 3 much more often than

* Formerly of the Department of Animal Science, Cornell University Ithaca, NY, USA.

REML or ML. If this is true, I certainly agree with minority and with Im *et al.* The fact is that BLUE and BLUP under my selection model are ML estimators of β and of the conditional mean of U.

I should now like to discuss how results compare under my selection model and under the model of the present authors. We agree partially regarding estimation when selection is on observable records. The authors' model clearly shows ignorability of selection in this case. Under my model, linear selection functions of Y must either be translation invariant or it must be true that $E(L'Y_s) = E(L'Y_u)$ when Y_s and Y_u refer to selection and to no selection, respectively. This difference is simply a consequence of different models.

We agree that if selection is on unobserved random variables, selection is not ignorable. A special case of this has been of interest to me. Base population animals have been selected on translation invariant linear functions of data, but these are not available for analysis. Assuming that such selection results in $E(U_b) \neq 0$ a simple modification of the regular mixed model equations leads to BLUE and BLUP, and presumably these modified equations could be used to derive REML estimation of the variances and covariances, Henderson (1988).

I believe that this final question is justified, namely, "What are the operating characteristics of the authors' estimators?" Likelihood methods for variance estimation have known desirable properties only in large samples. We need studies for various methods of bias, MSE, and maximization of selection progress using BLUP with estimated variances and covariances. Probably this can be done only through extensive simulation for a wide range of parameter values, selection intensity, etc.

The authors have made a valuable contribution to the problem of estimation in selection models. This paper should motivate further studies on this problem.

ADDITIONAL REFERENCE

Henderson C.R. (1988). A simple method to account for selected base populations. *J. Dairy Sci* 71, 3399-3404

COMMENT

R. Thompson *

This paper considers ways of constructing likelihoods for selected data, specifically taking account of the presence and absence of data following a method developed by Rubin (1976) and explained in the recent book by Little and Rubin (1987). Some of the likelihoods have been given previously without any formal recourse to ideas of missingness, for example extensions of case (a) Henderson *et al.* (1959), Curnow (1961) and Thompson (1973). Theses authors used a sequential approach to build up likelihoods that I find appealing. Using this approach it is easy to see that r is a function of y and so does not contribute any extra information on θ. To derive the same likelihood by differing routes is reassuring.

* AFRC Institute of Animal Physiology and Genetics Research, Edinburgh, UK.

It is valuable to know when selection is ignorable. I have always found it confusing that in extensions of case (a) a likelihood approach would say that selection on y_{1i}, is always ignorable but Henderson (1975) suggests that selection is only ignorable if selection is on a culling variate (w) that is translation invariant.

In an interesting paper the same 3 authors (Gianola *et al.*, 1988) have constructed the joint density of the data and random effects conditional on the culling variate (D_c). Inferences based on D_c suggest that selection can be ignored only if it is based on functions of the data that do not depend on the fixed effects. It would have been instructive to relate D_c to terms used in the present paper, as presumably r can be related to the culling variate and might help to answer 3 comments I have on the use of D_c.

First, Gianola *et al.* (1988) condition on w, the culling variate, by integrating over y and the random effects. I am not sure of the need to integrate over the random effects. One might sometimes want to consider repeated samples over (or conditioning on) all possible genetic material and only repeated over the same genetic material. Henderson (1988) has recently suggested a procedure that involves no integration over y of the random effects, *i.e.* conditioning on the observed value of w. What should one do?

Secondly, Gianola *et al.* (1988) highlighted differences between using D_c and Henderson's (1975) approach when selection is on random effects or residuals ($w = L'u$ or $L'e$). This case is artificial in the sense that random effects and residuals will never be known exactly. But if selection is on known random effects it scarcely seems necessary to predict them using only the data. It might be more interesting to compare the 2 predictions. Similarly, if $w = L'e$ is known, this known value could improve estimation and prediction of the other parameters.

Thirdly, if selection is on $w = L'y$, but is not translation invariant, presumably the authors technique, which is non-linear, should be more efficient than Henderson's approach. I wonder if the authors have quantitative information on this.

Finally there are cases when one wants to estimate parameters associated with equ.(4) and, (Robertson (1966)). There is discussion of this area in Little and Rubin (1987) and techniques developed by Foulley, Gianola and Thompson (1983) for quantitative and binary traits can sometimes be used.

ADDITIONAL REFERENCES

Foulley J.L., Gianola D. & Thompson R. (1983) Prediction of genetic merit from data on binary and quantitative variates with application to calving difficulty, birth weight and pelvic opening. *Génét. Sél. Évol.* 15, 401-424

Gianola D., Im S. & Fernando R.L. (1988) Prediction of breeding value under Henderson's selection model: a revisitation. *J. Dairy Sci.* 71, 2790-2798

Henderson C.R. (1988) Simple method to compute biases and mean squared error of linear estimators and predictors in a selection model assuming normality. *J. Dairy Sci.* 71, 3135-3142

Robertson A. (1966) A mathematical model of the culling process in dairy cattle. *Anim. Prod.* 8, 95-108

REJOINDER

S. Im, R.L. Fernando and D. Gianola

We thank the participants in this discussion and, in particular, Professor Henderson, whose comments were received by us a few weeks before his unexpected death, for their contributions to the theory of parameter estimation under selection. As Thompson points out in his comments, there is a controversy regarding selection based on observed data. Accordingly, we begin our rejoinder with a discussion on this problem. Then we adress the issues raised by the discussants.

Different methods of parameter estimation or prediction of breeding values developed to account for selection lead to different results on ignorability of selection based on observed data, as stated by Thompson. The repeated sampling developments of Henderson (1975) are made using the conditional distribution of Y_{obs} given R (the observed pattern of missing data) and require, as indicated by missing data theory (Rubin, 1976), stronger conditions for ignorability than likelihood based inferences. However, it should be noted that the latter inferences are based on the joint distribution of Y_{obs} and R instead of the conditional distribution mentioned above. Some papers (Gianola *et al.*, 1988; Goffinet, 1988) have considered selection from the conditional likelihood viewpoint and given conditions for ignoring it, and these are very restrictive and similar to those of Henderson (1975). Goffinet (1988) advocated the use of conditional likelihood for selection on observed data and found that it is ignorable only if the marginal distribution of R does not depend on the parameter θ. The crucial question to be answered is: should inferences be based on the conditional distribution of Y_{obs} given R?

In repeated sampling inferences, the statistical quality of an estimator is usually measured in terms of quantities (bias, variance) evaluated by averaging over all possible samples according to the randomness generated by the sampling process. According to this principle, inferences should be made unconditionally on the observed value of R. This is done in survey sampling theory where selection schemes do not depend on the response variable and the selection probabilities are known; see, for example, Gourieroux (1981). However, these conditions are not satisfied in animal breeding situations and, as noticed by Henderson (1975), the unconditional approach is rather intractable. Consequently a conditional analysis, while not fully efficient, may be useful.

From the likelihood viewpoint, the unconditional method should be preferred over the conditional method because the former leads to better estimates than the latter, in the sense of having smaller asymptotic variances of estimators. Conditional likelihood is usually considered as a device for obtaining a consistent estimate of the parameter of interest in the presence of infinitely many nuisance parameters (Kalbfleisch and Sprott, 1970). According to Andersen (1970), the conditional maximum likelihood estimator is consistent and asymptotically normally distributed but, in contrast to the maximum likelihood estimators, it will not in general be efficient even under regularity conditions.

When selection is based on observed data, a conditional analysis is disturbing because it implies that selection also affects the distribution of the data observed even before it has taken place. For example, in case (a), it would say that the

distribution of y_1, sampled at random, is affected by selection. Im (1989) highlighted difficulties that arise when applying BLUP under selection in this case, and considered estimation and prediction based on the unconditional likelihood. The conditional likelihood approach, which is less efficient, requires knowledge of the selection process and more complicated calculation, unless the marginal distribution of R does not depend on the parameters being estimated.

For selection problems dealt with in this paper, as well as in most animal breeding literature, the correct likelihood is given by the joint distribution of Y_{obs} and R. This may not be always the case. Consider, for example, situation (b) in Table I. We supposed that the unselected individuals were available for analysis, and used the information that they were not selected when deriving the likelihood. If they were not available, the actual likelihood would be a conditional one. In any selection problem, one should construct the correct likelihood and use it to make inferences.

We agree with Henderson that likelihood methods have known desirable properties only in large samples, but little is known about their small sample behavior. Simulation studies he indicated could be useful. We disagree that BLUE and BLUP under his selection model are ML estimators of β and of the conditional mean of U because the normality requirement is not met under selection, unless selection is translation invariant.

Thompson's comments are mostly concerned with another paper, Gianola *et al.* (1988), who considered prediction of breeding values by maximizing the joint distribution of the data and the random effects using the conditional selection scheme proposed by Henderson. For known variances and $\theta = (\beta, u)$, D_c would be the joint density of (Y_{obs}, U) given $R, f(y_{\text{obs}}|r, \beta, u)f(u)$. In Gianola *et al.* (1988), the selection process is defined as the modification of the joint density of (Y, U) into another density, due to a restriction in the sample space of W. Integration over y and u is needed for obtaining the joint density of (Y, U) conditional on $W \in R_s$ from that of (Y, U, W). The procedure suggested by Henderson (1988a) does not involve integration explicitly because it is developed using conditional means, variances and covariances. However, integration is required when calculating these conditional quantities from the joint density of (Y, U, W). It seems to us that Henderson's (1988a) procedure is not conditional on the observed value of W but, rather, on that $W \in R_s$. If it were so, then $H_s = \text{Var}(W_s) = \text{Var}(W|W) = 0$. But, in his example of cow culling, he simulated with $H_s \neq 0$ (p. 3139).

Selection on random effects or residuals (Henderson, 1975) is indeed artificial and, consequently, has no real practical interest. Gianola *et al.* (1988) studied this in order to compare the results with those of Henderson (1975). We are not sure of the need for further developing this type of selection. In his comments, Henderson gave a new and more realistic definition of selection on random effects. Namely, this selection is based on records correlated with U but not available for analysis. It might be interesting to compare different methods under this scheme.

We have no quantitative information on the efficiency of Henderson's approach when selection is on $L'y$, but is not translation invariant. This question deserves further study.

We agree with Thompson that techniques developed by Foulley *et al.* (1983) can sometimes be used to estimates parameters associated with equ.(4) when selection is not ignorable. The selection process must be completely specified and it is not

possible to handle situations in which animals are selected in an unspecified manner (Henderson, 1988b).

To end our rejoinder, we should like to introduce a practical and important question. Is it possible to relax the condition of normality required in Henderson's developments? This question should motivate some further studies on the selection problem.

ADDITIONAL REFERENCES

Andersen E.B. (1970) Asymptotic properties of conditional maximum likelihood estimators.*J.R. Statist. Soc. B* 32, 283-301

Goffinet B. (1988) A propos de l'estimation des paramètres en présence de sélection. *Biom. Praxim* 28, 49-60

Gourieroux C. (1981) *Théorie des sondages.* Economica, Paris

Henderson C.R. (1988a) Simple method to compute biases and mean squared error of linear estimators and predictors in a selection model assuming normality. *J. Dairy Sci.* 71, 3135-3142

Henderson C.R. (1988b) A simple method to account for selected base populations. *J. Dairy Sci.* 71, 3399-3404

Im S. (1989) On a mixed linear model when the data are subject to selection. *Biom. J.* in press.

Kalbfleisch J.D. & Sprott D.A. (1970) Application of likelihood methods to models involving large numbers of parameters (with discussion). *J. R. Statist. Soc. B* 32, 175-208