Listy Biometryczne – Biometrical Letters
Vol. 32 (1995), No. 2, 51-60

Estimation of parameters in linear model with the presence of selection

J. Hyánek,[1] K. Moliński[2] and T. Szwaczkowski[3]

1) Research Institute of Animal Production, 10400 Praha-Uhrineves, Czech Republic
2) Department of Mathematical and Statistical Methods, Agricultural University, Wojska Polskiego 28, 60-637 Poznań, Poland
3) Department of Genetics and Animal Breeding, Agricultural University, Wołyńska 33, 60-637 Poznań, Poland

Summary

The paper contains the theoretical considerations on an estimation of parameters in linear model when original data are selected by given truncated point, but the selection intensity is not known. The maximum likelihood estimators in truncated normal distribution are determined. In the next part of the study, the estimation of parameters in mixed model (under truncated normal distribution) is discussed.

1. Introduction

Prediction of the breeding value (BV) is one of the most important questions for breeders. On the basis of the best prediction it is possible to carry out a selection with maximal genetic gain. Quaas and Pollak (1980) developed the classical BLUP (Best Linear Unbiased Prediction) theory for the so-called animal model (AM). An advantage of the AM method for prediction of BV is the possibility to include into considerations not only the phenotypic value of all related animals but also the relationship matrix.

For the best linear unbiased prediction it is necessary to estimate the components of variance. Many papers have been devoted to this problem (Swalve and Van Vleck, 1987; Hill *et al.* 1983; Van Vleck, 1986; Vischer *et al.*, 1991). There are various methods for the estimation of the components of variance, but

Key words: maximum likelihood approach, mixed model, selection, truncated normal distribution.

the most frequently used is restricted maximum likelihood (REML) estimation. This quadratic estimation is translation invariant and therefore is unbiased not only in a random population, but even in the selected population (Van Vleck, 1985; Van der Werf, 1992).

In the conventional breeding program of livestock the populations are selected. For example in cow and sire evaluation programs selection of heifers is based on their complete lactation yields of milk, fat and protein. In this situation heifers are culling after the first complete lactation, i.e. the first lactation records are unselected. The data from the second and higher lactations are selected. Using multivariate AM in the case of repeated measurements the prediction of BV is unbiased even in this situation (Thompson, 1973). Another approach to the problem of estimation and prediction in the selected population was suggested by Henderson (1975). In the second variant it is possible to provide a selection of heifers on the basis of a part of lactation. Records from the first lactation are selected before the end of lactation and therefore the prediction of BV using Thompson and Henderson approach is biased.

In our paper we would like to present how to solve the case when original data are selected by a given truncation point, but the selection intensity is not known. At first, the maximum likelihood (ML) estimator in truncated normal distribution is determined. The system of equations is solved by quadratic approximation of selection intensity and a simple example is given for illustration.

2. Linear model. Normal distribution.

Let us consider a random variable y with normal distribution $N(\mu, \sigma^2, y_t)$ truncated at the point y_t. The density function of this distribution is:

$$f(y) = 0 \qquad \text{for} \quad y \le y_t \ ,$$

$$f(y) = \frac{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}}{1 - F\left(\frac{y_t - \mu}{\sigma}\right)} \qquad \text{for} \quad y > y_t \ .$$

ML estimators of μ and σ are a solution of the system of equations

$$\begin{aligned} \frac{\partial \ln L}{\partial \mu} &= 0 \\ \frac{\partial \ln L}{\partial \sigma} &= 0 \end{aligned} \ , \qquad (1)$$

where

$$\ln L = \sum_{j=1}^{n} \ln f(y_j) = \text{const} + \ln\sigma^{-n} - \frac{1}{2}\sum_{j=1}^{n}\left(\frac{y_j - \mu}{\sigma}\right)^2 + \ln\left[1 - F\left(\frac{y_t - \mu}{\sigma}\right)\right]^{-n} .$$

After substitution of partial derivatives $\ln L$ into (1) the system of equations for estimation of μ and σ is

$$\begin{aligned} \bar{y}_s &= \mu + i\sigma \\ s_o^2 &= \sigma^2 + i\sigma(y_t - \bar{y}_s) \end{aligned} , \tag{2}$$

where

$$i = \frac{f\left(\frac{y_t - \hat{\mu}}{\hat{\sigma}}\right)}{1 - F\left(\frac{y_t - \hat{\mu}}{\hat{\sigma}}\right)} ,$$

$$s_o^2 = \frac{1}{n}\sum_{j=1}^{n}(y_j - \bar{y}_s)^2 ,$$

and where

$$\bar{y}_s = \frac{1}{n}\sum_{j=1}^{n} y_j$$

is the mean of selected records.

It is possible to show that

$$\begin{aligned} E(\bar{y}_s) &= \mu + i\sigma , \\ E(s^2) &= \sigma^2 + i\sigma(y_t - \bar{y}_s) , \end{aligned} \tag{3}$$

where

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y}_s)^2 .$$

If expected values in (3) are substituted by the sample mean and variance then the system of equations for estimation of parameters μ and σ is nearly identical with the system of equations for ML estimators (2), only sample variance s^2 is used instead of s_o^2 in the ML system (2).

Approximating selection intensity i by a quadratic function:

$$i = \frac{f(u)}{1 - F(u)} \approx a + bu + cu^2 ,$$

one can show that the estimators of μ and σ are the solution of a system of equations:

$$\begin{aligned} \hat{\mu} &= \bar{y}_s - \left[a + b\left(\frac{y_t - \hat{\mu}}{\hat{\sigma}}\right) + c\left(\frac{y_t - \hat{\mu}}{\hat{\sigma}}\right)^2\right]\sigma \\ \hat{\sigma}^2 &= (\hat{\mu} - \bar{y}_s)(y_t - \bar{y}_s) + s^2 \end{aligned} \quad , \tag{4}$$

where $a = 0.867$, $b = 0.603$, $c = 0.071$.

If variance σ^2 is known a priori then the system (4) is reduced to one equation:

$$0.071\hat{\mu}^2 - [0.142y_t - 0.317\sigma]\hat{\mu} + 0.867\sigma^2 + \sigma(0.603y_t - \bar{y}_s) + 0.071y_t^2 = 0 \; . \tag{5}$$

By solving this quadratic equation we obtain the estimator

$$\hat{\mu} = y_t + \sigma\left(\sqrt{14.084\,\frac{\bar{y}_s - y_t}{\sigma} - 4.3989} - 2.7957\right) \; . \tag{6}$$

Example: Let us assume that the average milk yield per one lactation, greater than 7500 kg (truncation point $y_t = 7500$kg), is $\bar{y}_s = 8000$kg. It is clear, that $\bar{y}_s = 8000$kg is not an unbiased estimator of μ, but it is overestimated. If the variance of milk yield is known a priori, for example $\sigma^2 = 1600$, then after substitution into equation (6) an unbiased estimator of μ is approximately equal

$$\hat{\mu} = 7500 + 40\left[\sqrt{14.084\,\frac{8000 - 7500}{40} - 4.3989} - 2.7957\right] = 7912$$

If the variance is unknown, σ^2 and μ can be evaluated by solving mixed model equations.

3. Mixed linear model

3.1. General considerations

Consider the mixed linear model:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e} \; ,$$

where:

y is the $n \times 1$ vector of observations,

β is a $p \times 1$ vector of fixed unknown parameters,

u is a $q \times 1$ vector of random effects having normal distribution $N(\mathbf{0}; \sigma_u^2\mathbf{I})$,

e is a $n \times 1$ vector of random errors having normal distribution $N(\mathbf{0}; \sigma_e^2\mathbf{I})$,

X, **Z** are $n \times p$ and $n \times q$ incidence matrices, respectively,

with cov(**u**,**e**) = **0**. The mixed model equations (MME) for $\hat{\boldsymbol{\beta}}$ and $\tilde{\mathbf{u}}$ are

$$\begin{aligned} \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}'\mathbf{Z}\tilde{\mathbf{u}} &= \mathbf{X}'\mathbf{y} \ , \\ \mathbf{Z}'\mathbf{X}\hat{\boldsymbol{\beta}} + (\mathbf{Z}'\mathbf{Z} + \gamma^{-1}\mathbf{I})\tilde{\mathbf{u}} &= \mathbf{Z}'\mathbf{y} \ , \quad \text{for } \gamma = \frac{\sigma_u^2}{\sigma_e^2} \ . \end{aligned} \tag{7}$$

In a selected population the expected value of **y** is not **Xβ**, but

$$E(\bar{\mathbf{y}}_s) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Delta} \ .$$

It is possible to modify MME (7) by elimination of bias **Δ**

$$\begin{aligned} \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}'\mathbf{Z}\tilde{\mathbf{u}} &= \mathbf{X}'(\mathbf{y}_s - \boldsymbol{\Delta}) = \mathbf{X}'\mathbf{y}_s - \mathbf{X}'\boldsymbol{\Delta} \ , \\ \mathbf{Z}'\mathbf{X}\hat{\boldsymbol{\beta}} + (\mathbf{Z}'\mathbf{Z} + \gamma^{-1}\mathbf{I})\tilde{\mathbf{u}} &= \mathbf{Z}'(\mathbf{y}_s - \boldsymbol{\Delta}) = \mathbf{Z}'\mathbf{y}_s - \mathbf{Z}'\boldsymbol{\Delta} \ . \end{aligned} \tag{8}$$

After separating **β** and **u** it is possible to obtain:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \hat{\boldsymbol{\beta}}_s - \mathbf{W}^{-1}(\mathbf{X}' - \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z} + \gamma^{-1}\mathbf{I})^{-1}\mathbf{Z})\boldsymbol{\Delta} = \hat{\boldsymbol{\beta}}_s - \mathbf{W}^{-1}(\mathbf{X}'\mathbf{P}_{\gamma,z}\boldsymbol{\Delta}) \ , \\ \tilde{\mathbf{u}} &= \tilde{\mathbf{u}}_s - \mathbf{U}^{-1}(\mathbf{Z}' - \mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\boldsymbol{\Delta} = \tilde{\mathbf{u}}_s - \mathbf{U}^{-1}\mathbf{Z}'\mathbf{P}_x\boldsymbol{\Delta} \ , \end{aligned} \tag{9}$$

where

$$\begin{aligned} \mathbf{W} &= \mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z} + \gamma^{-1}\mathbf{I})^{-1}\mathbf{Z}'\mathbf{X} = \mathbf{X}'\mathbf{P}_{\gamma,z}\mathbf{X} \ , \\ \mathbf{U} &= \mathbf{Z}'\mathbf{Z} + \gamma^{-1}\mathbf{I} - \mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z} = \mathbf{Z}'\mathbf{P}_x\mathbf{Z} + \gamma^{-1}\mathbf{I} \ , \\ \mathbf{P}_{\gamma,z} &= \mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z} + \gamma^{-1}\mathbf{I})^{-1}\mathbf{Z}' \ , \\ \mathbf{P}_x &= \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \ , \\ \hat{\boldsymbol{\beta}}_s &= \mathbf{W}^{-1}\mathbf{X}'\mathbf{P}_{\gamma,z}\mathbf{y} \ , \\ \tilde{\mathbf{u}}_s &= \mathbf{U}^{-1}\mathbf{Z}'\mathbf{P}_x\mathbf{y} \end{aligned}$$

[$\boldsymbol{\beta}_s$ and $\mathbf{u}_s$ are solutions of system equations (7)].

If the matrix **X'X** is singular an alternative formula for the prediction of $\tilde{\mathbf{u}}$ can be used:

$$\begin{aligned} \tilde{\mathbf{u}} &= \tilde{\mathbf{u}}_s - (\mathbf{Z}'\mathbf{Z} + \gamma^{-1})^{-1}\mathbf{Z}'(\mathbf{I} + \mathbf{X}\mathbf{W}^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z} + \gamma^{-1}\mathbf{I})^{-1}\mathbf{Z}))\boldsymbol{\Delta} \\ &= \tilde{\mathbf{u}}_s - (\mathbf{Z}'\mathbf{Z} + \gamma^{-1})^{-1}\mathbf{Z}'(\mathbf{I} + \mathbf{X}(\mathbf{X}'\mathbf{P}_{\gamma,z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_{\gamma,z})\boldsymbol{\Delta} \ . \end{aligned}$$

From equations (9) it follows that the bias is dependent on matrix **Z'Z** and **X'X**. The above given formulae may be used if the value of **Δ** is known. Unfortunately,

Δ is often unknown. Hence we will present two approaches to solving this problem.

3.2. Thompson's approach

Thompson (1973) derived the estimation of the mean on the basis of a likelihood function. In this case we assume a two-dimensional normal distribution

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}; \mathbf{V} \right) ,$$

where

$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix} .$$

ML estimators of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are solutions of the system of equations:

$$\begin{aligned} \frac{\partial \ln L}{\partial \boldsymbol{\mu}_1} &= \mathbf{0} , \\ \frac{\partial \ln L}{\partial \boldsymbol{\mu}_2} &= \mathbf{0} , \end{aligned} \tag{10}$$

where

$$L = L(\mathbf{y}_1, \mathbf{y}_{2,s} - \mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{y}_{1s})$$

is the likelihood function of random vectors $\mathbf{y}_1$ and $\mathbf{y}_{2s} - \mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{y}_{1,s}$. Considering that (because of the factorization of the likelihood function)

$$\operatorname{cov}(\mathbf{y}_{1s}, \mathbf{y}_{2s} - \mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{y}_{1s}) = \mathbf{V}_{12,s} - \mathbf{V}_{11,s}\mathbf{V}_{11}^{-1}\mathbf{V}_{12,s} = 0 ,$$

the solution of the system (10) is computable.

If selection is carried out according to random vector $\mathbf{y}_1$ then parameters of selected random vector $\mathbf{y}_{2,s}$ are changed in the following way (Thompson, 1973):

$$E(\mathbf{y}_{2,s}) = \boldsymbol{\mu}_{2,s} = \boldsymbol{\mu}_2 + \mathbf{V}_{21}\mathbf{V}_{11}^{-1}(\boldsymbol{\mu}_{1s} - \boldsymbol{\mu}_1) ,$$

$$\operatorname{var}(\mathbf{y}_{2,s}) = \mathbf{V}_{22,s} = \mathbf{V}_{22} - \mathbf{V}_{21}\mathbf{V}_{11}^{-1}(\mathbf{V}_{11} - \mathbf{V}_{11,s})\mathbf{V}_{11}^{-1}\mathbf{V}_{12} ,$$

$$\operatorname{cov}(\mathbf{y}_{2s},\mathbf{y}_{1s}) = \mathbf{V}_{21,s} = \mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{V}_{11,s} ,$$

$$\operatorname{cov}(\mathbf{y}_{1s},\mathbf{y}_{2s}) = \mathbf{V}_{12,s} = \mathbf{V}_{11,s}\mathbf{V}_{11}^{-1}\mathbf{V}_{12} ,$$

$$\boldsymbol{\mu}_2 = \boldsymbol{\mu}_{2,s} - \mathbf{V}_{21}\mathbf{V}_{11}^{-1}(\boldsymbol{\mu}_{1s} - \boldsymbol{\mu}_1) ,$$

where $\boldsymbol{\mu}_{1s}$ is the mean of the selected random vector $\mathbf{y}_{1s}$ and $\mathbf{V}_{11,s}$ is the variance-covariance matrix of the random vector $\mathbf{y}_{1,s}$.

3.3. Henderson's approach

Henderson (1975) suggested another approach. Let us assume that selection is provided according to a random variable **w** (selection index). Prior to selection, the expected values of **y**, **u**, **e**, **w** equal:

$$E\begin{pmatrix}\mathbf{y}\\ \mathbf{u}\\ \mathbf{e}\\ \mathbf{w}\end{pmatrix} = \begin{pmatrix}\mathbf{X\beta}\\ \mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\end{pmatrix} \tag{11}$$

and their variance covariance matrix is

$$\text{var}\begin{pmatrix}\mathbf{y}\\ \mathbf{u}\\ \mathbf{e}\\ \mathbf{w}\end{pmatrix} = \begin{pmatrix}\mathbf{V} & \mathbf{ZG} & \mathbf{R} & \mathbf{B}\\ \mathbf{GZ'} & \mathbf{G} & \mathbf{0} & \mathbf{B_u}\\ \mathbf{R} & \mathbf{0} & \mathbf{R} & \mathbf{B_e}\\ \mathbf{B'} & \mathbf{B'_u} & \mathbf{B'_e} & \mathbf{H}\end{pmatrix} . \tag{12}$$

Estimator of $\boldsymbol{\beta}$ is a solution of the system of equations:

$$\mathbf{X'V^{-1}X\hat{\beta}} = \mathbf{X'V^{-1}y} .$$

Prediction of the random variable **u** can be achieved on the basis of the regression analysis:

$$\tilde{\mathbf{u}} - E(\tilde{\mathbf{u}}) = \text{cov}(\tilde{\mathbf{u}},\mathbf{y})\,(\text{var}(\mathbf{y}))^{-1}(\mathbf{y} - E(\mathbf{y})) .$$

Using $E(\mathbf{u})$, $E(\mathbf{y})$, $\text{cov}(\mathbf{u},\mathbf{y})$ and $\text{var}(\mathbf{y})$ given by (11) and (12), one can show that the predictor of **u** is equal to

$$\tilde{\mathbf{u}} = \mathbf{GZ'V^{-1}}(\mathbf{y} - \mathbf{X\hat{\beta}}) .$$

Henderson (1975) proved that this process is equivalent to the MME in the following form:

$$\begin{aligned}\mathbf{X'R^{-1}X\hat{\beta}} + \mathbf{X'R^{-1}Z\tilde{u}} &= \mathbf{X'R^{-1}y}\\ \mathbf{Z'R^{-1}X\hat{\beta}} + (\mathbf{Z'R^{-1}Z} + \mathbf{G})\tilde{\mathbf{u}} &= \mathbf{Z'R^{-1}y}\end{aligned} . \tag{13}$$

From the model equation $\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e}$ it follows that

$$\mathbf{B} = \text{cov}(\mathbf{y},\mathbf{w}) = \mathbf{Z}[\text{cov}(\mathbf{u},\mathbf{w})] + \text{cov}(\mathbf{e},\mathbf{w}) = \mathbf{ZB_u} + \mathbf{B_e} .$$

It is possible to derive expected values and the variance-covariance matrix of **y**, **u**, **e**, **w** in the model $\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e}$ after selection:

$$E\begin{pmatrix}\mathbf{y}\\ \mathbf{u}\\ \mathbf{e}\\ \mathbf{w}\end{pmatrix} = \begin{pmatrix}\mathbf{X\beta} + \mathbf{BH}^{-1}\mathbf{s}\\ \mathbf{B}_u\mathbf{H}^{-1}\mathbf{s}\\ \mathbf{B}_e\mathbf{H}^{-1}\mathbf{s}\\ \mathbf{s}\end{pmatrix} \tag{14}$$

and

$$\text{var}\begin{pmatrix}\mathbf{y}\\ \mathbf{u}\\ \mathbf{e}\\ \mathbf{w}\end{pmatrix} = \begin{pmatrix}\mathbf{V} - \mathbf{BH}_o\mathbf{B}' & \mathbf{ZG} - \mathbf{BH}_o\mathbf{B}'_u & \mathbf{R} - \mathbf{BH}_o\mathbf{B}'_e & \mathbf{BH}^{-1}\mathbf{H}_s\\ \mathbf{GZ}' - \mathbf{B}_u\mathbf{H}_o\mathbf{B}' & \mathbf{G} - \mathbf{B}_u\mathbf{H}_o\mathbf{B}'_u & \mathbf{B}_u\mathbf{H}_o\mathbf{B}'_e & \mathbf{B}_u\mathbf{H}^{-1}\mathbf{H}_s\\ \mathbf{R} - \mathbf{B}_e\mathbf{H}_o\mathbf{B}' & -\mathbf{B}_e\mathbf{H}_o\mathbf{B}'_u & \mathbf{R} - \mathbf{B}_e\mathbf{H}_o\mathbf{B}'_e & \mathbf{B}_e\mathbf{H}^{-1}\mathbf{H}_s\\ \mathbf{H}_s\mathbf{H}^{-1}\mathbf{B}' & \mathbf{H}_s\mathbf{H}^{-1}\mathbf{B}'_u & \mathbf{H}_s\mathbf{H}^{-1}\mathbf{B}'_e & \mathbf{H}_s\end{pmatrix}, \tag{15}$$

with $\mathbf{H}_o = \mathbf{H}^{-1}(\mathbf{H} - \mathbf{H}_s)\mathbf{H}^{-1}$.

After selection, covariance $\text{cov}(\mathbf{u};\mathbf{e}) \neq \mathbf{0}$, and it is not generally true that

$$\mathbf{V}_s = \mathbf{Z}'\mathbf{G}_s\mathbf{Z} + \mathbf{R}_s \ ,$$

where

$$\begin{aligned}\mathbf{V}_s &= \mathbf{V} - \mathbf{BH}_o\mathbf{B}' \ ,\\ \mathbf{G}_s &= \mathbf{G} - \mathbf{B}_u\mathbf{H}_o\mathbf{B}'_u \ ,\\ \mathbf{R}_s &= \mathbf{R} - \mathbf{B}_e\mathbf{H}_o\mathbf{B}'_e \ .\end{aligned}$$

Therefore it is not possible to use MME (13) with matrices $\mathbf{R}_s$, $\mathbf{G}_s$ instead of matrices $\mathbf{R}$ and $\mathbf{G}$.

First, it is necessary to compute estimators of $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{s}}$ from the system of equations:

$$\begin{bmatrix}\mathbf{X}'\mathbf{V}_s^{-1}\mathbf{X} & \mathbf{X}'\mathbf{V}_s^{-1}\mathbf{B}\\ \mathbf{B}'\mathbf{V}_s^{-1}\mathbf{X} & \mathbf{B}'\mathbf{V}_s^{-1}\mathbf{B}\end{bmatrix}\cdot\begin{bmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\mathbf{s}}\end{bmatrix} = \begin{bmatrix}\mathbf{X}'\mathbf{V}_s^{-1}\mathbf{y}\\ \mathbf{B}'\mathbf{V}_s^{-1}\mathbf{y}\end{bmatrix} \ .$$

Predictor $\tilde{\mathbf{u}}$ is determined on the basis of correlation with $\mathbf{y}$:

$$\tilde{\mathbf{u}} - E(\tilde{\mathbf{u}}) = \text{cov}(\mathbf{y},\tilde{\mathbf{u}})\,(\text{var}\,(\mathbf{y}))^{-1}(\mathbf{y} - E(\mathbf{y})) \ .$$

After substituting $E(\mathbf{u})$, $\text{cov}(\mathbf{y},\mathbf{u})$ and $\text{var}(\mathbf{y})$, $E(\mathbf{y})$ after selection from (14) and (15), it follows that

$$\tilde{\mathbf{u}} = \mathbf{B}_u\mathbf{H}^{-1}\hat{\mathbf{s}} + (\mathbf{GZ}' - \mathbf{B}_u\mathbf{H}_o\mathbf{B}')\mathbf{V}_s^{-1}(\mathbf{y} - \mathbf{X\beta} - \mathbf{B}\hat{\mathbf{s}}) \ .$$

4. Comments

If selection intensity is known, it is possible to compute $\hat{\boldsymbol{\beta}}$ and $\tilde{\mathbf{u}}$ very simply from a system of equations (8) or (9), where $\hat{\boldsymbol{\beta}}_s$ and $\tilde{\mathbf{u}}_s$ are solutions of the usual MME (7). The advantage of this approach is that in a simulation study one can see the extent of bias for various selection intensities and various incidence matrices **X** and **Z**. Alternatively, it is possible to use the Henderson approach or the Thompson ML estimate for multitrait AM. Henderson used the best prediction in the case of a given expected mean of variable w, characterizing selection. Thompson (1973) used multitrait AM, where repeated records are expressed as several traits on each animal.

The selection intensity is often not known. Then it is not possible to apply either the Henderson or Thompson approach. In the special case, when the truncation point is given, ML estimator provides unbiased estimation of BV as it is shown in section 2. In the general case it is necessary to incorporate the whole information about selection into the estimation of BV.

As mentioned above, the distribution parameters of both the observation vector and vectors of fixed and random effects change under selection. This influences the estimation and prediction (Henderson, 1984). However, assumptions concerning distribution parameters are also taken into account in the estimation of variance components. Henderson (1984) reported that conventional assumptions are correct in unselected populations, only. Generally, the variances are reduced in these populations. As already mentioned, under selection nonzero covariances between random effects (for instance random genetic and random error effects – see section 3.3) are generated. The magnitude of the bias is difficult to express for two reasons. Firstly, since selection intensity varies from, let us say, one herd to another, a different set of parameters would be needed for each herd. Secondly, correlation between **u** and **e** complicates the computations. More details concerning the biases of variance component estimators under a selection model are given by Schaeffer (1987). He concludes that the degree of the bias of the dispersion component estimators is connected with the definition of the so-called selection rule matrix.

Problems presented above are often ignored in genetic analyses of quantitative traits in selected populations. However, it must be stressed here, that a simulation study performed by Rothschild *et al.* (1979) indicated that estimation via MIVQUE with good priors, REML and ML may considerably control the bias caused by selection.

References

Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31**, 423-449.

Henderson, C.R. (1984). Applications of linear models in animal breeding. Univ. of Guelph, Ontario, Canada.

Hill, W.G., Edwards, M.R., Ahmed, M.K.A., Thompson, R. (1983). Heritability of milk yield and composition at different levels and variability of production. *Anim. Prod.* **36**, 59-68.

Quaas, R.L., Pollak, E.J. (1980). Mixed model methodology for farm beef cattle testing programs. *J. Anim. Sci.* **51**, 1277.

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.

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

Swalve, H., Van Vleck, L.D. (1987). Estimation of genetic (co)variances for milk yield in first three lactations using an animal model and restricted maximum likelihood. *J. Dairy Sci.* **70**, 842-849.

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

Van der Werf, J.H.J. (1992). Restricted maximum likelihood estimation of additive genetic variance when selected base animals are considered fixed. *J. Anim. Sci.* **70**, 1068-1076.

Van Vleck, L.D. (1985). Including records of daughters of selected bulls in estimation of sire components of variance. *J. Dairy Sci.* **68**, 2396-2407.

Van Vleck, L.D. (1986). Evaluation of dairy cattle breeding programs: specialised milk production. *Proceedings of the Third World Congress on Genetics Applied to Livestock Production*, Lincoln, Nebraska, Vol. IX, 141-152.

Visscher, P.M., Thompson, R., Hill, W.G. (1991). Estimation of genetic and environmental variances for fat yield in individual herds and an investigation into heterogeneity of variance between herds. *Livest. Prod. Sci.* **28**, 273-290.

Received 10 January 1995; revised 5 October 1995

Estymacja parametrów w modelu liniowym dla danych selekcjonowanych

Streszczenie

Praca zawiera teoretyczne rozważania dotyczące estymacji parametrów w modelu liniowym, w przypadku gdy oryginalne dane są selekcjonowane, a intensywność selekcji nie jest znana. Wyznaczono dla danych "obciętych" estymatory największej wiarogodności parametrów modelu. Dalsza część pracy zawiera dyskusję dotyczącą różnych podejść do problemu estymacji parametrów w modelach mieszanych.

Słowa kluczowe: metoda największej wiarogodności, model mieszany, selekcja, "obcięty" rozkład normalny.