DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS
40

THE RELIABILITY OF LINEAR MIXED MODELS IN GENETIC STUDIES

TANEL KAART

TARTU UNIVERSITY
PRESS

Faculty of Mathematics and Computer Science, University of Tartu, Tartu, Estonia

Dissertation is accepted for the commencement of the degree of Doctor of Philosophy (PhD) in mathematical statistics on November 18, 2005 by the Council of the Faculty of Mathematics and Computer Science, University of Tartu.

Supervisor:
Cand. Sc., docent Tõnu Möls
University of Tartu
Tartu, Estonia

Opponents:
PhD, professor Dietrich von Rosen
Swedish University of Agricultural Sciences
Uppsala, Sweden

PhD, professor Anita Dobek
Agricultural University of Poznan
Poznan, Poland

The public defence will take place on January 11, 2006.

Publication of the dissertation is financed by the Institute of Mathematical Statistics, University of Tartu (research project TMTMS1776).

ISBN 9949–11–212–5 (trükis)
ISBN 9949–11–213–3 (PDF)

Tartu Ülikooli Kirjastus
www.tyk.ee
Tellimus nr. 618

CONTENTS

SELECTED ORIGINAL PUBLICATIONS

1. Saveli, O., Kaasiku, U., Kaart, T. (1999). Breeding value of Estonian Holstein bulls depending on pedigree. *Animal Husbandry, Scientific articles*, **35**, 11–16.
2. Aland, A., Kaart, T., Praks, J. (2000). Keskkonna riskitegurite olulisuse selgitamine lüpsilehmade mastiiti haigestumises. *Veterinaarmeditsiin '2000*. ELÜ kirjastus, Tartu, 9–13.
3. Saveli, O., Bulitko, T., Kaart, T., Kaasiku, U., Kalamees, K., Kureoja, A., Orgmets, E., Pulk, H., Siiber, E., Uba, M. (2001). Eesti veisetõugude aretuskomponentide võrdlev hinnang ja kasutamine aretusprogrammides. *Agraarteadus*, **12**, 224–248.
4. Kaart, T. (2001). Ülevaade geneetiliste parameetrite hindamisel kasutatavatest mudelitest. *Eesti Põllumajandusülikooli Loomakasvatusinstituudi teadustöid 71*, EPMÜ Loomakasvatusinstituut, Tartu, 52–67.
5. Tänavots, A., Somelar, E., Viinalass, H., Värv, S., Kaart, T., Saveli, O., Eilart, K., Põldvere, A. (2001). Pork quality and porcine stress syndrome in Estonia. *Proceedings of the Latvian Academy of Sciences. Section B: Natural, Exact, and Applied Sciences*, **55**, 242–246.
6. Tänavots A., Kaart T., Saveli O.(2002). Heritability and Correlation of Meat and Fertility Traits in Pigs in Estonia. *Veterinarija ir Zootechnika*, **19**, 106–108.
7. Kärt, O., Rihma, E., Tölp, S., Kaart, T. (2003). Dry matter intake of the first-parity cows, bred in Estonia, at the begginning of lactation. *Veterinarija ir Zootechnika*, **22**, 53–57.
8. Kureoja, A., Kaart, T. (2004). Genetic and environmental influences on urea concentration in dairy cows' milk. In *Animal Breeding in the Baltics* (ed. Saveli, O., Kärt, O., etc), Tartu, Estonia, 42–47.
9. Värv, S., Viinalass, H., Kaart, T., Kantanen, J. (2004). Genetic differentiation among commercial and native cattle breeds. In *Animal Breeding in the Baltics* (ed. Saveli, O., Kärt, O., etc), Tartu, Estonia, 111–114.
10. Kaart, T. (2004). About the data designs for estimation of genetic parameters in animal breeding studies. *Acta et Commentationes Universitatis Tartuensis de Mathematica*, **8**, 113–121.
11. Kaart. T. (2005). A new approximation for the variance of the ANOVA estimate of the intraclass correlation coefficient. *Proceedings of the Estonian Academy of Science, Physics, Mathematics*, **54**, 243–254.
12. Aland A., Kaart, T., Poikalainen, V., Praks, J. (2005). Incidence of multifactorial diseases in estonian dairy herds. *Deutsche Tierärztliche Wochenschrift*, (accepted).

Publications [2], [3] and [4] are abstracted by AGRIS/CARIS database, publications [5]-[9] and [12] are abstracted by CAB International (UK) database, publication [10] is abstracted by Mathematical Reviews and Zentralblatt für Mathematik, publication [11] is abstracted by Mathematical Reviews.

Selected conference abstracts and proceedings

1. Kaart, T. (1996). Statistical analysis of heritability coefficient. In: *The 2nd Baltic Animal Breeding Conference*, Proceedings, Kaunas, Lithuania, 1996, 25–27.
2. Kaart, T. (1997). Probability of the estimate of heritability being negative or greater than one. In: *The 3rd Baltic Animal Breeding Conference*, Proceedings, Riia, Latvia, April 3–4, 1997, 57–59.
3. Kaart, T. (1998). Estimation of variance components and heritability coefficient with four different methods in case of different sire-daughter design. In: *The 4th Baltic Animal Breeding Conference*, Proceedings, Tartu, Estonia, 1998, 28–32.
4. Kaart, T. (1999). Multivariate mixed model in animal breeding. In: *The 6th Tartu Conference on Multivariate Statistics*, Abstracts, Tartu, Estonia, August 19–23, 1999, 25.
5. Kaart, T., Piirsalu, P. (2000). The complex analysis of genetic parameters in Estonian sheep breeds. In: *The 6th Baltic Animal Breeding Conference*, Proceedings, Jelgava, Latvia, April 27–28, 2000, 135–140.
6. Kureoja, A., T. Kaart. (2001). Interaction Between Genotype and Feeding-keeping Conditions of Estonian Red Cows. In: *The 7th Baltic Animal Breeding Conference*, Proceedings, Tartu, Estonia, April 17–18, 2001, 60–64.
7. Kaart, T. (2001). Reliability of mixed models as tools for finding major genes in genetic studies. In: *The 12th European Young Statisticians Meeting*, Abstracts, Liptovský Ján, Slovakia, September 4–8, 2001, 24.
8. Kiiman, H., Saveli, O., Kaart, T. (2002). The variations and heritabilities of measures of somatic cell count per lactation. In: *The 8th Baltic Animal Breeding and Genetics Conference*, Proceedings, Kaunas, Lithuania, May 5–8, 2002, 36–37.
9. Kureoja, A., Kaart, T. (2002). The use of ANOVA to evaluate sources of variation in reproductive performance in Estonian Red breed cattle. In: *The 8th Baltic Animal Breeding and Genetics Conference*, Proceedings, Kaunas, Lithuania, May 5–8, 2002, 39.
10. Pärna, E., Saveli, O., Kaart, T. (2002). Genetic improvement of production and functional traits in Estonian Holstein population. In: *The 8th Baltic Animal Breeding and Genetics Conference*, Proceedings, Kaunas, Lithuania, May 5–8, 2002, 52–53.
11. Pärna, E., Saveli, O., Kaart, T. (2002). Economic weights for production and functional traits of Estonian holstein populations. In: *Proceedings of the 7th World Congress on Genetics Applied to Livestock Production*, Montpellier, France, August 19–23, 2002, 323–326.
12. Aland, A., Kaart, T., Praks, J., Poikalainen, V. (2003). The effect of environmental risk factors on the occurrence of multifactorial diseases in Estonian dairy cattle. In: *The X International Congress in Animal Hygiene*, Proceedings, Vol 1, Mexico City, Mexico, February 23–27, 2003, 201–205.

13. Ling, K., Waldmann, A., Samarütel, J., Jaakson, H., Leesmäe, A., Kaart, T. (2004). Field trial on blood metabolites, body condition score (BCS) and their relation to the recurrence of ovarian cyclicity in Estonian Holstein dairy cows. In: *12th International Conference on Production Diseases in Farm Animals*, Program and Abstracts, Michigan State University, East Lansing, Michigan, USA, July 19–22, 2004, 32.

INTRODUCTION

The need to analyse data by linear mixed models occurs in many areas of research, especially in the analysis of biological field experiments. As pointed out by Searle, Casella and McCulloch (1992), the first applications of the random effects models appeared in the literature in the middle of the 19th century and concerned astronomy. The following development of the theory of linear models, and especially linear mixed models, is often connected with the applications in genetics.

The parallel development of genetics and linear (mixed) models started in the beginning of the 20th century when K. Pearson and F. Galton studied the selection index theory and the inheritance of continuous traits, respectively, and worked out the cornerstones of regression and correlation analysis. In 1918 R. A. Fisher reconciled the work of G. Mendel on the inheritance of discrete effects and the work of F. Galton on the continuous variation of metric traits, giving the present population genetic description of inheritance. In the same paper Fisher worked out the basis of the analysis of variance and some years later derived the first method for variance components estimation. The best-known scientists in developing and advertising the theory of mixed linear models in the second part of the 20th century, C. R. Henderson and S. R. Searle, were both working in animal breeding and genetics (Searle, 1998).

The connection between mixed model and genetics appears in the following.

A mixed model representing the relation between response variables and factors consists of two different types of effects: fixed and random. The first of them has a finite set of levels, all (potentially) represented in the data and an object of interest. The second type of effects have (usually) an infinite set of levels, all generated by some random process and being represented in the data by a random sample, the user is interested in both – in the observed and the unobserved levels – and in the variability of the studied variable explained by them.

In genetic analysis the response variable can be influenced by thousands of genes and their interactions, from which we can predict only a random sample occurring in the data. The prediction can be made by measuring the similarity between relatives as the effect of identical genes of one family. These effects are the values of the random process of Mendelian inheritance. Thus, the genetic effects exactly follow the definition of random effects.

Usually the analysed individuals are not homogeneous by non-genetic effects, and some supplementary effects – the influence of which is essential – are exactly recorded (sex and age, for example). Those effects are of interest only by their observed levels and are considered fixed effects.

The development in the mixed models theory is necessary for animal breeding, because better knowledge about the genetic determination of economically important traits needs better models to implement these new cognitions into the

selection of parents of the next generation. These better models mean more accurate selection, more intensive breeding, bigger production and finally, more money. In human populations there is no need to rank people and this is the reason for the scarce contributions into the development of mixed models theory by scientists working in human genetics. But this situation may change soon – the new models are needed to prevent or better predict the diseases using already available or only developed joint databases of genetic markers, human behaviour and diseases. Although in this dissertation models suitable for both these applications are considered, the main focus is on the animal genetics.

The following study focuses on the normally distributed data, which in the context of genetics means that the traits are quantitative in nature, and are affected by many genes and by environmental factors (blood pressure or survival time after infection in human medicine and the majority of production traits in animal breeding, for example). Also, to avoid redundant straggling, the study focuses only on single trait models and the methods of classical statistics.

Chapter 1 gives an overview of the mixed linear models, presenting the basic concepts and formulas without complicated derivations for model building. In addition, a theory for estimating the fixed effects and for predicting the realised values of random effects, as well as the variance components estimation theory is introduced. Approximated expressions of two-stage predictor's variances are extended to variance-covariance matrices by the author.

In Chapter 2 the basic models used in genetic parameter estimation are presented and their differences discussed. As in the literature, some of these models are presented using only statistical notation, but at the same time, since the right understanding of genetics behind statistics required exploiting the model parameters properly, both the genetic and statistical models are reduced to a similar scheme and are presented concurrently. Topics like half- and full-sib models, animal models, models with maternal effects and models with single genes effects are considered. The estimable effects and genetic parameters for each model are introduced.

In Chapter 3 the behaviour of estimated parameters is studied, based on the simplest mixed linear model – the one-way random model. The formulas for variances of predictors of random effects, for variances of estimators of variance components and intraclass correlation coefficient as well as for the probabilities of inadmissible estimates of the heritability coefficient are congregated or derived if needed both for balanced and unbalanced data sets. The effect of data structure and imbalance is examined based both on the theoretical results and simulation experiments.

In Chapter 4 some problems occurring when using the general mixed models with mathematically unstructured covariance structure – frequent in genetic studies but fairly unreasonable for standard mixed linear models theory – are considered. Propositions concerning the prediction of effects which have no data are proved. Also, the effect of variance-covariance structure on the accuracy of estimates and the dependence of the population genetic structure and

genetic model choice are discussed and illustrated with a real data example of lambs weaning weight.

All theoretical results derived by the author are presented with proof. Theoretical results known and published before are given without detailed proof, showing their exact source.

The parallel approach of the basic genetic models and their statistical analogues presented in Chapter 2 is with some modifications published in Kaart (2001). The probability of inadmissible heritability estimates (Chapter 3.2.4) is discussed in Kaart (1997), but here some extensions are made. The studies dealing with the effect of data design on the accuracy of genetic parameters estimates in balanced case presented in Chapter 3.3 are published in Kaart (2004) and discussed partly in Kaart (1998). The results regarding the sampling variance of intraclass correlation coefficient and its dependency on the data structure and imbalance (Chapters 3.4.2 and 3.5.2) are published in Kaart (2005). The real data example of the effect of genetic model choice and amount of used pedigree information (Chapter 4.4.2) follows the pilot study published in Kaart and Piirsalu (2000).

In the past years many books about the models used in estimating genetic parameters have been published. The basic polygenetic models, with emphasis on their practical resolve via matrix calculus but without deeper mathematical proof, are introduced by Henderson (1984) and Mrode (1996). The deeper genetic background can be found in Falconer and Mackay (1996) and Lynch and Walsh (1998). The excellent books on mathematical theory of mixed linear models are Searle, Casella and McCulloch (1992) and Khuri, Mathew and Sinha (1998), for example. The statistical methods used in animal breeding and genetics are quite mathematically discussed in Gianola and Hammond (1990). The overview of biometrical genetics with an emphasis on the molecular data without deeper discussion on mixed linear models is given by Weller (2001) for animal science and by Ott (1999), Sham (1998) and Lange (1997) for human genetics. There are some short compendious writings on statistical and genetic models in Estonian published by the author – Kaart (2001), plus more than 200 pages of lecture notes prepared by the author for courses ‘Statistical Models in Gene Technology’ and ‘Linear Models and Estimation of Breeding Values’ taught at the University of Tartu and at the Estonian Agricultural University, respectively.

CHAPTER 1
THE LINEAR MIXED MODEL: REVIEW

In the terminology of linear mixed models we shall follow closely the notation and conventions of Searle (1987) and Henderson (1984). All the proof and derivations of reviewed results can be found in these and many other books. However, some symbols will be modified to avoid the deceptive similarity with the notation in the other parts of the dissertation. In Section 1.2.4 the variance-covariance matrix of two-stage predictors will be derived extending the previously published results.

1.1. Matrix formulation

1.1.1. Some basic notation and definitions

In this dissertation matrices are marked with bold capital letters, the row and column vectors with bold minuscule letters and the elements of matrices with italic minuscule letters.

In some formulas the following, more detailed notation for a *matrix* **A** of order $p \times q$ is used:

$$\mathbf{A} = \{_{\mathrm{m}}\, a_{ij}\}_{i=1,\,j=1}^{p,\,q},$$

where *aij* is the element that is in the *i*th row and *j*th column of **A**, the subindex m indicates that the elements inside the braces are arrayed as a matrix.

This notation is extended to row and column vectors and to diagonal matrices with the use of r, c and d as follows. First, a *column vector* is

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_s \end{pmatrix} = \{_{\mathrm{c}}\, u_i\}_{i=1}^{s},$$

the c being used to show that it is a column vector. Similarly

$$\mathbf{u}' = \{_{\mathrm{r}}\, u_i\}_{i=1}^{s}$$

is a *row vector*. The apostrophe here and in the following indicates the *transposed matrix*. The *diagonal matrix* is

$$\begin{pmatrix} a_1 & 0 & \cdots & 0 \\ 0 & a_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_t \end{pmatrix} = \{_{\mathrm{d}}\, a_i\}_{i=1}^{t}.$$

The same notation is used for partitioned matrices.

The identity matrix, square matrix of ones, vector of ones and matrix of nulls are denoted as

$$\mathbf{I}_a = \{_{\mathrm{d}}\, 1\}_{i=1}^{a},\ \mathbf{J}_a = \{_{\mathrm{m}}\, 1\}_{i=1}^{a},\ \mathbf{1}_a = \{_{\mathrm{c}}\, 1\}_{i=1}^{a} \text{ and } \mathbf{0}_{a\times n} = \{_{\mathrm{m}}\, 0\}_{i=1,}^{a}{}_{j=1}^{n},$$

respectively, where the subindexes a and n show the dimension of the matrix or vector. If it is not necessary, the indexes showing the dimension of matrices are omitted to shorten notation.

The symbols $\oplus$ and $\otimes$ are used respectively for direct sum and direct product of matrices. For example

$$\mathbf{A}_1 \oplus \mathbf{A}_2 = \{_{\mathrm{d}}\, \mathbf{A}_i\}_{i=1}^{2}$$

and

$$\mathbf{A} \otimes \mathbf{B} = \{_{\mathrm{m}}\, a_{ij}\mathbf{B}\}_{i=1,}^{p}{}_{j=1}^{q}, \text{ where } \mathbf{A} = \{_{\mathrm{m}}\, a_{ij}\}_{i=1,}^{p}{}_{j=1}^{q}.$$

The *trace* operation is the sum of the diagonal elements of a square matrix and is denoted by the letters tr. For example the trace of $p \times p$-matrix $\mathbf{A}$ is

$$\mathrm{tr}(\mathbf{A}) = \textstyle\sum_{i=1}^{p} a_{ii}.$$

The number of linearly independent rows or columns is called the *rank* of a matrix and is denoted by the letter r, for example, $\mathrm{r}(\mathbf{A})$.

An *inverse* of a square matrix $\mathbf{A}$, denoted by $\mathbf{A}^{-1}$, is a matrix which when pre- or post-multiplied times the original matrix yields an identity matrix. That is,

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}, \text{ and } \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}.$$

A *general inverse*, denoted as $\mathbf{A}^{-}$, is a matrix that satisfies the following expression

$$\mathbf{A}\mathbf{A}^{-}\mathbf{A} = \mathbf{A}.$$

The scalar λ is called an *eigenvalue* of the $n \times n$-matrix $\mathbf{A}$ if

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$$

for some non-zero vector $\mathbf{x}$. Then the vector $\mathbf{x}$ is an *eigenvector* of the matrix $\mathbf{A}$ corresponding to the eigenvalue λ. If $\mathbf{A}$ is symmetric, then all eigenvalues are real. The eigenvalues are roots of the characteristic equation

$$|\mathbf{A} - \lambda\mathbf{I}| = 0.$$

An $n \times n$ matrix

$$\mathbf{P}_{\mathbf{X},\mathbf{W}} = \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}$$

is called a *projection matrix* with respect to an $n \times n$ symmetric positive definite matrix $\mathbf{W}$ for some $n \times p$ matrix $\mathbf{X}$, because it defines the orthogonal projector onto the column space of $\mathbf{X}$ with respect to the inner product matrix $\mathbf{W}$.

1.1.2. Matrix representation of the linear mixed model

Let $\mathbf{y}$ be the N-vector of observed values of the trait.

Definition 1.1. *We say that*

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{1.1}$$

defines the linear mixed model for $\mathbf{y}$, *where*
$\boldsymbol{\beta}$ *is a* $p \times 1$ *vector of factors known as fixed effects*,
$\mathbf{u} = \{_{\mathrm{c}}\, \mathbf{u}_i\}_{i=1}^{r}$ *is a* $q \times 1$ *vector of factors known as random effects (*$\mathbf{u}_i$ *is a* $q_i \times 1$ *vector of effects of the ith random factor)*,
$\mathbf{e}$ *is a* $N \times 1$ *vector of random residuals*,
$\boldsymbol{X}$ *and* $\mathbf{Z} = \{_{\mathrm{r}}\, \mathbf{Z}_i\}_{i=1}^{r}$ *are known design matrices of order* $N \times p$ *and* $N \times q$ *respectively*. ■

The design matrices $\mathbf{X}$ and $\mathbf{Z}$ describe the precise relationship between the elements of $\boldsymbol{\beta}$ and $\mathbf{u}$ with those of $\mathbf{y}$ ($\mathbf{Z}_i$ is a $N \times q_i$ design matrix which associates effects in $\mathbf{u}_i$ with $\mathbf{y}$).

The expectation of the random variables in the model is:

$$\mathrm{E}\begin{pmatrix} \mathbf{y} \\ \mathbf{u} \\ \mathbf{e} \end{pmatrix} = \begin{pmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \tag{1.2}$$

and the variance-covariance structure is represented as a block-diagonal matrix:

$$\mathrm{Var}\begin{pmatrix} \mathbf{u} \\ \mathbf{e} \end{pmatrix} = \begin{pmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{pmatrix}. \tag{1.3}$$

Consequently,

$$\mathrm{Var}(\mathbf{y}) = \mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R},$$

and

$$\mathrm{Cov}(\mathbf{u}, \mathbf{y}') = \mathbf{Z}\mathbf{G},$$

$$\mathrm{Cov}(\mathbf{e}, \mathbf{y}') = \mathbf{R}.$$

Traditional mixed model assumes that the random effects are independent; this means that the variance-covariance matrices are of the diagonal form:

$$\mathrm{Var}(\mathbf{u}) = \mathbf{G} = \bigoplus_{i=1}^{r} \mathbf{I}_{q_i}\sigma_i^2 = \{_{\mathrm{d}}\, \mathbf{I}_{q_i}\sigma_i^2\}_{i=1}^{r} \tag{1.4}$$

and

$$\mathrm{Var}(\mathbf{e}) = \mathbf{R} = \mathbf{I}_N\sigma_0^2 . \tag{1.5}$$

Defining $\mathbf{u}_0 = \mathbf{e}$ and $\mathbf{Z}_0 = \mathbf{I}_N$, the model (1.1) and $\mathrm{Var}(\mathbf{y})$ can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sum_{i=1}^{r}\mathbf{Z}_i\mathbf{u}_i + \mathbf{e} = \mathbf{X}\boldsymbol{\beta} + \sum_{i=0}^{r}\mathbf{Z}_i\mathbf{u}_i$$

and

$$\mathrm{Var}(\mathbf{y}) = \mathbf{V} = \sum_{i=0}^{r}\mathbf{Z}_i\mathbf{Z}_i'\sigma_i^2 . \tag{1.6}$$

The variances σ_i^2 are called *variance components* because they are the components of the variance of an individual observation.

1.2. Estimation and prediction

1.2.1. Estimation and prediction for known V

Estimation of fixed effects and prediction of realised values of random effects are usually the first major subtasks in linear mixed model analysis. All formulas derived for this purpose assume that the variance-covariance structure of random effects, denoted by the matrix **V**, is known. As discussed in Section 1.2.4 this simplifying assumption is motivated by practice.

In the case of fixed effects the goal is to estimate a set of functions of $\boldsymbol{\beta}$, say $\mathbf{L}'\boldsymbol{\beta}$, using a linear function of the observation vector, say $\mathbf{T}'\mathbf{y}$, where **T** has to be determined. It's easy to show (see, for example, Searle, 1987) that the *Best Linear Unbiased Estimator* (BLUE) of $\mathbf{L}'\boldsymbol{\beta}$ is of the form

$$\mathbf{T}'\mathbf{y} = \mathbf{L}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} = \mathbf{L}'\hat{\boldsymbol{\beta}} .$$

The generalized least square estimate or BLUE of $\boldsymbol{\beta}$ is then

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} . \tag{1.7}$$

If **V** is an identity matrix times a scalar, the generalized least squares estimator is equivalent to ordinary least squares estimator, and the least squares equation for $\hat{\boldsymbol{\beta}}$ can be written as

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} .$$

In the theory of mixed linear models the problem is to predict the function $\mathbf{L}'\boldsymbol{\beta} + \mathbf{M}'\mathbf{u}$. The *Best Linear Unbiased Predictor* (BLUP) of $\mathbf{L}'\boldsymbol{\beta} + \mathbf{M}'\mathbf{u}$ is

$$\mathbf{T}'\mathbf{y} = \mathbf{L}'\hat{\boldsymbol{\beta}} + \mathbf{M}'\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) ,$$

where $\hat{\boldsymbol{\beta}}$ is the BLUE of $\boldsymbol{\beta}$. The realised values of random effects (BLUP of **u**) are predictable from the equations

$$\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) . \tag{1.8}$$

In the normal case, BLUP has the smallest mean square error in the class of all linear, unbiased predictors. However, if $\mathbf{y}$ is not normally distributed, then non-linear predictors of function $\mathbf{L}'\boldsymbol{\beta} + \mathbf{M}'\mathbf{u}$ may exist with smaller mean square error than BLUP.

1.2.2. Mixed model equations

Henderson (1950) developed a set of equations known as the *Mixed Model Equations* (MME) for simultaneous computing of BLUP of $\mathbf{u}$ and BLUE of $\boldsymbol{\beta}$. The main advantage of these equations is that they do not require the inversion of $\mathbf{V}$ of order N, but only the inversion of matrix of a order $p+q$ (the total number of levels of fixed and random effects in the data) is required.

For model (1.1) with expectations (1.2) and variance-covariance structure (1.3) the MME is

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}. \tag{1.9}$$

For traditional mixed model with variance-covariance structure (1.4)-(1.6) the MME is reduced to

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \{\mathrm{d}\, \mathbf{I}_{q_i} \frac{\sigma_0^2}{\sigma_i^2}\}_{i=1}^{r} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{pmatrix}. \tag{1.10}$$

1.2.3. The variances of predictors and prediction errors

Let $\mathbf{Q} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}$ and

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}. \tag{1.11}$$

Then it can be shown (see, Henderson, 1975, for instance) that variances of predictors and prediction errors are

$$\mathrm{Var}(\mathbf{L}'\hat{\boldsymbol{\beta}} + \mathbf{M}'\hat{\mathbf{u}}) = \mathbf{L}'\mathbf{Q}\mathbf{L} + \mathbf{M}'\mathbf{G}\mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{G}\mathbf{M},$$

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \mathrm{Var}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-} = \mathbf{Q},$$

$$\mathrm{Var}(\hat{\mathbf{u}}) = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G} - \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G} = \mathbf{G}\mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{G}, \tag{1.12}$$

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}}') = \mathbf{0},$$

$$\mathrm{Var}(\hat{\mathbf{u}} - \mathbf{u}) = \mathbf{G} - \mathrm{Var}(\hat{\mathbf{u}}) \tag{1.13}$$

and

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}}' - \mathbf{u}') = \mathbf{0} - (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G} = -\mathbf{Q}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G}.$$

The mean square error of BLUP of $\mathbf{u}$ is

$$\mathrm{MSE}(\hat{\mathbf{u}}) = \mathrm{E}(\hat{\mathbf{u}}-\mathbf{u})'(\hat{\mathbf{u}}-\mathbf{u}) = \mathrm{tr}[\,\mathrm{Var}(\hat{\mathbf{u}}-\mathbf{u})] + \underbrace{\mathrm{E}(\hat{\mathbf{u}}-\mathbf{u})'\,\mathrm{E}(\hat{\mathbf{u}}-\mathbf{u})}_{=0} = \mathrm{tr}[\,\mathrm{Var}(\hat{\mathbf{u}}-\mathbf{u})]\,. \tag{1.14}$$

The variances of $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{u}}$ and $\hat{\mathbf{u}}-\mathbf{u}$ can also be obtained directly from the MME. Let a general inverse of the coefficient matrix of the MME (1.9) be

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{pmatrix}^{-} = \begin{pmatrix} \mathbf{C}_{XX} & \mathbf{C}_{XZ} \\ \mathbf{C}_{ZX} & \mathbf{C}_{ZZ} \end{pmatrix}$$

Then the variance-covariance matrix of predictors is

$$\mathrm{Var}\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{C}_{XX} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}-\mathbf{C}_{ZZ} \end{pmatrix},$$

and the variance-covariance matrix of prediction errors is

$$\mathrm{Var}\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}}-\mathbf{u} \end{pmatrix} = \begin{pmatrix} \mathbf{C}_{XX} & \mathbf{C}_{XZ} \\ \mathbf{C}_{ZX} & \mathbf{C}_{ZZ} \end{pmatrix}$$

(see, Henderson, 1984, for instance).

1.2.4. Two-stage estimators and predictors

As mentioned before the equations (1.7) and (1.8) like MME (1.9) yield to the best linear unbiased estimators of $\boldsymbol{\beta}$ and $\mathbf{u}$ only if the variance-covariance matrices of random variables are known without error. In practice this assumption almost never holds and $\mathbf{G}$ and $\mathbf{R}$ are replaced by their estimates resulted with the so-called two-stage or estimated estimators and predictors

$$\mathrm{EBLUE}(\boldsymbol{\beta}) = \tilde{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}$$

and

$$\mathrm{EBLUP}(\mathbf{u}) = \tilde{\mathbf{u}} = \hat{\mathbf{G}}\mathbf{Z}'\hat{\mathbf{V}}^{-1}(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}})\,.$$

Kakwani (1967) suggested the basic idea and Kackar and Harville (1981) showed that if $\mathbf{y}$ has a symmetric distribution and if $\hat{\mathbf{V}}$ is an even function of $\mathbf{y}$, then $\tilde{\boldsymbol{\beta}}$ is an unbiased estimate of $\hat{\boldsymbol{\beta}}$ and $\tilde{\mathbf{u}}$ is an unbiased estimate of $\hat{\mathbf{u}}$. For example, the normal distribution is symmetric and all the variance components estimation methods introduced in Section 1.3 satisfy the second condition.

Kackar and Harville (1984) observed that both $\mathrm{Var}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})$ and $\mathrm{Var}(\tilde{\mathbf{u}}-\mathbf{u})$ can be expressed as the sum of two variances:

$$\mathrm{Var}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) = \mathrm{Var}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) + \mathrm{Var}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}) \tag{1.15}$$

and

$$\mathrm{Var}(\tilde{\mathbf{u}}-\mathbf{u}) = \mathrm{Var}(\hat{\mathbf{u}}-\mathbf{u}) + \mathrm{Var}(\tilde{\mathbf{u}}-\hat{\mathbf{u}})\,, \tag{1.16}$$

respectively. As the second addends in formulas (1.15) and (1.16) depend on the variance components estimation methods and are not exactly expressed, several approximated variances are derived (Kackar and Harville, 1984; Kenward and Roger, 1997; Prasad and Rao, 1990). McCulloch and Searle (2001) aggregated previously published results and presented approximated expressions for scalars $\mathrm{Var}(\mathbf{l}'\tilde{\boldsymbol{\beta}} - \mathbf{l}'\boldsymbol{\beta})$ and $\mathrm{Var}(\mathbf{m}'_*\tilde{\mathbf{u}}_* - \mathbf{m}'_*\mathbf{u}_*)$:

$$\mathrm{Var}(\mathbf{l}'\tilde{\boldsymbol{\beta}} - \mathbf{l}'\boldsymbol{\beta}) \approx \mathbf{l}'\mathbf{Q}\mathbf{l} + \mathbf{l}'\mathbf{Q}\left\{\sum_{i=0}^{r}\sum_{j=0}^{r} d_{ij}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{V}^{-1}\mathbf{X}\right\}\mathbf{Q}\mathbf{l} \qquad (1.17)$$

and

$$\mathrm{Var}(\mathbf{m}'_*\tilde{\mathbf{u}}_* - \mathbf{m}'_*\mathbf{u}_*) \approx \mathbf{m}'_*\left[\mathbf{G}_* - \mathbf{G}_*\mathbf{Z}'_*\mathbf{P}\mathbf{Z}_*\mathbf{G}_*\right]\mathbf{m}_* + \sum_{i=0}^{r}\sum_{j=0}^{r} d_{ij}\left[(\mathbf{m}'_i - \mathbf{m}'_*\mathbf{G}_*\mathbf{Z}'_*\mathbf{P}\mathbf{Z}_i)\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j(\mathbf{m}_j - \mathbf{Z}_j'\mathbf{P}\mathbf{Z}_*\mathbf{G}_*\mathbf{m}_*)\right], \qquad (1.18)$$

where

$$\mathbf{D} = \{{}_{\mathrm{m}}d_{ij}\}_{i,j=0}^{r} = \{{}_{\mathrm{m}}\mathrm{Cov}_{\infty}(\hat{\sigma}_i^2, \hat{\sigma}_j^2)\}_{i,j=0}^{r} \qquad (1.19)$$

is an asymptotic variance-covariance matrix of estimated variance components,

$$\mathbf{u}'_* = \begin{pmatrix}\mathbf{e}' & \mathbf{u}'\end{pmatrix},\ \mathbf{Z}_* = \begin{pmatrix}\mathbf{Z}_0 & \mathbf{Z}\end{pmatrix},\ \mathbf{G}_* = \begin{pmatrix}\mathbf{R} & \mathbf{0}\\ \mathbf{0} & \mathbf{G}\end{pmatrix} \text{ and } \mathbf{m}'_* = \begin{pmatrix}\mathbf{m}'_0 & \mathbf{m}'_1 & \cdots & \mathbf{m}'_r\end{pmatrix}.$$

Approximately unbiased estimators of (1.17) and (1.18) are derived by Kenward and Roger (1987) and McCulloch and Searle (2001), respectively. They investigated the bias studying expected values of two-term Taylor series expansions of expressions (1.17) and (1.18) and resulted estimators of the form

$$\widehat{\mathrm{Var}}(\mathbf{l}'\tilde{\boldsymbol{\beta}} - \mathbf{l}'\boldsymbol{\beta}) \approx \mathbf{l}'\hat{\mathbf{Q}}\mathbf{l} + 2\mathbf{l}'\hat{\mathbf{Q}}\left\{\sum_{i=0}^{r}\sum_{j=0}^{r} d_{ij}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\hat{\mathbf{P}}\mathbf{Z}_j\mathbf{Z}_j'\hat{\mathbf{V}}^{-1}\mathbf{X}\right\}\hat{\mathbf{Q}}\mathbf{l}$$

and

$$\widehat{\mathrm{Var}}(\mathbf{m}'_*\tilde{\mathbf{u}}_* - \mathbf{m}'_*\mathbf{u}_*) \approx \mathbf{m}'_*\left[\hat{\mathbf{G}}_* - \hat{\mathbf{G}}_*\mathbf{Z}'_*\hat{\mathbf{P}}\mathbf{Z}_*\hat{\mathbf{G}}_*\right]\mathbf{m}_* + 2\sum_{i=0}^{r}\sum_{j=0}^{r} d_{ij}\left[(\mathbf{m}'_i - \mathbf{m}'_*\hat{\mathbf{G}}_*\mathbf{Z}'_*\hat{\mathbf{P}}\mathbf{Z}_i)\mathbf{Z}_i'\hat{\mathbf{P}}\mathbf{Z}_j(\mathbf{m}_j - \mathbf{Z}_j'\hat{\mathbf{P}}\mathbf{Z}_*\hat{\mathbf{G}}_*\mathbf{m}_*)\right], \qquad (1.20)$$

where $\hat{\mathbf{Q}}$, $\hat{\mathbf{P}}$ and $\hat{\mathbf{G}}_*$ are calculated with $\hat{\mathbf{V}}$, $\hat{\mathbf{G}}$ and $\hat{\mathbf{R}}$ replacing $\mathbf{V}$, $\mathbf{G}$ and $\mathbf{R}$, respectively. Note that stars in expressions (1.18) and (1.20) denoting the expanded vectors and matrices are erroneously missing in McCulloch and Searle (2001, p 164–171).

More theoretical discussion and proof concerning the accuracy of mean square errors of two-stage estimators and predictors are given by Das, Jiang and Rao (2004).

In the following the results of Kackar and Harville (1984) and McCulloch and Searle (2001) are extended for the variance-covariance matrix of a vector $\mathrm{Var}(\tilde{\mathbf{u}}-\mathbf{u})$.

Corollary 1.1. *In the traditional linear mixed model with variance-covariance structure* (1.4)-(1.6) *the* $\mathrm{Var}(\tilde{\mathbf{u}}-\mathbf{u})$ *is approximately expressed as*

$$\begin{aligned}\mathrm{Var}(\tilde{\mathbf{u}}-\mathbf{u}) &\approx \mathbf{G}-\mathbf{G}\mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{G} \\ &+\sum_{i=0}^{r}\mathrm{Var}(\hat{\sigma}_i^2)\left[(\mathbf{E}_i'-\mathbf{G}\mathbf{Z}'\mathbf{P}\mathbf{Z}_i)\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_i(\mathbf{E}_i-\mathbf{Z}_i'\mathbf{P}\mathbf{Z}\mathbf{G}\right],\end{aligned} \tag{1.21}$$

where $\mathbf{E}_0=\mathbf{0}_{N\times q}$ *and* $\mathbf{E}_i$, $i=1,\ldots,r$, *is a* $q_i\times q$ *block matrix with* i^{th} *column block equals to* $\mathbf{I}_{q_i}$ *and the rest zeros*:

$$\mathbf{E}_i=\begin{cases}\mathbf{0}_{N\times q}, i=0,\\ \left(\mathbf{0}_{\sum_{j=1}^{i-1}q_j}\ \ \mathbf{I}_{q_i}\ \ \mathbf{0}_{\sum_{j=i+1}^{r}q_j}\right), i>0.\end{cases}$$

Proof. To find the variance-covariance matrix of prediction errors instead of a single variance, the $N+q\times N+q$ block matrix $\mathbf{M}_*=\{{}_{\mathrm{m}}\mathbf{M}_{ij}\}_{i,j=0}^{r}$ instead of $N+q\times 1$ vector $\mathbf{m}_*$ must be used in expression (1.18). If we set the matrix $\mathbf{M}_*$ equal to the identity matrix $\mathbf{I}_{N+q}$, the approximated variance-covariance matrix of $\tilde{\mathbf{u}}_*-\mathbf{u}_*$ is expressed as

$$\begin{aligned}\mathrm{Var}(\tilde{\mathbf{u}}_*-\mathbf{u}_*) &\approx \mathbf{G}_*-\mathbf{G}_*\mathbf{Z}_*'\mathbf{P}\mathbf{Z}_*\mathbf{G}_* \\ &+\sum_{i=0}^{r}\mathrm{Var}(\hat{\sigma}_i^2)\left[(\mathbf{M}_i'-\mathbf{G}_*\mathbf{Z}_*'\mathbf{P}\mathbf{Z}_i)\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_i(\mathbf{M}_i-\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_*\mathbf{G}_*)\right],\end{aligned}$$

where $\mathbf{M}_*'=\left(\mathbf{M}_0'\ \ \mathbf{M}_1'\ \cdots\ \mathbf{M}_r'\right)$ and $\mathbf{M}_i=\left(\mathbf{0}_{\sum_{j=0}^{i-1}q_j}\ \ \mathbf{I}_{q_i}\ \ \mathbf{0}_{\sum_{j=i+1}^{r}q_j}\right)$.

Using partitions of $\mathbf{u}_*$, $\mathbf{G}_*$, $\mathbf{Z}_*$ and $\mathbf{M}_i$, and also considering that $\mathbf{M}_0=\left(\mathbf{I}_N\ \ \mathbf{0}_{N\times q}\right)$, we can write the expression of $\mathrm{Var}(\tilde{\mathbf{u}}_*-\mathbf{u}_*)$ after some matrix algebra as follows:

$$\begin{aligned}&\begin{pmatrix}\mathrm{Var}(\tilde{\mathbf{e}}-\mathbf{e}) & \mathrm{Cov}(\tilde{\mathbf{e}}-\mathbf{e},\tilde{\mathbf{u}}'-\mathbf{u}')\\ \mathrm{Cov}(\tilde{\mathbf{e}}-\mathbf{e},\tilde{\mathbf{u}}'-\mathbf{u}')' & \mathrm{Var}(\tilde{\mathbf{u}}-\mathbf{u})\end{pmatrix}\approx\begin{pmatrix}\mathbf{R} & \mathbf{0}\\ \mathbf{0} & \mathbf{G}\end{pmatrix}-\begin{pmatrix}\mathbf{R}\mathbf{Z}_0'\mathbf{P}\mathbf{Z}_0\mathbf{R} & \mathbf{R}\mathbf{Z}_0'\mathbf{P}\mathbf{Z}\mathbf{G}\\ \mathbf{G}\mathbf{Z}'\mathbf{P}\mathbf{Z}_0\mathbf{R} & \mathbf{G}\mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{G}\end{pmatrix}\\ &+\mathrm{Var}(\hat{\sigma}_0^2)\\ &\times\begin{pmatrix}(\mathbf{I}_N-\mathbf{R}\mathbf{Z}_0'\mathbf{P}\mathbf{Z}_0)\mathbf{Z}_0'\mathbf{P}\mathbf{Z}_0(\mathbf{I}_N-\mathbf{Z}_0'\mathbf{P}\mathbf{Z}_0\mathbf{R}) & (\mathbf{I}_N-\mathbf{R}\mathbf{Z}_0'\mathbf{P}\mathbf{Z}_0)\mathbf{Z}_0'\mathbf{P}\mathbf{Z}_0(\mathbf{0}_{N\times q}-\mathbf{Z}_0'\mathbf{P}\mathbf{Z}\mathbf{G})\\ (\mathbf{0}_{q\times N}-\mathbf{G}\mathbf{Z}'\mathbf{P}\mathbf{Z}_0)\mathbf{Z}_0'\mathbf{P}\mathbf{Z}_0(\mathbf{I}_N-\mathbf{Z}_0'\mathbf{P}\mathbf{Z}_0\mathbf{R}) & (\mathbf{0}_{q\times N}-\mathbf{G}\mathbf{Z}'\mathbf{P}\mathbf{Z}_0)\mathbf{Z}_0'\mathbf{P}\mathbf{Z}_0(\mathbf{0}_{N\times q}-\mathbf{Z}_0'\mathbf{P}\mathbf{Z}\mathbf{G})\end{pmatrix}\\ &+\sum_{i=1}^{r}\mathrm{Var}(\hat{\sigma}_i^2)\\ &\times\begin{pmatrix}(\mathbf{0}_N-\mathbf{R}\mathbf{Z}_0'\mathbf{P}\mathbf{Z}_i)\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_i(\mathbf{0}_N-\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_0\mathbf{R}) & (\mathbf{0}_N-\mathbf{R}\mathbf{Z}_0'\mathbf{P}\mathbf{Z}_i)\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_i(\mathbf{E}_i-\mathbf{Z}_i'\mathbf{P}\mathbf{Z}\mathbf{G})\\ (\mathbf{E}_i'-\mathbf{G}\mathbf{Z}'\mathbf{P}\mathbf{Z}_i)\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_i(\mathbf{0}_N-\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_0\mathbf{R}) & (\mathbf{E}_i'-\mathbf{G}\mathbf{Z}'\mathbf{P}\mathbf{Z}_i)\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_i(\mathbf{E}_i-\mathbf{Z}_i'\mathbf{P}\mathbf{Z}\mathbf{G})\end{pmatrix},\end{aligned}$$

where $\mathbf{E}_i=\left(\mathbf{0}_{\sum_{j=1}^{i-1}q_j}\ \ \mathbf{I}_{q_i}\ \ \mathbf{0}_{\sum_{j=i+1}^{r}q_j}\right)$.

To finish the proof we must look at the lower right block of the derived block matrix, which represents the sampling variance of two-stage predictors $\tilde{\mathbf{u}}$. Defining $\mathbf{E}_0 = \mathbf{0}_{N\times q}$ we got that

$$\begin{aligned}\mathrm{Var}(\tilde{\mathbf{u}} - \mathbf{u}) \approx\ & \mathbf{G} - \mathbf{G}\mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{G} \\ & + \mathrm{Var}(\hat{\sigma}_0^2)\left[(\mathbf{E}_0' - \mathbf{G}\mathbf{Z}'\mathbf{P}\mathbf{Z}_0)\mathbf{Z}_0'\mathbf{P}\mathbf{Z}_0(\mathbf{E}_0 - \mathbf{Z}_0'\mathbf{P}\mathbf{Z}\mathbf{G}\right] \\ & + \sum_{i=1}^{r}\mathrm{Var}(\hat{\sigma}_i^2)\left[(\mathbf{E}_i' - \mathbf{G}\mathbf{Z}'\mathbf{P}\mathbf{Z}_i)\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_i(\mathbf{E}_i - \mathbf{Z}_i'\mathbf{P}\mathbf{Z}\mathbf{G}\right] \\ =\ & \mathbf{G} - \mathbf{G}\mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{G} + \sum_{i=0}^{r}\mathrm{Var}(\hat{\sigma}_i^2)\left[(\mathbf{E}_i' - \mathbf{G}\mathbf{Z}'\mathbf{P}\mathbf{Z}_i)\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_i(\mathbf{E}_i - \mathbf{Z}_i'\mathbf{P}\mathbf{Z}\mathbf{G}\right],\end{aligned}$$

which establishes the proof. ■

1.2.5. Variance component estimation

Nowadays several difference variance components estimation methods are usually included in statistical packages. The reason for their plurality is that there is no one method best for all cases. For discussion concerning comparisons of the different algorithms look, for example, at Searle, et al (1992). The results of simulation studies performed to compare different variance components estimation methods are also present in Swallow and Monahan (1984) and Kaart (1998).

For simple models, usually the Analysis Of Variance (ANOVA) method, generally known as Henderson's Method III (Henderson, 1953), is applied. Method III uses the reductions in sums of squares by fitting the submodels of the full model. The reductions are expressed as quadratic forms and the variance components estimators are obtained by equating these quadratic forms to their expected values, which are functions of the unknown variance components. This approach is used in Section 3 of the present dissertation studying the one-way random model.

The other group of variance components estimation methods are the so-called criteria-based methods. The three main criteria which determine the quadratic forms used in the estimation of the variance components are the unbiasedness of the estimates, the translation invariance of the quadratic forms in relation to the fixed effects, and the minimum variance (or norm, or mean square) of the estimates. The methods satisfying these criteria were first published independently by C. R. Rao and by L. R. LaMotte in 1970's. Rao derived the variance components estimators both for the normally and the non-normally distributed **y**, called Minimum Variance Quadratic Unbiased Estimators (MIVQUE or MINVAR; Rao, 1970), and Minimum Norm Quadratic Unbiased Estimators (MINQUE; Rao, 1971). LaMotte also derived MIVQUE (LaMotte, 1970). When **y** is normally distributed, then these two methods, MIVQUE and MINQUE, are equivalent. The main shortages of the criteria-based methods are that the esti-

mators are quite laborious to find and greatly depend on the pre-assigned values.

The most applied variance components estimation methods today are the Maximum Likelihood (ML) method and the Restricted (or residual) Maximum Likelihood (REML) method. The ML approach in variance components estimation was first described by Hartley and Rao in 1967 (Hartley, Rao, 1967). Contrary to the ANOVA method, the ML estimation needs the underlying probability distribution of the data. A natural choice is the multivariate normal distribution. Assume that the *density function* of the data vector $\mathbf{y}$ is expressed as

$$f(\mathbf{y}) = \frac{\exp[-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})]}{(2\pi)^{N/2}\,|\,\mathbf{V}\,|^{1/2}}.$$

The corresponding *likelihood function* as a function of the parameters $\boldsymbol{\beta}$ and $\mathbf{V}$ is

$$L = L(\boldsymbol{\beta}, \mathbf{V}\,|\,\mathbf{y}) = \frac{\exp[-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})]}{(2\pi)^{N/2}\,|\,\mathbf{V}\,|^{1/2}}.$$

The idea of ML estimation is to find the $\boldsymbol{\beta}$ and $\mathbf{V}$ that maximize the likelihood (1.11) subject to them falling within the parameter space. Maximization of $L(\boldsymbol{\beta}, \mathbf{V}\,|\,\mathbf{y})$ can be achieved by maximizing the *log-likelihood function*

$$l = \ln L = -\tfrac{1}{2}N\ln(2\pi) - \tfrac{1}{2}\ln(|\,\mathbf{V}\,|) - \tfrac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}).$$

As the ML estimation does not take into account the information (degrees of freedom) lost in estimating fixed effects, then the REML method avoids this property by using the likelihood function of a set of *error contrasts* denoted by $\mathbf{K}'\mathbf{y}$ instead of the likelihood function of $\mathbf{y}$. The rows of $\mathbf{K}'$ are determined so that $\mathbf{K}'\mathbf{X} = \mathbf{0}$ and $\mathbf{K}'$ has the full row rank: $\mathrm{r}(\mathbf{K}) = N - \mathrm{r}(\mathbf{X})$. This method was derived for general mixed models in Patterson and Thompson (1971).

Using error contrasts, the maximized log-likelihood function is

$$l_{\mathrm{R}} = l(\mathbf{V}\,|\,\mathbf{K}'\mathbf{y}) = -\tfrac{1}{2}[N - \mathrm{r}(\mathbf{X})]\ln(2\pi) - \tfrac{1}{2}\ln(|\,\mathbf{K}'\mathbf{V}\mathbf{K}\,|) - \tfrac{1}{2}\mathbf{y}'\mathbf{K}(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}\mathbf{K}'\mathbf{y}. \tag{1.22}$$

Taking the derivative of the log-likelihood function with respect to σ_i^2 for $i = 0, 1, ..., r$ is

$$\frac{\partial l_{\mathrm{R}}}{\partial \sigma_i^2} = -\tfrac{1}{2}\,\mathrm{tr}[(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}\mathbf{K}'\mathbf{Z}_i\mathbf{Z}_i'\mathbf{K}] + \tfrac{1}{2}\mathbf{y}'\mathbf{K}(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}\mathbf{K}'\mathbf{Z}_i\mathbf{Z}_i'\mathbf{K}(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}\mathbf{K}'\mathbf{y}.$$

Taking the derivatives equal to zero and using the identity

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1} = \mathbf{K}(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}\mathbf{K}'$$

gives the REML equations

$$\{_{\mathrm{c}}\,\mathrm{tr}(\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i')\}_{i=0}^{r} = \{_{\mathrm{c}}\,\mathbf{y}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{y})\}_{i=0}^{r}. \tag{1.23}$$

An alternative form of the equations (1.23) is

$$\{_{\mathrm{m}}\,\mathrm{tr}(\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j')\}_{i,j=0}^{r}\boldsymbol{\sigma}^2 = \{_{\mathrm{c}}\,\mathbf{y}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{y})\}_{i=0}^{r}. \tag{1.24}$$

One attractive feature of ML and REML estimation is that the *asymptotic dispersion matrix* of the estimators is always available. It equals to the inverse of the *information matrix*, the elements of which are defined as the negative expected values of second derivatives with respect to $\boldsymbol{\sigma}^2 = \{_{\mathrm{c}}\sigma_i^2\}_{i=0}^{r}$ (and $\boldsymbol{\beta}$, in ML). Therefore the information matrix corresponding to the REML method is

$$\mathrm{I}(\boldsymbol{\sigma}^2) = -\mathrm{E}\left[\partial^2 l_{\mathrm{R}}/\partial\boldsymbol{\sigma}^2\partial(\boldsymbol{\sigma}^2)'\right] = \tfrac{1}{2}[\{_{\mathrm{m}}\,\mathrm{tr}(\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j')\}_{i,j=0}^{r}].$$

Asymptotic variance of REML estimators is therefore of the form:

$$\mathrm{Var}(\hat{\sigma}^2_{\mathrm{REML}}) \approx 2[\{_{\mathrm{m}}\,\mathrm{tr}(\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j')\}_{i,j=0}^{r}]^{-1}. \tag{1.25}$$

There are no analytical expressions available for the variance component estimators produced by ML or REML. In both cases one needs to use numerical iterative techniques to derive the estimates. The simplest algorithm is the *EM algorithm*, which guarantees that the iterations will always remain in the parameter space. On the other hand, the convergence is very, very slow, and it may not converge in some cases.

Usually, to increase the speed of convergence, *gradient methods* in the numerical analysis are used to maximize a non-linear function. There are two main algorithms used in variance components estimations. These are the *Newton-Raphson algorithm* and the *method of scoring algorithm*.

In animal breeding, where the dimensions of variance-covariance matrices are equal or even bigger than the number of observations, there are two other algorithms used with the REML method. Both of these methods avoid the inversing of high dimensional matrices. One of them is called the *derivative free REML* (DF REML), proposed by Smith and Gaser (1986) and Graser, Smith and Tier (1987), and transforms the log-likelihood function into the form which enables to find its maximum based on the Gaussian elimination of properly constructed coefficient matrices. The other method, called the *average information REML* (AI REML), is by nature a gradient method that finds the iteration step direction based on the average of the observed and expected information matrices using, again, the Gaussian elimination of properly constructed coefficient matrices (Johnson and Thompson, 1995).

For discussion concerning the comparison of EM, Newton-Rhapson and the method of scoring algorithms look at Searle, et al (1992, p 312), for example. The algorithms used in the animal breeding programs are compared in Hofer (1998).

CHAPTER 2
THE APPLICATIONS OF THE MIXED MODEL IN ESTIMATING GENETIC PARAMETERS

2.1. Introduction

Every phenotypic value of an individual is determined by environmental and genetic factors. In population genetics the basic model reflecting this fact is of the form

$$P = \bar{P} + G + E, \tag{2.1}$$

where P, $\bar{P}$, G and E are the observed phenotypic value, the average phenotypic value in the examined population and the unknown genetic and environmental effects, respectively. Presently there exist many modifications of this basic model where both genetic and environmental effects are presented as the sums of different influences.

There are two different ways to write down the models used in genetics. One includes the so-called *genetic models* or the *models in genetic notation*, where the genetic effect is divided into so many parts as is known from the genetic science to be relevant for current analysis. That way the genetic model forms the ideal model for the researcher.

The other model is the *mathematical* (or *statistical*) *model* or the *model in statistical notation*, which is based on the genetic model and on the really measurable effects in keeping with analysed data structure. In many situations the terms estimable from the statistical model do not match with the genetic model parameters – some genetic model parameters can be expressed as functions of the mathematical model parameters and some can not be uniquely separated based on the mathematical model parameters.

Understanding of both the genetic and mathematical models is required to collect the data with proper structure, to select the right mathematical model for data analysis and to interpret the statistical results properly.

As the genetic effect usually collects the effects of many single genes, and in different studies many different individuals are analysed, it is common to consider the genetic effect to be random and to measure the magnitude of the genetic effect via corresponding variance components.

The variance-covariance structure of observed values and studied effects also forms the basis of distinguishing between models. From a statistician's viewpoint the models can be separated into two groups: models with a simple diagonal variance-covariance structure and models with a generally unstructured (in the mathematical sense) variance-covariance structure.

Taking into account the source of genetic influence, additional subgrouping should be done. The basic applications of the traditional linear mixed model with

a simple diagonal variance-covariance structure in biometrical genetics are *half-* and *full-sib models* examining polygenetic effects inherited only from parents. The applications of models with a general variance-covariance structure are the so-called *animal models* and their extensions evaluating polygenetic effects inherent to the analysed individual's self and involving all known pedigree information; and models representing an integration between molecular and biometrical genetics, allowing to isolate single genes and to elucidate their phenotypic actions based on the sequential exploitation of models over all genome, reflecting the inheritance of single alleles via elements of the variance-covariance matrix.

All these models have many supplementary variants and are applicable in a one- or multidimensional situation.

The primary parameters describing the effects of random factors are the variance components. To measure the magnitude of random effects, variance components ratios to the total variance, called *intraclass correlation coefficients*, are used. In the following chapters the intraclass correlation coefficient is noted as ρ.

In genetic studies the percentage of genetic contribution on the observed trait, called *heritability coefficient* and denoted as h^2, is calculated. It is shown in the next sections that depending on the genetic model, the intraclass correlation is representing different types or proportions of heritabilities.

Based on known or estimated values of variance components or heritabilities, the realized values of random genetic effects are estimable. In animal breeding, where most of the traits the animals are selected for, are assumed to be affected by many genes, all with a small effect, the relevant effects are polygenetic ones, measuring the part of the observed phenotypic values heritable from the parents to the progeny. These polygenetic heritable effects are called *breeding values*. Besides the polygenetic effects, also the effects on the level of single chromosomal regions or even based on the single nucleotides can be estimated by specific models – this has already been experimented in animal breeding for some years and has given one supplement analysis method for human genetics as well.

Fixed environmental effects are usually not of interest in genetic studies. These are included in the model only to consider the potential non-genetic differences between studied individuals. Due to this, fixed effects are avoided in genetic models and are not specified in theoretical mathematical models.

In the following the basic genetic models and their statistical analogues are introduced, the commonly estimated genetic parameters are formulated, and the problems and genetic background are shortly discussed. All considered models are – with minor modifications – also discussed in Kaart (2001).

In the present chapter we shall denote the effects in genetic models with italic capital letters and the effects in statistical models with italic minuscule letters. To show the identity of the genetic parameters separable from the genetic set-up of the model and the estimable effects (or their functions) of the statistical model, the sign $\hat{=}$ is used.

2.2. The half- and full-sib models for estimating polygenetic effects

2.2.1. The half-sib model

The earliest applications of mixed models in animal breeding, especially in dairy cattle, were based on the half-sib model. In this model the analysed animals are grouped only by one parent, the genetic contribution of the second parent is considered inappreciable. As sires in animal husbandry commonly have a large number of progeny, the evaluated parent is the sire, and the model is also called the *sire model*.

As half of the alleles in the individual genotype come from the sire and half from the mother, the genotypic effect G in the model (2.1) can be substituted with the sum $\frac{1}{2}A_S + \frac{1}{2}A_D + MS$, where A_S and A_D are breeding values (sums of allelic values) of the sire and the dam respectively, and MS is the Mendelian sampling effect, accounting for the sampling that occurs in the formation of gametes at meiosis and describing the deviation of the individual additive genetic value from the average additive genetic value of its parents. All non-additive genetic effects like the dominance and epistatic effects, for example, are dealt as parts of the environmental effect because their influence is not ascertainable in half-sib analysis and is said to be remote.

The simplest sire model is assuming that all sires are mated randomly to dams and dams are mated to only one sire and have just one progeny with a measured record. The sire model in genetic notation for jth progeny of ith sire is

$$P_{ij} = \bar{P} + \tfrac{1}{2}A_{S_i} + \tfrac{1}{2}A_{D_{ij}} + MS_{ij} + E_{ij}\,, \tag{2.2}$$

where P_{ij}, MS_{ij} and E_{ij} are the observed phenotypic value, unknown Mendelian sampling effect and environmental effect corresponding to the jth progeny of ith sire, respectively, $\bar{P}$ is the average phenotypic value in the examined population, A_{S_i} and $A_{D_{ij}}$ are unknown breeding values of ith sire and dam of jth progeny of ith sire, respectively.

The phenotypic variance is expressed as a sum of two components, additive genetic and environmental variance. By the assumption of model (2.2), the extended formula for phenotypic variance is of the form

$$\begin{aligned}\sigma_P^2 &= \mathrm{Var}(\tfrac{1}{2}A_{S_i}) + \mathrm{Var}(\tfrac{1}{2}A_{D_{ij}}) + \mathrm{Var}(MS_{ij}) + \mathrm{Var}(E_{ij}) \\ &= \tfrac{1}{4}\sigma_A^2 + \tfrac{1}{4}\sigma_A^2 + \tfrac{1}{2}\sigma_A^2 + \sigma_E^2\,.\end{aligned} \tag{2.3}$$

So, the Mendelian sampling generates one half of the additive genetic variance and the phenotypic variance includes only one quarter of the additive genetic variance caused by the sire.

In the statistical model the phenotypic value y_{ij} of jth progeny of ith the sire is presented as the sum of fixed effects (population mean, age, sex, birthplace and so on) expressed via the product of fixed effects vector $\boldsymbol{\beta}$ and an incidence vec-

tor $\mathbf{x}_{ij}$ relating records of observed progeny to the fixed effects, the random effect of *i*th sire s_i and the random residual effect e_{ij} :

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + s_i + e_{ij} \,. \tag{2.4}$$

The last term e_{ij} includes the possible effect of the dam and also the Mendelian sampling. Comparing the parts of random effects in models (2.2) and (2.4), it follows that statistically estimable are only the realized values of random sire effects (in genetic terms the *transmission abilities* of sires):

$$s_i \triangleq \tfrac{1}{2} A_{S_i} \,.$$

As the independence of genotype and environment, the independence of records, similar environmental variance, and the unrelated sires are assumed, then

$$\mathrm{Var}\begin{pmatrix}\mathbf{s}\\ \mathbf{e}\end{pmatrix} = \begin{pmatrix}\mathbf{I}_a\sigma_s^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_N\sigma_e^2\end{pmatrix},$$

where *N* denotes the number of individuals and *a* the number of sires. On these conditions the phenotypic variance is expressed as a simple sum of variance components caused by the sire and random environment:

$$\sigma_y^2 = \sigma_s^2 + \sigma_e^2 \,.$$

The usual intraclass correlation coefficient calculated from sire model is expressed as

$$\rho = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2} \,. \tag{2.5}$$

In genetic studies the fraction of total phenotypic variance attributable to additive genetic differences among individuals, known as the heritability coefficient, is of interest. Based on the genetic model, the heritability coefficient is defined as

$$h^2 = \sigma_A^2 / \sigma_P^2 \,. \tag{2.6}$$

As the variance caused by sires constitutes only one quarter of the whole additive genetic variance: $\sigma_s^2 \triangleq \tfrac{1}{4}\sigma_A^2$, then the heritability coefficient is calculable from the statistical model by the formula

$$h^2 = \frac{4\sigma_s^2}{\sigma_s^2 + \sigma_e^2} \,. \tag{2.7}$$

It is evident that $h^2 = 4\rho$ in the sire model.

Although the genetic assumptions of the sire model are never absolutely satisfied, this model is widely used, especially in pilot studies in populations where the exact pedigrees are not known or values of the interested traits are registered only in the last generation. The main reason is the mathematical simplicity of the model, which makes it possible to prepare appropriate datasets with minimal effort and to quickly get preliminary results from statistical analysis. Due to this

the half-sib model is the basic application of the traditional linear mixed model in estimating genetic parameters. The accuracy of the genetic parameters estimators found by the sire model is discussed in Chapter 3.

2.2.2. The full-sib model

The other frequently discussed application of the traditional linear mixed model in population genetics is the full-sib model. In this situation, both the sire and the dam are known, but as in the sire model, neither have any records of their own. Usually dams are nested within sires, and there can be only one record per offspring.

For the full-sib model it is necessary to express that the same full sib group resemble each other for at least three reasons. These are: the same sire, the same dam, and the same prenatal and often also postnatal environments, which together are denoted as the *common environment*. The relevant genetic model describing an observation on the *k*th individual P_{ijk} in a full sib group with sire effect S_i and dam effect D_{ij}, subject to an environmental effect common to all the full sibs $E_{C_{ij}}$ and subject to individual specific Mendelian sampling effect MS_{ijk} and environmental effect E_{ijk}, is

$$P_{ijk} = \overline{P} + \tfrac{1}{2} A_{S_i} + \tfrac{1}{2} A_{D_{ij}} + MS_{ijk} + E_{C_{ij}} + E_{ijk} .$$

The phenotypic variance is presented respectively as

$$\begin{aligned}\sigma_P^2 &= \mathrm{Var}(\tfrac{1}{2} A_{S_i}) + \mathrm{Var}(\tfrac{1}{2} A_{D_{ij}}) + \mathrm{Var}(MS_{ijk}) + \mathrm{Var}(E_{C_{ij}}) + \mathrm{Var}(E_{ijk}) \\ &= \tfrac{1}{4}\sigma_A^2 + \tfrac{1}{4}\sigma_A^2 + \tfrac{1}{2}\sigma_A^2 + \sigma_{E_C}^2 + \sigma_E^2 .\end{aligned}$$

The equation of the model in statistical notation is

$$y_{ijk} = \mathbf{x}'_{ijk}\boldsymbol{\beta} + s_i + d_{ij} + e_{ijk} .$$

The only estimable random effects are the sire effects s_i and the dam effects d_{ij}. Thereby, $s_i \triangleq \tfrac{1}{2} A_{S_i}$, as in the sire model, and the dam effect contains one half of the additive genetic effect plus a common environmental effect – $d_{ij} \triangleq \tfrac{1}{2} A_{D_{ij}} + E_{C_{ij}}$.

Since the random effects are again assumed to be mutually independent and the sires and dams unrelated, the variance-covariance structure of the random effects is expressed as

$$\mathrm{Var}\begin{pmatrix}\mathbf{s}\\ \mathbf{d}\\ \mathbf{e}\end{pmatrix} = \begin{pmatrix}\mathbf{I}_a\sigma_s^2 & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_b\sigma_d^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{I}_N\sigma_e^2\end{pmatrix},$$

where N denotes the number of individuals, a the number of sires and b the number of dams per sire. The phenotypic variance is expressed as the sum of the variance components caused by the sire, the dam and the random environment:

$$\sigma_y^2 = \sigma_s^2 + \sigma_d^2 + \sigma_e^2 .$$

Many different proportions of variance components are possible to find in the full-sib analysis, but only two of them are commonly used in genetic studies. The prime of these proportions is, again, the heritability coefficient, which can be calculated in three different ways. The most reliable estimate is got from the sire component with the equation

$$h^2 = \frac{4\sigma_s^2}{\sigma_s^2 + \sigma_d^2 + \sigma_e^2} .$$

The other possible formulas $h^2 = 2(\sigma_s^2 + \sigma_d^2)/\sigma_y^2$ and $h^2 = 4\sigma_d^2/\sigma_y^2$ produce biased estimates, because in both of them the numerator is involving non-genetic variance caused by the common environment – $\sigma_d^2 \hat{=} \frac{1}{4}\sigma_A^2 + \sigma_{EC}^2$.

The second estimable proportion is measuring the part of the variability caused by the common environment in the total variance:

$$c^2 = \frac{\sigma_d^2 - \sigma_s^2}{\sigma_s^2 + \sigma_d^2 + \sigma_e^2} \hat{=} \frac{\sigma_{EC}^2}{\sigma_P^2} .$$

In the genetic form of the full-sib model, sometimes the dominance effect – common to all the full sibs and also measuring the effect of allele pairs – is pointed out. But as in the statistical model, it is – analogously to the common environmental effect – included in the dam effect and is not separately estimable, its influence is said to be remote.

All the mathematical theory evolved for the half-sib model is easily expanded to the full-sib model. Due to this and due to the higher genetic complexity of the full-sib models when compared with the sire model, in theoretical studies the latter is used as the application of the traditional linear mixed model in genetic studies.

2.3. The animal model

2.3.1. The simple animal model

In genetic notation the animal model is the phenotypic value measured on the *i*th individual expressed as the sum of the average phenotypic value in the examined population, the additive genetic effect (the breeding value in animal breeding) of the concerned individual and the environmental effect:

$$P_i = \bar{P} + A_i + E_i .$$

Thus, the only difference from the basic genetic model (2.1) is in the use of the additive genetic effect instead of the total genetic merit.

The statistical model includes the same addends as the genetic model, all known fixed environmental effects that influenced the individual's i phenotype are included in first term expressed as $\mathbf{x}_i'\boldsymbol{\beta}$, terms a_i and e_i incorporates all individual's i additive genetic effects and residual individual-specific effects, respectively:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + a_i + e_i . \tag{2.8}$$

The analogous congruity holds for variances, too. The phenotypic variance is the sum of the additive genetic variance and the environmental residual variance, and is expressed based on the statistical model (2.8) as

$$\sigma_y^2 = \sigma_a^2 + \sigma_e^2 . \tag{2.9}$$

From the model (2.8) it is clear that based on the traditional mixed model with covariance structure (1.3)–(1.6) assuming the independence of the observations, it is not possible to separate the additive genetic and random residual effects. In this situation all realised values of the random effects predicted after considering the known fixed effects denote the sum of the genetic and residual effects, and are useless for practical applications.

The additional information needed for determining the proportion of the additive genetic effect for each individual concurs with using the relationships between analysed individuals. From the fact that two individuals are related, it follows that the covariance between them caused by genetic factors differs from zero. It is common in genetic studies that all individuals with unknown genealogy are assumed to be sampled from a single population with an average genetic effect of zero and the common variance σ_a^2 and are expected to be not related. Such individuals constitute the *base population* and are called in animal breeding as *foundation animals*.

As previously, the basic model examines only the additive genetic part of the genetic variation between individuals and the effects of alleles' interaction in one locus (dominance) and genes' interaction (epistatis) are said to be remote, which, considering only the polygenetic effect, seems to be non-restrictive.

The *additive genetic relationship* between two individuals is the probability that those two individuals have alleles in common which are *identical by descent*. These probabilities are considered in the model (2.8)–(2.9) through the *additive genetic relationship matrix* (or *numerator relationship matrix*). This matrix, usually noted as **A**, is symmetric and has one row and column for each individual in the examined pedigree. Its diagonal element, a_{ii}, represents twice (each gene has two copies) the probability that two gametes taken at random from individual i will carry identical alleles by descent and is equal to $1 + F_i$, where F_i is the inbreeding coefficient of individual i (Wright, 1922). The offdiagonal element, a_{ij}, equals to the numerator of the coefficient of the relationship (Wright, 1922) between individuals i and j, and has the value ½ for full-sibs and parent-progeny pairs, ¼ for half-sibs and grandparent-progeny pairs and so on.

Based on the facts that the additive genetic relationship between two individuals is the average of the relationships between one of them and the parents of the other, and that the inbreeding of an individual is half the additive genetic relationship between its parents, Henderson (1976) described a recursive method for computing the additive genetic relationship matrix. By this method, individuals in the pedigree are coded 1 to *N*, the number of them in the pedigree, and are ordered in such a way that the parents precede their progeny, firstly. Then the following rules are exerted to compute **A**:

- ✓ if both parents of individual *i* are unknown and are assumed unrelated, then

$$a_{ij} = a_{ji} = 0 \text{ for } j = 1 \text{ to } (i-1), \text{ and } a_{ii} = 1;$$

- ✓ if only one parent (for example sire *s*) is known and assumed unrelated to the mate, then

$$a_{ij} = a_{ji} = 0.5\, a_{js} \text{ for } j = 1 \text{ to } (i-1), \text{ and } a_{ii} = 1;$$

- ✓ if both parents (*s* and *d*) of individual *i* are known, then

$$a_{ij} = a_{ji} = 0.5\,(a_{js} + a_{jd}) \text{ for } j = 1 \text{ to } (i-1), \text{ and } a_{ii} = 1 + 0.5\, a_{sd}.$$

The defined relationship matrix is used on the determining of the variance-covariance structure of random effects in the model (2.8) as

$$\mathrm{Var}\begin{pmatrix}\mathbf{a}\\ \mathbf{e}\end{pmatrix} = \begin{pmatrix}\mathbf{A}\sigma_a^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_N\sigma_e^2\end{pmatrix}. \tag{2.10}$$

The only calculable proportion from the animal model based on the equation (2.9) is the heritability coefficient:

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2},$$

which equals to the intraclass correlation coefficient.

As reviewed in Chapter 1.2, the estimation of the realised values of random effects and variance components need the inverse of $\mathrm{Var}(\mathbf{y}) = \mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$ or $\mathrm{Var}(\mathbf{a}) = \mathbf{G} = \mathbf{A}\sigma_a^2$. Both of these inverses include the inversion of the additive genetic relationship matrix **A**, which has in a mathematical sense quite a general structure and has mostly too big a dimension to employ straight inversion procedures.

In 1976, Henderson (Henderson, 1976) presented the first strategy for calculating $\mathbf{A}^{-1}$ without the direct inversion of **A** ignoring inbreeding. In the same year Quaas (Quaas, 1976) elaborated this method for general relationships between individuals. Based on these studies the inversion of the additive genetic relationship matrix is based on the identity $\mathbf{A} = \mathbf{TDT}'$, where **T** is a lower triangular matrix that traces the flow of genes from one generation to the other, and **D** is a diagonal matrix measuring the proportion of the additive genetic effect not explained by the pedigree. As $\mathbf{A}^{-1} = (\mathbf{T}^{-1})'\mathbf{D}^{-1}\mathbf{T}^{-1}$, and $\mathbf{T}^{-1}$ and $\mathbf{D}^{-1}$ are calculable by recursive algorithms similar to the ones presented for the calculation

of **A**, the inversion of **A** is quite easy to obtain and the setting up of matrix **A** self is unnecessary. The faster algorithm based on the modified calculations of the matrices **T** and **D** was presented by Meuwissen and Luo (1992).

Especially advantageous for animal breeding are the results presented by Sorensen and Kennedy (1983) and Kennedy and Sorensen (1988) – that the animal model by the use of the full relationship matrix can accommodate the change in genetic mean and variance caused by inbreeding and selection. Due to this and due to the development of the estimation methods and the increase of the computing power, the animal model became the basic model for genetic evaluation of animals since the end of the 1980's. In Estonia the animal model was first implemented in 1996 (Reents, et al, 1996).

One possible situation where the simple animal model can fail appears if the individuals in the base population are sampled from different genetic groups. Then the hypothesis of a single population with an average genetic effect of zero and the common variance σ_a^2 did not hold and the modified animal model including the additional genetic group effect can be used (Westell, et al, 1988).

The other more complex situations occur when there should be other relevant heritable genetic effects – except for the additive genetic effect – in the model. Two widely applied variants of models of that kind are presented in the following sections.

There are several properties of the animal model not considered in the traditional mixed model theory, but commonly used in practical animal breeding and genetic studies. Many of these, strange properties at first glance, are discussed in Henderson (1984) and Searle (1992) – like the situation where the number of unknown parameters exceeds the number of observations or the models with non-null covariances between random genetic effects and random errors. In Sections 4.2 it is proved that adding individuals without records on observed traits into the model and estimating the additive genetic effects for them does not change the estimators reviewed in Chapter 1.

2.3.2. The maternal effect animal model

Nowadays the economical interests in animal husbandry – maximum intensification of the breeding process and abbreviation of the generation interval – necessitate the selection among animals as early as possible. But as the traits measured on juveniles depend a lot on the mother and/or on the pre- and postnatal conditions, then the simple animal model discussed previously is not precise enough. Thus, the more and more complicated so-called maternal effect animal models were developed to reflect the extra genetic effect of the dam.

Good generalizations on the theoretical works on the subject of maternal genetic effects were done by Willham (1972) and by Lynch and Walsh (1998). Based on those studies, the phenotypic value measured on individual i can be

viewed as the sum of two components. The first component, the so-called direct phenotypic value is a function of the direct expression of individual's genotype and general environmental effects. Since in population genetics the interest is basically focused only on the additive genetic effect of individual *i*, A_i, then the direct phenotypic value of individual *i*, P_i^o, is expressed as $P_i^o = \bar{P} + A_i + E_i$, where upper index *o* denotes a direct effect, $\bar{P}$ is the population mean phenotype and E_i represents general environmental and non-additive direct genetic effects. The second component, the so-called maternal effect, $P_{D_i}^m$, is an indirect effect of the maternal phenotype of individual's *i* dam, D_i, and can also have genetic and environmental components: $P_{D_i}^m = A_{D_i}^m + E_{C_i}$, where upper index *m* denotes a maternal effect, $A_{D_i}^m$ is the maternal additive genetic effect measuring the inheritable genetic ability of the dam to provide a suitable environment, and E_{C_i} is the environmental effect common to all progeny of dam D_i.

Generally the phenotypic value measured on individual *i*, noted as P_i, can be expressed as follows:

$$P_i = \bar{P} + A_i + A_{D_i}^m + E_{C_i} + E_i .$$

The reciprocal references between effects can be better understood from Figure 2.1. As previously, it is assumed that the genetic and environmental effects are independently distributed with no interaction between them. But the genetic covariance may exist between direct additive and maternal additive genetic effects, for example the genes with direct effect on the body size of the dam may also affect the characters influencing the provisioning of offspring.

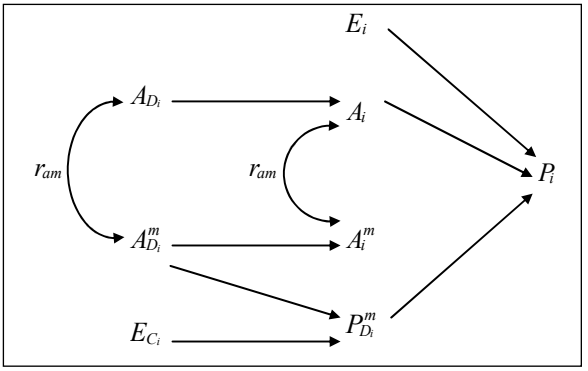


Figure 2.1. The path diagram representing the determination of the phenotype P_i of an individual *i* by direct additive genetic effect A_i, direct environmental effect E_i, maternal additive genetic effect $A_{D_i}^m$ and common environmental effect E_{C_i}; r_{am} is the additive genetic correlation between direct and maternal effects.

The statistical model includes the same effects as the genetic model, only the indexes are different, allowed to express the maternal additive genetic effect for all individuals:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + a_i + m_j + c_k + e_i \,. \tag{2.11}$$

Here it is relevant to understand that albeit for both, the direct additive genetic effect a_i and the maternal additive genetic effect m_j, the indexes vary in the same range – $i, j = 1, ..., N$, the estimates are different based on different incidence matrices relating records to the effects. It is better understood from the model written down in the matrix notation:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \mathbf{W}\mathbf{m} + \mathbf{S}\mathbf{c} + \mathbf{e} \,.$$

Here design matrices **Z** and **W** have one column for each individual in the pedigree, thereby **Z** has in each column at least one 1 corresponding to individual self and columns corresponding to individuals without record contain only zeros, **W** has 1's only in columns corresponding to dams with progeny having records and all other columns consist only of zeros; matrix **S** relating each record to the common environmental effect based on the mother (or litter) is the usual design matrix having one row for each individual and one column for each effect.

To avoid the coincidence of genetic and environmental effects, the additional information by additive genetic relationships between individuals expressed in numerator relationship matrix **A** is used. Then the variance-covariance structure of random effects in model (2.11) is

$$\mathrm{Var}\begin{pmatrix}\mathbf{a}\\ \mathbf{m}\\ \mathbf{c}\\ \mathbf{e}\end{pmatrix} = \begin{pmatrix}\mathbf{A}\sigma_a^2 & \mathbf{A}\sigma_{am} & \mathbf{0} & \mathbf{0}\\ \mathbf{A}\sigma_{am} & \mathbf{A}\sigma_m^2 & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{I}_b\sigma_c^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}_N\sigma_e^2\end{pmatrix},$$

and the phenotypic variance is expressed based on statistical model (2.11) as

$$\sigma_y^2 = \sigma_a^2 + \sigma_m^2 + 2\sigma_{am} + \sigma_c^2 + \sigma_e^2 \,.$$

There are four functions of the variance-covariance components in the maternal trait animal model that have sense in genetics. At first the heritability,

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2} \,.$$

Then the so-called maternal heritability, showing the part of the total variance explained by the maternal additive genetic effect:

$$m^2 = \frac{\sigma_m^2}{\sigma_y^2} \,,$$

the proportion of the common environment variance in the total variance:

$$c^2 = \frac{\sigma_c^2}{\sigma_y^2} \,,$$

and finally, the correlation between the direct and maternal additive genetic effects:

$$r_{am} = \frac{\sigma_{am}}{\sqrt{\sigma_a^2 \cdot \sigma_m^2}} .$$

While the application of the best linear unbiased prediction (1.8) to models with maternal effects was first presented by Quaas and Pollak already in 1981 (Quaas, Pollak, 1981), in Estonia the maternal effect animal model was first implemented in 2000 for estimating the heritabilities and breeding values on ewe's litter size and lambs weight (Kaart, Piirsalu, 2000).

2.3.3. The mixed models for detecting single genes

If based on the measured records and relationships between individuals it is possible to estimate only the polygenetic effects, then the availability of the mass of molecular marker data made it possible to estimate the effects of single chromosomal regions. The models needed for this are nothing completely new, however – the genetic background and the complicacy of the models are largely different.

The chromosomal regions having a putatively measurable effect on a studied trait are called the *major genes* or the *major alleles* (alleles are the different forms of genes). The other often-used term is the *QTL* (Quantitative Trait Loci), which means the location of a gene with some measure effect on the quantitative phenotype and is, at the moment, also the concept synonym to the gene itself.

There are three basic sample designs used in "gene hunting". The clearest picture on the transmission of genetic material and its reflection in the phenotype can be derived by using experimental crosses of pure lines where the segregation of genotypes is accurately observed. An overview of statistical methods applied in this situation is given by Broman (2001). Except for all kinds of experimental crosses widely used in animal and plant breeding to explain the genetic part of the phenotypic diversity, the sample sizes and thus the reliability of analyses in that kind of experiments can't be large at all (especially in humans and wild animals).

The second design is based on fixed relative groups. The most common analysis uses the sib pairs, and the first method proposed for this scheme was the Haseman-Elston regression (Haseman, Elston, 1972). By this the squared difference of trait values between sibs is regressed on the number of alleles shared identical by descent at a test locus, and the null hypothesis – that the regression coefficient is 0 – is tested against the alternative hypothesis that the regression coefficient is negative. Its simplicity is the biggest advantage of this method. At the same time the Haseman-Elston regression and also its modified

variant, which uses the cross-product of sib pair values instead of their squared difference (Elston, et al, 2000), have a relatively small power in detecting single genes in situations where the more general pedigree of studied individuals is known and when the analysed trait is complex (influenced by many genes plus environment) in nature.

The mixed models are applicable for detecting the effects of single genes on traits influenced also by fixed environmental and random polygenetic effects in any population structure. The first introduction of an appropriate model in animal breeding was presented by Fernando and Grossman (1989). The following development of the model was parallelly done for applications in human genetics and in animal breeding (Almasy and Blangero, 1998, Meuwissen and Goddard, 2000).

For detecting the major genes or the QTL, data on molecular genetic level are needed. As the exact determining of the nucleotide sequences in the genotypes of the observed individuals is, in the present day, still inconceivable and unnecessary (for example, in the human genome there is, supposedly, informative material of only about 5%, the rest is junk), the specific sequences of base pairs with a unique physical location in the genome and with sufficient variation between individuals, called molecular *markers*, are used. If a decade ago the common markers were *microsatellites* – DNA variants due to tandem repetition of a short DNA sequence, then these days the analyses are based on the *SNP* (Single Nucleotide Polymorphism – a DNA sequence variations due to change in a single nucleotide).

Similar mixed models can be used to separately or simultaneously model two genetic phenomena, the properties of which are very different. These two phenomena are the linkage of genes and the linkage disequilibrium between genetic units. The first of them is based on *linkage analysis* and the second is based on *association analysis*.

Linkage refers to the tendency of certain genes to be inherited together, due to their close proximity in a chromosome. This means that if marker alleles show familiar co-segregation with the phenotype, then the gene causative of the phenotypic difference is located near the investigated marker. The central idea of linkage analysis based on the mixed model is to identify loci making a significant contribution to the population variance of the trait, by the use of allele-sharing probabilities derived from genotyped marker loci. As the linkage analysis uses only the re-combinations occurred between alleles within the dataset, which typically contains two to three generations, there is no effect to look at in dense marker maps – there will be few re-combinations between adjacent markers during these two to three generations.

Linkage Disequilibrium (LD, although *allelic association*) means that two alleles at different loci occur together within an individual more often than would be predicted by random chance. LD can be generated by mutation, selection, population admixture, finite population size and migration. If the LD is caused by mutation, the mutant allele will be in very strong linkage with alleles

at a number of marker alleles. The association analysis is based on identifying an identity by descent region that flanks the putative gene using all recombinations since the mutation occurred.

Hereby, the linkage disequilibrium mapping methods seem more useful for precise estimation of the major gene position, while linkage mapping is more useful for a genome-wide scan for locating major genes.

Both the linkage and association analysis methods assume that the difference of individual's *i* phenotypic value, P_i, from average phenotypic value in the examined population, $\bar{P}$, is determined by the additive major gene effects received from the sire and dam, Q_i^p and Q_i^m, respectively, by the random polygenetic effect not explained by the genetic markers, A_i, and by the random residual E_i:

$$P_i = \bar{P} + Q_i^p + Q_i^m + A_i + E_i \tag{2.12}$$

The same model in statistical and matrix notation is

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Za} + \mathbf{ZTq} + \mathbf{e}, \tag{2.13}$$

where **y** is a $N \times 1$ vector of observations, **β** is a vector of fixed effects, **X** is a design matrix relating fixed effects to records, **Z** is a $N \times N$ matrix relating records to the individuals, **a** is a $N \times 1$ vector of random additive polygenetic effects, **T** is a $N \times 2N$ matrix relating each individual to its two QTL alleles, **q** is a $2N \times 1$ vector of individuals two random QTL allelic effects and **e** is a vector of residuals.

In the basic model the independence of the genetic and environmental effects and the QTL and polygenetic effects is assumed. Due to this, the total phenotypic variance is expressed as a simple sum of variance components corresponding to the random effects in model (2.13):

$$\sigma_y^2 = \sigma_a^2 + \sigma_q^2 + \sigma_e^2 .$$

The variance-covariance structure of the random effects in the model (2.13) is expressed as

$$\mathrm{Var}\begin{pmatrix}\mathbf{a}\\ \mathbf{q}\\ \mathbf{e}\end{pmatrix} = \begin{pmatrix}\mathbf{A}\sigma_a^2 & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{Q}\sigma_q^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{I}_n\sigma_e^2\end{pmatrix},$$

where **A** is the additive genetic relationship matrix containing the mean probabilities that individuals share alleles identical by descent across the entire genome, and **Q** contains the probabilities for individual QTL alleles being identically by descent conditional on the marker genotypes, linkage phase and putative position of the QTL.

The basic population genetic proportion, the heritability coefficient, is estimable from the ratio

$$h^2 = \frac{\sigma_a^2 + \sigma_q^2}{\sigma_a^2 + \sigma_q^2 + \sigma_e^2} .$$

In addition, the part of the phenotypic variance explained by the QTL effect, $h_q^2 = \sigma_q^2 / (\sigma_a^2 + \sigma_q^2 + \sigma_e^2)$, and the part of the phenotypic variance explained by the complementary additive polygenetic effect are distinguished, $h_a^2 = \sigma_a^2 / (\sigma_a^2 + \sigma_q^2 + \sigma_e^2)$.

Since potentially the quantitative trait may be influenced by more than one QTL, the model (2.12) is easily elaborated to form

$$P = \bar{P} + \sum_i Q_i^p + \sum_i Q_i^m + A + E ,$$

where Q_i^p and Q_i^m denote the *i*-th putative QTL-effect inherited respectively from the sire and the dam.

CHAPTER 3
THE ACCURACY OF THE ANOVA ESTIMATES IN THE ONE-WAY RANDOM MODEL

The one-way random model is the simplest mixed model. Due to its relative simplicity this model is a suitable instrument for studying the behaviour of the estimates of mixed model parameters. Its application in genetics is usually the sire model (2.4), which is due to the mainly sire-based breeding still widely used in the present day. The heritability coefficient in this chapter measures the genetic variability caused by sires and equals four times the intraclass correlation coefficient as stated by expressions (2.5) and (2.7).

All proved results are derived by the author, the other expressions are supplied with references. The results presented in chapter 3.2.4 are discussed in Kaart (1997), but here some extensions are made. The results presented in chapter 3.3 are partly published in Kaart (2004) and the results presented in paragraphs 3.4.2 and 3.5.2 are published in Kaart (2005).

3.1. The one-way random model and its properties

3.1.1. Model and predictors

Consider the mixed linear model

$$y_{ij} = \mu + u_i + e_{ij} ,$$

or in matrix notation

$$\mathbf{y} = \mathbf{1}_N \mu + \mathbf{Z}\mathbf{u} + \mathbf{e} , \tag{3.1}$$

where
$\mathbf{y}$ is the $N \times 1$ vector of observed values,
μ is the only fixed effect in the model (mean),
$\mathbf{1}_N = \mathbf{X}$ and $\mathbf{Z} = \{_{\mathrm{d}}\, \mathbf{1}_{n_i}\}_{i=1}^{a}$ are known design matrices of order $N \times 1$ and $N \times a$ associating fixed and random effects with $\mathbf{y}$,
$\mathbf{u}' = (u_1 \ \ldots \ u_a)'$ is a vector of random effects,
$\mathbf{e}$ is a $N \times 1$ vector of random residuals.

The number of levels in the random factor is traditionally marked as a, and the number of objects per i-th level in the one-way model is denoted by n_i, $\sum_{i=1}^{a} n_i = N$.

In the following we use the traditional notation of variance components for the one-way random model, substituting σ_u^2 and σ_e^2 for σ_1^2 and σ_0^2, applied in the general theory in Chapter 1, respectively. The expectations and the variance-covariance structure are then represented as

$$\mathrm{E}(\mathbf{y}) = \mu , \quad \mathrm{Var}(\mathbf{u}) = \sigma_u^2 \mathbf{I}_a , \quad \mathrm{Var}(\mathbf{e}) = \sigma_e^2 \mathbf{I}_N , \quad \mathrm{Cov}(\mathbf{u}, \mathbf{e}') = \mathbf{0}$$

and

$$\mathbf{V} = \mathrm{Var}(\mathbf{y}) = \{_{\mathrm{d}}\, \sigma_u^2 \mathbf{J}_{n_i} + \sigma_e^2 \mathbf{I}_{n_i}\}_{i=1}^{a}. \tag{3.2}$$

If we assume variance components σ_u^2 and σ_e^2 known, the best linear unbiased predictor of **u** is

$$\hat{\mathbf{u}} = \sigma_u^2 \mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{1}_N \hat{\mu}),$$

or component-wise written

$$\hat{u}_i = \frac{\sigma_u^2}{\sigma_e^2 + n_i \sigma_u^2} \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu}),$$

where

$$\hat{\mu} = \sum_{i=1}^{a} \frac{\sum_{j=1}^{n_i} y_{ij}}{\sigma_e^2 + n_i \sigma_u^2} \Big/ \sum_{i=1}^{a} \frac{n_i}{\sigma_e^2 + n_i \sigma_u^2}.$$

If the data set is balanced, that is, $n_i = n$ for $i = 1, \ldots, a$, then the last formula is simplifying to a well known equation

$$\hat{\mu} = \sum_{i=1}^{N} y_i \big/ N.$$

In further studies the estimation of mean is left without additional attention, because in genetics the interest is focused mainly on the prediction of random effects and on the estimation of variance components.

3.1.2. The ANOVA estimators of variance components

As discussed in Section 1.3, there are many different methods for variance components estimation which give different results in unbalanced data. For models with simple covariance structure, where analytical results are possible and relatively easy to derive, the traditional ANOVA-method is frequently used.

It is assumed that effects u_i and e_{ij} in model (3.1) are independently and normally distributed so that

$$u_i \sim N(0, \sigma_u^2) \text{ and } e_{ij} \sim N(0, \sigma_e^2) \ (i = 1, 2, \ldots, a;\ j = 1, 2, \ldots, n_i). \tag{3.3}$$

The ANOVA table corresponding to model (3.1) is shown in Table 3.1 where A is denoting the random factor with effects u_i, $y_{i\cdot} = \sum_{j=1}^{n_i} y_{ij}$ and $y_{\cdot\cdot} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} y_{ij}$.

Tabel 3.1. The ANOVA table for model (3.1).

Source	Sum of Squares	d.f.	Mean Squares	Expected Mean Squares
A	$SS(u) = \sum_{i=1}^{a} \frac{y_{i\cdot}^2}{n_i} - \frac{y_{\cdot\cdot}^2}{N}$	$a-1$	$MS(u) = SS(u)/a-1$	$\left(\frac{N - \frac{1}{N}\sum_{i=1}^{a} n_i^2}{a-1} \right) \sigma_u^2 + \sigma_e^2$
Error	$SS(e) = \sum_{i=1}^{a} \sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{a} \frac{y_{i\cdot}^2}{n_i}$	$N-a$	$MS(e) = SS(e)/N-a$	σ_e^2

In matrix notation the sums of squares are expressed as quadratic forms:

$$SS(u) = \mathbf{y}'\left[\left\{ {}_{\mathrm{d}}\, \mathbf{J}_{n_i}/n_i \right\}_{i=1}^{a} - \mathbf{J}_N/N \right]\mathbf{y} = \mathbf{y}'\mathbf{Q}_1\mathbf{y} \tag{3.4}$$

and

$$SS(e) = \mathbf{y}'\left[\mathbf{I}_N - \left\{ {}_{\mathrm{d}}\, \mathbf{J}_{n_i}/n_i \right\}_{i=1}^{a} \right]\mathbf{y} = \mathbf{y}'\mathbf{Q}_2\mathbf{y} \ .$$

The ANOVA estimators of variance components σ_u^2 and σ_e^2 are obtained by equating the mean squares with their expected values and are expressed as

$$\hat{\sigma}_u^2 = \frac{1}{d}\left[MS(u) - MS(e) \right] \tag{3.5}$$

and

$$\hat{\sigma}_e^2 = MS(e)\,, \tag{3.6}$$

where

$$d = \frac{1}{a-1}\left(N - \frac{1}{N}\sum_{i=1}^{a} n_i^2 \right). \tag{3.7}$$

If the data set is balanced, that is, $n_i = n$ for $i = 1,\ldots,a$, then $d = n$ and

$$\hat{\sigma}_u^2 = \frac{1}{n}\left[MS(u) - MS(e) \right]. \tag{3.8}$$

The estimator $\hat{\rho}$ of the intraclass correlation coefficient ρ, which measures the magnitude of random effects, is calculated as the ratio of variances:

$$\hat{\rho} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2} = \frac{MS(u) - MS(e)}{MS(u) + (d-1)MS(e)}\,. \tag{3.9}$$

In the half-sib model (2.4) the intraclass correlation coefficient estimates one quarter of the heritability coefficient, measuring the magnitude of the additive genetic effect:

$$\hat{h}^2 = \frac{4\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2} = 4\hat{\rho}\,, \tag{3.10}$$

and the random effect *u* represents the random parent (mainly sire) effect.

3.1.3. The distributional properties of the ANOVA estimators

It is well known that the sums of squares corresponding to the main effect and error term, $SS(u)$ and $SS(e)$, respectively, are independent and that

$$SS(e)/\sigma_e^2 \sim \chi^2_{N-a} \tag{3.11}$$

(see, for example, Searle, Casella and McCulloch, 1992, p 73). If the data set is balanced, then it also holds that

$$SS(u)/(n\sigma_u^2+\sigma_e^2) \sim \chi^2_{a-1}$$

and

$$\frac{MS(u)}{n\sigma_u^2+\sigma_e^2} \Big/ \frac{MS(e)}{\sigma_e^2} \sim F_{a-1,N-a}\,. \tag{3.12}$$

In the case of unbalanced data, the two latter distributions are not true. However, as noted in Khuri, Mathew and Sinha (1998), the formulas derived by Johnson and Kotz (1970) can be used to express the quadratic form (3.4) as a linear combination of independent central Chi-squared variables of the form

$$SS(u) = \mathbf{y}'\mathbf{Q}_1\mathbf{y} \sim \sum_{i=1}^{s} \lambda_i \chi^2_{m_i}\,, \tag{3.13}$$

where $\lambda_1, \lambda_2, ..., \lambda_s$ are the distinct non-zero eigenvalues of $\mathbf{Q}_1\mathbf{V}$ with multiplicities $m_1, m_2, ..., m_s$, respectively, and $\mathbf{V}$ is the variance-covariance matrix of observed values defined with the equation (3.2). As further operations with the linear combination of independent central Chi-squared variables of the form (3.13) are complicated, in the following paragraphs an approximation, based on Satterthwaite's procedure (1941) and presented by Khuri, Mathew and Sinha (1998), is used of the form

$$\sum_{i=1}^{s} \lambda_i \chi^2_{m_i} \approx \lambda \chi^2_m\,, \tag{3.14}$$

where

$$\lambda = \frac{\sum_{i=1}^{s} m_i \lambda_i^2}{\sum_{i=1}^{s} m_i \lambda_i} \tag{3.15}$$

and

$$m = \frac{\left(\sum_{i=1}^{s} m_i \lambda_i\right)^2}{\sum_{i=1}^{s} m_i \lambda_i^2}\,. \tag{3.16}$$

As discussed in Khuri, Mathew and Sinha (1998, p 58–59), the approximation is exact when the data set is balanced, that is, $n_i = n$ for $i = 1, \ldots, a$.

3.1.4. The measure of imbalance

If a data set is unbalanced, usually the question arises how to measure this imbalance. Having some kind of a numerical measure of the degree of imbalance makes it possible to compare designs and study the effect of data imbalance on the properties of parameters estimators.

In this chapter the measure of data design imbalance first introduced by Ahrens and Pincus (1981) is used. Following the notations of Khuri, Mathew and Sinha (1998, p 62–63), let $\mathbf{D} = \{n_1, n_2, \ldots, n_a\}$ denote the associated design. Then the measure of **D**'s imbalance is defined as

$$\nu(\mathbf{D}) = 1 \Big/ a \sum_{i=1}^{a} \left(\frac{n_i}{N} \right)^2 . \qquad (3.17)$$

Hence, $1/a < \nu(\mathbf{D}) \le 1$. The lower bound follows from the fact that $\sum_{i=1}^{a} n_i^2 < N^2$ for $a > 1$. The measure $\nu(\mathbf{D})$ attains its maximum value 1 if and only if the design **D** is balanced.

Since there are several different designs corresponding to the fixed measure of data imbalance, it is not possible to find the exact expression between data imbalance and the accuracy of parameters estimators. One possible variant to study this potential dependency is to use simulations.

To generate designs for the one-way model with a specified degree of imbalance, data size and number of groups, the algorithm proposed by Khuri, Mathew and Sinha (1998, p 76–80) was realised in SAS Interactive Matrix Language (IML; SAS Institute Inc., 1999) by the author. The algorithm calculates groups sizes n_1, n_2, …, n_a as solutions to the equations

$$\sum_{i=1}^{a} n_i = N \text{ and } \sum_{i=1}^{a} n_i^2 = N^2 / a\,\nu(\mathbf{D}) , \qquad (3.18)$$

finding the intersection of the hyperplane with the hypersphere whose representatives are the equations (3.18), respectively. As the equations (3.18) may not have exact integer solutions in general, the approximate solution would be needed. Based on the Lemma 3.6.1 presented in Khuri, Mathew and Sinha (1998, p 77), the optimal integer group sizes are achieved, taking first a_0 group sizes equal to $[n_i]$ and the remaining $a - a_0$ group sizes equal to $[n_i] + 1$, where $[n_i]$ is the greatest integer in n_i, for $i = 1, 2, ..., a$, and $a_0 = \Sigma_i (n_i - [n_i])$.

Examples of the approximate designs generated for data size $N = 360$ and corresponding to specified numbers of groups (a) and data set imbalances (ν) are presented in Table 3.2. The same data size, similar group sizes and data set imbalances are used in paragraphs 3.4 and 3.5 in studying the relationship between data imbalance and the accuracy of parameters estimators.

3.1.5. Some results on traces, eigenvalues, projection matrices and $c_1\mathbf{I}_n + c_2\mathbf{J}_n$ matrices useful in studying the properties of the ANOVA estimators

In this paragraph several standard properties of traces, eigenvalues and projection matrices are presented that are used later in Sections 3.2 and 3.4. Also, the characteristics of $c_1\mathbf{I}_n + c_2\mathbf{J}_n$ type matrices typically operated in one-way ANOVA are derived.

Tabel 3.2. Examples of approximate designs with the specific data set imbalance ν generated for data size $N = 360$ and fixed number of groups a.

ν	a	$\mathbf{D} = \{n_1, n_2, \ldots, n_a\}$
0.3	4	$\{8,12,14,326\}$; $\{2,12,24,322\}$; $\{10,10,13,327\}$; $\{2,13,41,304\}$
	15	$\{1,3,7,8,10,10,10,\underbrace{11,\ldots,11}_{5},25,108,123\}$; $\{9,11,14,14,14,\underbrace{15,\ldots,15}_{7},16,16,161\}$
	24	$\{6,\underbrace{8,\ldots,8}_{12},\underbrace{9,\ldots,9}_{8},13,60,113\}$; $\{1,2,7,7,8,8,9,\underbrace{10,\ldots,10}_{12},11,11,15,44,118\}$;
	90	$\{1,1,1,2,2,\underbrace{3,\ldots,3}_{63},\underbrace{4,\ldots,4}_{16},5,5,7,9,15,59\}$; $\{\underbrace{3,\ldots,3}_{66},\underbrace{4,\ldots,4}_{21},8,10,80\}$
0.6	4	$\{18,67,68,207\}$; $\{43,45,62,210\}$; $\{26,34,95,205\}$; $\{46,49,50,215\}$
	15	$\{9,\underbrace{19,\ldots,19}_{6},\underbrace{20,\ldots,20}_{7},97\}$; $\{1,2,15,\underbrace{16,\ldots,16}_{4},17,17,26,26,32,35,41,84\}$
	24	$\{8,\underbrace{10,\ldots,10}_{6},\underbrace{11,\ldots,11}_{13},16,26,47,60\}$; $\{2,7,\underbrace{10,\ldots,10}_{14},11,11,14,17,24,34,47,53\}$
	90	$\{1,\underbrace{2,\ldots,2}_{4},\underbrace{3,\ldots,3}_{28},\underbrace{4,\ldots,4}_{54},9,9,33\}$; $\{2,\underbrace{3,\ldots,3}_{38},\underbrace{4,\ldots,4}_{49},17,31\}$
0.9	4	$\{41,103,104,112\}$; $\{66,77,80,137\}$; $\{50,76,114,120\}$; $\{38,107,107,108\}$
	15	$\{1,22,23,\underbrace{24,\ldots,24}_{5},\underbrace{25,\ldots,25}_{6},44\}$; $\{15,16,\underbrace{22,\ldots,22}_{5},23,23,\underbrace{24,\ldots,24}_{4},25,52\}$
	24	$\{11,12,13,13,\underbrace{14,\ldots,14}_{16},15,17,18,37\}$; $\{13,\underbrace{14,\ldots,14}_{21},15,38\}$
	90	$\{\underbrace{2,3,\ldots,3}_{10},\underbrace{4,\ldots,4}_{78},16\}$; $\{1,2,2,\underbrace{3,\ldots3}_{12},\underbrace{4,\ldots4}_{72},8,10,13\}$; $\{1,\underbrace{3,\ldots,3}_{13},\underbrace{4,\ldots,4}_{74},12,12\}$

Proof of the following useful properties of trace operator is trivial and can be partly found in Searle (1982) and Harville (1997), for example.

Proposition 3.1. *For trace operation the following properties hold*.

(i) $\mathrm{tr}(k\mathbf{A}) = k\,\mathrm{tr}(\mathbf{A})$;

(ii) $\mathrm{tr}(\mathbf{A} + \mathbf{B}) = \mathrm{tr}(\mathbf{A}) + \mathrm{tr}(\mathbf{B})$;

(iii) $\mathrm{tr}(\mathbf{A}') = \mathrm{tr}(\mathbf{A})$;

(iv) $\mathrm{tr}(\mathbf{ABC}) = \mathrm{tr}(\mathbf{CAB}) = \mathrm{tr}(\mathbf{BCA})$;

(v) $\mathrm{tr}(\mathbf{ABC}) = \mathrm{tr}(\mathbf{BAC}) = \mathrm{tr}(\mathbf{ACB})$, *if* **A**, **B** *and* **C** *are symmetric*;

(vi) $\mathrm{tr}(\mathbf{J}_n\mathbf{A}) = \sum_{i,j=1}^{n} a_{ij}$, *where* $\mathbf{A} = \{_{\mathrm{m}}\, a_{ij}\}_{i,j=1}^{n}$;

(vii) $\mathrm{tr}\{_{\mathrm{d}}\, \mathbf{D}_i\}_{i=1}^{a} = \sum_{i=1}^{a} \mathrm{tr}(\mathbf{D}_i)$. ■

The following properties of eigenvalues are collected from matrix books by Harville (1997) and Schott (1997).

Proposition 3.2. *Let* $\mathbf{A}$ *represent an* $n \times n$ *matrix. Then for its eigenvalues the following properties hold.*

(i) *If* $\mathbf{A}$ *is a real symmetric matrix, then all its eigenvalues are real.*

(ii) *If* λ *represents an eigenvalue of* $\mathbf{A}$*, then for every positive integer* k*,* λ^k *is an eigenvalue of* $\mathbf{A}^k$*.*

(iii) *If* $\lambda_1, \lambda_2, ..., \lambda_s$ *are distinct eigenvalues of* $\mathbf{A}$ *with multiplicities* $m_1, m_2, ..., m_s$*, respectively, then* $\operatorname{tr}(\mathbf{A}) = \sum_{i=1}^{s} m_i \lambda_i$*.*

(iv) *If* $\mathbf{A}$ *is a non-negative definite, then all its eigenvalues are non-negative, if* $\mathbf{A}$ *is a positive definite, then all its eigenvalues are positive.* ■

Proof of the following and many other useful properties concerning projection matrices can be found in Harville (1997, p 260–264).

Proposition 3.3. *Let* $\mathbf{X}$ *be an arbitrary* $n \times p$ *matrix and* $\mathbf{W}$ *a* $n \times n$ *symmetric positive definite matrix. Then for the* projection *matrix* $\mathbf{P}_{\mathbf{X},\mathbf{W}} = \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}$ *the following properties hold.*

(i) $\mathbf{W}\mathbf{P}_{\mathbf{X},\mathbf{W}}$ *and* $\mathbf{P}_{\mathbf{X},\mathbf{W}}\mathbf{W}^{-1}$ *are symmetric*;

(ii) $\mathbf{P}'_{\mathbf{X},\mathbf{W}}\mathbf{W}\mathbf{P}_{\mathbf{X},\mathbf{W}} = \mathbf{W}\mathbf{P}_{\mathbf{X},\mathbf{W}}$;

(iii) $(\mathbf{I} - \mathbf{P}_{\mathbf{X},\mathbf{W}})'\mathbf{W}(\mathbf{I} - \mathbf{P}_{\mathbf{X},\mathbf{W}}) = \mathbf{W}(\mathbf{I} - \mathbf{P}_{\mathbf{X},\mathbf{W}})$. ■

The following results formulate the basis for simplifying the general expressions about estimators and predictors and their accuracy in the model (3.1).

Proposition 3.4. *For structured matrices in the form* $c_1\mathbf{I}_n + c_2\mathbf{J}_n$ *the following properties hold.*

(i) $(b_1\mathbf{I}_n + b_2\mathbf{J}_n)(c_1\mathbf{I}_n + c_2\mathbf{J}_n) = b_1c_1\mathbf{I}_n + (b_1c_2 + b_2c_1 + nb_2c_2)\mathbf{J}_n$;

(ii) $(c_1\mathbf{I}_n + c_2\mathbf{J}_n)^{-1} = \frac{1}{c_1}\mathbf{I}_n - \frac{c_2}{c_1(c_1 + nc_2)}\mathbf{J}_n$ for $c_1 \neq 0$ and $c_1 \neq -nc_2$;

(iii) $(c_1\mathbf{I}_n + c_2\mathbf{J}_n)^k = c_1^k\mathbf{I}_n + \frac{1}{n}\left[(c_1 + nc_2)^k - c_1^k\right]\mathbf{J}_n$;

(iv) $\operatorname{tr}\left[(c_1\mathbf{I}_n + c_2\mathbf{J}_n)^k\right] = (n-1)c_1^k + (c_1 + nc_2)^k$.

Proof. Properties (i) and (ii) can be found in several matrix books (Searle, 1982, p 322, for example). Proof to (i) is straightforward. Proof to (ii) is given in Nahtman (2004, Lemma 1.5.1, p 21), for example.

(iii) To prove the third statement of the theorem, induction is used. If $k = 1$ then $c_1^1\mathbf{I}_n + \frac{1}{n}[(c_1 + nc_2)^1 - c_1^1]\mathbf{J}_n = c_1\mathbf{I}_n + c_2\mathbf{J}_n$. If $k = 2$ then

$$c_1^2\mathbf{I}_n + \tfrac{1}{n}[(c_1+nc_2)^2 - c_1^2]\mathbf{J}_n = c_1^2\mathbf{I}_n + \tfrac{1}{n}(2nc_1c_2 + n^2c_2^2)\mathbf{J}_n$$
$$= c_1^2\mathbf{I}_n + 2c_1c_2\mathbf{J}_n + nc_2^2\mathbf{J}_n = (c_1\mathbf{I}_n + c_2\mathbf{J}_n)^2.$$

So, the statement is true for $k=1$ and $k=2$.

Suppose the statement is true for $(c_1\mathbf{I}_n + c_2\mathbf{J}_n)^{k-1}$, then we shall show that it is also true for $(c_1\mathbf{I}_n + c_2\mathbf{J}_n)^k$. In proof we use the formula from statement (i).

$$\begin{aligned}(c_1\mathbf{I}_n + c_2\mathbf{J}_n)^k &= (c_1\mathbf{I}_n + c_2\mathbf{J}_n)^{k-1} \times (c_1\mathbf{I}_n + c_2\mathbf{J}_n)\\ &= c_1^{k-1}\mathbf{I}_n + \frac{1}{n}\left[(c_1+nc_2)^{k-1} - c_1^{k-1}\right]\mathbf{J}_n \times (c_1\mathbf{I}_n + c_2\mathbf{J}_n)\\ &= c_1^k\mathbf{I}_n + \left[c_1^{k-1}c_2 + \frac{c_1(c_1+nc_2)^{k-1}}{n} - \frac{c_1^k}{n} + c_2(c_1+nc_2)^{k-1} - c_1^{k-1}c_2\right]\mathbf{J}_n\\ &= c_1^k\mathbf{I}_n + \frac{(c_1+nc_2)^{k-1}(c_1+nc_2) - c_1^k}{n}\mathbf{J}_n\\ &= c_1^k\mathbf{I}_n + \frac{1}{n}\left[(c_1+nc_2)^k - c_1^k\right]\mathbf{J}_n,\end{aligned}$$

and thus the formula (iii) is proved via induction.

(iv) As the trace of sum equals to the sum of traces, we separate the expression $(c_1\mathbf{I}_n + c_2\mathbf{J}_n)^k$ into two parts based on statement (iii) and apply the trace operation to both of them:

$$\begin{aligned}\mathrm{tr}\left[(c_1\mathbf{I}_n + c_2\mathbf{J}_n)^k\right] &= \mathrm{tr}(c_1^k\mathbf{I}_n) + \mathrm{tr}\left\{\tfrac{1}{n}\left[(c_1+nc_2)^k - c_1^k\right]\mathbf{J}_n\right\}\\ &= nc_1^k + (c_1+nc_2)^k - c_1^k = (n-1)c_1^k + (c_1+nc_2)^k.\end{aligned}$$ ■

Corollary 3.1. *The inverse of variance-covariance matrix of observed values* (3.2) *is expressed as*

$$\mathbf{V}^{-1} = \left\{\frac{1}{\sigma_e^2}\left(\mathbf{I}_{n_i} - \frac{\sigma_u^2}{\sigma_e^2 + n_i\sigma_u^2}\mathbf{J}_{n_i}\right)\right\}_{i=1}^{a}. \qquad (3.19)$$

■

3.2. The accuracy of the estimates and predictors in balanced data

3.2.1. The sampling variances of variance components

Based on the distributional properties of the ANOVA estimators discussed in Section 3.1.3, the sampling variances of the ANOVA estimators of variance components are expressed as

$$\mathrm{Var}(\hat{\sigma}_u^2) = \frac{2}{n^2}\left[\frac{(n\sigma_u^2 + \sigma_e^2)^2}{a-1} + \frac{\sigma_e^4}{a(n-1)}\right], \tag{3.20}$$

and

$$\mathrm{Var}(\hat{\sigma}_e^2) = \frac{2\sigma_e^4}{N-a}. \tag{3.21}$$

The unbiased estimators of the sampling variances of the ANOVA estimators of variance components are expressed as

$$\widehat{\mathrm{Var}}(\hat{\sigma}_u^2) = \frac{2}{n^2}\left[\frac{(n\hat{\sigma}_u^2 + \hat{\sigma}_e^2)^2}{a+1} + \frac{\hat{\sigma}_e^4}{a(n-1)+2}\right], \tag{3.22}$$

and

$$\widehat{\mathrm{Var}}(\hat{\sigma}_e^2) = \frac{2\hat{\sigma}_e^4}{N-a+2}. \tag{3.23}$$

The derivation of these formulas can be found in several text-books, for example in Searle, Casella and McCulloch (1992, p 63–64).

In real data analysis usually the standard deviations of estimated parameters – instead of sampling variances – are calculated. The motivation for this is that the standard deviations are easier to interpret and they are also the basis for the accuracy and significance testing of parameters estimates. The square root of estimated sampling variance is also called the standard error (Weisstein, 2004, for example).

But usually it is not perceived that it matters whether the sampling variance or the standard error is found in studying the accuracy of estimates. Visscher (1998) discussed that often expressions which give biased estimates to sampling variances, result from taking their square root with unbiased estimates to standard errors, and the opposite. This can especially be a problem in the case of small sample sizes.

The following simulation study with SAS IML (SAS Institute Inc., 1999) was carried out by the author to explore the possible bias in standard errors and sampling variances of the variance components depending on the used formulas. Both the standard deviations and variances of the estimated variance components were found in three ways. Firstly, the empirical variances of $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$ were found, based on 10000 replicated samples. Secondly, the predicted variances were calculated by formulas (3.20) and (3.21), using the true population values. Thirdly, the average estimated biased variances were calculated by formulas (3.20) and (3.21), and the average estimated unbiased variances were calculated by formulas (3.22) and (3.23), with $\hat{\sigma}_u^2$ substituted for σ_u^2 and $\hat{\sigma}_e^2$ for σ_e^2, respectively. The estimates of the standard deviations of the variance components estimators were calculated in similar ways, applying the square root to expressions (3.20)-(3.23), if these were used.

The data size $N = 360$ was used as a reasonable number for small practical experiments and also giving a possibility to divide data into groups with equal integer size in different ways. Number of groups *a* used in simulations were 4, 15, 24 and 90. Random effects were generated by normal distributions (3.3).

Without loss of generality, a $\sigma_e^2 = 1$ was used throughout. In the generation of random effects u_i, the intraclass correlation coefficient ρ was taken equal to 0.0125, 0.0625, 0.15 and 0.8. Corresponding values of variance component σ_u^2 are calculable by expression $\sigma_u^2 = \rho\sigma_e^2/(1-\rho)$ and are approximately equal to 0.01266, 0.06667, 0.17647 and 4.0; corresponding heritability coefficient h^2 values are by expression (3.10) equal to 0.05, 0.25, 0.6 and 3.2 (the last is not a possible value of heritability in genetic studies based on sire model, but is correct in the mathematical sense; if the genetic model states the intraclass correlation coefficient equal to proportion of whole genetic variability, then $\rho = h^2$ and value $\rho = 0.8$ is acceptable also in genetic sense).

Simulations results dealing with variances of $\hat{\sigma}_u^2$ are presented in Table 3.3.

Table 3.3. The observed, predicted and estimated sampling variances of the variance component σ_u^2 in the case of different true population values and number of groups (sires) a. $N = 360$, $\sigma_e^2 = 1$.

σ_u^2 (h^2)	a	$\mathrm{E}(\hat{\sigma}_u^2)$	$\mathrm{Var}(\hat{\sigma}_u^2)^*$	$\mathrm{Var}(\hat{\sigma}_u^2 \mid \sigma_{e,u}^2)^{\#}$	$\widehat{\mathrm{Var}}(\hat{\sigma}_u^2 \mid \hat{\sigma}_{e,u}^2)^{\square}$	$\widehat{\mathrm{Var}}_b(\hat{\sigma}_u^2 \mid \hat{\sigma}_{e,u}^2)^{\square}$
0.01266	4	0.0127	0.3893×10^{-3}	0.3773×10^{-3}	0.3840×10^{-3}	0.6395×10^{-3}
(0.05)	15	0.0125	0.4351×10^{-3}	0.4317×10^{-3}	0.4298×10^{-3}	0.4898×10^{-3}
	24	0.0130	0.5771×10^{-3}	0.5736×10^{-3}	0.5770×10^{-3}	0.6251×10^{-3}
	90	0.0124	0.2033×10^{-2}	0.2013×10^{-2}	0.2008×10^{-2}	0.2046×10^{-2}
0.06667	4	0.0662	0.3907×10^{-2}	0.4034×10^{-2}	0.3952×10^{-2}	0.6586×10^{-2}
(0.25)	15	0.0662	0.1694×10^{-2}	0.1687×10^{-2}	0.1678×10^{-2}	0.1916×10^{-2}
	24	0.0666	0.1602×10^{-2}	0.1572×10^{-2}	0.1573×10^{-2}	0.1707×10^{-2}
	90	0.0669	0.2684×10^{-2}	0.2716×10^{-2}	0.2712×10^{-2}	0.2766×10^{-2}
0.17647	4	0.1774	2.4568×10^{-2}	2.3459×10^{-2}	2.4044×10^{-2}	4.0072×10^{-2}
(0.6)	15	0.1758	0.6669×10^{-2}	0.6808×10^{-2}	0.6752×10^{-2}	0.7715×10^{-2}
	24	0.1765	0.5216×10^{-2}	0.5167×10^{-2}	0.5171×10^{-2}	0.5618×10^{-2}
	90	0.1774	0.4545×10^{-2}	0.4550×10^{-2}	0.4563×10^{-2}	0.4659×10^{-2}
4.0	4	3.9327	10.1416	10.726	10.2777	17.1296
(3.2)	15	3.9763	2.2899	2.3336	2.3042	2.6333
	24	3.9931	1.4007	1.4381	1.4306	1.5550
	90	4.0014	0.4131	0.4064	0.4068	0.4159

* Observed sampling variances $\mathrm{Var}(\hat{\sigma}_u^2)$ were found based on 10000 replicated samples.
# Predicted sampling variances $\mathrm{Var}(\hat{\sigma}_u^2 \mid \sigma_{e,u}^2)$ were calculated by formula (3.20).
¤ Estimated sampling variances $\widehat{\mathrm{Var}}(\hat{\sigma}_u^2 \mid \hat{\sigma}_{e,u}^2)$ and $\widehat{\mathrm{Var}}_b(\hat{\sigma}_u^2 \mid \hat{\sigma}_{e,u}^2)$ were calculated by formulas (3.22) and (3.20), respectively, with $\hat{\sigma}_u^2$ substituted for σ_u^2 and $\hat{\sigma}_e^2$ substituted for σ_e^2 in (3.20), and based on 10000 replicated samples.

Simulations results concerning standard deviations of $\hat{\sigma}_u^2$ are presented in Table 3.4. Results showing $\hat{\sigma}_e^2$ are not presented because these are similar to those in Tables 3.3 and 3.4, only the differences in accuracy are less notable.

Results in table 3.3 indicate that the expression (3.22), giving an unbiased estimator to the sampling variance of the ANOVA estimator of the variance component σ_u^2, is correct. Using the standard formula (3.20) with $\hat{\sigma}_u^2$ substituted for

σ_u^2 and $\hat{\sigma}_e^2$ substituted for σ_e^2 will lead to overestimated sampling variance. The bias is especially large in the case of a small number of groups and a small intra-class correlation (the latter means all possible heritabilities from the sire model); in the case of a large intraclass correlation (the values of which are impossible in the sire model) the bias is bigger for a large number of groups.

Results in table 3.4 indicate that taking the square root from the expression (3.22) will result in an underestimated standard error of σ_u^2. To get a more precise estimate, the formula (3.20) with $\hat{\sigma}_u^2$ substituted for σ_u^2 and $\hat{\sigma}_e^2$ substituted for σ_e^2 should be used in calculating the standard deviation of the estimates of σ_u^2. The latter is a standard practice for example in the commercial statistical package SAS, where asymptotic standard errors are calculated, which in the case of the ANOVA method means that the square roots of the expressions (3.20) and (3.21) are used (SAS, 1999).

Table 3.4. The observed, predicted and estimated standard deviations of the variance component estimate $\hat{\sigma}_u^2$ in the case of different true population values and number of groups (sires) a. $N = 360$, $\sigma_e^2 = 1$.

σ_u^2 (h^2)	a	$\mathrm{E}(\hat{\sigma}_u^2)$	$\sigma(\hat{\sigma}_u^2)^*$	$\sigma(\hat{\sigma}_u^2 \mid \sigma_{e,u}^2)^{\#}$	$\hat{\sigma}(\hat{\sigma}_u^2 \mid \hat{\sigma}_{e,u}^2)^{\square}$	$\hat{\sigma}_b(\hat{\sigma}_u^2 \mid \hat{\sigma}_{e,u}^2)^{\square}$
0.01266	4	0.0127	0.0197	0.0194	0.0152	0.0195
(0.05)	15	0.0125	0.0209	0.0208	0.0195	0.0208
	24	0.0130	0.0240	0.0240	0.0231	0.0241
	90	0.0124	0.0451	0.0449	0.0445	0.0449
0.06667	4	0.0662	0.0625	0.0635	0.0489	0.0631
(0.25)	15	0.0662	0.0412	0.0411	0.0383	0.0410
	24	0.0666	0.0400	0.0397	0.0381	0.0397
	90	0.0669	0.0518	0.0521	0.0517	0.0522
0.17647	4	0.1774	0.1567	0.1532	0.1192	0.1539
(0.6)	15	0.1758	0.0817	0.0825	0.0770	0.0823
	24	0.1765	0.0722	0.0719	0.0690	0.0719
	90	0.1774	0.0674	0.0675	0.0669	0.0676
4.0	4	3.9327	3.1846	3.2751	2.4943	3.2201
(3.2)	15	3.9763	1.5133	1.5276	1.4206	1.5186
	24	3.9931	1.1835	1.1992	1.1483	1.1972
	90	4.0014	0.6427	0.6375	0.6307	0.6377

* Observed standard deviations $\sigma(\hat{\sigma}_u^2)$ were found based on 10000 replicated samples.

Predicted standard deviations $\sigma(\hat{\sigma}_u^2 \mid \sigma_{e,u}^2)$ were calculated as square roots of formula (3.20).

¤ Estimated standard deviations $\hat{\sigma}(\hat{\sigma}_u^2 \mid \hat{\sigma}_{e,u}^2)$ and $\hat{\sigma}_b(\hat{\sigma}_u^2 \mid \hat{\sigma}_{e,u}^2)$ were calculated as square roots of formulas (3.22) and (3.20), respectively, with $\hat{\sigma}_u^2$ substituted for σ_u^2 and $\hat{\sigma}_e^2$ substituted for σ_e^2 in (3.20), and based on 10000 replicated samples.

Thus, there is a big difference in the accuracy of the estimators of the variability of $\hat{\sigma}_u^2$ depending on whether the sampling variance or standard error is found and which formula is used. Expression giving an unbiased estimate to sampling

variance gives a biased estimate to the standard error in estimating variance components.

3.2.2. The sampling variance of the intraclass correlation coefficient

For the sampling variance of the intraclass correlation coefficient there does not exist an exact formula even in the balanced case. Usually an approximate formula is used:

$$\mathrm{Var}(\hat{\rho}) \approx \frac{2[1+(n-1)\rho]^2(1-\rho)^2}{n(n-1)(a-1)}, \tag{3.24}$$

derived by Osborne and Paterson (1952) using a first-order Taylor-series expansion of equality (3.9) with n replacing d. Zerbe and Goldgar (1980) derived an alternative formula based on the F-ratio (3.12). They used a first-order Taylor series expansion of variance of non-linear function of parameter w estimator of the form

$$\mathrm{Var}[f(\hat{w})] \approx \left[\partial f(w)/\partial w\right]^2 \mathrm{Var}(\hat{w}), \tag{3.25}$$

where the derivative is evaluated at the mean of $\hat{w}$, and reached to the following final formula:

$$\mathrm{Var}(\hat{\rho}) \approx 2[1+(n-1)\rho]^2(1-\rho)^2 \frac{[a(n-1)]^2(an-3)}{n^2(a-1)[a(n-1)-2]^2[a(n-1)-4]}. \tag{3.26}$$

To examine the accuracy of the expressions (3.24) and (3.26), similarly to the previous section, a simulation study was performed. The compared parameters were the variances and standard deviations of the estimators of the intraclass correlation coefficient. Both these parameters were found in five different ways. At first the observed variability of the estimators was found based on 10000 replicated samples. The predicted sample variances and standard errors were calculated both by Osborne's and Paterson's approximation and by Zerbe's and Goldgar's approximation. The estimated sample variances and standard errors are averages of the parameters found on each simulation by expressions (3.24) or (3.26) (or the square root of them) with $\hat{\rho}$ substituted for ρ.

The simulations results concerning the variances of $\hat{\rho}$ are presented in Table 3.5 and the simulations results showing the standard deviations of $\hat{\rho}$ are presented in Table 3.6. As in the sire model, the sampling variance of the heritability coefficient equals to 16 times the sampling variance of the intraclass correlation coefficient (for the standard error the corresponding coefficient is 4), then all the following conclusions apply also in the case of estimation of heritabilities.

Table 3.5. The observed, predicted and estimated sampling variances of the intraclass correlation coefficient ρ in the case of different true population values and number of groups (sires) a. $N = 360$, $\sigma_e^2 = 1$.

ρ (h^2)	a	E($\hat{\rho}$)	Var($\hat{\rho}$)*	$\mathrm{Var}_{OP}(\hat{\rho}\|\rho)^{\#}$	$\widehat{\mathrm{Var}}_{OP}(\hat{\rho}\|\hat{\rho})^{¤}$	$\mathrm{Var}_{ZG}(\hat{\rho}\|\rho)^{\#}$	$\widehat{\mathrm{Var}}_{ZG}(\hat{\rho}\|\hat{\rho})^{¤}$
0.0125	4	0.0124	0.3625×10^{-3}	0.3622×10^{-3}	0.5572×10^{-3}	0.3674×10^{-3}	0.5651×10^{-3}
(0.05)	15	0.0123	0.4165×10^{-3}	0.4183×10^{-3}	0.4568×10^{-3}	0.4246×10^{-3}	0.4637×10^{-3}
	24	0.0122	0.5734×10^{-3}	0.5575×10^{-3}	0.5832×10^{-3}	0.5662×10^{-3}	0.5924×10^{-3}
	90	0.0121	0.1961×10^{-2}	0.1966×10^{-2}	0.1954×10^{-2}	0.2008×10^{-2}	0.1996×10^{-2}
0.0625	4	0.0595	0.2682×10^{-2}	0.3150×10^{-2}	0.3856×10^{-2}	0.3195×10^{-2}	0.3911×10^{-2}
(0.25)	15	0.0612	0.1304×10^{-2}	0.1351×10^{-2}	0.1405×10^{-2}	0.1372×10^{-2}	0.1426×10^{-2}
	24	0.0617	0.1272×10^{-2}	0.1280×10^{-2}	0.1306×10^{-2}	0.1230×10^{-2}	0.1326×10^{-2}
	90	0.0624	0.2274×10^{-2}	0.2321×10^{-2}	0.2303×10^{-2}	0.2371×10^{-2}	0.2353×10^{-2}
0.15	4	0.1382	0.0987×10^{-2}	0.1238×10^{-1}	0.1199×10^{-1}	0.1256×10^{-1}	0.1216×10^{-1}
(0.6)	15	0.1462	0.3440×10^{-2}	0.3703×10^{-2}	0.3637×10^{-2}	0.3758×10^{-2}	0.3692×10^{-2}
	24	0.1477	0.2782×10^{-2}	0.2875×10^{-2}	0.2835×10^{-2}	0.2920×10^{-2}	0.2879×10^{-2}
	90	0.1497	0.2798×10^{-2}	0.2845×10^{-2}	0.2811×10^{-2}	0.2906×10^{-2}	0.2872×10^{-2}
0.8	4	0.7048	0.3529×10^{-1}	0.1735×10^{-1}	0.2260×10^{-1}	0.1760×10^{-1}	0.2292×10^{-1}
(3.2)	15	0.7799	0.4974×10^{-2}	0.3896×10^{-2}	0.4455×10^{-2}	0.3955×10^{-2}	0.4522×10^{-2}
	24	0.7885	0.2772×10^{-2}	0.2465×10^{-2}	0.2687×10^{-2}	0.2504×10^{-2}	0.2729×10^{-2}
	90	0.7969	0.9042×10^{-3}	0.8659×10^{-3}	0.8937×10^{-3}	0.8847×10^{-3}	0.9131×10^{-3}

* Observed sampling variances $\mathrm{Var}(\hat{\rho})$ were found based on 10000 replicated samples.
# Predicted sampling variances $\mathrm{Var}_{OP}(\hat{\rho}|\rho)$ and $\mathrm{Var}_{ZG}(\hat{\rho}|\rho)$ were calculated by formulas (3.24) and (3.26), respectively.
¤ Estimated sampling variances $\widehat{\mathrm{Var}}_{OP}(\hat{\rho}|\hat{\rho})$ and $\widehat{\mathrm{Var}}_{ZG}(\hat{\rho}|\hat{\rho})$ were calculated by formulas (3.24) and (3.26), respectively, with $\hat{\rho}$ substituted for ρ, and based on 10000 replicated samples.

Comparing the observed variances $\mathrm{Var}(\hat{\rho})$ with the predicted variances $\mathrm{Var}_{OP}(\hat{\rho}|\rho)$ and $\mathrm{Var}_{ZG}(\hat{\rho}|\rho)$ in Table 3.5, the following conclusion can be made. For admissible heritability values both Osborne's and Paterson's approximation and Zerbe's and Goldgar's approximation are quite exact. Only in the case of small number of groups (sires) the approximated predicted variances are a little overestimated, being more imprecise using Zerbe's and Goldgar's approximation (3.26). For large heritability values and especially for large intraclass correlation values (corresponding to inadmissible heritabilities) the approximated predicted sampling variances are underestimated, being most imprecise for small number of groups. The same conclusions were made by Visscher (1998) and by Donner and Koval (1983).

Conclusions dealing with predicted standard deviations $\sigma_{OP}(\hat{\rho}|\rho)$ and $\sigma_{ZG}(\hat{\rho}|\rho)$ presented in Table 3.6 are similar to those concerning sample variances. For small intraclass correlations both approximations are quite precise,

being a little overestimated in Zerbe's and Goldgar's expression. The standard deviations for large intraclass correlation coefficients are underestimated.

Table 3.6. The observed, predicted and estimated standard deviations of the intraclass correlation coefficient estimate $\hat{\rho}$ in the case of different true population values and number of groups (sires) a. $N = 360$, $\sigma_e^2 = 1$.

ρ (h^2)	a	$E(\hat{\rho})$	$\sigma(\hat{\rho})^*$	$\sigma_{OP}(\hat{\rho}\|\rho)^\#$	$\hat{\sigma}_{OP}(\hat{\rho}\|\hat{\rho})^\square$	$\sigma_{ZG}(\hat{\rho}\|\rho)^\#$	$\hat{\sigma}_{ZG}(\hat{\rho}\|\hat{\rho})^\square$
0.0125	4	0.0124	0.0190	0.0190	0.0187	0.0192	0.0188
(0.05)	15	0.0123	0.0204	0.0204	0.0202	0.0206	0.0204
	24	0.0122	0.0239	0.0236	0.0235	0.0238	0.0237
	90	0.0121	0.0443	0.0443	0.0440	0.0448	0.0445
0.0625	4	0.0595	0.0518	0.0561	0.0518	0.0565	0.0523
(0.25)	15	0.0612	0.0361	0.0368	0.0359	0.0370	0.0361
	24	0.0617	0.0357	0.0358	0.0352	0.0360	0.0355
	90	0.0624	0.0477	0.0482	0.0479	0.0487	0.0484
0.15	4	0.1382	0.0994	0.1113	0.0965	0.1121	0.0972
(0.6)	15	0.1462	0.0587	0.0609	0.0587	0.0613	0.0591
	24	0.1477	0.0527	0.0536	0.0524	0.0540	0.0528
	90	0.1497	0.0529	0.0533	0.0530	0.0539	0.0535
0.8	4	0.7048	0.1879	0.1317	0.1430	0.1327	0.1440
(3.2)	15	0.7799	0.0705	0.0624	0.0652	0.0629	0.0657
	24	0.7885	0.0526	0.0487	0.0510	0.0500	0.0514
	90	0.7969	0.0301	0.0294	0.0297	0.0297	0.0300

* Observed standard deviations $\sigma(\hat{\rho})$ were found based on 10000 replicated samples.
$^\#$ Predicted standard deviations $\sigma_{OP}(\hat{\rho}|\rho)$ and $\sigma_{ZG}(\hat{\rho}|\rho)$ were calculated as square roots of formulas (3.24) and (3.26), respectively.
$^\square$ Estimated standard errors $\hat{\sigma}_{OP}(\hat{\rho}\,|\,\hat{\rho})$ and $\hat{\sigma}_{ZG}(\hat{\rho}|\hat{\rho})$ were calculated as square roots of formulas (3.24) and (3.26), respectively, with $\hat{\rho}$ substituted for ρ, and based on 10000 replicated samples.

The accuracy of the expressions for the sample variances and standard deviations of the intraclass correlation coefficient estimates in the case of admissible heritabilities gives good opportunities for studies in data designs, and for additional simulations.

In real data analyses the population values are not known and sample variances are estimated by formulas (3.24) or (3.26) with $\hat{\rho}$ substituted for ρ. In all cases the estimated sampling variances are larger than the corresponding theoretical approximations. As can be seen from Table 3.5, the sampling variances of the intraclass correlation coefficient are overestimated for admissible heritability values, especially in the case of small number of groups (sires). For bigger intraclass correlation values the sampling variances are underestimated, but less than with the theoretical approximations assuming population parameters

known. In the case of big number of groups, the estimated variances of $\hat{\rho}$ got with Zerbe's and Goldgar's approximation quite precise.

Unlike the estimated sampling variances, the standard errors of the estimated intraclass correlation coefficients in the case of admissible heritability coefficient values are quite precise. If the Osborne's and Paterson's approximation gives somewhat underestimated values to the estimated $\sigma(\hat{\rho})$, then the Zerbe's and Goldgar's approximation is more precise. For large intraclass correlation values and small number of groups the standard errors of $\hat{\rho}$ are underestimated. As the last situation is impossible in genetic applications like the sire model, then the use of the approximated standard errors of heritabilities is justified.

3.2.3. The mean square errors of predictors

The following theorem gives an exact expression for the variance-covariance matrix of prediction errors in the balanced one-way ANOVA model, allowing to study the effect of data design.

Theorem 3.1. *In the one-way balanced random model under the normality assumptions the variance-covariance matrix of prediction errors and mean square error of predictors are expressed by the following formulas*:

$$\mathrm{Var}(\hat{\mathbf{u}}-\mathbf{u}) = \frac{\sigma_u^2\sigma_e^2}{\sigma_e^2+n\sigma_u^2}\mathbf{I}_a + \frac{n(\sigma_u^2)^2}{a(\sigma_e^2+n\sigma_u^2)}\mathbf{J}_a = \left\{ \underset{\mathrm{m}}{\frac{\delta_{ij}a\sigma_u^2\sigma_e^2+n(\sigma_u^2)^2}{a(\sigma_e^2+n\sigma_u^2)}} \right\}_{i,j=1}^{a}, \tag{3.27}$$

where $\delta_{ij} = \begin{cases} 1, & i=j \\ 0, & i\neq j \end{cases}$, *and*

$$\mathrm{MSE}(\hat{\mathbf{u}}) = \frac{\sigma_u^2(a\sigma_e^2+n\sigma_u^2)}{\sigma_e^2+n\sigma_u^2}. \tag{3.28}$$

Proof. By formulas (1.12) and (1.13) the variance-covariance matrix of prediction errors is expressed as

$$\mathrm{Var}(\hat{\mathbf{u}}-\mathbf{u}) = \mathbf{G}-\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G}+\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G}. \tag{3.29}$$

Design matrices $\mathbf{X}$ and $\mathbf{Z}$ corresponding to model (3.1) are expressed as follows: $\mathbf{X}=\mathbf{1}_N$ and $\mathbf{Z}=\{_{\mathrm{d}}\mathbf{1}_n\}_{i=1}^{a}$. Variance-covariance matrix of random effects $\mathbf{u}$ is $\mathbf{G}=\sigma_u^2\mathbf{I}_a$ and the inverse of variance-covariance matrix $\mathbf{V}$ (3.2) is by Corollary 3.1 equal to

$$\mathbf{V}^{-1} = \mathbf{I}_a \otimes \frac{1}{\sigma_e^2}\left(\mathbf{I}_n - \frac{\sigma_u^2}{\sigma_e^2+n\sigma_u^2}\mathbf{J}_n\right) = \left\{ \underset{\mathrm{d}}{\frac{1}{\sigma_e^2}\left(\mathbf{I}_n - \frac{\sigma_u^2}{\sigma_e^2+n\sigma_u^2}\mathbf{J}_n\right)} \right\}_{i=1}^{a}.$$

After some matrix algebra we get that the two first terms in (3.29) add up to

$$\mathbf{G}-\mathbf{GZ}'\mathbf{V}^{-1}\mathbf{ZG}=\sigma_u^2\mathbf{I}_a-\frac{n(\sigma_u^2)^2}{\sigma_e^2+n\sigma_u^2}\mathbf{I}_a=\frac{\sigma_u^2\sigma_e^2}{\sigma_e^2+n\sigma_u^2}\mathbf{I}_a\,. \tag{3.30}$$

For the third component in the expression (3.29) we have first that the middle term $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-$ is scalar because the design matrix $\mathbf{X}$ is $N\times 1$ vector. The invertible matrix $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$ has a form

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}=\mathbf{1}'_N\left\{\mathrm{d}\,\frac{1}{\sigma_e^2}\left(\mathbf{I}_n-\frac{\sigma_u^2}{\sigma_e^2+n\sigma_u^2}\mathbf{J}_n\right)\right\}_{i=1}^{a}\mathbf{1}_N=\frac{1}{\sigma_e^2}\left(N-\frac{Nn\sigma_u^2}{\sigma_e^2+n\sigma_u^2}\right)=\frac{N}{\sigma_e^2+n\sigma_u^2}\,,$$

from which we have that

$$(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-=\frac{\sigma_e^2+n\sigma_u^2}{N}\,. \tag{3.31}$$

The other terms in the third component of expression (3.29) can be presented as follows:

$$\begin{aligned}(\mathbf{GZ}'\mathbf{V}^{-1}\mathbf{X})(\mathbf{X}'\mathbf{V}^{-1}\mathbf{ZG})&=\left(\frac{n\sigma_u^2}{\sigma_e^2+n\sigma_u^2}\mathbf{1}_a\right)\left(\frac{n\sigma_u^2}{\sigma_e^2+n\sigma_u^2}\mathbf{1}'_a\right)\\&=\frac{(n\sigma_u^2)^2}{(\sigma_e^2+n\sigma_u^2)^2}\mathbf{J}_a\,.\end{aligned} \tag{3.32}$$

If we substitute the expressions (3.30), (3.31) and (3.32) into (3.29) we get that

$$\begin{aligned}\mathrm{Var}(\hat{\mathbf{u}}-\mathbf{u})&=\frac{\sigma_u^2\sigma_e^2}{\sigma_e^2+n\sigma_u^2}\mathbf{I}_a+\frac{\sigma_e^2+n\sigma_u^2}{N}\times\frac{(n\sigma_u^2)^2}{(\sigma_e^2+n\sigma_u^2)^2}\mathbf{J}_a\\&=\frac{\sigma_u^2\sigma_e^2}{\sigma_e^2+n\sigma_u^2}\mathbf{I}_a+\frac{n(\sigma_u^2)^2}{a(\sigma_e^2+n\sigma_u^2)}\mathbf{J}_a\,,\end{aligned}$$

which is exactly the formula (3.27) in the theorem statement.

The mean square error of predictors equals to the sum of the diagonal elements of the matrix (3.27):

$$\mathrm{MSE}(\hat{\mathbf{u}})=\sum_{i=1}^{a}\left(\frac{\sigma_u^2\sigma_e^2}{\sigma_e^2+n\sigma_u^2}+\frac{n(\sigma_u^2)^2}{a(\sigma_e^2+n\sigma_u^2)}\right)=\frac{\sigma_u^2(a\sigma_e^2+n\sigma_u^2)}{\sigma_e^2+n\sigma_u^2}\,.$$

The latter is equal to the expression (3.28) in the theorem statement, which ends the proof. ■

Corollary 3.2. *In the one-way balanced random model under the normality assumptions the mean square error of predictor $\hat{u}_i$ is expressed as*

$$\mathrm{MSE}(\hat{u}_i)=\frac{\sigma_u^2(a\sigma_e^2+n\sigma_u^2)}{a(\sigma_e^2+n\sigma_u^2)} \tag{3.33}$$

and naive estimated as

$$\widehat{\mathrm{MSE}}(\hat{u}_i) = \frac{\hat{\sigma}_u^2(a\hat{\sigma}_e^2 + n\hat{\sigma}_u^2)}{a(\hat{\sigma}_e^2 + n\hat{\sigma}_u^2)}\,. \tag{3.34}$$

■

Taking into account the sampling variance of the estimators of the variance components the approximated formula for the mean square error of two-stage predictors is given in the following theorem.

Theorem 3.2. *In the one-way balanced random model under the normality assumptions the variance-covariance matrix of two-stage prediction errors and the mean square error of two-stage predictors are expressed by the following formulas*:

$$\begin{aligned}\mathrm{Var}(\tilde{\mathbf{u}} - \mathbf{u}) \approx{}& \mathrm{Var}(\hat{\mathbf{u}} - \mathbf{u}) + \mathrm{Var}(\hat{\sigma}_e^2) \times (\sigma_u^2)^2 \mathbf{Z}'\mathbf{V}^{-2}\mathbf{P}\mathbf{Z} \\ &+ \mathrm{Var}(\hat{\sigma}_u^2) \times (\mathbf{I}_a - \sigma_u^2\mathbf{Z}'\mathbf{P}\mathbf{Z})\mathbf{Z}'\mathbf{P}\mathbf{Z}(\mathbf{I}_a - \sigma_u^2\mathbf{Z}'\mathbf{P}\mathbf{Z})\end{aligned} \tag{3.35}$$

and

$$\mathrm{MSE}(\tilde{\mathbf{u}}) \approx \mathrm{MSE}(\hat{\mathbf{u}}) + \frac{n(a-1)}{(\sigma_e^2 + n\sigma_u^2)^3}\left[(\sigma_u^2)^2\,\mathrm{Var}(\hat{\sigma}_e^2) + (\sigma_e^2)^2\,\mathrm{Var}(\hat{\sigma}_u^2)\right]. \tag{3.36}$$

Proof. To derive the expression for $\mathrm{Var}(\tilde{\mathbf{u}} - \mathbf{u})$ corresponding to model (3.1), we proceed from general expression (1.21) derived in Corollary 1.1. As in the model (3.1) $\mathbf{G} = \sigma_u^2\mathbf{I}_a$ and $\mathbf{Z}_0 = \mathbf{I}_N$, we get that

$$\begin{aligned}\mathrm{Var}(\tilde{\mathbf{u}} - \mathbf{u}) \approx{}& \mathrm{Var}(\hat{\mathbf{u}} - \mathbf{u}) + \mathrm{Var}(\hat{\sigma}_e^2) \times (\sigma_u^2)^2 \mathbf{Z}'\mathbf{P}\mathbf{P}\mathbf{P}\mathbf{Z} \\ &+ \mathrm{Var}(\hat{\sigma}_u^2) \times (\mathbf{I}_a - \sigma_u^2\mathbf{Z}'\mathbf{P}\mathbf{Z})\mathbf{Z}'\mathbf{P}\mathbf{Z}(\mathbf{I}_a - \sigma_u^2\mathbf{Z}'\mathbf{P}\mathbf{Z}).\end{aligned} \tag{3.37}$$

The matrix $\mathbf{P}$ in this approximation is given in formula (1.11) and can be expressed also as the following matrix product: $\mathbf{P} = \mathbf{V}^{-1} \times [\mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}]$. Here the part $\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}$ can be referred as projection matrix $\mathbf{P}_{\mathbf{X},\mathbf{V}^{-1}}$. Then by the Proposition 3.3 we have that

$$\begin{aligned}&[\mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}]' \times \mathbf{V}^{-1} \times [\mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}] \\ &= \mathbf{V}^{-1} \times [\mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}].\end{aligned}$$

As $\mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}$ is symmetric in the balanced case, then $\mathbf{P}\mathbf{P}\mathbf{P} = \mathbf{V}^{-1}\mathbf{V}^{-1}\mathbf{P}$. The substitution of the last identity into (3.37) results in expression (3.35) in the theorem statement.

To derive $\mathrm{MSE}(\tilde{\mathbf{u}})$, we first note that

$$\begin{aligned}\mathrm{MSE}(\tilde{\mathbf{u}}) &= \mathrm{tr}[\mathrm{Var}(\tilde{\mathbf{u}} - \mathbf{u})] = \mathrm{tr}[\mathrm{Var}(\hat{\mathbf{u}} - \mathbf{u})] + \mathrm{tr}[\mathrm{Var}(\tilde{\mathbf{u}} - \hat{\mathbf{u}})] \\ &= \mathrm{MSE}(\hat{\mathbf{u}}) + \mathrm{tr}[\mathrm{Var}(\tilde{\mathbf{u}} - \hat{\mathbf{u}})].\end{aligned}$$

Using (3.35) and partition of $\mathbf{P}$, the trace of $\mathrm{Var}(\tilde{\mathbf{u}} - \hat{\mathbf{u}})$ can be divided into eight terms, which after applying the first five trace properties from Proposition 3.1 are expressed as follows:

$$\begin{aligned}\mathrm{tr}[\mathrm{Var}(\tilde{\mathbf{u}}-\hat{\mathbf{u}})] \approx\ & \mathrm{Var}(\hat{\sigma}_e^2)\times(\sigma_u^2)^2\times\left\{\mathrm{tr}(\mathbf{ZZ}'\mathbf{V}^{-3})-(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\,\mathrm{tr}\left[\mathbf{ZZ}'\mathbf{XX}'\mathbf{V}^{-4}\right]\right\}\\ &+\mathrm{Var}(\hat{\sigma}_u^2)\times\left\{\mathrm{tr}(\mathbf{ZZ}'\mathbf{V}^{-1})-(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\,\mathrm{tr}(\mathbf{ZZ}'\mathbf{XX}'\mathbf{V}^{-2})\right.\\ &-2\sigma_u^2\times\left[\mathrm{tr}(\mathbf{ZZ}'\mathbf{ZZ}'\mathbf{V}^{-2})-(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\,\mathrm{tr}(\mathbf{ZZ}'\mathbf{ZZ}'\mathbf{XX}'\mathbf{V}^{-3})\right]\\ &\left.+(\sigma_u^2)^2\times\left[\mathrm{tr}(\mathbf{ZZ}'\mathbf{ZZ}'\mathbf{ZZ}'\mathbf{V}^{-3})-(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\,\mathrm{tr}(\mathbf{ZZ}'\mathbf{ZZ}'\mathbf{ZZ}'\mathbf{XX}'\mathbf{V}^{-4})\right]\right\}.\end{aligned} \tag{3.38}$$

In deriving this expression also the facts that $\mathbf{ZZ}'$, $\mathbf{XX}'$ and $\mathbf{V}^{-1}$ are symmetric, and that $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}$ is scalar, were used.

As $\mathbf{X}=\mathbf{1}_N$ and $\mathbf{Z}=\{\mathrm{d}\,\mathbf{1}_n\}_{i=1}^a$, then $\mathbf{XX}'=\mathbf{J}_N$, $\mathbf{ZZ}'=\{\mathrm{d}\,\mathbf{J}_n\}_{i=1}^a$, $\mathbf{ZZ}'\mathbf{ZZ}'=\{\mathrm{d}\,n\mathbf{J}_n\}_{i=1}^a$ and $\mathbf{ZZ}'\mathbf{XX}'=\{\mathrm{d}\,n\mathbf{J}_n\}_{i=1}^a$. Analogously we get that $\mathbf{ZZ}'\mathbf{ZZ}'\mathbf{ZZ}'\mathbf{XX}'=\{\mathrm{d}\,n^3\mathbf{J}_n\}_{i=1}^a$.

Next we notice that according to the expression of $\mathbf{V}^{-1}$ and by the form of the products of design matrices, all arguments of the trace operations in expression (3.38) are block diagonal matrices with block dimensions $n\times n$. By statement (vii) in Proposition 3.1 we can take trace separately from all diagonal blocks and then sum the results. Denoting the exponent of n in products of design matrices by j and exponent of $\mathbf{V}^{-1}$ in (3.38) by h, we can represent each diagonal block of matrices under trace operation in a form

$$n^j\mathbf{J}_n\times\left[\frac{1}{\sigma_e^2}\left(\mathbf{I}_n-\frac{\sigma_u^2}{\sigma_e^2+n\sigma_u^2}\mathbf{J}_n\right)\right]^h,$$

which can be more compactly expressed as $n^j\mathbf{J}_n\times[(\mathbf{I}_n-g\mathbf{J}_n)/\sigma_e^2]^h$, using notation

$$g=\sigma_u^2/(\sigma_e^2+n\sigma_u^2).$$

By the statement (vi) in Proposition 3.1, the trace of such a matrix equals to $n^j/(\sigma_e^2)^h$ times sum of elements of $(\mathbf{I}_n-g\mathbf{J}_n)^h$. Following the statement (iii) in Proposition 3.4 we get that matrix $(\mathbf{I}_n-g\mathbf{J}_n)^h$ has n diagonal elements equal to $1+[(1-ng)^h-1]/n$ and $n(n-1)$ offdiagonal elements equal to $[(1-ng)^h-1]/n$. So, the sum of elements of matrix $(\mathbf{I}_n-g\mathbf{J}_n)^h$ is expressed as

$$n\left[1+\frac{(1-ng)^h-1}{n}\right]+n(n-1)\left[\frac{(1-ng)^h-1}{n}\right]=n(1-ng)^h.$$

In summary we get, for example, that the first trace in (3.38) equals to

$$\mathrm{tr}(\mathbf{ZZ}'\mathbf{V}^{-3})=\sum_{i=1}^a n(1-ng)^3/(\sigma_e^2)^3=an(1-ng)^3/(\sigma_e^2)^3.$$

According to the formula (3.31) we know that $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}=(\sigma_e^2+n\sigma_u^2)/N$. Denoting this scalar with c:

$$c=(\sigma_e^2+n\sigma_u^2)/N,$$

we get the second addend in the expression (3.38) as follows:

$$(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\,\mathrm{tr}(\mathbf{ZZ}'\mathbf{XX}'\mathbf{V}^{-4})=c\sum_{i=1}^a n^2(1-ng)^4/(\sigma_e^2)^4=c\,an^2(1-ng)^4/(\sigma_e^2)^4.$$

The other terms in (3.38) can be expressed similarly and finally resulted in the following expression:

$$\begin{aligned}\mathrm{tr}[\mathrm{Var}(\tilde{\mathbf{u}}-\hat{\mathbf{u}})] \approx \mathrm{Var}(\hat{\sigma}_e^2)\times(\sigma_u^2)^2\times\left[an(1-ng)^3/(\sigma_e^2)^3 - c\,an^2(1-ng)^4/(\sigma_e^2)^4\right]\\ +\mathrm{Var}(\hat{\sigma}_u^2)\times\left\{an(1-ng)/\sigma_e^2 - c\,an^2(1-ng)^2/(\sigma_e^2)^2\right.\\ -2\sigma_u^2\times\left[an^2(1-ng)^2/(\sigma_e^2)^2 - c\,an^3(1-ng)^3/(\sigma_e^2)^3\right]\\ \left.+(\sigma_u^2)^2\times\left[an^3(1-ng)^3/(\sigma_e^2)^3 - c\,an^4(1-ng)^4/(\sigma_e^2)^4\right]\right\}.\end{aligned}$$

After reducing and bringing front similar terms we got approximated $\mathrm{tr}[\mathrm{Var}(\tilde{\mathbf{u}}-\hat{\mathbf{u}})]$ of the form

$$\mathrm{tr}[\mathrm{Var}(\tilde{\mathbf{u}}-\hat{\mathbf{u}})] \approx \frac{an(1-ng)}{\sigma_e^2}\left[1-\frac{cn(1-ng)}{\sigma_e^2}\right]\times\left[g^2\,\mathrm{Var}(\hat{\sigma}_e^2)+(1-ng)^2\,\mathrm{Var}(\hat{\sigma}_u^2)\right].$$

To simplify the last formula we note that $1-ng=\sigma_e^2/(\sigma_e^2+n\sigma_u^2)$ and

$$1-\frac{cn(1-ng)}{\sigma_e^2}=1-\frac{(\sigma_e^2+n\sigma_u^2)}{a\sigma_e^2}\times\frac{\sigma_e^2}{\sigma_e^2+n\sigma_u^2}=\frac{a-1}{a}.$$

Now we substitute these expressions and get that

$$\mathrm{tr}[\mathrm{Var}(\tilde{\mathbf{u}}-\hat{\mathbf{u}})] \approx \frac{n(a-1)}{(\sigma_e^2+n\sigma_u^2)^3}\left[(\sigma_u^2)^2\,\mathrm{Var}(\hat{\sigma}_e^2)+(\sigma_e^2)^2\,\mathrm{Var}(\hat{\sigma}_u^2)\right].$$

So,

$$\mathrm{MSE}(\tilde{\mathbf{u}}) \approx \mathrm{MSE}(\hat{\mathbf{u}}) + \frac{n(a-1)}{(\sigma_e^2+n\sigma_u^2)^3}\left[(\sigma_u^2)^2\,\mathrm{Var}(\hat{\sigma}_e^2)+(\sigma_e^2)^2\,\mathrm{Var}(\hat{\sigma}_u^2)\right],$$

which establishes the theorem. ■

The second approximation in the next corollary follows from expression (1.20).

Corollary 3.3. *In the one-way balanced random model under the normality assumptions the mean square error of two-stage predictor $\tilde{u}_i$ is expressed as*

$$\mathrm{MSE}(\tilde{u}_i) \approx \mathrm{MSE}(\hat{u}_i) + \frac{n(a-1)}{a(\sigma_e^2+n\sigma_u^2)^3}\left[(\sigma_u^2)^2\,\mathrm{Var}(\hat{\sigma}_e^2)+(\sigma_e^2)^2\,\mathrm{Var}(\hat{\sigma}_u^2)\right] \quad (3.39)$$

and unbiasedly estimated approximately as

$$\widehat{\mathrm{MSE}}(\tilde{u}_i) \approx \widehat{\mathrm{MSE}}(\hat{u}_i) + \frac{2n(a-1)}{a(\hat{\sigma}_e^2+n\hat{\sigma}_u^2)^3}\left[(\hat{\sigma}_u^2)^2\,\mathrm{Var}(\hat{\sigma}_e^2)+(\hat{\sigma}_e^2)^2\,\mathrm{Var}(\hat{\sigma}_u^2)\right]. \quad (3.40)$$

■

To study the accuracy of the expressions (3.33), (3.34), (3.39) and (3.40), and to examine the effect of using estimated variance components instead of their true population values, similarly to the previous sections a simulation study was performed. The compared parameters were the observed mean square errors of

predictors $\hat{u}_i$ and $\tilde{u}_i$ based on 10000 replicated samples; the predicted mean square errors of predictors $\hat{u}_i$ and $\tilde{u}_i$ calculated by formulas (3.33) and (3.39), respectively; and the average estimated mean square errors based on 10000 replicated samples. In calculating the estimated $\mathrm{MSE}(\tilde{u}_i)$ by expression (3.40), sampling variances of variance components σ_u^2 and σ_e^2 were found by formulas (3.22) and (3.23), respectively. Simulations results showing the mean square errors of random effect u_i are presented in Table 3.7.

Based on simulations results presented in Table 3.7, the formula (3.33), giving the mean square errors of random effects u_i (denoted as $\mathrm{MSE}(\hat{u}_i \mid \sigma_{e,u}^2)$ in the table), is correct – in all cases the observed and predicted mean square errors are identical. This allows performing precise studies in data designs. In the point of genetic applications, where population genetic parameters like heritabilities and intraclass correlations are often assumed known, the correctness of the mean square errors means precise estimates to the accuracy of random genetic effects.

Table 3.7. The observed, predicted and estimated mean square errors of random effect u_i in the case of different true population values and number of groups (sires) a. Population size $N = 360$ and error variance $\sigma_e^2 = 1$ were used.

ρ (h^2)	a	$\mathrm{MSE}(\hat{u}_i)^*$	$\mathrm{MSE}(\hat{u}_i \mid \sigma_{e,u}^2)^\#$	$\widehat{\mathrm{MSE}}(\hat{u}_i \mid \hat{\sigma}_{e,u}^2)^\square$	$\mathrm{MSE}(\tilde{u}_i)^*$	$\mathrm{MSE}(\tilde{u}_i \mid \sigma_{e,u}^2)^\#$	$\widehat{\mathrm{MSE}}(\tilde{u}_i \mid \hat{\sigma}_{e,u}^2)^\square$
0.0125	4	0.0076	0.0076	-0.0001	0.0157	0.0102	3.9413
(0.05)	15	0.0099	0.0099	0.0046	0.0151	0.0143	0.0141
	24	0.0107	0.0107	0.0055	0.0162	0.0156	0.0153
	90	0.0121	0.0121	0.0045	0.0193	0.0189	0.0190
0.0625	4	0.0240	0.0238	0.0215	0.0261	0.0246	0.0297
(0.25)	15	0.0284	0.0284	0.0257	0.0311	0.0306	0.0302
	24	0.0348	0.0347	0.0314	0.0381	0.0376	0.0373
	90	0.0529	0.0528	0.0471	0.0588	0.0581	0.0581
0.15	4	0.0515	0.0520	0.0511	0.0526	0.0523	0.0805
(0.6)	15	0.0432	0.0432	0.0419	0.0446	0.0443	0.0441
	24	0.0536	0.0537	0.0519	0.0554	0.0553	0.0551
	90	0.1043	0.1043	0.1002	0.1087	0.1081	0.1079
0.8	4	1.0067	1.0083	0.9915	1.0068	1.0083	0.9916
(3.2)	15	0.3022	0.3052	0.3035	0.3023	0.3052	0.3036
	24	0.2303	0.2295	0.2291	0.2304	0.2296	0.2293
	90	0.2763	0.2771	0.2768	0.2768	0.2771	0.2768

* Observed mean square errors $\mathrm{MSE}(\hat{u}_i)$ and $\mathrm{MSE}(\tilde{u}_i)$ were found based on 10000 replicated samples.

Predicted mean square errors $\mathrm{MSE}(\hat{u}_i \mid \sigma_{e,u}^2)$ and $\mathrm{MSE}(\tilde{u}_i \mid \sigma_{e,u}^2)$ were calculated by formulas (3.33) and (3.39), respectively.

□ Estimated mean square errors $\widehat{\mathrm{MSE}}(\hat{u}_i \mid \hat{\sigma}_{e,u}^2)$ and $\widehat{\mathrm{MSE}}(\tilde{u}_i \mid \hat{\sigma}_{e,u}^2)$ were calculated by formulas (3.34) and (3.40), respectively, with $\hat{\sigma}_u^2$ substituted for σ_u^2 and $\hat{\sigma}_e^2$ substituted for σ_e^2, and based on 10000 replicated samples.

If the estimated variance components (or their ratios) are used in prediction, then the additional variability should be taken into account and the approximated expression of the mean square error of a two-stage prediction might be used. Simulations show that in the case of admissible heritability values the observed $\mathrm{MSE}(\tilde{u}_i)$ is underestimated using approximation (3.39) resulted with $\mathrm{MSE}(\tilde{u}_i \mid \sigma^2_{e,u})$ in Table 3.7. Inaccuracy is largest for small heritabilities and small number of groups, which is logical, as the approximated formulas for $\mathrm{MSE}(\tilde{u}_i)$ are derived, assuming large sample size and group number.

In real data analysis it is standard practice that the accuracy of calculated predictors $\tilde{u}_i$ is characterised via the estimated $\mathrm{MSE}(\hat{u}_i)$. Simulations showed that due to such action the real mean square error of predictor $\tilde{u}_i$ is underestimated, especially in the case of small values of the intraclass correlation coefficient.

The estimated mean square errors of two-stage predictors $\tilde{u}_i$ calculated with (3.40) are in the case of not too small number of groups and in the range of admissible heritability values also somewhat underestimated. But this bias is small compared with the differences between the observed $\mathrm{MSE}(\tilde{u}_i)$ and estimated $\mathrm{MSE}(\hat{u}_i)$ which are equated in practical studies. In the case of big values of the intraclass correlation coefficient there seems to be no need for additional addend, considering the variability of the variance components estimators in the expression of the $\mathrm{MSE}(\tilde{u}_i)$. Both estimated mean square errors of the predictors $\hat{u}_i$ and $\tilde{u}_i$ are very imprecise and variable in the case of small number of groups and small heritabilities. Then the $\mathrm{MSE}(\hat{u}_i)$ is strongly underestimated, resulting even in a negative average estimated mean square error of $\hat{u}_i$; at the same time the $\mathrm{MSE}(\tilde{u}_i)$ is overestimated, being in several cases – due to the inappropriate estimates of variance components and their variances – more than 1000 times bigger than the observed $\mathrm{MSE}(\tilde{u}_i)$.

Similarly to the variance components and heritability estimation where the accuracy of estimates is usually examined based on standard errors instead of sampling variances, also in the prediction of random effects the standard deviations of prediction errors $\sqrt{\mathrm{MSE}(\hat{u}_i)} = \sigma(\hat{u}_i - u_i)$ and $\sqrt{\mathrm{MSE}(\tilde{u}_i)} = \sigma(\tilde{u}_i - u_i)$ are used instead of variances $\mathrm{MSE}(\hat{u}_i) = \mathrm{Var}(\hat{u}_i - u_i)$ and $\mathrm{MSE}(\tilde{u}_i) = \mathrm{Var}(\tilde{u}_i - u_i)$. The problem in studying the standard deviations is that the estimated variances of prediction errors calculated by formulas (3.34) and (3.40) can be negative, not allowing to take the square root and to find the standard deviations. The probability to get such inadmissible estimates is discussed in the following paragraph. In simulation studies the estimated standard deviations of prediction errors were found in two ways – at first the negative estimates were taken equal to zero and secondly the negative estimates were omitted and only positive or zero estimates were used. Wang et al (1992), who studied the negative estimates of variance components, called the first type of estimators *concentrated estimators* and second type of estimators *truncated estimators*. The observed standard errors were found also in two ways – using all generated samples and using

only samples corresponding to non-negative estimated mean square error of predictor $\hat{u}_i$.

The simulations results concerning the standard deviations of prediction errors $\sigma(\hat{u}_i - u_i)$ are presented in Table 3.8 and the simulations results dealing with the standard deviations of prediction errors $\sigma(\tilde{u}_i - u_i)$ are presented in Table 3.9.

Study of the standard deviations of prediction errors (Table 3.8 and Table 3.9) shows that the predicted standard deviations of prediction errors calculated by theoretical formulas and assuming the parameters real values known give over-estimated values to the observed standard errors.

Table 3.8. The observed, predicted and estimated standard deviations of prediction errors in the case of different true population values and number of groups (sires) a. Population size $N = 360$ and error variance $\sigma_e^2 = 1$ were used.

ρ (h^2)	a	$\sigma(\hat{u}_i - u_i)^*$	$\sigma_+(\hat{u}_i - u_i)^*$	$\sigma(\hat{u}_i - u_i \mid \sigma^2_{e,u})^{\#}$	$\hat{\sigma}_0(\hat{u}_i - u_i \mid \hat{\sigma}^2_{e,u})^{¤}$	$\hat{\sigma}_+(\hat{u}_i - u_i \mid \hat{\sigma}^2_{e,u})^{¤}$
0.0125	4	0.0816	0.0818	0.0872	0.0640	0.0914
(0.05)	15	0.0980	0.0978	0.0995	0.0763	0.1094
	24	0.1026	0.1024	0.1035	0.0830	0.1212
	90	0.1096	0.1097	0.1098	0.1011	0.1699
0.0625	4	0.1396	0.1394	0.1543	0.1367	0.1465
(0.25)	15	0.1655	0.1655	0.1685	0.1563	0.1598
	24	0.1845	0.1845	0.1863	0.1724	0.1763
	90	0.2293	0.2294	0.2298	0.2038	0.2244
0.15	4	0.1967	0.1966	0.2279	0.2100	0.2124
(0.6)	15	0.2029	0.2029	0.2079	0.2033	0.2034
	24	0.2287	0.2287	0.2318	0.2268	0.2269
	90	0.3220	0.3220	0.3229	0.3136	0.3140
0.8	4	0.8091	0.8091	1.0041	0.9192	0.9194
(3.2)	15	0.4770	0.4770	0.5524	0.5435	0.5435
	24	0.4367	0.4367	0.4791	0.4759	0.4759
	90	0.5219	0.5219	0.5264	0.5257	0.5257

* Observed standard deviations of prediction errors $\sigma(\hat{u}_i - u_i)$ and $\sigma_+(\hat{u}_i - u_i)$ were found based on 10000 replicated samples and based only on samples with non-negative estimates of $\mathrm{MSE}(\hat{u}_i)$, respectively.

Predicted standard deviations $\sigma(\hat{u}_i - u_i \mid \sigma^2_{e,u})$ were calculated as square roots of formula (3.33).

¤ Estimated standard deviations $\hat{\sigma}_0(\hat{u}_i - u_i \mid \hat{\sigma}^2_{e,u})$ and $\hat{\sigma}_+(\hat{u}_i - u_i \mid \hat{\sigma}^2_{e,u})$ were calculated as square roots of formula (3.34), based on 10000 replicated samples and based only on non-negative estimates of $\mathrm{MSE}(\hat{u}_i)$, respectively, with $\hat{\sigma}^2_u$ substituted for σ^2_u and $\hat{\sigma}^2_e$ substituted for σ^2_e, and based on 10000 replicated samples.

Table 3.9. The observed, predicted and estimated standard deviations of two-stage prediction errors in the case of different true population values and number of groups (sires) a. Population size $N = 360$ and error variance $\sigma_e^2 = 1$ were used.

ρ (h^2)	a	$\sigma(\tilde{u}_i - u_i)^*$	$\sigma_+(\tilde{u}_i - u_i)^*$	$\sigma(\tilde{u}_i - u_i \mid \sigma^2_{e,u})^{\#}$	$\hat{\sigma}_0(\tilde{u}_i - u_i \mid \hat{\sigma}^2_{e,u})^{\square}$	$\hat{\sigma}_+(\tilde{u}_i - u_i \mid \hat{\sigma}^2_{e,u})^{\square}$
0.0125	4	0.1062	0.1062	0.1010	0.1495	0.1495
(0.05)	15	0.1181	0.1110	0.1195	0.1119	0.1262
	24	0.1242	0.1158	0.1250	0.1387	0.1165
	90	0.1351	0.1262	0.1375	0.1334	0.1836
0.0625	4	0.1459	0.1459	0.1569	0.1519	0.1519
(0.25)	15	0.1728	0.1723	0.1748	0.1718	0.1727
	24	0.1929	0.1925	0.1939	0.1905	0.1914
	90	0.2413	0.2390	0.2411	0.2304	0.2417
0.15	4	0.1994	0.1994	0.2287	0.2176	0.2176
(0.6)	15	0.2060	0.2060	0.2105	0.2091	0.2091
	24	0.2326	0.2325	0.2352	0.2341	0.2341
	90	0.3286	0.3286	0.3287	0.3264	0.3265
0.8	4	0.8091	0.8091	1.0041	0.9193	0.9194
(3.2)	15	0.4771	0.4771	0.5525	0.5436	0.5436
	24	0.4368	0.4368	0.4792	0.4761	0.4761
	90	0.5223	0.5223	0.5264	0.5257	0.5257

* Observed standard deviations of prediction errors $\sigma(\tilde{u}_i - u_i)$ and $\sigma_+(\tilde{u}_i - u_i)$ were found based on 10000 replicated samples and based only on samples with non-negative estimates of $\mathrm{MSE}(\tilde{u}_i)$, respectively.

Predicted standard deviations $\sigma(\tilde{u}_i - u_i \mid \sigma^2_{e,u})$ were calculated by square root of formula (3.39).

□ Estimated standard deviations $\hat{\sigma}_0(\tilde{u}_i - u_i \mid \hat{\sigma}^2_{e,u})$ and $\hat{\sigma}_+(\tilde{u}_i - u_i \mid \hat{\sigma}^2_{e,u})$ were calculated as square roots of formula (3.40), based on 10000 replicated samples and based only on non-negative estimates of $\mathrm{MSE}(\tilde{u}_i)$, respectively, with $\hat{\sigma}^2_u$ substituted for σ^2_u and $\hat{\sigma}^2_e$ substituted for σ^2_e, and based on 10000 replicated samples.

Due to the fact that the estimated mean square errors can be negative, two types of estimators of standard errors were calculated. As inadmissible estimates are a problem especially in the case of the intraclass correlation values near to zero, then the differences between the concentrated and truncated estimators appear also if the influences of the random factor are small. The truncated estimators of $\sigma(\hat{u}_i - u_i)$ are overestimated, being quite precise in the case of average heritability values, the concentrated estimators of $\sigma(\hat{u}_i - u_i)$ are underestimated in the case of small intraclass correlations and overestimated in the case of big intraclass correlations. Since in calculating the mean square errors of two-stage predictors $\tilde{u}_i$ an additional, always positive (shown in the next paragraph) term is added, then the chance to get a negative estimated mean square error is smaller and the difference between concentrated and truncated estimators of $\sigma(\tilde{u}_i - u_i)$ is not notable. Both of these estimators are quite precise in the case of average

heritability values and are overestimated in the case of very small or big values of the intraclass correlations coefficient.

3.2.4. The inadmissible estimates

The inadmissible estimates are those outside the permissible limits, which are $[0, \infty]$ for variance components and mean square errors, $[0, 1]$ for intraclass correlation coefficients and heritabilities and $[-1, 1]$ for genetic correlations. The possibility to get negative estimates of variance components, mean square errors and intraclass correlation coefficients is an undesirable feature of the ANOVA estimation. From more complicated methods, applicable in general linear mixed model and reviewed in Section 1.2.5, negative variance components estimates can be obtained with the Henderson III method and the MIVQUE-method.

An excellent – but unfortunately almost without any algebraic derivations and proof – discussion on this problem is given by Henderson (1984). The first study dealing with the probability of inadmissible estimates of variance components was published by Gill and Jensen (1968). Hill and Thompson (1978) extended the theory to the multivariate case, deriving the formula for probability to get non-positive definite covariance matrices with ANOVA. The detailed discussion concerning negative variance components can be found in Searle, Casella and McCulloch (1992) and in Khuri, Mathew and Sinha (1998).

The estimates of the intraclass correlation coefficient cannot be bigger than one. But, depending on the genetic model and the genetic application of the intraclass correlation coefficient, the estimated heritability coefficient can be bigger than one. As this is not an inadmissible estimate in the mathematical sense but only in genetics, the problem of too big estimates is left without attention. Also, the possibility to get negative estimates of mean square errors is not discussed yet.

In the following, the sire model (2.4) in notation (3.1) is used and the probability to get an inadmissible estimate of heritability with the ANOVA method in the balanced case is derived. The proof of the following theorem and the results of the simulation studies, checking the correctness of derived formulas, are also presented in Kaart (1997).

Theorem 3.3. *In the additive genetic sire model under the normality assumptions and in balanced data the probability to get the inadmissible heritability estimates is*

$$\mathrm{P}(\hat{h}^2 < 0) + \mathrm{P}(\hat{h}^2 > 1) = \mathrm{P}\left[\mathrm{F}_{a(n-1),\,a-1} > 1 + n\tau\right] + \mathrm{P}\left[\mathrm{F}_{a(n-1),\,a-1} < (n/3+1)^{-1}(1+n\tau)\right], \tag{3.41}$$

where a is the number of sires, n is the number of analysed progeny on each sire, τ is the ratio of between- and within-sires variance components, $\tau = \sigma_u^2/\sigma_e^2$, and $\mathrm{F}_{a(n-1),(a-1)}$ is the random variable with F-distribution having $a(n-1)$ and $a-1$ degrees of freedom.

Proof. In the 1-way balanced classification (half-sib analysis) the estimates of variance components are represented with equations (3.8) and (3.6). As $\hat{\sigma}_e^2$ is always positive, then from the equation (3.10) we have that $\hat{h}^2 < 0$ if $\hat{\sigma}_u^2 < 0$ and $\hat{h}^2 > 1$ if $\hat{\sigma}_u^2/\hat{\sigma}_e^2 > 1/3$. Knowing that in balanced data the mean squares are proportional to χ^2-distribution, the probability of the negative estimate of heritability is expressed as follows:

$$\begin{aligned} \mathrm{P}(\hat{h}^2 < 0) &= \mathrm{P}(\hat{\sigma}_u^2 < 0) = \mathrm{P}\left[MS(u) < MS(e)\right] \\ &= \mathrm{P}\left[\frac{\chi^2_{a-1}(\sigma_e^2 + n\sigma_u^2)}{a-1} < \frac{\chi^2_{a(n-1)}(\sigma_e^2)}{a(n-1)}\right] \\ &= \mathrm{P}\left[\frac{\chi^2_{a-1}}{a-1} \times \frac{a(n-1)}{\chi^2_{a(n-1)}} \times \frac{\sigma_e^2 + n\sigma_u^2}{\sigma_e^2} < 1\right] \\ &= \mathrm{P}\left[\left(\frac{n\sigma_u^2}{\sigma_e^2} + 1\right)\mathrm{F}_{(a-1),\,a(n-1)} < 1\right] = \mathrm{P}\left[\mathrm{F}_{a(n-1),a-1} > 1 + n\tau\right], \end{aligned} \tag{3.42}$$

where χ^2_{a-1} and $\mathrm{F}_{(a-1),a(n-1)}$ denote the random variables with χ^2-distribution and F-distribution, respectively. The last equality is true due to the fact that the reciprocal of the random variable with $F_{(a-1),a(n-1)}$ distribution has the $F_{a(n-1),a-1}$ distribution.

Similarly,

$$\begin{aligned} \mathrm{P}(\hat{h}^2>1) &= \mathrm{P}\left(\frac{\hat{\sigma}_u^2}{\hat{\sigma}_e^2} > \frac{1}{3}\right) = \mathrm{P}\left[\frac{MS(u)}{MS(e)} > \frac{n}{3} + 1\right] \\ &= \mathrm{P}\left[\left(\frac{n\sigma_u^2}{\sigma_e^2} + 1\right)\mathrm{F}_{a-1,\,a(n-1)} > \frac{n}{3} + 1\right] = \mathrm{P}\left[\mathrm{F}_{a(n-1),\,a-1} < \frac{3(1+n\tau)}{n+3}\right], \end{aligned} \tag{3.43}$$

which establishes the theorem. ■

Using the relationship between the heritability coefficient and the ratio of variance components of the form $\tau = h^2/(4-h^2)$, the probability to get an inadmissible estimate of heritability can be expressed also as

$$\mathrm{P}(\hat{h}^2 < 0) + \mathrm{P}(\hat{h}^2 > 1) = \mathrm{P}\left[\mathrm{F}_{a\text{-}1,a(n\text{-}1)} < \frac{4-h^2}{h^2(n-1)+4}\right] + \mathrm{P}\left[\mathrm{F}_{a(n-1),\,a-1} < \frac{(n+3)(4-h^2)}{3h^2(n-1)+12}\right].$$

In studying the negative estimates of the $\mathrm{MSE}(\hat{u}_i)$ it follows from expression (3.33) that

$$\mathrm{P}\left[\widehat{\mathrm{MSE}}(\hat{u}_i) < 0\right] = \mathrm{P}\left[\frac{\hat{\sigma}_u^2(a\hat{\sigma}_e^2 + n\hat{\sigma}_u^2)}{a(\hat{\sigma}_e^2 + n\hat{\sigma}_u^2)} < 0\right] = \mathrm{P}(\hat{\sigma}_u^2 < 0),$$

because $\mathrm{P}(a\hat{\sigma}_e^2 + n\hat{\sigma}_u^2 < 0) = \mathrm{P}[MS(u) < (1-a)MS(e)] = 0$ and $\mathrm{P}(\hat{\sigma}_e^2 + n\hat{\sigma}_u^2 < 0) = \mathrm{P}[MS(u) < MS(e)] = 0$. From expression (3.40) follows, that

$$\begin{aligned}
&\mathrm{P}\left[\widehat{\mathrm{MSE}}(\tilde{u}_i) < 0\right] \\
&\quad \approx \mathrm{P}\left\{\widehat{\mathrm{MSE}}(\hat{u}_i) + \frac{2n(a-1)}{a(\hat{\sigma}_e^2 + n\hat{\sigma}_u^2)^3}\left[(\hat{\sigma}_u^2)^2 \operatorname{Var}(\hat{\sigma}_e^2) + (\hat{\sigma}_e^2)^2 \operatorname{Var}(\hat{\sigma}_u^2)\right] < 0\right\} \\
&\quad < \mathrm{P}\left[\widehat{\mathrm{MSE}}(\hat{u}_i) < 0\right],
\end{aligned}$$

because

$$\mathrm{P}\left\{\frac{2n(a-1)}{a(\hat{\sigma}_e^2 + n\hat{\sigma}_u^2)^3}\left[(\hat{\sigma}_u^2)^2 \operatorname{Var}(\hat{\sigma}_e^2) + (\hat{\sigma}_e^2)^2 \operatorname{Var}(\hat{\sigma}_u^2)\right] < 0\right\} = 0 .$$

Thus, the estimate of the $\mathrm{MSE}(\hat{u}_i)$ is negative if and only if $\hat{\sigma}_u^2$ is negative, but the estimate of the $\mathrm{MSE}(\tilde{u}_i)$ can be positive even if $\hat{\sigma}_u^2$ is negative. As the latter is very deceptive and incomprehensible, and the probability to get the estimate of the $\mathrm{MSE}(\tilde{u}_i)$ negative depends not only on the data design and values of variance components, but also on the variance components estimation methods and the sampling variances of the variance components estimates, then it is natural to treat all estimates corresponding to negative $\hat{\sigma}_u^2$ as inadmissible.

In the following a simulation study was carried out by the rules described in Section 3.2.1. The purpose of simulations is to control the accuracy of the probabilities (3.42) and (3.43), and to find out the magnitude of the chance to get a non-negative estimate of the $\mathrm{MSE}(\tilde{u}_i)$ even if $\hat{\sigma}_u^2 < 0$. The results of the simulations are presented in Table 3.10. The probabilities of inadmissible estimates are not presented for intraclass correlation coefficient value 0.8, because the corresponding population value of the heritability coefficient $h^2 = 3.6$ is already inadmissible in itself and the probability of negative estimates is irrelevant (for example, $\mathrm{P}(\hat{\sigma}_u^2 < 0)$ is bigger than 0.0001 only in case of very small number of groups).

Simulations show the compatibility of the observed probabilities (based on 10000 replicated samples) and theoretical predicted probabilities. The chance to get a non-negative estimate of the $\mathrm{MSE}(\tilde{u}_i)$ even if $\hat{\sigma}_u^2 < 0$ is bigger in the case of small number of groups, in the case of big number of groups from $\hat{\sigma}_u^2 < 0$ the negative estimate of $\mathrm{MSE}(\tilde{u}_i)$ follows more often.

Table 3.10. The observed and predicted probabilities of inadmissible estimates in the case of different true population values and number of groups (sires) *a*. Population size $N = 360$ and error variance $\sigma_e^2 = 1$ were used.

ρ (h^2)	a	$P(\hat{\sigma}_u^2 < 0)^*$	$P(\hat{\sigma}_u^2 < 0 \mid \sigma_{u,e}^2)^\#$	$P(\hat{h}^2 > 1)^*$	$P(\hat{h}^2 > 1 \mid \sigma_{u,e}^2)^\#$	$P\left[\widehat{\mathrm{MSE}}(\hat{u}_i) < 0\right]^*$
0.0125	4	0.2949	0.2948	0.0000	0.0000	0.0000
(0.05)	15	0.3022	0.2950	0.0000	0.0000	0.1129
	24	0.3264	0.3208	0.0000	0.0000	0.1062
	90	0.4047	0.3994	0.0000	0.0000	0.2732
0.0625	4	0.0667	0.0658	0.0033	0.0045	0.0000
(0.25)	15	0.0217	0.0212	0.0000	0.0000	0.0049
	24	0.0236	0.0246	0.0000	0.0000	0.0050
	90	0.0918	0.0953	0.0001	0.0000	0.0467
0.15	4	0.0196	0.0189	0.1402	0.1402	0.0000
(0.6)	15	0.0004	0.0005	0.0481	0.0503	0.0000
	24	0.0005	0.0003	0.0328	0.0328	0.0002
	90	0.0012	0.0018	0.0306	0.0298	0.0002

* Observed probabilities were found based on 10000 replicated samples.

\# Predicted probabilities $P(\hat{\sigma}_u^2 < 0 \mid \sigma_{u,e}^2)$ and $P(\hat{h}^2 > 1 \mid \sigma_{u,e}^2)$ were calculated by formulas (3.42) and (3.43), respectively.

3.3. The effect of data structure

3.3.1. The effect of data structure on $\mathrm{Var}(\hat{\sigma}_u^2)$ and $\mathrm{Var}(\hat{\rho})$

The number of objects per group which minimizes $\mathrm{Var}(\hat{\sigma}_u^2)$ has been derived already by Hammersley in 1948 (Hammersley, 1948) considering *n* as a continuous argument and studying the derivatives of equation (3.20). The minimum $\mathrm{Var}(\hat{\sigma}_u^2)$ is obtained by considering

$$n = \frac{N(\tau + 1) + 1}{N\tau + 2} \tag{3.44}$$

observations per group. The same group size guarantees the minimum of $\sigma(\hat{\sigma}_u^2)$.

To illustrate the optimal criterion and to study the effect of using non-optimal designs, the simulation study is performed. Data size $N = 360$ is used similarly to previous paragraphs. As standard errors are accuracy parameters used usually in practice and also the differences between different designs are better noticeable in studying standard errors, here and below in this paragraph the patterns of standard deviations instead of the patterns of variances are examined. Calculations base on the formula (3.20), because the modelling results presented in Table 3.4 show the accuracy of predictors of $\sigma(\hat{\sigma}_u^2)$ got in such way. The number of observations per group minimizing the standard error of σ_u^2, is found by

formula (3.44). To control the optimum criterion, in fixed values of the intraclass correlation the optimal integer numbers of observations per group is found by simulations. The pattern of $\sigma(\hat{\sigma}_u^2)$ and optimum numbers of observations per group are shown in Figure 3.1.

It appears that optimum criterion given by equation (3.44) agrees with simulation results. About the behaviour of $\sigma(\hat{\sigma}_u^2)$ the following conclusions are drawn. The standard errors of σ_u^2 increase drastically if the intraclass correlation coefficient values come close to its upper limit 1. In the case of average or bigger magnitude of random effects measured by the intraclass correlation, the optimal group size $n = 2$. The optimal group size increases fast only in the case of small values of the intraclass correlation. But this is exactly the situation occurring in genetic applications like the sire model, where the admissible values of the intraclass correlations lie in the interval $0 \le \rho \le 0.25$.

In the following the pattern of $\sigma(\hat{\sigma}_u^2)$ in different true heritability values is found and presented in Figure 3.2. Also both the continuous and integer optimum numbers of daughters per sire are calculated by formula (3.44) and by simulations, respectively. The pattern of $\sigma(\hat{\sigma}_u^2)$ shown on Figure 3.2 corresponds to the modelling results presented in Table 3.4. The basic conclusion from the simulation study about $\sigma(\hat{\sigma}_u^2)$ is that the deficiency of sires increases the inaccuracy of the estimates of the variance components when the effect of sires, measured via heritability, is large.

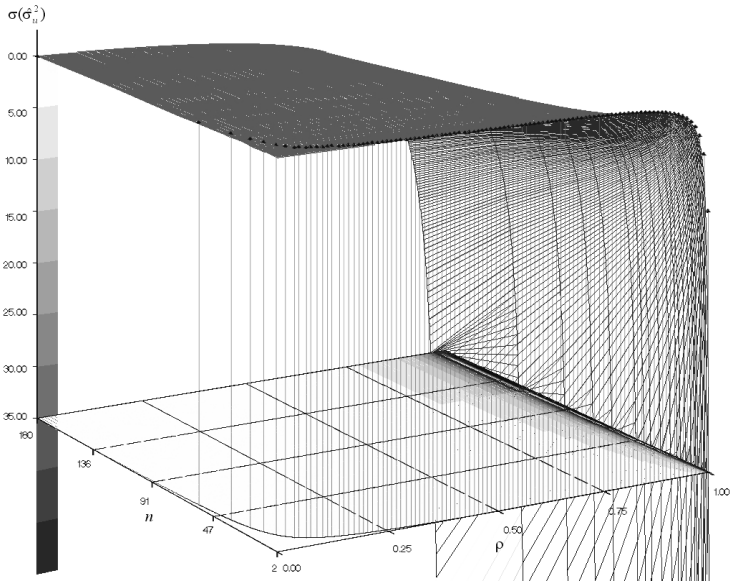


Figure 3.1. The pattern of $\sigma(\hat{\sigma}_u^2)$ and optimal number of observations per group in different true intraclass correlation values ($N = 360$, $\sigma_e^2 = 1$).

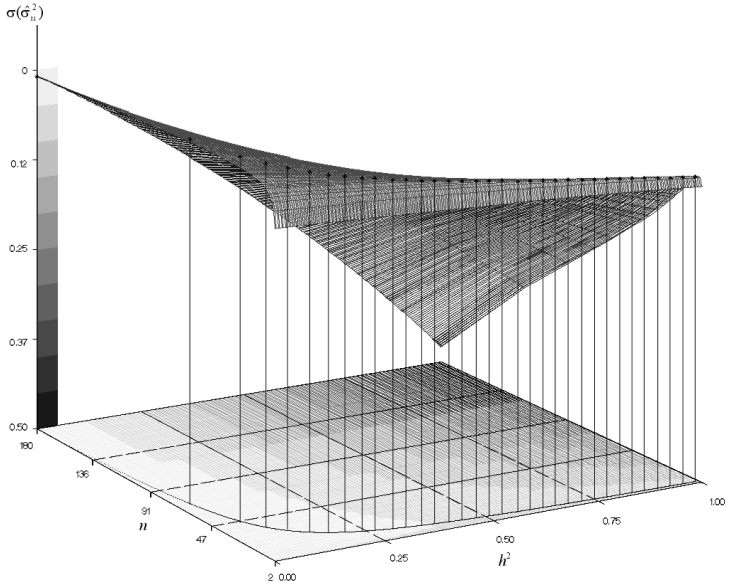


Figure 3.2. The pattern of $\sigma(\hat{\sigma}_u^2)$ and optimal number of daughters per sire (vertical arrows for integer numbers and dotted line on *xy*-plane for continuous numbers) in different true heritability values ($N = 360$, $\sigma_e^2 = 1$).

The modelling results concerning the standard deviations of the estimators of the intraclass correlation coefficients and the heritability coefficients showed that the most accurate estimates are got using the optimum designs for variance components estimation, the differences from the optimum group sizes calculated by (3.44) appear only in decimal places. The patterns of $\sigma(\hat{\rho})$ and $\sigma(\hat{h}^2) = 4\sigma(\hat{\rho})$, based on the sampling variance expression (3.26), are shown in Figures 3.3 and 3.4, respectively,. Also the number of observations per group minimizing studied standard errors is found by simulations based on expression (3.26).

The patterns of $\sigma(\hat{\rho})$ and $\sigma(\hat{h}^2)$ shown in Figures 3.3 and 3.4 correspond to the modelling results presented in Table 3.6. The conclusions based on these tables and figures are the following. Although the patterns of $\sigma(\hat{\sigma}_u^2)$ and $\sigma(\hat{\rho})$ are very different, the optimum designs are the same. The deficiency of groups increases the inaccuracy of estimates of intraclass correlation coefficients when the effect of studied factor is of average level, in genetic studies this means that the deficiency of sires has the worst effect in case of large heritabilities. A small number of groups (sires), even with a big number of observations (daughters), may cause dramatic loss of accuracy. For small and large intraclass correlation coefficient values the estimates are more accurate and depend less on the design. Of course, very small number of observations per group should be avoided. Compendium for genetic studies is that as the heritability of trait increases, the accuracy of estimate decreases, the optimum number of sires increases and the optimal number of daughters per sire decreases.

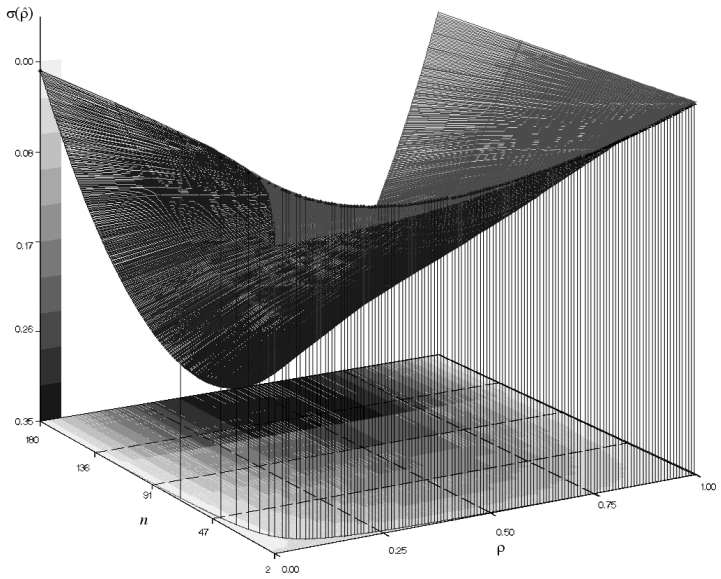


Figure 3.3. Pattern of $\sigma(\hat{\rho})$ and optimal number of daughters per sire (vertical arrows for integer numbers and dotted line on *xy*-plane for continuous numbers) in different true intraclass correlation values ($N = 360$, $\sigma_e^2 = 1$).

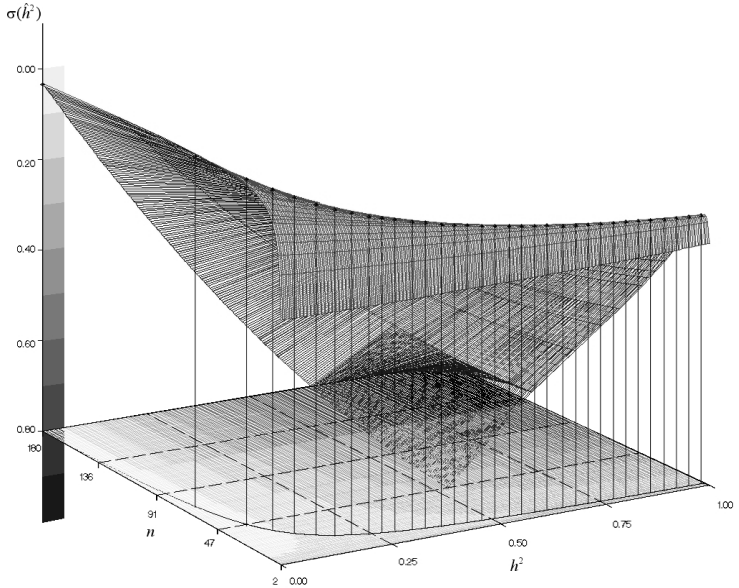


Figure 3.4. Pattern of $\sigma(\hat{h}^2)$ and optimal number of daughters per sire (vertical arrows for integer numbers and dotted line on *xy*-plane for continuous numbers) in different true heritability values ($N = 360$, $\sigma_e^2 = 1$).

3.3.2. The effect of data structure on $\mathrm{MSE}(\hat{u}_i)$ and $\mathrm{MSE}(\tilde{u}_i)$

It is natural to suppose for balanced data that both the number of groups a and number of observations per group n are bigger than one. Then it is obvious that $\mathrm{MSE}(\hat{\mathbf{u}})$ is minimized when the number of groups is minimal, $a = 2$, and the number of observations per group is maximal, $n = N/2$.

The optimal number of observations per group minimizing $\mathrm{MSE}(\hat{u}_i)$ is found considering n as a continuous argument and studying the derivatives of equation (3.33).

Theorem 3.4. *Considering the number of objects per group as a continuous argument the minimum value of* $\mathrm{MSE}(\hat{u}_i)$ *is obtained if the size of groups is expressed as*

$$n = \frac{-1 + \sqrt{1 + N\tau}}{\tau}. \tag{3.45}$$

Proof. Mean square error of predictor $\hat{u}_i$ given by (3.33) can be rewritten as a function of group size n in the form

$$\mathrm{MSE}(\hat{u}_i) = \frac{\sigma_u^2(N\sigma_e^2 + n^2\sigma_u^2)}{N(\sigma_e^2 + n\sigma_u^2)}.$$

To find the value of n which minimizes $\mathrm{MSE}(\hat{u}_i)$ we study the derivative of last expression:

$$\frac{\partial \mathrm{MSE}(\hat{u}_i)}{\partial n} = \frac{(\sigma_u^2)^2(2n\sigma_e^2 + n^2\sigma_u^2 - N\sigma_e^2)}{N(\sigma_e^2 + n\sigma_u^2)}.$$

We assume that both σ_u^2 and σ_e^2 differ from null. Now we equate the derivative to zero and get: $\tau n^2 + 2n - N = 0$, where $\tau = \sigma_u^2/\sigma_e^2$. The only positive solution to this quadratic is expressed as $n = (-1 + \sqrt{1 + N\tau})/\tau$, which establishes the theorem.■

As $\sigma(\hat{u}_i - u_i) = \sqrt{\mathrm{MSE}(\hat{u}_i)}$ and $\mathrm{MSE}(\hat{u}_i) \geq 0$, then the derived criterion of optimality applies also for standard deviations of prediction errors.

To follow the optimal group size depending on the magnitude of random effects and to examine the accuracy of estimates besides optimal designs, the pattern of $\sigma(\hat{u}_i - u_i)$ was found, and both continuous and integer optimum numbers of observations per group were calculated by formula (3.45) and by simulations, respectively (Figure 3.5). The pattern was drawn for the data size $N = 360$ and for error variance equal to one.

Standard errors of two-stage predictors depend in addition to data design and variance components values also on the sampling variance of variance components. For this reason it is complicated to find exact and simple expression for optimal group size. Simulation studies in range of intraclass correlation coefficient values showed that the pattern of $\sigma(\tilde{u}_i - u_i)$ is similar to this presented in Figure 3.5, and also the optimal group sizes are analogous.

Similarly to variance components estimation, the accuracy of prediction of random effects decreases drastically if the intraclass correlation coefficient comes near to its upper limit. The inaccuracy is the biggest in case of small number of groups with big number of observations.

As in genetic applications only the small values of intraclass correlations are meaningful and also the difference between $\sigma(\hat{u}_i - u_i)$ and $\sigma(\tilde{u}_i - u_i)$ is better noticeable in case of smaller magnitude of random effects u_i, then additional patterns of $\sigma(\hat{u}_i - u_i)$ and $\sigma(\tilde{u}_i - u_i)$ are drawn for admissible values of heritability (Figure 3.6).

It follows that there is not a big difference in the optimal designs for the random effects prediction according to the estimated or true population values of variance components. For small heritability values the estimates are more accurate and do not depend so much on the design. When the heritability of a trait increases, then the accuracy of estimates decreases and the optimum number of sires increases. Compared with variance components and heritability estimation (Figures 3.2 and 3.4) the number of daughters per sire, needed for the most accurate prediction of random effects, is bigger. The poorest combination is large heritability and a small number of daughters per sire, but also a very big number of daughters per sire compared with the number of sires is not good.

It appears that the usual assumption of animal breeders to increase the number of daughters per sire for increasing the accuracy of estimates of genetic parameters is not the best way. Surely, increasing the number of daughters per sire will increase the accuracy of estimates, but increasing the number of sires has a bigger effect (shortly is this misconception also noted by Searle et al, 1992, p 68–69).

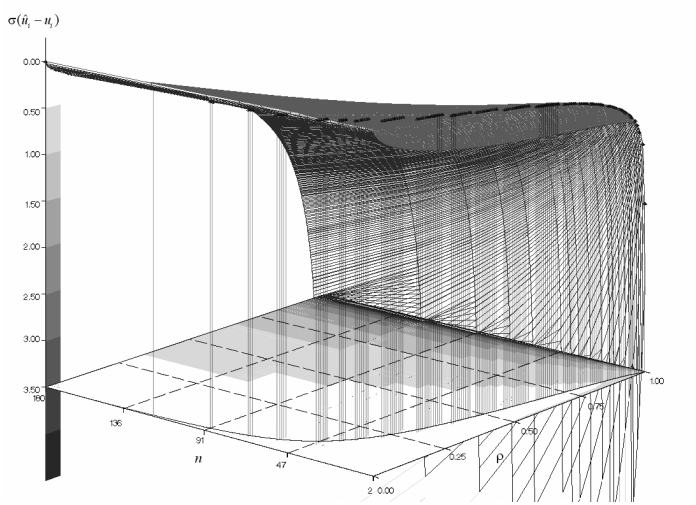


Figure 3.5. The pattern of $\sigma(\hat{u}_i - u_i)$ and the optimal number of observations per group in different true intraclass correlation values ($N = 360$, $\sigma_e^2 = 1$).

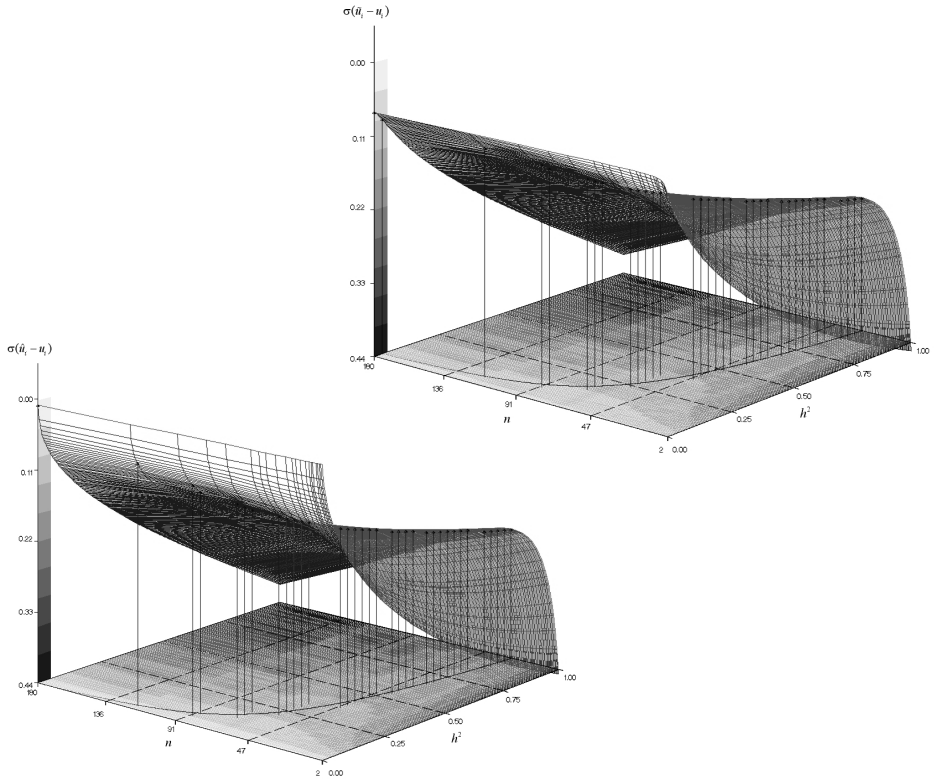


Figure 3.6. The patterns of $\sigma(\hat{u}_i - u_i)$ and $\sigma(\tilde{u}_i - u_i)$ and the optimal number of daughters per sire (vertical arrows for integer numbers and dotted line on *xy*-plane for continuous numbers) in different true heritability values ($N = 360$, $\sigma_e^2 = 1$).

3.3.3. The effect of data structure on the probability of inadmissible estimates

An undesirable feature of ANOVA estimation is the possibility to get negative estimates of variance components. As discussed in Section 3.2.4, from negative $\hat{\sigma}_u^2$ follow also negative estimates of intraclass correlation coefficient and heritability coefficient, and also negative mean square error of random effects. Following the effect of data structure on probability to get negative variance component estimate is studied by simulations. The pattern of $\mathrm{P}(\hat{\sigma}_u^2 < 0)$ calculated by equation (3.42) is shown in Figure 3.7. Also the trajectory of optimal group sizes and possible discrete numbers of observations per group minimizing $\mathrm{P}(\hat{\sigma}_u^2 < 0)$ are found assuming fixed data size $N = 360$.

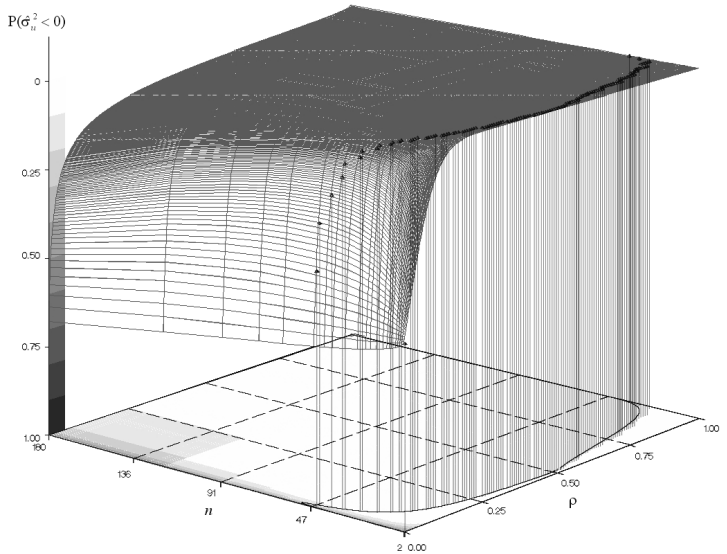


Figure 3.7. The pattern of $P(\hat{\sigma}_u^2 < 0)$ and the optimal number of observations per group (vertical arrows for integer numbers and dotted line on *xy*-plane for continuous numbers) in different true intraclass correlation values ($N = 360$, $\sigma_e^2 = 1$).

The results are expected – probability of negative variance component estimate depends not so much on the data structure than the magnitude of random effects u_i in real population. If the real value of σ_u^2 is close to zero (then also ρ is close to zero), then the probability of getting negative estimate increases. As the probability of inadmissible estimates is bigger when accuracy of estimates is poor, then it is obvious that group sizes minimizing $P(\hat{\sigma}_u^2 < 0)$ are similar to group sizes minimizing $\sigma(\hat{\sigma}_u^2)$. The smallest probability of negative $\hat{\sigma}_u^2$ is achieved with enough big number of groups, whereby the necessary number of groups increases when the population value of σ_u^2 increases. In case of intraclass correlation coefficient values bigger than 0.5 the optimal number of groups increases again, but as in these situations the probability to get negative $\hat{\sigma}_u^2$ is close to zero, then there is no difference between optimal or near to optimal design (for example, when $\rho = 0.5$ then $P(\hat{\sigma}_u^2 < 0) = 8.9 \times 10^{-16}$ in case of 90 groups and 4 observations per group, and $P(\hat{\sigma}_u^2 < 0) = 9.5 \times 10^{-10}$ in case of 20 groups and 18 observations per group).

In genetic applications not only negative but also too big positive estimates are inadmissible. In the following, based on equation (3.41), the pattern of probabilities is found for heritability estimates that are negative or greater than one. Also both continuous and integer numbers of daughters per sire minimizing $P(\hat{h}^2<0) + P(\hat{h}^2>1)$ are presented (Figure 3.8).

As the heritability coefficient values lying in interval $(0,1)$, correspond to intraclass correlation coefficients lying in interval $(0,0.25)$, and $P(\hat{h}^2<0) = P(\hat{\sigma}_u^2<0)$, then it is obvious that the left-side part of $P(\hat{h}^2<0) + P(\hat{h}^2>1)$'s pat-

tern is similar to the part of $P(\hat{\sigma}_u^2 < 0)$'s pattern corresponding to small values of intraclass correlation coefficient values. The increase in probabilities of inadmissible heritability estimates comparing with probabilities of negative variance component estimates concur when the h^2 values get bigger than 0.5 , because at this point the chance to get heritability estimate bigger than one has already considerable effect. The optimal numbers of sires and daughters per sire differs from designs minimizing $\sigma(\hat{\sigma}_u^2)$ and $P(\hat{\sigma}_u^2 < 0)$ only when the real h^2 value is big enough (for example approximately 0.9, when $N = 360$). This is because in case of h^2 values near to one the optimal design in relation to $P(\hat{h}^2>1)$ changes from maximal number of sires to maximum number of daughters per sire. But as such big heritability values are not reality for polygenetic traits, there is usually no need to worry about heritability estimates bigger than one.

Searle et al (1992, p 67–68) studied the negative variance components estimates and illustrated calculations with figures about contour lines of $P(\hat{\sigma}_u^2 < 0)$. Analogous contour lines for $P(\hat{h}^2<0) + P(\hat{h}^2>1) = 0.5, 0.3, 0.1, 0.05, 0.01$ and 0.001 in different true h^2 values plotted on (a, n) co-ordinates, ranging from 2 to 360 for the number of sires *a* and from 2 to 100 for the number of daughters per sire *n*, are presented on Figure 3.9. Similarity of contours on Figure 3.9 and those presented by Searle et al indicate that the probability to get a too big heritability estimate has almost no extra effect in data size studies.

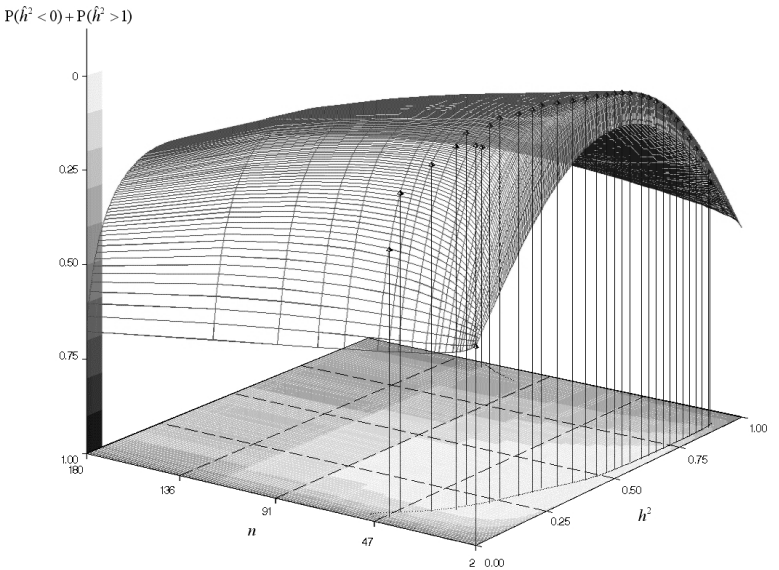


Figure 3.8. The pattern of $P(\hat{h}^2<0) + P(\hat{h}^2>1)$ and the optimal number of daughters per sire (vertical arrows for integer numbers and dotted line on *xy*-plane for continuous numbers) in different true heritability values ($N = 360$, $\sigma_e^2 = 1$).

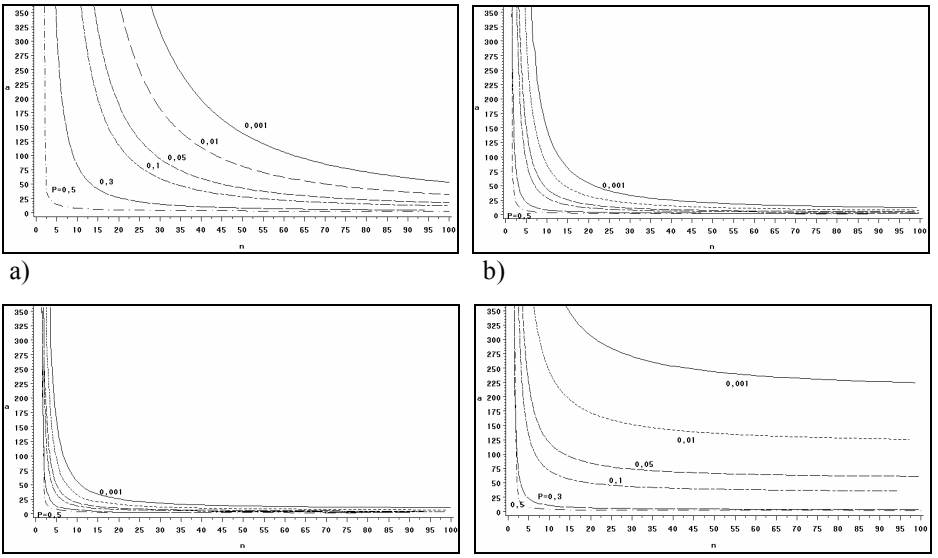


Figure 3.9. The contours of $P(\hat{h}^2<0)+P(\hat{h}^2>1)=0.5$, 0.3, 0.1, 0.05, 0.01 and 0.001 plotted on (a, n) co-ordinates for a) $\tau=0.01$ $(h^2=0.0396)$; b) $\tau=0.05$ $(h^2=0.1905)$; c) $\tau=0.1$ $(h^2=0.3636)$; d) $\tau=0.25$ $(h^2=0.8)$.

3.4. The accuracy of estimates and predictors in unbalanced data

3.4.1. The sampling variances of variance components

Define $S_2=\sum_{i=1}^{a} n_i^2$ and $S_3=\sum_{i=1}^{a} n_i^3$,

$$h_1=\frac{2}{N-a},\ h_2=\frac{-2N(a-1)}{(N-a)(N^2-S_2)},\ h_3=\frac{2N^2(N-1)(a-1)}{(N-a)(N^2-S_2)^2},$$

$$h_4=\frac{4N}{N^2-S_2},\ h_5=\frac{2(N^2S_2+S_2^2-2NS_3)}{(N^2-S_2)^2}.$$

Then the sampling variances of ANOVA estimators of variance components in unbalanced one-way random model under the normality assumptions are expressed as

$$\mathrm{Var}(\hat{\sigma}_u^2)=h_5\sigma_u^4+h_4\sigma_u^2\sigma_e^2+h_3\sigma_e^4, \tag{3.46}$$

and

$$\mathrm{Var}(\hat{\sigma}_e^2)=\frac{2\sigma_e^4}{N-a}.$$

Unbiased estimators to these sampling variances are expressed as

$$\widehat{\mathrm{Var}}(\hat{\sigma}_u^2) = \frac{1}{1+h_5}\left(h_5\hat{\sigma}_u^4 + h_4\hat{\sigma}_u^2\hat{\sigma}_e^2 + \frac{h_3 - h_2h_4}{1+h_1}\hat{\sigma}_e^4\right), \tag{3.47}$$

and

$$\widehat{\mathrm{Var}}(\hat{\sigma}_e^2) = \frac{2\hat{\sigma}_e^4}{N-a+2}.$$

Derivation of these formulas can be found for example in Searle, Casella and McCulloch (1992, p 74–75).

In the following a simulation study was carried out to investigate the accuracy of estimated variance component $\hat{\sigma}_u^2$ standard error expressions, got as square roots of formulas (3.46) and (3.47), with variance components estimates substituted for their unknown population values.

As it appears in balanced designs studies with 360 individuals, there is not a big difference between designs with average number of groups ($a = 15, 24$), then in the following only one average design with 20 groups was modelled. The other studied designs had 4 and 90 groups. Data sets with three different imbalances ($\nu = 0.3$, 0.6, 0.9) were generated assuming the same fixed parameters values as in balanced data studies ($N = 360$, $\sigma_e^2 = 1$, $\rho = 0.0125$, 0.0625, 0.15 and 0.8). Due to the very computer intensive calculations, only 1000 simulations were made with each of combinations of parameters a, ρ and υ values. This modelling size enabled to get an idea of the general tendencies. The compared parameters were (a) the observed standard deviation $\sigma(\hat{\sigma}_u^2)$ of the estimated variance component; (b) the predicted standard deviation $\sigma(\hat{\sigma}_u^2 \mid \sigma_{e,u}^2)$ of the estimated variance component calculated as the square root of expression (3.46); (c) the estimated standard deviation $\hat{\sigma}(\hat{\sigma}_u^2 \mid \hat{\sigma}_{e,u}^2)$ of the estimated variance component calculated as the square root of expression (3.47); (d) the estimated standard deviation $\hat{\sigma}_b(\hat{\sigma}_u^2 \mid \hat{\sigma}_{e,u}^2)$ of the estimated variance component calculated as the square root of expression (3.46) with $\hat{\sigma}_u^2$ substituted for σ_u^2 and $\hat{\sigma}_e^2$ substituted for σ_e^2. The simulations results are presented in Table 3.11.

The simulation study does not call in question the rightness of theoretical expressions. Several discrepancies between observed and predicted standard errors of σ_u^2 are apparently caused by the relatively small number of simulations made. In the case of standard errors $\sigma(\hat{\sigma}_u^2)$ the same conclusions can be made as in the balanced case (Paragraph 3.2.1) – expression (3.47), giving unbalanced estimates to sampling variance of σ_u^2, produce underestimated values to standard error of $\hat{\sigma}_u^2$, and the expression (3.46), giving biased estimates to sampling variance of σ_u^2, is preferred in estimating $\sigma(\hat{\sigma}_u^2)$.

3.4.2. The sampling variance of the intraclass correlation coefficient

Approximate formula for the sampling variance of intraclass correlation coefficient in unbalanced data was published by Swinger et al (1964) and has the following form:

$$\mathrm{Var}(\hat{\rho}) \approx \frac{2(N-1)(1-\rho)^2[1+(d-1)\rho]^2}{d^2(N-a)(a-1)}. \tag{3.48}$$

Derivation of this formula is based on the approximate formula for the variance of the ratio of two random variables:

$$\mathrm{Var}(y/x) \approx \left[\mathrm{E}(y)/\mathrm{E}(x)\right]^2 \left\{\mathrm{Var}(y)/[\mathrm{E}(y)]^2 + \mathrm{Var}(x)/\left[\mathrm{E}(x)\right]^2 - 2\,\mathrm{Cov}\left[y,x/\mathrm{E}(y)\,\mathrm{E}(x)\right]\right\},$$

applied to the estimate of intraclass correlation coefficient expressed through sums of squares.

Next an alternative expression for $\mathrm{Var}(\hat{\rho})$ is derived based on approximations (3.25) and (3.14).

Theorem 3.5. *In the one-way random model under the normality assumptions* (3.3) *the variance of the intraclass correlation coefficient estimate can approximately be expressed as*

$$\mathrm{Var}(\hat{\rho}) \approx \frac{2m\lambda^2(N-a)^2(N-a+m-2)(1-\rho)^4}{d^2(a-1)^2(N-a-2)^2(N-a-4)(\sigma_e^2)^2}, \tag{3.49}$$

where d is a coefficient (3.7), *m and* λ *are defined with formulas* (3.15) *and* (3.16), *respectively*.

Proof. Let $\hat{w} = MS(u)/MS(e)$, then based on (3.9)

$$\hat{\rho} = \frac{\hat{w}-1}{\hat{w}+d-1} = f(\hat{w}). \tag{3.50}$$

Table 3.11. The observed, predicted and estimated standard errors of variance component σ_u^2 in the case of different true population values, data set imbalance ν and number of groups a ($N = 360$, $\sigma_e^2 = 1$) found based on 1000 replicated samples.

σ_u^2 (h^2)	ν	a	$\mathrm{E}(\hat{\sigma}_u^2)$	$\sigma(\hat{\sigma}_u^2)^*$	$\sigma(\hat{\sigma}_u^2 \mid \sigma_{e,u}^2)^{\#}$	$\hat{\sigma}(\hat{\sigma}_u^2 \mid \hat{\sigma}_{e,u}^2)^{\square}$	$\hat{\sigma}_b(\hat{\sigma}_u^2 \mid \hat{\sigma}_{e,u}^2)^{\square}$
0.0127	0.3	4	0.0169	0.0512	0.0478	0.0403	0.0578
(0.05)		20	0.0127	0.0257	0.0256	0.0241	0.0271
		90	0.0133	0.0459	0.0461	0.0457	0.0471
	0.6	4	0.0122	0.0217	0.0224	0.0167	0.0231
		20	0.0132	0.0229	0.0234	0.0224	0.0240
		90	0.0142	0.0480	0.0452	0.0452	0.0459
	0.9	4	0.0119	0.0191	0.0199	0.0149	0.0195
		20	0.0121	0.0219	0.0227	0.0215	0.0227
		90	0.0143	0.0457	0.0449	0.0448	0.0453
0.0667	0.3	4	0.0702	0.1026	0.1009	0.0757	0.1090
(0.25)		20	0.0660	0.0484	0.0501	0.0449	0.0505
		90	0.0677	0.0542	0.0549	0.0541	0.0557
	0.6	4	0.0660	0.0744	0.0734	0.0529	0.0732
		20	0.0673	0.0433	0.0433	0.0408	0.0439
		90	0.0648	0.0537	0.0530	0.0523	0.0531
	0.9	4	0.0694	0.0657	0.0654	0.0518	0.0677
		20	0.0692	0.0407	0.0406	0.0392	0.0414
		90	0.0652	0.0522	0.0523	0.0517	0.0523
0.1765	0.3	4	0.1662	0.2005	0.2138	0.1426	0.2054
(0.6)		20	0.1762	0.1025	0.1052	0.0937	0.1055
		90	0.1720	0.0756	0.0772	0.0748	0.0770
	0.6	4	0.1810	0.1814	0.1790	0.1325	0.1841
		20	0.1779	0.0843	0.0862	0.0807	0.0868
		90	0.1727	0.0697	0.0707	0.0692	0.0703
	0.9	4	0.1741	0.1513	0.1580	0.1192	0.1560
		20	0.1704	0.0739	0.0776	0.0715	0.0756
		90	0.1761	0.0672	0.0680	0.0672	0.0680
4.0	0.3	4	4.0768	3.5356	3.3844	2.6354	3.4478
(3.2)		20	4.0732	2.2399	2.0698	1.8687	2.1019
		90	4.0458	1.0732	1.0286	1.0088	1.1551
	0.6	4	4.0644	3.8438	3.8424	2.8152	3.9027
		20	3.8841	1.4780	1.6049	1.4489	1.5592
		90	3.9678	0.7425	0.7828	0.7634	0.7768
	0.9	4	3.8422	4.0376	4.1816	2.7758	3.9957
		20	3.9571	1.4097	1.3781	1.2908	1.3635
		90	3.9877	0.6852	0.6660	0.6559	0.6640

* Observed standard errors $\sigma(\hat{\sigma}_u^2)$ were found based on 1000 replicated samples.

Predicted standard errors $\sigma(\hat{\sigma}_u^2 \mid \sigma_{e,u}^2)$ were calculated as square roots of formula (3.46).

¤ Estimated standard deviations $\hat{\sigma}(\hat{\sigma}_u^2 \mid \hat{\sigma}_{e,u}^2)$ and $\hat{\sigma}_b(\hat{\sigma}_u^2 \mid \hat{\sigma}_{e,u}^2)$ were calculated as square roots of formulas (3.47) and (3.46), respectively, with $\hat{\sigma}_u^2$ substituted for σ_u^2 and $\hat{\sigma}_e^2$ substituted for σ_e^2 .

Following (3.11), (3.13) and (3.14) we get that $\hat{w}$ is approximately distributed as follows

$$\hat{w} \sim \frac{m\lambda}{(a-1)\sigma_e^2} F_{m,N-a},$$

where $F_{m,N-a}$ is the F-distribution having m and $N-a$ degrees of freedom. Here, depending on the context, F denotes both the distribution and random variable with the F-distribution. Because of

$$\mathrm{Var}(\mathrm{F}_{m,N-a}) = 2(N-a)^2(N-a+m-2)\Big/\Big[m(N-a-2)^2(N-a-4)\Big]$$

we have

$$\mathrm{Var}(\hat{w}) \approx \frac{(m\lambda)^2}{(a-1)^2\sigma_e^4} \times \frac{2(N-a)^2(N-a+m-2)}{m(N-a-2)^2(N-a-4)}.$$

From (3.50) it follows

$$\frac{\partial f(\hat{w})}{\partial \hat{w}} = \frac{d}{(\hat{w}+d-1)^2},$$

or, because $\hat{w} = [1+(d-1)\hat{\rho}]/(1-\hat{\rho})$,

$$\frac{\partial f(\hat{w})}{\partial \hat{w}} = \frac{(1-\hat{\rho})^2}{d},$$

and, based on approximation (3.25),

$$\begin{aligned}\mathrm{Var}(\hat{\rho}) &\approx \frac{(1-\rho)^4}{d^2} \times \frac{(m\lambda)^2}{(a-1)^2\sigma_e^4} \times \frac{2(N-a)^2(N-a+m-2)}{m(N-a-2)^2(N-a-4)} \\ &= \frac{2m\lambda^2(N-a)^2(N-a+m-2)(1-\rho)^4}{d^2(a-1)^2(N-a-2)^2(N-a-4)\sigma_e^4},\end{aligned}$$

which completes the proof of Theorem 3.5. ■

Similarly to the previous paragraph, a simulation study was carried out to investigate the accuracy of estimated intraclass correlation coefficient standard deviation estimators. The compared parameters were (a) the observed standard deviation $\sigma(\hat{\rho})$ of estimated intraclass correlation coefficient; (b) the predicted standard deviation $\sigma_S(\hat{\rho}\,|\,\rho)$ of the intraclass correlation coefficient estimate calculated as square root of approximation (3.48); (c) the predicted standard deviation $\sigma_K(\hat{\rho}\,|\,\rho)$ of the intraclass correlation coefficient estimate calculated as square root of approximation (3.49); (d) the estimated standard deviation $\hat{\sigma}_S(\hat{\rho}\,|\,\hat{\rho})$ of the intraclass correlation coefficient estimate calculated as square root of approximation (3.48) with $\hat{\rho}$ substituted for ρ; (e) the estimated standard deviation $\hat{\sigma}_K(\hat{\rho}\,|\,\hat{\rho})$ of the intraclass correlation coefficient estimate calculated as square root of approximation (3.49) with $\hat{\rho}$ substituted for ρ. The simulations results are presented in Table 3.12.

Table 3.12. The observed, predicted and estimated standard errors of the intraclass correlation coefficient ρ in the case of different true population values, data set imbalance ν and number of groups a ($N = 360$, $\sigma_e^2 = 1$) found based on 1000 replicated samples.

ρ (h^2)	ν	a	$E(\hat{\rho})$	$\sigma(\hat{\rho})^*$	$\sigma_S(\hat{\rho}\|\rho)^\#$	$\hat{\sigma}_S(\hat{\rho}\|\hat{\rho})^¤$	$\sigma_K(\hat{\rho}\|\rho)^\#$	$\hat{\sigma}_K(\hat{\rho}\|\hat{\rho})^¤$
0.0125	0.3	4	0.0147	0.0468	0.0460	0.0459	0.0471	0.0531
(0.05)		20	0.0122	0.0249	0.0257	0.0261	0.0254	0.0264
		90	0.0133	0.0459	0.0517	0.0459	0.0460	0.0466
	0.6	4	0.0118	0.0209	0.0210	0.0201	0.0221	0.0221
		20	0.0129	0.0225	0.0286	0.0240	0.0234	0.0239
		90	0.0140	0.0471	0.0446	0.0444	0.0451	0.0453
	0.9	4	0.0116	0.0183	0.0193	0.0183	0.0196	0.0188
		20	0.0118	0.0214	0.0371	0.0232	0.0225	0.0222
		90	0.0139	0.0449	0.0485	0.0529	0.0449	0.0448
0.0625	0.3	4	0.0587	0.0794	0.0806	0.0730	0.0897	0.0849
(0.25)		20	0.0608	0.0418	0.0381	0.0371	0.0451	0.0442
		90	0.0631	0.0496	0.0636	0.0597	0.0511	0.0514
	0.6	4	0.0585	0.0592	0.0580	0.0523	0.0653	0.0591
		20	0.0620	0.0378	0.0616	0.0627	0.0397	0.0389
		90	0.0602	0.0491	0.0579	0.0600	0.0494	0.0491
	0.9	4	0.0620	0.0543	0.0563	0.0536	0.0582	0.0554
		20	0.0641	0.0361	0.0780	0.0739	0.0373	0.0371
		90	0.0607	0.0479	0.0906	0.0855	0.0488	0.0484
0.15	0.3	4	0.1231	0.1210	0.1312	0.1054	0.1563	0.1250
(0.6)		20	0.1453	0.0706	0.0628	0.0598	0.0777	0.0739
		90	0.1448	0.0585	0.0740	0.0715	0.0606	0.0597
	0.6	4	0.1377	0.1097	0.1126	0.0959	0.1309	0.1113
		20	0.1477	0.0613	0.0777	0.0750	0.0639	0.0620
		90	0.1462	0.0549	0.0825	0.0805	0.0561	0.0555
	0.9	4	0.1364	0.0979	0.1114	0.0959	0.1156	0.0994
		20	0.1472	0.0594	0.0882	0.0841	0.0639	0.0619
		90	0.1490	0.0528	0.0816	0.0821	0.0543	0.0539
0.8	0.3	4	0.6614	0.2230	0.1407	0.1555	0.1691	0.1879
(3.2)		20	0.7728	0.0821	0.0765	0.1010	0.0844	0.0894
		90	0.7950	0.0406	0.0598	0.0589	0.0442	0.0446
	0.6	4	0.6903	0.2034	0.1395	0.1496	0.1554	0.1685
		20	0.7770	0.0640	0.1722	0.1703	0.0657	0.0697
		90	0.7945	0.0332	0.0572	0.0612	0.0350	0.0355
	0.9	4	0.7042	0.1922	0.1433	0.1534	0.1370	0.1476
		20	0.7823	0.0640	0.4972	0.5041	0.0944	0.0991
		90	0.7965	0.0309	0.0294	0.0297	0.0308	0.0311

* Observed standard deviations $\sigma(\hat{\rho})$ were found based on 1000 replicated samples.

Predicted standard deviations $\sigma_S(\hat{\rho}|\rho)$ and $\sigma_K(\hat{\rho}|\rho)$ were calculated as square roots of formulas (3.48) and (3.49), respectively.

¤ Estimated standard deviations $\hat{\sigma}_S(\hat{\rho}|\hat{\rho})$ and $\hat{\sigma}_K(\hat{\rho}|\hat{\rho})$ were calculated as square roots of formulas (3.48) and (3.49), respectively, with $\hat{\rho}$ substituted for ρ .

Results in Table 3.12 show that in case of small values of ρ, both approximations (3.48) and (3.49) give quite similar results that seem to be unbiased. For large intraclass correlation coefficient values the formula (3.9) underestimates the real values of ρ. The estimates of $\sigma(\hat{\rho})$ got with expressions (3.48) and (3.49) underestimate the real $\sigma(\hat{\rho})$ when the number of groups is small and the intraclass correlation coefficient is large. The difference from simulated values is smaller by using expression (3.49).

3.4.3. The mean square errors of predictors

The following theorem gives an exact expression for variance-covariance matrix of prediction errors, allowing to study the effect of data imbalance.

Theorem 3.6. *In one-way random model under the normality assumptions the variance-covariance matrix of prediction errors is expressed as*

$$\begin{aligned}\mathrm{Var}(\hat{\mathbf{u}}-\mathbf{u}) &= \left\{ \underset{\mathrm{d}}{\frac{\sigma_u^2\sigma_e^2}{\sigma_e^2+n_i\sigma_u^2}} \right\}_{i=1}^{a} \\ &+ \frac{\sigma_e^2}{N-\sum_{i=1}^{a}\left[n_i^2\sigma_u^2/(\sigma_e^2+n_i\sigma_u^2) \right]} \left\{ \underset{\mathrm{m}}{\frac{n_i n_j(\sigma_u^2)^2}{(\sigma_e^2+n_i\sigma_u^2)(\sigma_e^2+n_j\sigma_u^2)}} \right\}_{i,j=1}^{a}.\end{aligned} \tag{3.51}$$

Proof. We derive the expression in theorem statement similarly to proof of Theorem 3.1. By formulas (1.12) and (1.13) the variance-covariance matrix of prediction errors is expressed as

$$\mathrm{Var}(\hat{\mathbf{u}}-\mathbf{u}) = \mathbf{G}-\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G}+\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G}. \tag{3.52}$$

Based on expressions of design matrix $\mathbf{Z}$ and variance-covariance matrix $\mathbf{G}$ corresponding to model (3.1) and the inverse of variance-covariance matrix $\mathbf{V}$ defined in Corollary 3.1 we get after some matrix algebra that

$$\mathbf{G}-\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G} = \sigma_u^2\mathbf{I}_a - \left\{ \underset{\mathrm{d}}{\frac{n_i(\sigma_u^2)^2}{\sigma_e^2+n_i\sigma_u^2}} \right\}_{i=1}^{a} = \left\{ \underset{\mathrm{d}}{\frac{\sigma_u^2\sigma_e^2}{\sigma_e^2+n_i\sigma_u^2}} \right\}_{i=1}^{a}. \tag{3.53}$$

The scalar $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$ has a form

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} = \mathbf{1}_N'\left\{ \underset{\mathrm{d}}{\frac{1}{\sigma_e^2}\left(\mathbf{I}_{n_i} - \frac{\sigma_u^2}{\sigma_e^2+n_i\sigma_u^2}\mathbf{J}_{n_i} \right)} \right\}_{i=1}^{a} \mathbf{1}_N = \frac{1}{\sigma_e^2}\left(N - \sum_{i=1}^{a}\frac{n_i^2\sigma_u^2}{\sigma_e^2+n_i\sigma_u^2} \right)$$

from which we have that

$$(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-} = \frac{\sigma_e^2}{N-\sum_{i=1}^{a}\left[n_i^2\sigma_u^2/(\sigma_e^2+n_i\sigma_u^2) \right]}. \tag{3.54}$$

The other terms in the third component of expression (3.52) can be presented as follows:

$$(\mathbf{GZ}'\mathbf{V}^{-1}\mathbf{X})(\mathbf{X}'\mathbf{V}^{-1}\mathbf{ZG}) = \left\{ \underset{\mathrm{c}}{\frac{n_i\sigma_u^2}{\sigma_e^2 + n_i\sigma_u^2}} \right\}_{i=1}^{a} \left\{ \underset{\mathrm{r}}{\frac{n_i\sigma_u^2}{\sigma_e^2 + n_i\sigma_u^2}} \right\}_{i=1}^{a} = \left\{ \underset{\mathrm{m}}{\frac{n_i n_j(\sigma_u^2)^2}{(\sigma_e^2 + n_i\sigma_u^2)(\sigma_e^2 + n_j\sigma_u^2)}} \right\}_{i,j=1}^{a}. \tag{3.55}$$

If we substitute the expressions (3.53), (3.54) and (3.55) into (3.52), we get the formula in the theorem statement. ■

Corollary 3.4. *In the one-way random model under the normality assumptions the mean square error of predictors is expressed as*

$$\mathrm{MSE}(\hat{\mathbf{u}}) = \sum_{i=1}^{a} \frac{\sigma_u^2\sigma_e^2}{\sigma_e^2 + n_i\sigma_u^2} + \frac{\sigma_e^2}{N - \sum_{i=1}^{a}\left[n_i^2\sigma_u^2/(\sigma_e^2 + n_i\sigma_u^2) \right]} \sum_{i=1}^{a} \frac{n_i^2(\sigma_u^2)^2}{(\sigma_e^2 + n_i\sigma_u^2)^2} \tag{3.56}$$

and naive estimated as

$$\widehat{\mathrm{MSE}}(\hat{\mathbf{u}}) = \sum_{i=1}^{a} \frac{\hat{\sigma}_u^2\hat{\sigma}_e^2}{\hat{\sigma}_e^2 + n_i\hat{\sigma}_u^2} + \frac{\hat{\sigma}_e^2}{N - \sum_{i=1}^{a}\left[n_i^2\hat{\sigma}_u^2/(\hat{\sigma}_e^2 + n_i\hat{\sigma}_u^2) \right]} \sum_{i=1}^{a} \frac{n_i^2(\hat{\sigma}_u^2)^2}{(\hat{\sigma}_e^2 + n_i\hat{\sigma}_u^2)^2}. \tag{3.57}$$

Proof. By expression (1.14) the mean square error of predictors equals to the sum of diagonal elements of matrix (3.51). Applying the trace properties (i), (ii) and (vii) listed in Proposition 3.1 we reach the desired formula (3.56). ■

In the next theorem the approximated formula of mean square error of two-stage predictors taking into account the sampling variance of estimators of variance components is presented.

Theorem 3.7. *In the one-way random model under the normality assumptions the mean square error of two-stage predictors is expressed by the following formula*:

$$\begin{aligned} \mathrm{MSE}(\tilde{\mathbf{u}}) \approx\ & \mathrm{MSE}(\hat{\mathbf{u}}) \\ & + \mathrm{Var}(\hat{\sigma}_e^2) \times (\sigma_u^2)^2 \times (k_{13} - 3ck_{24} + 2c^2k_{12}k_{23} + c^2k_{13}k_{22} - c^3k_{12}^2k_{22}) \\ & + \mathrm{Var}(\hat{\sigma}_u^2) \times \big[k_{11} - ck_{22} - 2\sigma_u^2(k_{22} - 2ck_{33} + c^2k_{22}^2) \\ & \qquad + (\sigma_u^2)^2(k_{33} - 3ck_{44} + 3c^2k_{22}k_{33} - c^3k_{22}^3) \big], \end{aligned} \tag{3.58}$$

where $k_{fh} = \sum_{i=1}^{a} n_i^f/(\sigma_e^2 + n_i\sigma_u^2)^h$, $c = \dfrac{\sigma_e^2}{N - \sum_{i=1}^{a}(n_i^2 g_i)}$ *and* $g_i = \dfrac{\sigma_u^2}{\sigma_e^2 + n_i\sigma_u^2}$.

Proof. First we note, that $\mathrm{MSE}(\tilde{\mathbf{u}}) = \mathrm{MSE}(\hat{\mathbf{u}}) + \mathrm{tr}[\mathrm{Var}(\tilde{\mathbf{u}} - \hat{\mathbf{u}})]$. To derive $\mathrm{tr}[\mathrm{Var}(\tilde{\mathbf{u}} - \hat{\mathbf{u}})]$, the general expression of $\mathrm{Var}(\tilde{\mathbf{u}} - \hat{\mathbf{u}})$ of the form (3.37) should be used, because in unbalanced case the projection matrix $\mathbf{P}_{\mathbf{X},\mathbf{V}^{-1}} =$

$\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}$ is not symmetric and the simplification $\mathbf{PPP}=\mathbf{V}^{-1}\mathbf{V}^{-1}\mathbf{P}$ used in balanced case (Theorem 3.2) is not valid.

According to the trace properties (i) and (ii) listed in Proposition 3.1 is the $\mathrm{tr}[\mathrm{Var}(\tilde{\mathbf{u}}-\hat{\mathbf{u}})]$ expressed as

$$\begin{aligned}\mathrm{tr}[\mathrm{Var}(\tilde{\mathbf{u}}-\hat{\mathbf{u}})] &\approx \mathrm{Var}(\hat{\sigma}_e^2)\times(\sigma_u^2)^2\times\mathrm{tr}(\mathbf{Z}'\mathbf{PPPZ})\\ &+\mathrm{Var}(\hat{\sigma}_u^2)\times[\mathrm{tr}(\mathbf{Z}'\mathbf{PZ})-2\sigma_u^2\,\mathrm{tr}(\mathbf{Z}'\mathbf{PZZ}'\mathbf{PZ})+(\sigma_u^2)^2\,\mathrm{tr}(\mathbf{Z}'\mathbf{PZZ}'\mathbf{PZZ}'\mathbf{PZ})].\end{aligned} \tag{3.59}$$

Using the definition of matrix $\mathbf{P}$ (1.11) and the facts that $\mathbf{ZZ}'\mathbf{V}^{-1}$, $\mathbf{V}^{-1}\mathbf{ZZ}'$ and $\mathbf{V}^{-1}\mathbf{XX}'\mathbf{V}^{-1}$ are symmetric and that general inverse $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}$ is a scalar, all matrices under traces operators in the last expression can be modified and divided into parts. For example, the $\mathrm{tr}(\mathbf{Z}'\mathbf{PPPZ})$ can be expressed as

$$\begin{aligned}\mathrm{tr}(\mathbf{Z}'\mathbf{PPPZ}) &= \mathrm{tr}(\mathbf{ZZ}'\mathbf{V}^{-3})-3\times(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\times\mathrm{tr}(\mathbf{ZZ}'\mathbf{XX}'\mathbf{V}^{-4})\\ &+2\times[(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}]^2\times\mathrm{tr}(\mathbf{ZZ}'\mathbf{XX}'\mathbf{V}^{-2}\mathbf{XX}'\mathbf{V}^{-3})\\ &+(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\times\mathrm{tr}(\mathbf{ZZ}'\mathbf{XX}'\mathbf{V}^{-3}\mathbf{XX}'\mathbf{V}^{-2})\\ &-[(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}]^3\times\mathrm{tr}(\mathbf{ZZ}'\mathbf{XX}'\mathbf{V}^{-2}\mathbf{XX}'\mathbf{V}^{-2}\mathbf{XX}'\mathbf{V}^{-2}).\end{aligned} \tag{3.60}$$

From Proposition 3.4 and Corollary 3.1 follows that

$$\mathbf{V}^{-h}=\left\{\mathrm{d}\left[\frac{1}{\sigma_e^2}\left(\mathbf{I}_{n_i}-\frac{\sigma_u^2}{\sigma_e^2+n_i\sigma_u^2}\mathbf{J}_{n_i}\right)\right]^h\right\}_{i=1}^{a}=\left\{\mathrm{d}\,\frac{1}{(\sigma_e^2)^h}\left[\mathbf{I}_{n_i}-\frac{(\sigma_e^2+n_i\sigma_u^2)^h-(\sigma_e^2)^h}{n_i(\sigma_e^2+n_i\sigma_u^2)^h}\mathbf{J}_{n_i}\right]\right\}_{i=1}^{a}.$$

This equality together with matrix products $\mathbf{XX}'=\mathbf{J}_N$, $\mathbf{ZZ}'=\{\mathrm{d}\,\mathbf{J}_{n_i}\}_{i=1}^{a}$ and $\mathbf{ZZ}'\mathbf{XX}'=\{\mathrm{d}\,n_i\mathbf{J}_{n_i}\}_{i=1}^{a}$ enable to express all traces in (3.60) as functions of sums in the form

$$k_{fh}=\sum_{i=1}^{a}n_i^f\big/(\sigma_e^2+n_i\sigma_u^2)^h\,,$$

where f and h denote different positive integer exponents.

For example, in calculating $\mathrm{tr}(\mathbf{ZZ}'\mathbf{V}^{-3})$ we first note, that $\mathbf{ZZ}'\mathbf{V}^{-3}$ is a block diagonal matrix with $n_i\times n_i$ blocks equal to

$$\mathbf{J}_{n_i}\times\left[\frac{1}{\sigma_e^2}\left(\mathbf{I}_{n_i}-\frac{\sigma_u^2}{\sigma_e^2+n_i\sigma_u^2}\mathbf{J}_{n_i}\right)\right]^3,$$

which can be more compactly expressed as $\mathbf{J}_{n_i}\times[(\mathbf{I}_{n_i}-g_i\mathbf{J}_{n_i})/\sigma_e^2]^3$, using notation $g_i=\sigma_u^2\big/(\sigma_e^2+n_i\sigma_u^2)$. By the statement (vii) in Proposition 3.1 we can take trace separately from all diagonal blocks and then sum the results. As each diagonal block is expressed as product where the first component is $\mathbf{J}_{n_i}$, then according to the statement (vi) in Proposition 3.1, the trace of such product matrix equals to the sum of elements of the second component $[(\mathbf{I}_{n_i}-g_i\mathbf{J}_{n_i})/\sigma_e^2]^3$. Following the statement (iii) in Proposition 3.4 we get, that matrix $(\mathbf{I}_{n_i}-g_i\mathbf{J}_{n_i})^3$ has n_i diagonal elements equal to $1+[(1-n_ig_i)^3-1]\big/n_i$ and $n_i(n_i-1)$ offdiagonal elements equal to $[(1-n_ig_i)^3-1]\big/n_i$. So, the sum of elements of matrix $[(\mathbf{I}_{n_i}-g_i\mathbf{J}_{n_i})/\sigma_e^2]^3$ is expressed as

$$\frac{n_i}{(\sigma_e^2)^3}\left[1+\frac{(1-n_ig_i)^3-1}{n_i}\right]+\frac{n_i(n_i-1)}{(\sigma_e^2)^3}\left[\frac{(1-n_ig_i)^3-1}{n_i}\right]=\frac{n_i(1-n_ig_i)^3}{(\sigma_e^2)^3}=\frac{n_i}{(\sigma_e^2+n_i\sigma_u^2)^3}.$$

And so,

$$\mathrm{tr}(\mathbf{ZZ'V}^{-3})=\sum_{i=1}^{a}\frac{n_i}{(\sigma_e^2+n_i\sigma_u^2)^3}=k_{13}\,.$$

The other terms in (3.60) can be similarly expressed and finally result in the following expression:

$$\mathrm{tr}(\mathbf{Z'PPPZ})=k_{13}-3ck_{24}+2c^2k_{12}k_{23}+c^2k_{13}k_{22}-c^3k_{12}^2k_{22}\,.$$

Here the scalar c denotes the general inverse expressed in (3.54).

Applying the statements of Propositions 3.1 and 3.3 to the other traces in (3.59), yields after tedious algebra to the desired expression of mean square error of two-stage predictors of the form (3.58). ■

Corollary 3.5. *In the one-way random model under the normality assumptions the mean square error of two-stage predictors is unbiasedly estimated approximately as*

$$\begin{aligned}\widehat{\mathrm{MSE}}(\tilde{\mathbf{u}}) \approx\ & \widehat{\mathrm{MSE}}(\hat{\mathbf{u}})\\ &+2\,\mathrm{Var}(\hat\sigma_e^2)\times(\hat\sigma_u^2)^2\times(\hat k_{13}-3\hat c\hat k_{24}+2\hat c^2\hat k_{12}\hat k_{23}+\hat c^2\hat k_{13}\hat k_{22}-\hat c^3\hat k_{12}^2\hat k_{22})\\ &+2\,\mathrm{Var}(\hat\sigma_u^2)\times\big[\hat k_{11}-\hat c\hat k_{22}-2\hat\sigma_u^2(\hat k_{22}-2\hat c\hat k_{33}+\hat c^2\hat k_{22}^2)\\ &\qquad+(\hat\sigma_u^2)^2(\hat k_{33}-3\hat c\hat k_{44}+3\hat c^2\hat k_{22}\hat k_{33}-\hat c^3\hat k_{22}^3)\big],\end{aligned}\tag{3.61}$$

where $\hat k_{fh}=\sum_{i=1}^{a}n_i^f/(\hat\sigma_e^2+n_i\hat\sigma_u^2)^h$, $\hat c=\dfrac{\hat\sigma_e^2}{N-\sum_{i=1}^{a}(n_i^2\hat g_i)}$ *and* $\hat g_i=\dfrac{\hat\sigma_u^2}{\hat\sigma_e^2+n_i\hat\sigma_u^2}$. ■

Results of simulation studies controlling derived formulas are presented in Table 3.13. As previously, the standard deviations of prediction errors were analysed instead of mean square errors, and based on studies made with balanced data (Paragraph 3.2.3) only concentrated estimators of standard errors, noted as $\sigma_0(\hat{\mathbf{u}}-\mathbf{u})$ and $\sigma_0(\tilde{\mathbf{u}}-\mathbf{u})$, were used.

The predicted standard errors of prediction errors $\sigma(\hat{\mathbf{u}}-\mathbf{u}\,|\,\sigma_{e,u}^2)$, are overestimated, and the bias is the biggest in case of large intraclass correlation coefficient values. The predicted standard errors of two-stage prediction errors $\sigma(\tilde{\mathbf{u}}-\mathbf{u}\,|\,\sigma_{e,u}^2)$, are underestimated in case of small values of ρ, and overestimated in case of large values of ρ. These results are similar to those, got in balanced case, and in connection with this the formulas derived for unbalanced data sets are suitable for further analyses.

Table 3.13. The observed, predicted and estimated standard deviations of the prediction errors and two-stage prediction errors in the case of different true population values, data set imbalance ν and number of groups a ($N = 360$, $\sigma_e^2 = 1$) found based on 1000 replicated samples.

ρ (h^2)	ν	a	$\sigma(\hat{\mathbf{u}}-\mathbf{u})^*$	$\sigma(\hat{\mathbf{u}}-\mathbf{u}\|\sigma_{e,u}^2)^\#$	$\hat{\sigma}_0(\hat{\mathbf{u}}-\mathbf{u}\|\hat{\sigma}_{e,u}^2)^\square$	$\sigma(\tilde{\mathbf{u}}-\mathbf{u})^*$	$\sigma(\tilde{\mathbf{u}}-\mathbf{u}\|\sigma_{e,u}^2)^\#$	$\hat{\sigma}_0(\tilde{\mathbf{u}}-\mathbf{u}\|\hat{\sigma}_{e,u}^2)^\square$
0.0125	0.3	4	0.1928	0.2076	0.1978	0.4147	0.3171	2.2944
(0.05)		20	0.4631	0.4665	0.1610	0.9449	0.5555	89.1258
		90	1.0380	1.0473	1.0076	1.7068	1.2866	2.5472
	0.6	4	0.1714	0.1845	0.1519	0.4978	0.2191	1.8843
		20	0.4495	0.4562	0.3591	0.7278	0.5520	1.3221
		90	1.0430	1.0423	1.0174	2.4044	1.2964	7.6049
	0.9	4	0.1642	0.1762	0.1475	0.3184	0.2055	0.7478
		20	0.4434	0.4452	0.2637	0.5817	0.5587	0.6098
		90	1.0370	1.0444	1.0007	1.4299	1.3049	3.4221
0.0625	0.3	4	0.3723	0.4060	0.3367	0.6045	0.4610	0.6007
(0.25)		20	0.8720	0.8860	0.7951	0.9203	0.9372	0.9961
		90	2.1994	2.2233	1.9754	2.3647	2.3182	2.8693
	0.6	4	0.3161	0.3362	0.2873	0.3626	0.3514	0.4180
		20	0.7995	0.8338	0.7711	0.8546	0.8765	0.9243
		90	2.1835	2.2005	1.9134	2.3235	2.2993	2.5090
	0.9	4	0.2786	0.3125	0.2849	0.3193	0.3190	0.6979
		20	0.7583	0.8059	0.7578	0.7995	0.8441	0.8408
		90	2.1652	2.2110	1.9408	2.2898	2.2886	2.1781
0.15	0.3	4	0.5430	0.5785	0.4651	0.6196	0.6187	1.0964
(0.6)		20	1.1514	1.1612	1.1038	1.1849	1.2004	1.1704
		90	3.1559	3.1928	3.0424	3.2308	3.2478	3.1751
	0.6	4	0.4348	0.4903	0.4435	0.4511	0.4998	0.4699
		20	1.0077	1.0631	1.0336	1.0374	1.0812	1.0731
		90	3.1015	3.1390	3.0129	3.1693	3.1774	3.1243
	0.9	4	0.3915	0.4592	0.4203	0.4006	0.4614	0.4569
		20	0.9419	1.0208	0.9900	0.9923	1.0328	1.0158
		90	3.0656	3.1001	3.0039	3.1314	3.1359	3.1048
0.8	0.3	4	1.6903	2.0775	1.8276	1.7123	2.0756	1.8352
(3.2)		20	2.2025	2.4563	2.4512	2.2108	2.4506	2.4266
		90	5.4605	5.5195	5.4999	5.4719	5.5069	5.5009
	0.6	4	1.6488	2.0287	1.8545	1.6567	2.0250	1.8526
		20	2.0542	2.3969	2.3471	2.0742	2.3286	2.2792
		90	5.1898	5.2764	5.2653	5.1957	5.2607	5.2524
	0.9	4	1.6122	2.0153	1.8659	1.6174	2.0091	1.8590
		20	1.9140	2.5574	2.5059	1.9804	2.2879	2.2351
		90	5.0987	5.0961	5.0792	5.1042	5.1006	5.0884

* Observed standard errors $\sigma(\hat{\mathbf{u}}-\mathbf{u})$ and $\sigma(\tilde{\mathbf{u}}-\mathbf{u})$ were found based on 1000 replicated samples.

Predicted standard errors $\sigma(\hat{\mathbf{u}}-\mathbf{u}|\sigma_{e,u}^2)$ and $\sigma(\tilde{\mathbf{u}}-\mathbf{u}|\sigma_{e,u}^2)$ were calculated as square roots of formulas (3.56) and (3.58), respectively.

□ Estimated standard deviations $\hat{\sigma}_0(\hat{\mathbf{u}}-\mathbf{u}|\hat{\sigma}_{e,u}^2)$ and $\hat{\sigma}_0(\tilde{\mathbf{u}}-\mathbf{u}|\hat{\sigma}_{e,u}^2)$ were calculated as square roots of formulas (3.57) and (3.61), respectively; in case of negative mean square error estimates the estimated standard deviations were equated to zero.

Studying the estimated standard deviations of prediction errors, the differences between these and corresponding observed accuracy parameters are more visible. The standard deviations of the prediction error $\hat{\mathbf{u}} - \mathbf{u}$ are mainly overestimated in case of large values of intraclass correlation coefficient and underestimated in case of small values of ρ. But this bias is not too big, especially compared with estimated standard deviations of two-stage prediction error $\tilde{\mathbf{u}} - \mathbf{u}$. The estimated standard deviations of two-stage predictors $\tilde{\mathbf{u}} - \mathbf{u}$ are without appreciable bias in case of average values of ρ, but are overestimated in case of small or large values of ρ (especially in case of small number of groups). There are several extremely large estimators of $\sigma_0(\tilde{\mathbf{u}} - \mathbf{u})$ in case of small values of ρ and a. On the average the variability of estimated standard deviations of prediction errors $\tilde{\mathbf{u}} - \mathbf{u}$ is 10 times bigger than the corresponding accuracy parameter of prediction errors $\hat{\mathbf{u}} - \mathbf{u}$. For estimated mean square errors of predictors (not shown here) the difference between variability of estimated MSE($\hat{\mathbf{u}}$) and estimated MSE($\tilde{\mathbf{u}}$) lies approximately in the interval (10000, 10^{16}), measured in times. In standard error scale this means the difference between 100 and 10^8 times (the difference is bigger compared to variability of $\sigma_0(\tilde{\mathbf{u}} - \mathbf{u})$ because the left-side negative part of estimated mean square errors distribution is also taken into account).

Thereby it seems that using the standard error of two-stage estimators in case of small values of intraclass correlation coefficient and numbers of groups is not advisable. But in case of not too small values of ρ, when the real values of variance components are not known, the use of derived formulas (3.58) and (3.61) is legitimate.

3.4.4. The inadmissible estimates of heritability

In the next theorem the approximated probability to get an inadmissible estimate of heritability with ANOVA method in unbalanced case is derived.

Theorem 3.8. *In the additive genetic sire model under the normality assumptions and in unbalanced data the probability to get the inadmissible heritability estimates is approximately expressed as*

$$\mathrm{P}(\hat{h}^2 < 0) + \mathrm{P}(\hat{h}^2 > 1) \approx \mathrm{P}\left[\mathrm{F}_{m,N-a} < \frac{(a-1)\sigma_e^2}{m\lambda} \right] + \mathrm{P}\left[\mathrm{F}_{m,N-a} > \frac{(d+3)(a-1)\sigma_e^2}{3m\lambda} \right], \tag{3.62}$$

where a is the number of sires, d is a coefficient (3.7), *m and λ are defined with formulas* (3.15) *and* (3.16), *respectively, and* $\mathrm{F}_{m,a(n-1)}$ *means random variable with F-distribution with m and $a(n-1)$ degrees of freedom.*

Proof. It is obvious that $\hat{h}^2 < 0$ if $\hat{\sigma}_u^2 < 0$ and $\hat{h}^2 > 1$ if $\hat{\sigma}_u^2 / \hat{\sigma}_e^2 > 1/3$ (Theorem 3.3). The approximated probability of negative $\hat{\sigma}_u^2$ is derived in Khuri, Mathew and Sinha (1998, p 58). Based on distributional properties of mean squares (3.11) and (3.13), and applying the approximation (3.14), the probability of negative $\hat{h}^2$ is expressed as follows:

$$\begin{aligned} \mathrm{P}(\hat{h}^2<0) &= \mathrm{P}(\hat{\sigma}_u^2<0) = \mathrm{P}\left[MS(u)<MS(e)\right] \\ &= \mathrm{P}\left[\frac{1}{a-1}\sum_{i=1}^{s}\lambda_i\chi_{m_i}^2 < \frac{\sigma_e^2}{N-a}\chi_{N-a}^2\right] \\ &\approx \mathrm{P}\left[\frac{\lambda\chi_m^2}{a-1} < \frac{\sigma_e^2\chi_{N-a}^2}{N-a}\right] = \mathrm{P}\left[\mathrm{F}_{m,N-a} < \frac{(a-1)\sigma_e^2}{m\lambda}\right]. \end{aligned} \tag{3.63}$$

Analogous the probability of $\hat{h}^2>1$ is approximated in the following way:

$$\begin{aligned} \mathrm{P}(\hat{h}^2>1) &= \mathrm{P}\left[\frac{\mathrm{MS}(u)}{\mathrm{MS}(e)} > \frac{d}{3}+1\right] \\ &\approx \mathrm{P}\left[\frac{\lambda\chi_m^2/(a-1)}{\sigma_e^2\chi_{N-a}^2/N-a} < \frac{d}{3}+1\right] = \mathrm{P}\left[\mathrm{F}_{m,N-a} > \frac{(d+3)(a-1)\sigma_e^2}{3m\lambda}\right]. \end{aligned} \tag{3.64}$$

The two derived expressions together give us the approximation in the theorem statement. ■

Results of simulation studies controlling derived formulas are presented in Table 3.14. There is no clear bias in predicted probabilities of $\hat{\sigma}_u^2<0$ and $\hat{h}^2>1$. Similarly to balanced data studies, the probability to get negative estimate of mean square error of two-stage predictors is smaller than the probability to get negative estimate of σ_u^2.

Table 3.14. The observed and predicted probabilities of the inadmissible estimates in the case of different true population values, data set imbalance ν and number of groups a ($N=360$, $\sigma_e^2=1$) found based on 1000 replicated samples.

ρ (h^2)	ν	a	$\mathrm{P}(\hat{\sigma}_u^2<0)^*$	$\mathrm{P}(\hat{\sigma}_u^2<0\|\sigma_{u,e}^2)^\#$	$\mathrm{P}(\hat{h}^2>1)^*$	$\mathrm{P}(\hat{h}^2>1\|\sigma_{u,e}^2)^\#$	$\mathrm{P}\left[\widehat{\mathrm{MSE}}(\tilde{\mathbf{u}})<0\right]^*$
0.0125	0.3	4	0.451	0.4968	0.000	0.0003	0.236
(0.05)		20	0.334	0.3381	0.000	0.0000	0.154
		90	0.405	0.4023	0.000	0.0000	0.225
	0.6	4	0.341	0.3449	0.000	0.0000	0.106
		20	0.287	0.3186	0.000	0.0000	0.080
		90	0.393	0.4002	0.000	0.0000	0.253
	0.9	4	0.298	0.3038	0.000	0.0000	0.048
		20	0.313	0.3116	0.000	0.0000	0.172
		90	0.397	0.3995	0.000	0.0000	0.297
0.0625	0.3	4	0.255	0.2891	0.029	0.0260	0.130
(0.25)		20	0.042	0.0637	0.000	0.0002	0.010
		90	0.092	0.1078	0.001	0.0002	0.033
	0.6	4	0.099	0.1210	0.014	0.0095	0.027
		20	0.023	0.0343	0.000	0.0000	0.004
		90	0.101	0.0992	0.000	0.0001	0.062
	0.9	4	0.050	0.0750	0.004	0.0053	0.005
		20	0.021	0.0246	0.000	0.0000	0.009
		90	0.093	0.0960	0.000	0.0001	0.070

ρ (h^2)	ν	a	$P(\hat{\sigma}_u^2<0)^*$	$P(\hat{\sigma}_u^2<0\|\sigma_{u,e}^2)^\#$	$P(\hat{h}^2>1)^*$	$P(\hat{h}^2>1\|\sigma_{u,e}^2)^\#$	$P\left[\widehat{\mathrm{MSE}}(\hat{\mathbf{u}})<0\right]^*$
0.15	0.3	4	0.137	0.1685	0.159	0.1798	0.054
(0.6)		20	0.004	0.0105	0.082	0.0869	0.000
		90	0.002	0.0054	0.033	0.0446	0.001
	0.6	4	0.033	0.0533	0.160	0.1571	0.012
		20	0.001	0.0016	0.061	0.0561	0.001
		90	0.002	0.0027	0.036	0.0344	0.001
	0.9	4	0.022	0.0237	0.132	0.1439	0.001
		20	0.000	0.0005	0.036	0.0421	0.000
		90	0.003	0.0019	0.027	0.0306	0.001

* Observed probabilities were found based on 1000 replicated samples.
# Predicted probabilities $P(\hat{\sigma}_u^2<0|\sigma_{u,e}^2)$ and $P(\hat{h}^2>1|\sigma_{u,e}^2)$ were calculated by formulas (3.63) and (3.64), respectively.

3.5. The effect of data imbalance

3.5.1. The effect of data imbalance on $\mathrm{Var}(\hat{\sigma}_u^2)$

Theorem 3.2.1 in Khuri et al (1998, p 56–57) proves that $\mathrm{Var}(\hat{\sigma}_u^2)$ attains a minimum for all σ_u^2 and σ_e^2 if and only if the data set is balanced.

For unbalanced data the criteria of optimal design is not clear. Anderson and Crump (1967) concluded that the design with closest number of classes should be used to get the optimum. Norell (2001, 2003) showed that this suggestion not always yields to the minimum of $\mathrm{Var}(\hat{\sigma}_u^2)$. Khuri et al (1998, p 81–87) established an approximate empirical relationship between $\mathrm{Var}(\hat{\sigma}_u^2)$ and $\nu(\mathbf{D})$, and showed that the sampling variance of $\hat{\sigma}_u^2$ increases as imbalance increases.

To visualise the effect of data imbalance on the accuracy of variance component estimate $\hat{\sigma}_u^2$, the modelling experiments were implemented. Similarly to paragraph 3.3 the standard deviations of $\hat{\sigma}_u^2$, as parameters found usually in practice, were calculated by formula (3.46) in case of different combinations of data imbalance and intraclass correlation coefficients. Since there are different data designs corresponding to a given imbalance, then with specified value of $\nu(\mathbf{D})$, on average five designs were generated and the average value of $\sigma(\hat{\sigma}_u^2)$ was used to characterise the effect of the corresponding imbalance. Figure 3.10 shows the dependence of standard deviation of estimated σ_u^2 on $\nu(\mathbf{D})$ and ρ, keeping the data size $N=360$, number of groups $a=20$, and error variance $\sigma_e^2=1$.

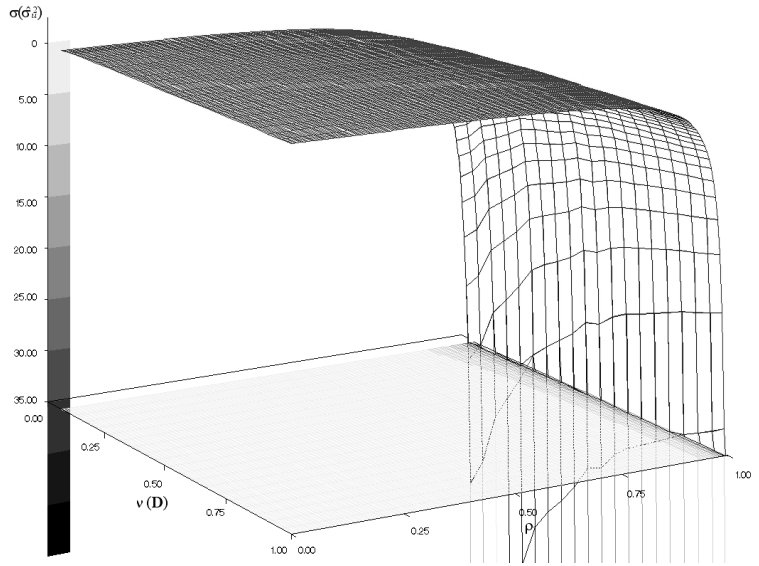


Figure 3.10. The patterns of $\sigma(\hat{\sigma}_u^2)$ in different data set imbalances ν(**D**) and true intraclass correlation coefficient values ρ for $N = 360$, $\sigma_e^2 = 1$ and numbers of groups $a = 20$.

Since similarly to balanced data case (Figure 3.1), the accuracy of variance components estimation decreases quickly if the intraclass correlation coefficient comes close to its upper limit, then the additional modelling experiments were implemented with intraclass correlation coefficient values varying from 0 to 0.25 (in case of half-sib model this covers all admissible values of heritability). Figure 3.11 shows the dependence of standard deviation of estimated σ_u^2 on ν(**D**) and $h^2 = \frac{1}{4}\rho$, keeping the data size $N = 360$, number of groups $a = 4, 20, 90$, and error variance $\sigma_e^2 = 1$.

The modelling results show that even a quite notable increase of data imbalance does practically not decrease the accuracy of intraclass correlation coefficient estimate. Yet in the case of very imbalanced data, the standard deviation of $\hat{\rho}$ quickly increases along with the imbalance. Comparing modelling experiments implemented with different numbers of groups, it is obvious that the influence of data imbalance on the accuracy of intraclass correlation coefficient estimators is stronger when the number of groups is small. It is also visible that in case of very unbalanced design, the accuracy of $\hat{\rho}$ decreases more drastically for small values of intraclass correlation coefficient. The most imprecise estimates are got for large values of ρ and h^2.

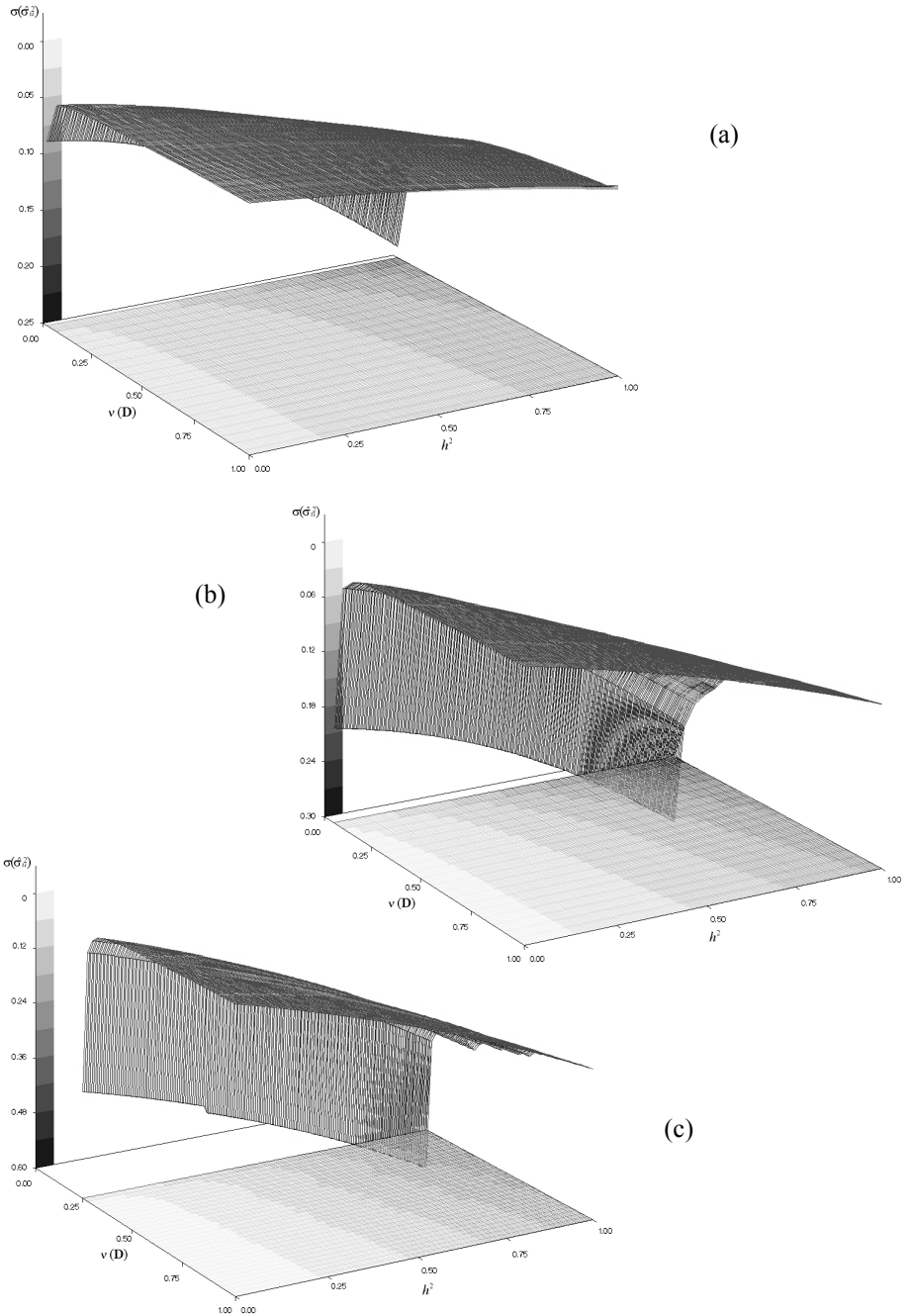


Figure 3.11. The patterns of $\sigma(\hat{\sigma}_u^2)$ in different data set imbalances $\nu(\mathbf{D})$ and true heritability coefficient values h^2 for $N = 360$, $\sigma_e^2 = 1$ and numbers of groups (a) $a = 90$, (b) $a = 20$, (c) $a = 4$.

3.5.2. The effect of data imbalance on $\mathrm{Var}(\hat{\rho})$

In the following it is proved that $\mathrm{Var}(\hat{\rho})$ expressed by (3.49) attains a minimum for all σ_u^2 and σ_e^2 if and only if the data set is balanced.

Theorem 3.9. *For fixed values of N, a, σ_u^2 and σ_e^2, $\mathrm{Var}(\hat{\rho})$ attains a minimum if and only if the data set is balanced.*

Proof. Rewrite the expression (3.49) in the form

$$\mathrm{Var}(\hat{\rho}) \approx \left[\frac{m\lambda^2(N-a-2)}{d^2(a-1)^2} + \frac{(m\lambda)^2}{d^2(a-1)^2} \right] \times \frac{2(N-a)^2(1-\rho)^4}{(N-a-2)^2(N-a-4)(\sigma_e^2)^2}, \tag{3.65}$$

where only the first part depends on the design.

From formulas (3.15) and (3.16), and general properties of eigenvalues (Proposition 3.2) we have

$$m\lambda = \textstyle\sum_{i=1}^{s} m_i\lambda_i = \mathrm{tr}(\mathbf{Q}_1\mathbf{V}) \text{ and } m\lambda^2 = \textstyle\sum_{i=1}^{s} m_i\lambda_i^2 = \mathrm{tr}\left[(\mathbf{Q}_1\mathbf{V})^2\right]. \tag{3.66}$$

From formula (3.4) we have

$$\mathrm{tr}(\mathbf{Q}_1\mathbf{V}) = \left(N - \tfrac{1}{N}\textstyle\sum_{i=1}^{a} n_i^2\right)\sigma_u^2 + (a-1)\sigma_e^2 .$$

As we also have the expression for d of the form (3.7), we can write

$$\mathrm{tr}(\mathbf{Q}_1\mathbf{V}) = (a-1)(\sigma_e^2 + d\sigma_u^2),$$

from which it follows that

$$d = \frac{\mathrm{tr}(\mathbf{Q}_1\mathbf{V})}{\sigma_u^2(a-1)} - \frac{\sigma_e^2}{\sigma_u^2} = \frac{1}{\sigma_u^2(a-1)}\left[\mathrm{tr}(\mathbf{Q}_1\mathbf{V}) - (a-1)\sigma_e^2\right]. \tag{3.67}$$

Based on expressions (3.66) and (3.67) we have for the first addend in the square brackets of (3.65) that

$$\begin{aligned} \frac{m\lambda^2(N-a-2)}{d^2(a-1)^2} &= \frac{\sigma_u^4(N-a-2)\,\mathrm{tr}\left[(\mathbf{Q}_1\mathbf{V})^2\right]}{\left[\mathrm{tr}(\mathbf{Q}_1\mathbf{V}) - (a-1)\sigma_e^2\right]^2} \\ &= \sigma_u^4(N-a-2) \times \frac{\mathrm{tr}\left[(\mathbf{Q}_1\mathbf{V})^2\right]}{\left[\mathrm{tr}(\mathbf{Q}_1\mathbf{V})\right]^2} \times \left[1 - \frac{(a-1)\sigma_e^2}{\mathrm{tr}(\mathbf{Q}_1\mathbf{V})}\right]^{-2}. \end{aligned}$$

Here the first term does not depend on design. The second term $\mathrm{tr}[(\mathbf{Q}_1\mathbf{V})^2]/[\mathrm{tr}(\mathbf{Q}_1\mathbf{V})]^2$ has its minimum value equal to the reciprocal of the rank of $\mathbf{Q}_1$, which is equal to $a-1$, if and only if $n_i = n$ for all i (this can be shown by making use of Theorem 9.1.22 in Graybill, 1983, p 303). For the third term we have that

$$\left[1 - \frac{(a-1)\sigma_e^2}{\mathrm{tr}(\mathbf{Q}_1\mathbf{V})}\right]^{-2} = \left[1 - \frac{\sigma_e^2}{\sigma_e^2 + d\sigma_u^2}\right]^{2} = \left(1 + \frac{\sigma_e^2}{d\sigma_u^2}\right)^{2},$$

which has its minimum, equal to $(1+\sigma_e^2/n\sigma_u^2)^2$, if and only if $n_i = n$ for all i, because d is at its maximum, namely $d = n$, if and only if $n_i = n$ for all i.

Similarly, we have for the second addend in the square brackets of (3.65) that

$$\frac{(m\lambda)^2}{d^2(a-1)^2} = \frac{\sigma_u^4 \operatorname{tr}\left[(\mathbf{Q}_1\mathbf{V})\right]^2}{\left[\operatorname{tr}(\mathbf{Q}_1\mathbf{V})\right]^2}\left(1+\frac{\sigma_e^2}{d\sigma_u^2}\right)^2 = \sigma_u^4\left(1+\frac{\sigma_e^2}{d\sigma_u^2}\right)^2.$$

The last expression here has its minimum $\sigma_u^4(1+\sigma_e^2/n\sigma_u^2)^2$ if and only if $n_i = n$ for all i. We therefore conclude that $\mathrm{Var}(\hat{\rho})$ is at its minimum, which is given by the formula

$$\begin{aligned}\mathrm{Var}(\hat{\rho}) &\approx \left[\frac{\sigma_u^4(N-a-2)}{(a-1)}\left(1+\frac{\sigma_e^2}{n\sigma_u^2}\right)^2 + \sigma_u^4\left(1+\frac{\sigma_e^2}{n\sigma_u^2}\right)^2\right] \\ &\quad\times\frac{2(N-a)^2(1-\rho)^4}{(N-a-2)^2(N-a-4)\sigma_e^4} \\ &= \frac{2(\sigma_e^2+n\sigma_u^2)(1-\rho)^4(N-a)^2(N-3)}{\sigma_e^4 n^2(a-1)(N-a-2)^2(N-a-4)} \\ &= \frac{2[1+(n-1)\rho]^2(1-\rho)^2(N-a)^2(N-3)}{n^2(a-1)(N-a-2)^2(N-a-4)},\end{aligned} \tag{3.68}$$

if and only if the data set is balanced. This completes the proof of Theorem 3.9. ■

Note that the minimum of $\mathrm{Var}(\hat{\rho})$ expressed by (3.68) and corresponding to balanced data has the same form as approximated $\mathrm{Var}(\hat{\rho})$ derived by Zerbe and Goldgar (1980) assuming balanced data and presented earlier in (3.26).

To visualize the effect of data imbalance on the accuracy of estimated intraclass correlation coefficient, the modelling experiments were used. The standard deviations of intraclass correlation coefficient estimate were calculated by formula (3.49) in case of different combinations of data imbalance and intraclass correlation coefficients. Since there are different data designs corresponding to a given imbalance, then with specified value of $\nu(\mathbf{D})$, on average five designs were generated and average value of $\sigma(\hat{\rho})$ was used as representative to corresponding imbalance. Figure 3.12 shows the dependence of standard deviation of estimated intraclass correlation coefficient on $\nu(\mathbf{D})$ and ρ, keeping the data size $N = 360$, number of groups $a = 4, 20, 90$, and error variance $\sigma_e^2 = 1$ (note that the axis of intraclass correlation coefficient values is drawn in the opposite order to better visualize the pattern of $\sigma(\hat{\rho})$).

The modelling results show that even quite a notable increase of data imbalance does practically not decrease the accuracy of intraclass correlation coefficient estimate. Yet in the case of very imbalanced data the standard deviation of $\hat{\rho}$ increases quickly along with the imbalance. Comparing modelling experiments implemented with different numbers of groups, it is obvious that the influence of data imbalance on the accuracy of intraclass correlation coefficient estimators is stronger when the number of groups is small. It is also visible that

in case of very unbalanced design the accuracy of $\hat{\rho}$ decreases more drastically for small values of intraclass correlation coefficient. The most imprecise estimates are got for average values of ρ.

3.5.3. The effect of data imbalance on MSE($\hat{u}_i$) and MSE($\bar{u}_i$)

For balanced design it is clear that the MSE($\hat{\mathbf{u}}$) has its minimum value if the number of groups is minimal. Due to this the effect of data design in balanced case was studied only for MSE($\hat{u}_i$). In unbalanced case the MSE($\hat{u}_i$) depends on the number of observations in group i and has the smallest value for the biggest n_i. The data imbalance has effect only on MSE($\hat{\mathbf{u}}$). In following theorem it is proved that the mean square error of predictors attains a minimum for all σ_u^2 and σ_e^2 if and only if the data set is balanced.

Theorem 3.10. *For fixed values of N and a,* MSE($\hat{\mathbf{u}}$) *attains a minimum for all* σ_u^2 *and* σ_e^2 *if and only if the data set is balanced.*

Proof. In the following three well-known inequalities are applied. These are
(i) the inequality involving arithmetic and harmonic means of the form

$$\frac{x_1 + x_2 + \cdots + x_n}{n} \geq \frac{n}{1/x_1 + 1/x_2 + \cdots + 1/x_n},$$

from which it follows that $1/x_1 + 1/x_2 + \cdots + 1/x_n \geq n^2/(x_1 + x_2 + \cdots + x_n)$, and where equality holds if and only if $x_1 = x_2 = \cdots = x_n$;
(ii) Cauchy-Schwarz inequality of the form

$$(x_1y_1 + x_2y_2 + \cdots + x_ny_n)^2 \leq (x_1^2 + x_2^2 + \cdots + x_n^2)(y_1^2 + y_2^2 + \cdots + y_n^2),$$

where equality holds if and only if $x_1/y_1 = x_2/y_2 = \cdots = x_n/y_n$, and from which it follows that $\sum_{i=1}^n x_i^2 \geq \frac{1}{n}(\sum_{i=1}^n x_i)^2$;
(iii) Chebyshev's inequality of the form

$$(x_1 + x_2 + \cdots + x_n)(y_1 + y_2 + \cdots + y_n) \leq n(x_1y_1 + x_2y_2 + \cdots + x_ny_n),$$

from which it follows that $\sum_{i=1}^n x_iy_i \geq \frac{1}{n} \times \sum_{i=1}^n x_i \times \sum_{i=1}^n y_i$.

For the first addend in formula (3.56) we have by the inequality (i)

$$\sum_{i=1}^{a} \frac{1}{\sigma_e^2 + n_i\sigma_u^2} \geq \frac{a^2}{\sum_{i=1}^a (\sigma_e^2 + n_i\sigma_u^2)} = \frac{a^2}{a\sigma_e^2 + N\sigma_u^2}, \tag{3.69}$$

where equality holds if and only if $n_i = n$ for all $i = 1, 2, ..., a$.

For the sum in the second addend of (3.56) it holds according to the inequality (iii) that

$$\sum_{i=1}^{a} \frac{n_i^2}{(\sigma_e^2 + n_i\sigma_u^2)^2} \geq \frac{1}{a} \sum_{i=1}^{a} n_i^2 \sum_{i=1}^{a} \frac{1}{(\sigma_e^2 + n_i\sigma_u^2)^2}. \tag{3.70}$$

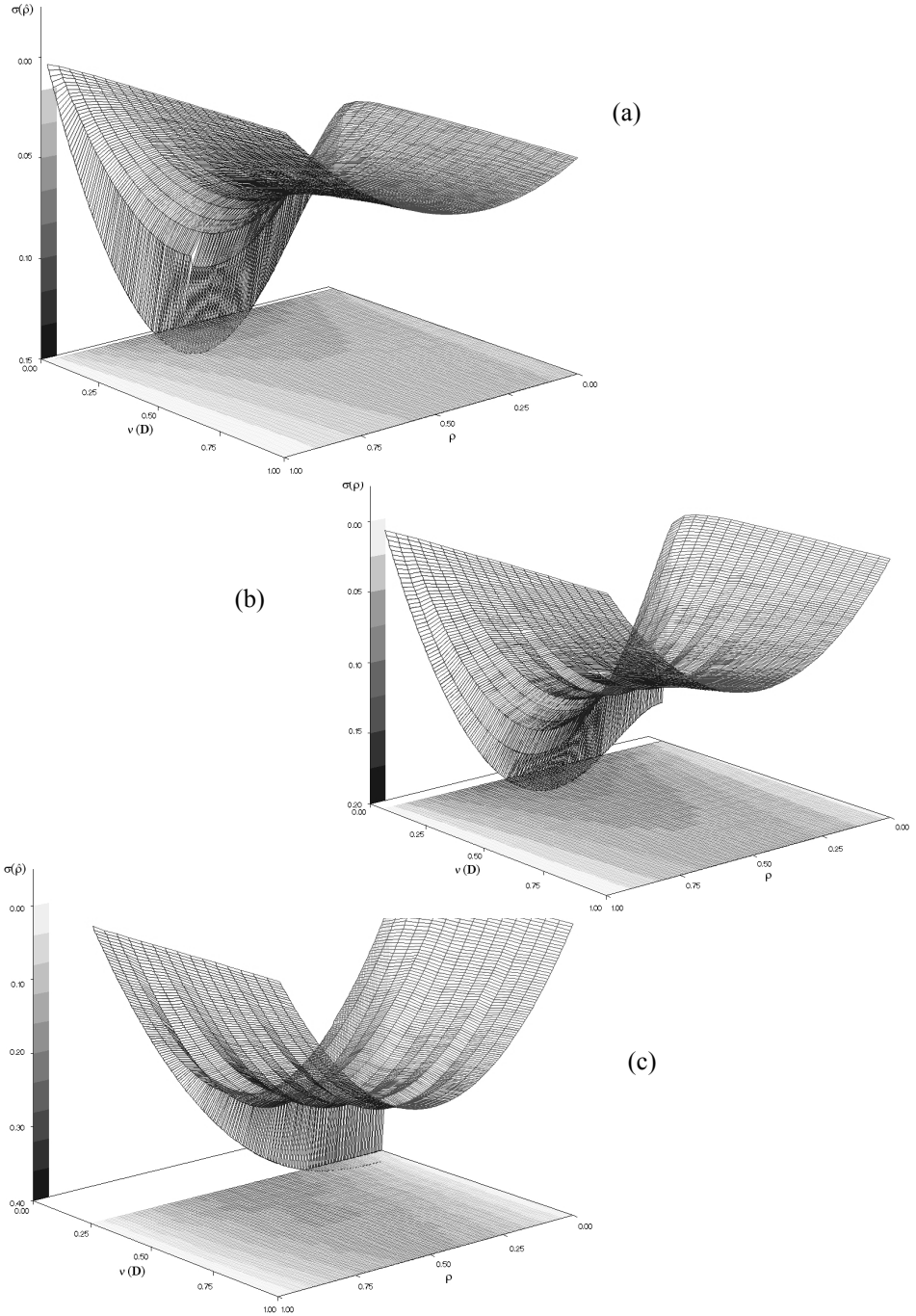


Figure 3.12. The patterns of $\sigma(\hat{\rho})$ in different data set imbalances ν(**D**) and true intra-class correlation coefficient values ρ for $N = 360$, $\sigma_e^2 = 1$ and numbers of groups (a) $a = 90$, (b) $a = 20$, (c) $a = 4$.

For the first sum in (3.70) we have by the Cauchy-Schwarz inequality

$$\sum_{i=1}^{a} n_i^2 \geq \frac{1}{a}\left(\sum_{i=1}^{a} n_i\right)^2 = \frac{1}{a}N^2,$$

where equality holds if and only if $n_i = n$ for all $i = 1,2,...,a$. For the second sum in (3.70) we have, after sequential applying of the Cauchy-Schwarz inequality and inequality (3.69), that it attains its minimum value $a^3/(a\sigma_e^2 + N\sigma_u^2)^2$ if and only if $n_i = n$ for all $i = 1,2,...,a$.

We therefore conclude that MSE($\hat{\mathbf{u}}$) is at its minimum, which is given by formula (3.28), if and only if the data set is balanced. ■

For mean square error of two-stage predictors MSE($\tilde{\mathbf{u}}$) of the form (3.58), it is easy to show that in case of balanced data the expression (3.58) simplifies to the previously derived form (3.36). All auxiliary variables in (3.58), noted as k_{fh} and c, attain their minimum values if and only if the data set is balanced, that is $n_i = n$ for all i. To show this the same three inequalities used in the proof of Theorem 3.10, must be applied. For variable k_{fh} in (3.58) the following inequality holds:

$$k_{fh} = \sum_{i=1}^{a} \frac{n_i^f}{(\sigma_e^2 + n_i\sigma_u^2)^h} \underset{\text{(iii)}}{\geq} \frac{1}{a} \times \sum_{i=1}^{a} n_i^f \times \sum_{i=1}^{a} \frac{1}{(\sigma_e^2 + n_i\sigma_u^2)^h}. \tag{3.71}$$

The lower limit of the first sum in the last product can be expressed as

$$\sum_{i=1}^{a} n_i^f \underset{\text{(ii)}}{\geq} \frac{1}{a^{f-1}}\left(\sum_{i=1}^{a} n_i\right)^f = \frac{N^f}{a^{f-1}},$$

and the lower limit of the second sum in the right side of (3.71) can be expressed as

$$\begin{aligned}\sum_{i=1}^{a} \frac{1}{(\sigma_{e2}^2 + n_i\sigma_u^2)^h} &= \sum_{i=1}^{a}\left[\frac{1}{(\sigma_{e2}^2 + n_i\sigma_u^2)}\right]^h \underset{\text{(iii)}}{\geq} \frac{1}{a^{h-1}}\left[\sum_{i=1}^{a}\frac{1}{(\sigma_e^2 + n_i\sigma_u^2)}\right]^h \\ &\underset{\text{(i)}}{\geq} \frac{1}{a^{h-1}}\left[\frac{a^2}{\sum_{i=1}^{a}(\sigma_e^2 + n_i\sigma_u^2)}\right]^h \\ &= \frac{a^{h+1}}{\left(a\sigma_e^2 + \sigma_u^2\sum_{i=1}^{a} n_i\right)^h} = \frac{a^{h+1}}{\left(a\sigma_e^2 + \sigma_u^2 N\right)^h}.\end{aligned}$$

Hence, the lower limit of variable k_{fh} is written of the form

$$k_{fh} \geq \frac{a^{h+1}N^f}{a^{f-1}\left(a\sigma_e^2 + \sigma_u^2 N\right)^h},$$

where equality holds if and only if the data set is balanced, that is $n_i = n$ for all $i = 1,2,...,a$ and thus $N = an$. Then $k_{fh} = an^f/(\sigma_e^2 + n\sigma_u^2)^h$. Similarly the lower limit of variable c in (3.58) is expressed as

$$c = \frac{\sigma_e^2}{N - \sum_{i=1}^{a}\left[n_i^2\sigma_u^2 / (\sigma_e^2 + n_i\sigma_u^2)\right]} \geq \frac{a\sigma_e^2 + N\sigma_u^2}{aN},$$

where equality holds if and only if $n_i = n$ for all i. Then $c = (\sigma_e^2 + n\sigma_u^2)/N$.

To visualize the effect of data imbalance on the standard deviations of prediction error, the modelling experiments were implemented. The standard deviations of prediction error were calculated as square roots of formula (3.56) in case of different combinations of data imbalance and intraclass correlation coefficients. Similarly to previous modelling experiments in paragraph 3.5, with specified value of $\nu(\mathbf{D})$, on average five designs were generated and average value of $\sigma(\hat{\mathbf{u}} - \mathbf{u})$ was used as representative to corresponding imbalance. Figure 3.13 shows the dependence of standard deviation of prediction error on $\nu(\mathbf{D})$ and ρ, keeping the data size $N = 360$, number of groups $a = 4, 20, 90$, and error variance $\sigma_e^2 = 1$.

Also, the analogous modelling experiments were used to study the effect of data imbalance on the standard deviations of second-stage prediction error. To calculate $\sigma(\tilde{\mathbf{u}} - \mathbf{u})$ in case of different combinations of data imbalance and intraclass correlation coefficient values, a square root of formula (3.58) was used. The patterns of the standard deviations of the second-stage prediction error are shown in Figure 3.14.

The patterns of $\sigma(\hat{\mathbf{u}} - \mathbf{u})$ and $\sigma(\tilde{\mathbf{u}} - \mathbf{u})$ are quite similar. More visible differences occur in case of small intraclass correlation coefficients, where the standard deviations of second-stage prediction errors are larger. The dependency on the data imbalance is imaginary compared to the data design effect and value of intraclass correlation coefficient. Yet in the case of very imbalanced data, the standard deviation of prediction error increases quickly along with the imbalance, being more conspicuous when the number of groups is big.

3.5.4 The effect of data imbalance on the probability of the inadmissible estimates

The effect of data imbalance on the approximated probability of negative $\hat{\sigma}_u^2$ of the form

$$\mathrm{P}(\hat{\sigma}_u^2 < 0) \approx \mathrm{P}\left[\mathrm{F}_{m,N-a} < (a-1)\sigma_e^2 / m\lambda\right] \tag{3.72}$$

is discussed in Khuri, Mathew and Sinha (1998, p 59). They concluded that imbalance causes a reduction in the value of *m*, which is the numerator's number of degrees of freedom of the *F*-variate in (3.72), and is defined with expression (3.16). They also showed that *m* is equal to its maximum value $a - 1$, if the data set is balanced.

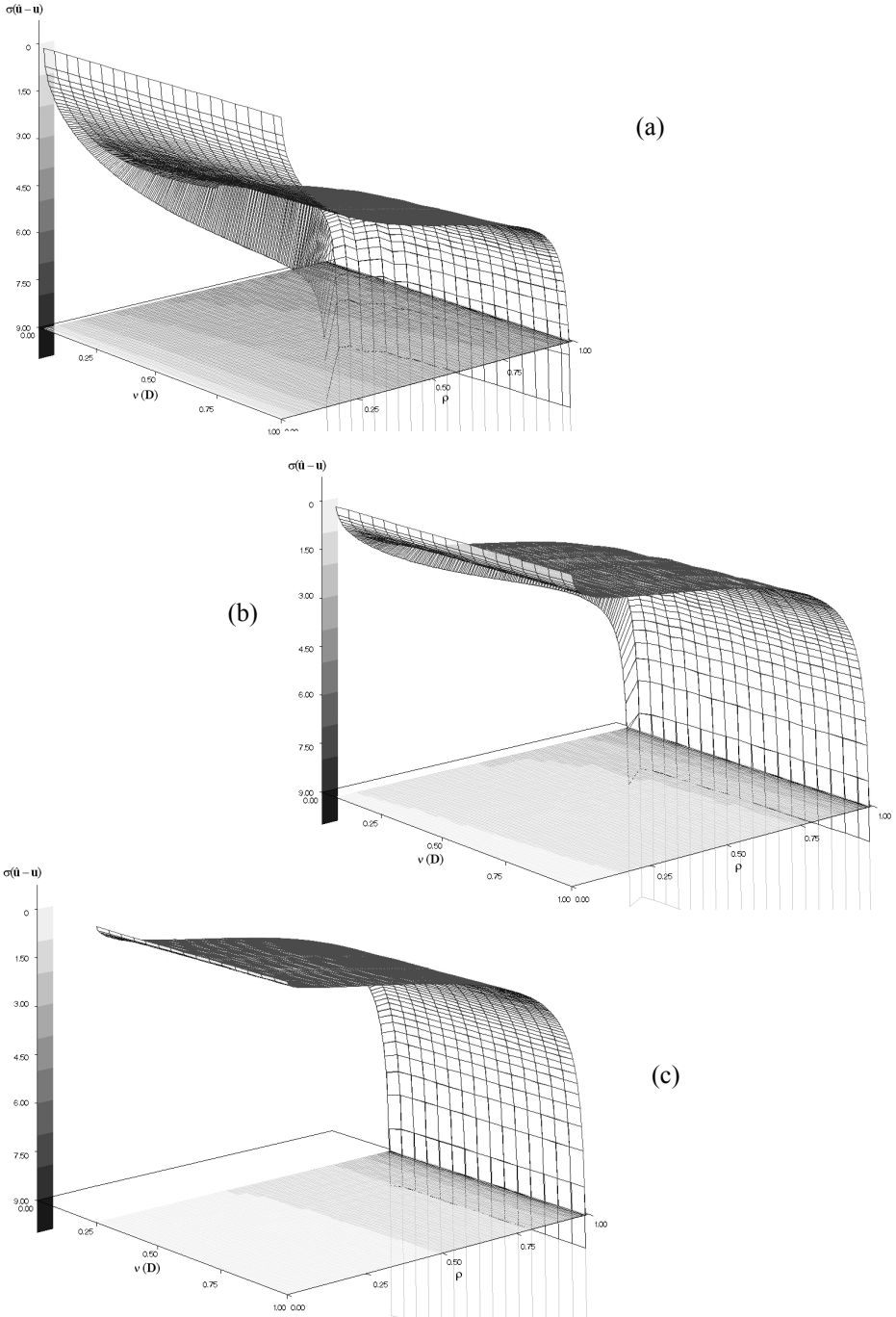


Figure 3.13. The patterns of $\sigma(\hat{\mathbf{u}} - \mathbf{u})$ in different data set imbalances $\nu(\mathbf{D})$ and true intraclass correlation coefficient values ρ for $N = 360$, $\sigma_e^2 = 1$ and numbers of groups (a) $a = 90$, (b) $a = 20$, (c) $a = 4$.

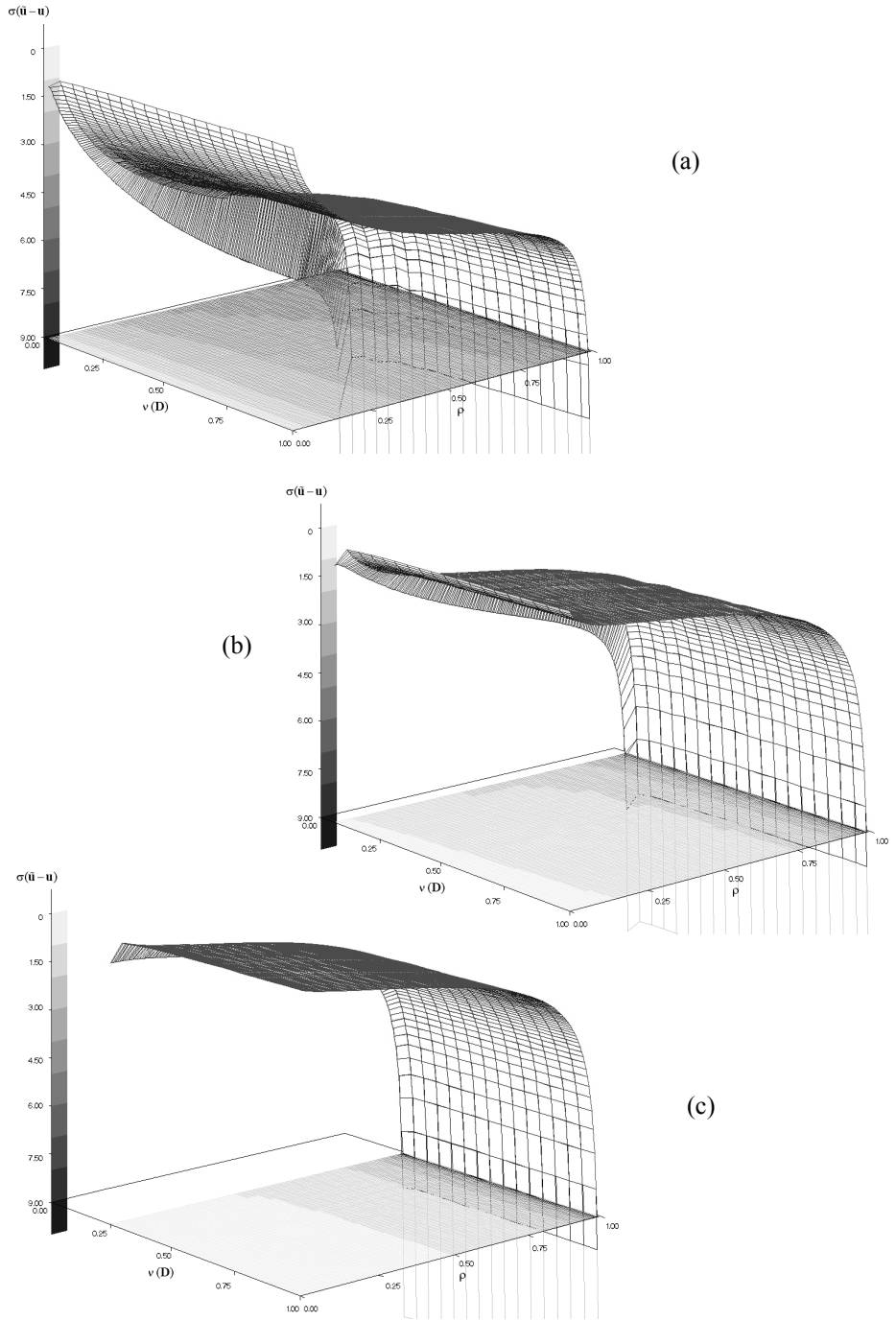


Figure 3.14. The patterns of $\sigma(\tilde{\mathbf{u}} - \mathbf{u})$ in different data set imbalances $\nu(\mathbf{D})$ and true intraclass correlation coefficient values ρ for $N = 360$, $\sigma_e^2 = 1$ and numbers of groups (a) $a = 90$, (b) $a = 20$, (c) $a = 4$.

By making use of the fact $m\lambda = \sum m_i\lambda_i = \mathrm{tr}(\mathbf{Q}_1\mathbf{V})$, and the expressions of $\mathrm{tr}(\mathbf{Q}_1\mathbf{V})$ derived in proof of Theorem 3.9, it is easy to see that imbalance causes a reduction in the value of $m\lambda$, and $m\lambda$ is equal to its maximum value $(a-1)(n\sigma_u^2+\sigma_e^2)$, if the data set is balanced. So, data imbalance causes an enlargement of the argument of the distribution function (3.72).

As a result, this implies that for balanced data set with $n_i = n \quad (i = 1,2,...,a)$ the approximated probability of negative $\hat{\sigma}_u^2$ of the form (3.72) reduces to the expression (3.42).

Singh (1989), who derived the exact expression of $\mathrm{P}(\hat{\sigma}_u^2 < 0)$ using an infinite weighted sum of incomplete beta functions, concluded based on numerical studies that imbalance increases the probability of a negative value of $\hat{\sigma}_u^2$.

Studying the effect of data imbalance on probability of inadmissible heritability estimate it follows, that the first term in approximation (3.62) $\mathrm{P}(\hat{h}^2 < 0) = \mathrm{P}(\hat{\sigma}_u^2 < 0)$ and the second term can be rewritten as

$$\mathrm{P}(\hat{h}^2 > 1) \approx \mathrm{P}\left[\mathrm{F}_{N-a,m} < \frac{3m\lambda}{(d+3)(a-1)\sigma_e^2} \right]. \tag{3.73}$$

Here the data imbalance causes a reduction in the value of *m*, which is the denominator's number of degrees of freedom of the *F*-variate in (3.73). Also, imbalance causes an enlargement of the argument of the distribution function (3.73). The last is true because (based on expressions derived in proof of Theorem 3.9)

$$\frac{3m\lambda}{(d+3)(a-1)\sigma_e^2} = \frac{3\sigma_u^2}{\sigma_e^2} \times \left[1 - \frac{(a-1)(\sigma_e^2 - 3\sigma_u^2)}{\mathrm{tr}(\mathbf{Q}_1\mathbf{V})} \right]^{-2}$$

and value of $\mathrm{tr}(\mathbf{Q}_1\mathbf{V})$ reduces if the imbalance increases.

Note that in balanced case the approximation (3.73) reduces to the previously derived form (3.43).

To visualize the effect of data imbalance on $\mathrm{P}(\hat{\sigma}_u^2 < 0)$ and $\mathrm{P}(\hat{h}^2<0) + \mathrm{P}(\hat{h}^2>1)$, the modelling experiments were implemented as in the previous sections of the current paragraph. Figures 3.15 and 3.16 show the dependence of probability of a negative value of $\hat{\sigma}_u^2$ and probability of an inadmissible estimate of $\hat{h}^2$, respectively, on $\nu(\mathbf{D})$ and ρ, keeping the data size $N = 360$, number of groups $a = 4, 20, 90$, and error variance $\sigma_e^2 = 1$.

The patterns of $\mathrm{P}(\hat{\sigma}_u^2 < 0)$ and $\mathrm{P}(\hat{h}^2<0) + \mathrm{P}(\hat{h}^2>1)$ show that – similarly to previous accuracy studies – there is no big necessity to worry about the data imbalancedness, especially in case of big number of groups (sires). The problems with inadmissible estimates may occur mainly if the real parameters values are too small or too big, and the data set contains a large number of groups with 1 or 2 observations.

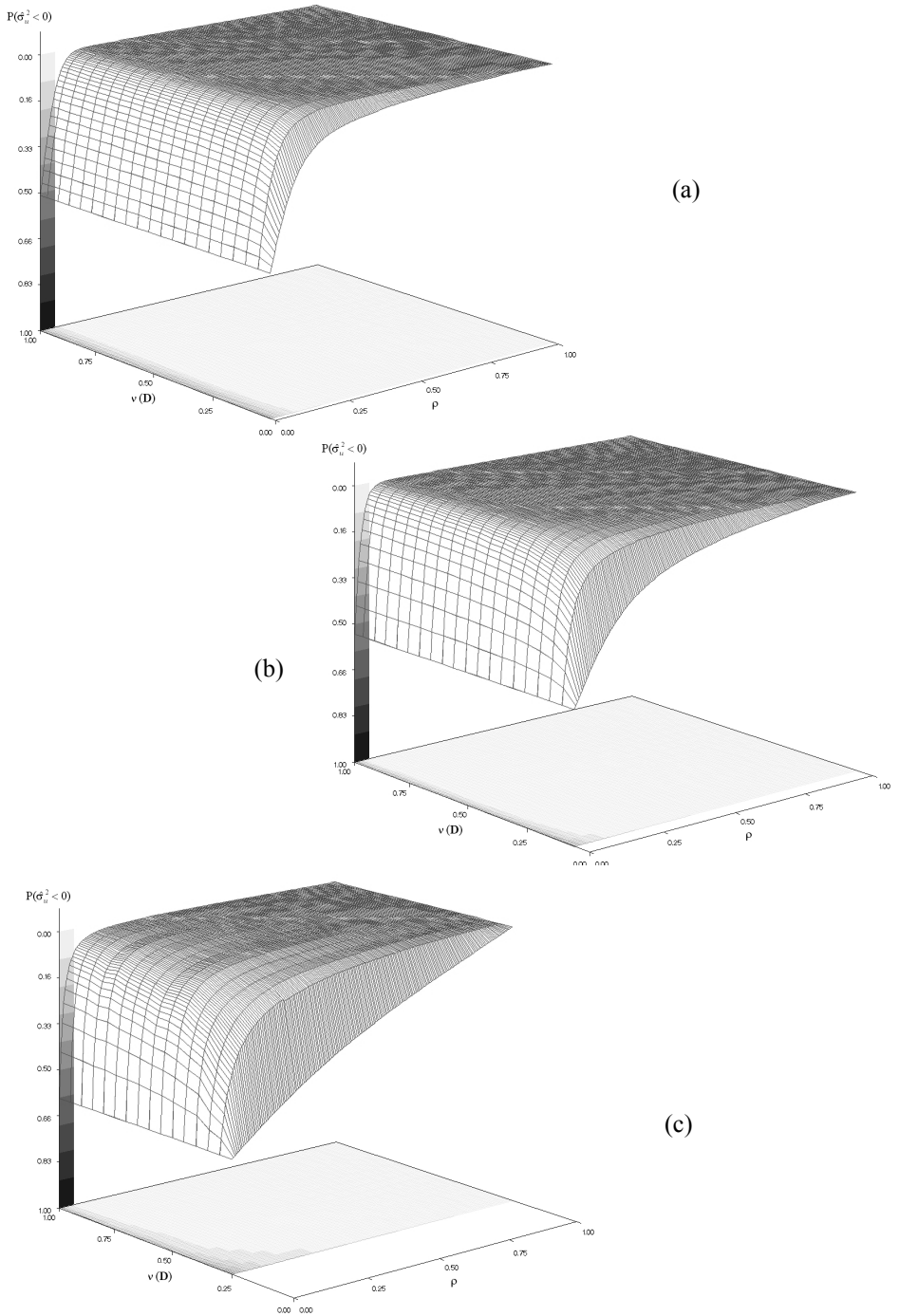


Figure 3.15. The patterns of $P(\hat{\sigma}_u^2 < 0)$ in different data set imbalances $\nu(\mathbf{D})$ and true intraclass correlation coefficient values ρ for $N = 360$, $\sigma_e^2 = 1$ and numbers of groups (a) $a = 90$, (b) $a = 20$, (c) $a = 4$.

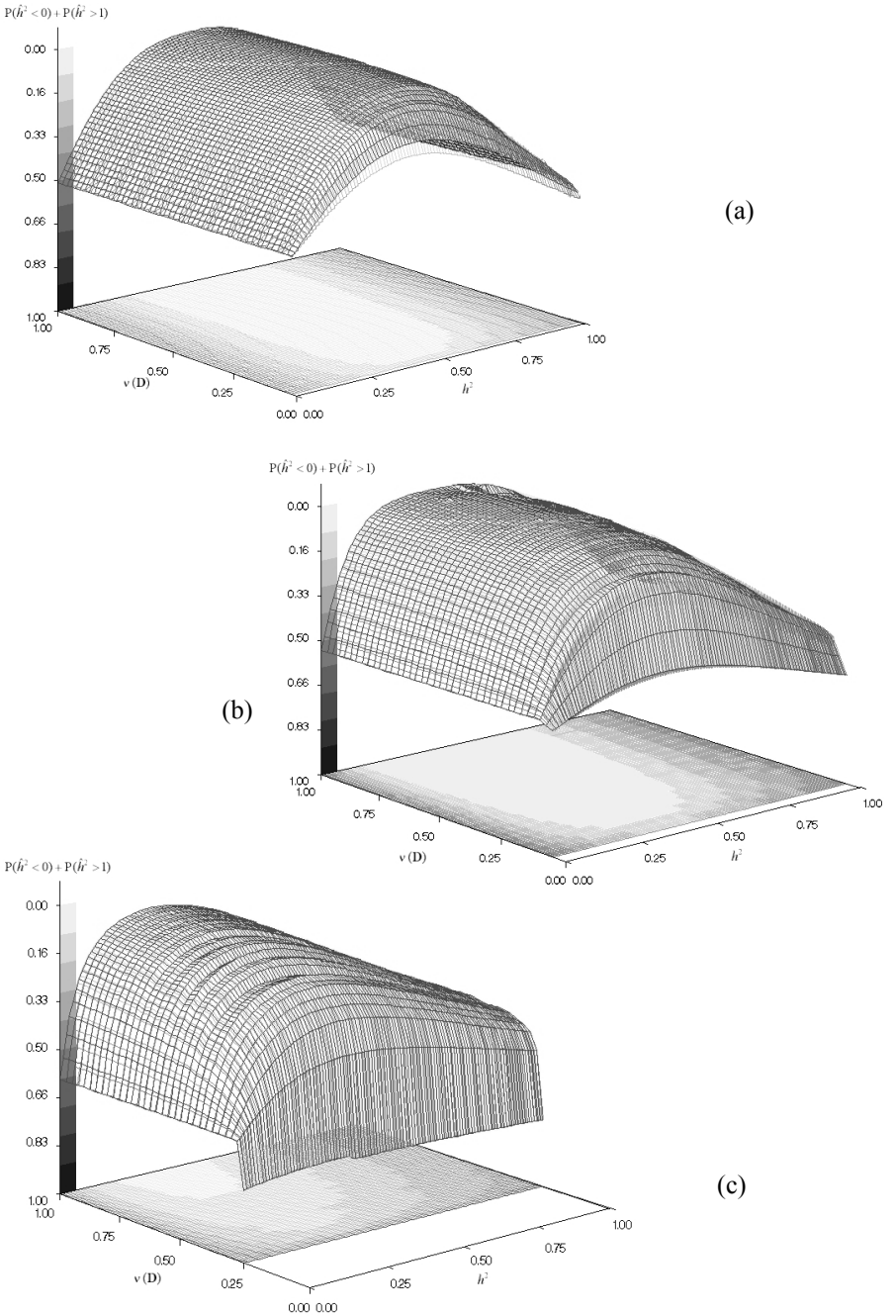


Figure 3.16. The patterns of $P(\hat{h}^2<0)+P(\hat{h}^2>1)$ in different data set imbalances $\nu(\mathbf{D})$ and true heritability coefficient values h^2 for $N=360$, $\sigma_e^2=1$ and numbers of groups/sires (a) $a=90$, (b) $a=20$, (c) $a=4$.

CHAPTER 4
THE ACCURACY OF THE ESTIMATES IN THE GENERAL LINEAR MIXED MODEL

4.1. Introduction

The applications of linear mixed models with general variance-covariance structure in genetic studies are animal models (2.8), (2.11) and models with single genes effects (2.13) containing wide pedigree information given in the structure of variance-covarince matrix. Estimation and prediction theory for these models is overviewed in Chapter 1.2. As the estimation of variance-covariance components from unstructured variance-covariance matrices is natural in the REML method (Chapter 1.2.5), this method is the main technique applied in general linear mixed models.

Due to the potential complexity of genetic studies there are several unreasonable situations for traditional mixed linear models theory, but they are a usual set-up for genetics and especially for animal breeding studies. Many of these are at first glance strange properties, like the models where the number of terms to be estimated (predicted) exceeds the number of observations or the models with non-null covariances between random genetic effects and random errors, are discussed in Henderson (1984) and Searle (1998). In Section 4.2 it is proved that adding into the model individuals without records on observed traits and estimating the additive genetic effects for them do not change the estimators and the accuracy of estimates reviewed in Chapter 1.

In Sections 4.3 and 4.4 the effects of pedigree structure on the accuracy of estimates and the effect of choice of genetic model are discussed based on short modelling experiments and real data study. The results of the first Estonian Black Face Sheep Database analyses, followed also by the model comparison study presented in Section 4.4.2 are published in Kaart and Piirsalu (2000).

4.2. The effect of predicting the non-measured effects

In animal husbandry it is quite frequent to predict the realised values of random genetic effects which have no data. The purpose is to select the best animals for parents of the next generation by their genetic potential, which may not always be expressed in their phenotype. For example, the potential milk production of bulls from animal model (2.8) or motherability of rams by maternal animal model (2.11).

The linear mixed model used in described situations is shortly examined in Henderson (1984, p 42). More detailed discussion is given in Searle (1998, p 74–77).

Let $\mathbf{y}$ be the vector of observed values of the trait and $\mathbf{0}$ be the vector of nulls corresponding to the individuals without records.

Definition 4.1. *We say that*

$$\begin{pmatrix}\mathbf{y}\\ \mathbf{0}\end{pmatrix}=\begin{pmatrix}\mathbf{X}\\ \mathbf{X}_0\end{pmatrix}\boldsymbol{\beta}+\begin{pmatrix}\mathbf{Z} & \mathbf{0}\\ \mathbf{0} & \mathbf{Z}_0\end{pmatrix}\begin{pmatrix}\mathbf{u}\\ \mathbf{u}_0\end{pmatrix}+\begin{pmatrix}\mathbf{e}\\ \mathbf{0}\end{pmatrix} \tag{4.1}$$

defines the linear mixed model for $(\mathbf{y}'\ \mathbf{0}')'$*, where*
$\boldsymbol{\beta}$ *is a vector of fixed effects,*
$\mathbf{u}$ *and* $\mathbf{u}_0$ *are vectors of random effects with and without data, respectively,*
$\mathbf{e}$ *is a vector of random residuals,*
$\mathbf{X}$, $\mathbf{X}_0$, $\mathbf{Z}$ *and* $\mathbf{Z}_0$ *are known design matrices which describe the precise relationship between the elements of* $\boldsymbol{\beta}$ *and* $(\mathbf{u}'\ \mathbf{u}_0')'$ *with those of* $(\mathbf{y}'\ \mathbf{0}')'$*, respectively.* ■

The variance-covariance structure of random effects (excluded the random residual term $\mathbf{e}$) is expressed as a block-matrix:

$$\mathrm{Var}\begin{pmatrix}\mathbf{u}\\ \mathbf{u}_0\\ \mathbf{y}\end{pmatrix}=\begin{pmatrix}\mathbf{G} & \mathbf{C} & \mathbf{G}\mathbf{Z}'\\ \mathbf{C}' & \mathbf{G}_0 & \mathbf{C}'\mathbf{Z}'\\ \mathbf{Z}\mathbf{G} & \mathbf{Z}\mathbf{C} & \mathbf{V}\end{pmatrix}. \tag{4.2}$$

Let

$$\mathbf{X}^{¤}=\begin{pmatrix}\mathbf{X}\\ \mathbf{X}_0\end{pmatrix},\ \mathbf{Z}^{¤}=\begin{pmatrix}\mathbf{Z} & \mathbf{0}\\ \mathbf{0} & \mathbf{Z}_0\end{pmatrix},\ \mathbf{y}^{¤}=\begin{pmatrix}\mathbf{y}\\ \mathbf{0}\end{pmatrix}\ \text{and}\ \mathbf{V}^{¤}=\mathrm{Var}\begin{pmatrix}\mathbf{y}\\ \mathbf{0}\end{pmatrix}=\begin{pmatrix}\mathbf{V} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}.$$

We denote the estimates from model (4.1) with hat $\hat{}$ instead of traditional notation $\hat{}$.

The following theorem summarises the basic estimation and prediction equations for model (4.1).

Theorem 4.1. *In the linear mixed model* (4.1) *estimators and predictors of parameters* $\boldsymbol{\beta}$ *and* $\mathbf{u}$*, and their accuracy are independent on the additional random effects which have no data. The realised values of random effects* $\mathbf{u}_0$ *are predictable from equation*

$$\hat{\mathbf{u}}_0=\mathbf{C}'\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})=\mathbf{C}'\mathbf{G}^{-1}\hat{\mathbf{u}},$$

and the corresponding variances and covariances of predictors and prediction errors are

$$\mathrm{Var}(\hat{\mathbf{u}}_0)=\mathbf{C}'\mathbf{Z}'\mathbf{V}^{-1}\mathbf{W}\mathbf{Z}\mathbf{C},$$

$$\mathrm{Cov}(\hat{\mathbf{u}},\hat{\mathbf{u}}_0')=\mathrm{Cov}(\hat{\mathbf{u}},\hat{\mathbf{u}}_0')=\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{W}\mathbf{Z}\mathbf{C},$$

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}_0')=\mathrm{Cov}(\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}_0')=\mathbf{0},$$

$$\mathrm{Var}(\hat{\mathbf{u}}_0-\mathbf{u}_0)=\mathbf{G}_0-\mathrm{Var}(\hat{\mathbf{u}}_0),$$

$$\mathrm{Cov}(\hat{\mathbf{u}},\hat{\mathbf{u}}_0'-\mathbf{u}_0')=\mathrm{Cov}(\hat{\mathbf{u}},\hat{\mathbf{u}}_0'-\mathbf{u}_0')=\mathbf{0},$$

$$\mathrm{Cov}\left(\tilde{\boldsymbol{\beta}}, \tilde{\mathbf{u}}_0' - \mathbf{u}_0'\right) = \mathrm{Cov}\left(\hat{\boldsymbol{\beta}}, \tilde{\mathbf{u}}_0' - \mathbf{u}_0'\right) = -\mathbf{QX}'\mathbf{V}^{-1}\mathbf{ZC},$$

where $\mathbf{Q} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-$ *and* $\mathbf{W} = \mathbf{I} - \mathbf{XQX}'\mathbf{V}^{-1}$.

Proof. Suppose that the model describing the relationship between data and factors that may have an effect on our data is (4.1) and the variance-covariance matrix between random effects is (4.2).

We now apply the estimation formula (1.7) to the linear mixed model set-up (4.1). The estimator $\tilde{\boldsymbol{\beta}}$ is

$$\begin{aligned}
\tilde{\boldsymbol{\beta}} &= (\mathbf{X}^{¤\prime}(\mathbf{V}^{¤})^{-1}\mathbf{X}^{¤})^{-}\mathbf{X}^{¤\prime}(\mathbf{V}^{¤})^{-1}\mathbf{y}^{¤} \\
&= \left[\begin{pmatrix} \mathbf{X}' & \mathbf{X}_0' \end{pmatrix} \begin{pmatrix} \mathbf{V}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{X} \\ \mathbf{X}_0 \end{pmatrix} \right]^{-} \begin{pmatrix} \mathbf{X}' & \mathbf{X}_0' \end{pmatrix} \begin{pmatrix} \mathbf{V}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} \\
&= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} = \hat{\boldsymbol{\beta}}.
\end{aligned}$$

Likewise, applying to model (4.1) the prediction formula (1.8) gives predictors $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{u}}_0$ as

$$\begin{aligned}
\begin{pmatrix} \tilde{\mathbf{u}} \\ \tilde{\mathbf{u}}_0 \end{pmatrix} &= \mathrm{Cov}\left(\mathbf{u}^{¤}, \mathbf{y}^{¤\prime}\right)\left(\mathbf{V}^{¤}\right)^{-}\left(\mathbf{y}^{¤} - \mathbf{X}^{¤}\tilde{\boldsymbol{\beta}}\right) \\
&= \mathrm{Cov}\left[\begin{pmatrix} \mathbf{u} \\ \mathbf{u}_0 \end{pmatrix}, \begin{pmatrix} \mathbf{y}' & \mathbf{0} \end{pmatrix} \right] \begin{pmatrix} \mathbf{V} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}^{-} \begin{pmatrix} \mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}} \\ \mathbf{0} - \mathbf{X}_0\tilde{\boldsymbol{\beta}} \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{GZ}' & \mathbf{0} \\ \mathbf{C}'\mathbf{Z}' & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{V}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}} \\ \mathbf{0} - \mathbf{X}_0\tilde{\boldsymbol{\beta}} \end{pmatrix} = \begin{pmatrix} \mathbf{GZ}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \\ \mathbf{C}'\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \end{pmatrix}.
\end{aligned}$$

Thus,

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

and

$$\tilde{\mathbf{u}}_0 = \mathbf{C}'\mathbf{Z}'\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}\right) = \mathbf{C}'\mathbf{G}^{-1}\tilde{\mathbf{u}} = \mathbf{C}'\mathbf{G}^{-1}\hat{\mathbf{u}}.$$

To study the variances and covariances of predictors and prediction errors we use the matrices $\mathbf{Q} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-$ and $\mathbf{W} = \mathbf{I} - \mathbf{XQX}'\mathbf{V}^{-1}$. Then

$$\tilde{\boldsymbol{\beta}} = \mathbf{QX}'\mathbf{V}^{-1}\mathbf{y},\ \tilde{\mathbf{u}} = \mathbf{GZ}'\mathbf{V}^{-1}\mathbf{Wy} \text{ and } \tilde{\mathbf{u}}_0 = \mathbf{C}'\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Wy}.$$

From the results on generalized inverses of $\mathbf{X}$ we have that

$$\mathbf{XQX}'\mathbf{V}^{-1}\mathbf{X} = \mathbf{X},$$

and therefore,

$$\mathbf{WX} = (\mathbf{I} - \mathbf{XQX}'\mathbf{V}^{-1})\mathbf{X} = \mathbf{X} - \mathbf{XQX}'\mathbf{V}^{-1}\mathbf{X} = \mathbf{X} - \mathbf{X} = \mathbf{0}. \qquad (4.3)$$

As the estimators and predictors of parameters $\boldsymbol{\beta}$ and $\mathbf{u}$ are independent from $\mathbf{u}_0$, then the equations for predictors variances and prediction errors also do not include the terms related with $\mathbf{u}_0$. The variance of predictor $\tilde{\mathbf{u}}$ is expressed as

$$\begin{aligned}\mathrm{Var}(\tilde{\mathbf{u}}) &= \mathbf{GZ}'\mathbf{V}^{-1}\mathbf{W}\,\mathrm{Var}(\mathbf{y})\mathbf{W}'\mathbf{V}^{-1}\mathbf{ZG} \\ &= \mathbf{GZ}'\mathbf{V}^{-1}\mathbf{WVW}'\mathbf{V}^{-1}\mathbf{ZG} \\ &= \mathbf{GZ}'\mathbf{V}^{-1}\mathbf{ZG} - \mathbf{GZ}'\mathbf{V}^{-1}\mathbf{XQX}'\mathbf{V}^{-1}\mathbf{ZG}.\end{aligned}$$

The covariance between $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mathbf{u}}$ is zero due to the equality (4.3):

$$\begin{aligned}\mathrm{Cov}\left(\tilde{\boldsymbol{\beta}}, \tilde{\mathbf{u}}'\right) &= \mathbf{QX}'\mathbf{V}^{-1}\,\mathrm{Var}(\mathbf{y})\mathbf{W}'\mathbf{V}^{-1}\mathbf{ZG} \\ &= \mathbf{QX}'\mathbf{W}'\mathbf{V}^{-1}\mathbf{ZG} = \mathbf{0}.\end{aligned}$$

The variances and covariances of prediction errors for parameters $\boldsymbol{\beta}$ and $\mathbf{u}$ are

$$\begin{aligned}\mathrm{Var}\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) &= \mathrm{Var}\left(\tilde{\boldsymbol{\beta}}\right) + \mathrm{Var}\left(\boldsymbol{\beta}\right) - \mathrm{Cov}\left(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}\right) - \mathrm{Cov}\left(\boldsymbol{\beta}, \tilde{\boldsymbol{\beta}}\right) \\ &= \mathrm{Var}\left(\tilde{\boldsymbol{\beta}}\right) = \mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right) = \mathbf{Q},\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(\tilde{\mathbf{u}} - \mathbf{u}) &= \mathrm{Var}(\tilde{\mathbf{u}}) + \mathrm{Var}(\mathbf{u}) - \mathrm{Cov}(\tilde{\mathbf{u}}, \mathbf{u}) - \mathrm{Cov}(\mathbf{u}, \tilde{\mathbf{u}}) \\ &= \mathbf{G} - \mathrm{Var}(\tilde{\mathbf{u}}) = \mathbf{G} - \mathrm{Var}(\hat{\mathbf{u}}), \text{ because } \mathrm{Cov}(\tilde{\mathbf{u}}, \mathbf{u}) = \mathrm{Var}(\tilde{\mathbf{u}}),\end{aligned}$$

$$\begin{aligned}\mathrm{Cov}\left(\tilde{\boldsymbol{\beta}}, \tilde{\mathbf{u}}' - \mathbf{u}'\right) &= \mathrm{Cov}\left(\tilde{\boldsymbol{\beta}}, \tilde{\mathbf{u}}\right) - \mathrm{Cov}\left(\tilde{\boldsymbol{\beta}}, \mathbf{u}\right) \\ &= \mathrm{Cov}\left(\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}}\right) - \mathrm{Cov}\left(\hat{\boldsymbol{\beta}}, \mathbf{u}\right) = \mathbf{0} - \mathbf{QX}'\mathbf{V}^{-1}\mathbf{ZG}.\end{aligned}$$

These results are the same as reported in Chapter 1.2.3.

Similarly, the variance of predictor $\tilde{\mathbf{u}}_0$ is expressed as

$$\begin{aligned}\mathrm{Var}(\tilde{\mathbf{u}}_0) &= \mathbf{C}'\mathbf{Z}'\mathbf{V}^{-1}\mathbf{W}\,\mathrm{Var}(\mathbf{y})\mathbf{W}'\mathbf{V}^{-1}\mathbf{ZC} \\ &= \mathbf{C}'\mathbf{Z}'\mathbf{V}^{-1}\mathbf{ZC} - \mathbf{C}'\mathbf{Z}'\mathbf{V}^{-1}\mathbf{XQX}'\mathbf{V}^{-1}\mathbf{ZC}.\end{aligned}$$

The covariance between $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mathbf{u}}_0$ is

$$\mathrm{Cov}\left(\tilde{\boldsymbol{\beta}}, \tilde{\mathbf{u}}_0'\right) = \mathbf{QX}'\mathbf{V}^{-1}\,\mathrm{Var}(\mathbf{y})\mathbf{W}'\mathbf{V}^{-1}\mathbf{ZC} = \mathbf{0}$$

and the covariance between $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{u}}_0$ is

$$\begin{aligned}\mathrm{Cov}(\tilde{\mathbf{u}}, \tilde{\mathbf{u}}_0') &= \mathbf{GZ}'\mathbf{V}^{-1}\mathbf{W}\,\mathrm{Var}(\mathbf{y})\mathbf{W}'\mathbf{V}^{-1}\mathbf{ZC} \\ &= \mathbf{GZ}'\mathbf{V}^{-1}\mathbf{ZC} - \mathbf{GZ}'\mathbf{V}^{-1}\mathbf{XQX}'\mathbf{V}^{-1}\mathbf{ZC}.\end{aligned}$$

The variance of prediction error is

$$\mathrm{Var}(\tilde{\mathbf{u}}_0 - \mathbf{u}_0) = \mathbf{G}_0 - \mathrm{Var}(\tilde{\mathbf{u}}_0)$$

and the suitable covariances are

$$\begin{aligned}\mathrm{Cov}(\tilde{\mathbf{u}}, \tilde{\mathbf{u}}_0' - \mathbf{u}_0') &= \mathrm{Cov}(\tilde{\mathbf{u}}, \tilde{\mathbf{u}}_0') - \mathrm{Cov}(\tilde{\mathbf{u}}, \mathbf{u}_0') \\ &= \mathbf{GZ}'\mathbf{V}^{-1}\mathbf{W}\,\mathrm{Var}(\mathbf{y})\mathbf{W}'\mathbf{V}^{-1}\mathbf{ZC} - \mathbf{GZ}'\mathbf{V}^{-1}\mathbf{W}\,\mathrm{Cov}(\mathbf{y}, \mathbf{u}_0') \\ &= \mathbf{GZ}'\mathbf{V}^{-1}\mathbf{ZC} - \mathbf{GZ}'\mathbf{V}^{-1}\mathbf{XQX}'\mathbf{V}^{-1}\mathbf{ZC} \\ &\quad -(\mathbf{GZ}'\mathbf{V}^{-1}\mathbf{ZC} - \mathbf{GZ}'\mathbf{V}^{-1}\mathbf{XQX}'\mathbf{V}^{-1}\mathbf{ZC}) \\ &= \mathbf{0},\end{aligned}$$

$$\mathrm{Cov}\left(\tilde{\boldsymbol{\beta}}, \tilde{\mathbf{u}}_0' - \mathbf{u}_0'\right) = \mathrm{Cov}\left(\tilde{\boldsymbol{\beta}}, \tilde{\mathbf{u}}_0'\right) - \mathrm{Cov}\left(\tilde{\boldsymbol{\beta}}, \mathbf{u}_0'\right) = -\mathbf{QX}'\mathbf{V}^{-1}\mathbf{ZC}.$$

■

The following theorem proves that the REML equations used in estimating variance components and the information matrix (and so the approximate variance of estimates) do not depend on the additional random effects which have no data.

Theorem 4.2 *The REML equations used in estimating variance components in the mixed linear model and the approximate variances of estimates do not depend on the additional random effects which have no data.*

Proof. At first we note that the matrix $\mathbf{P}$ (1.11) is for the mixed model set-up (4.1) expressed as

$$\begin{aligned}\mathbf{P}^{\text{¤}} &= \begin{pmatrix}\mathbf{V}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix} - \begin{pmatrix}\mathbf{V}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{X}\\ \mathbf{X}_0\end{pmatrix}\left[(\mathbf{X}' \;\; \mathbf{X}_0')\begin{pmatrix}\mathbf{V}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{X}\\ \mathbf{X}_0\end{pmatrix}\right]^{-}(\mathbf{X}' \;\; \mathbf{X}_0')\begin{pmatrix}\mathbf{V}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\\ &= \begin{pmatrix}\mathbf{V}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix} - \begin{pmatrix}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\\ &= \begin{pmatrix}\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix} = \begin{pmatrix}\mathbf{P} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}.\end{aligned}$$

The left-hand side of the REML equations (1.23) is rewritable as

$$\begin{aligned}\{_{\mathrm{c}}\,\mathrm{tr}(\mathbf{P}^{\text{¤}}\mathbf{Z}_i^{\text{¤}}\mathbf{Z}_i^{\text{¤}\prime})\}_{i=0}^{q} &= \left\{ \mathrm{tr}\left[\begin{pmatrix}\mathbf{P} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{Z}_i & \mathbf{0}\\ \mathbf{0} & \mathbf{Z}_{i0}\end{pmatrix}\begin{pmatrix}\mathbf{Z}_i' & \mathbf{0}\\ \mathbf{0} & \mathbf{Z}_{i0}'\end{pmatrix}\right]\right\}_{i=0}^{q} = \left\{ \mathrm{tr}\begin{pmatrix}\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i' & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\right\}_{i=0}^{q}\\ &= \{_{\mathrm{c}}\,\mathrm{tr}(\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i')\}_{i=0}^{q}\end{aligned}$$

and the right-hand side of these equations for our mixed model set-up is

$$\begin{aligned}\{_{\mathrm{c}}\,\mathbf{y}^{\text{¤}\prime}\mathbf{P}^{\text{¤}}\mathbf{Z}_i^{\text{¤}}\mathbf{Z}_i^{\text{¤}\prime}\mathbf{P}^{\text{¤}}\mathbf{y}^{\text{¤}}\}_{i=0}^{q} &= \left\{ (\mathbf{y}' \;\; \mathbf{0})\begin{pmatrix}\mathbf{P} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{Z}_i & \mathbf{0}\\ \mathbf{0} & \mathbf{Z}_{i0}\end{pmatrix}\begin{pmatrix}\mathbf{Z}_i' & \mathbf{0}\\ \mathbf{0} & \mathbf{Z}_{i0}'\end{pmatrix}\begin{pmatrix}\mathbf{P} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{y}\\ \mathbf{0}\end{pmatrix}\right\}_{i=0}^{q}\\ &= \{_{\mathrm{c}}\,\mathbf{y}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{y}\}_{i=0}^{q}.\end{aligned}$$

Applying the approximated formula for $\mathrm{Var}(\hat{\sigma}^2_{\mathrm{REML}})$ in (1.25) to our set-up gives the variances and covariances of variance components estimates as

$$\begin{aligned}\mathrm{Var}(\hat{\sigma}^2_{\mathrm{REML}}) &\approx 2\left[\left\{ {}_{\mathrm{m}}\,\mathrm{tr}(\mathbf{P}^{\text{¤}}\mathbf{Z}_i^{\text{¤}}\mathbf{Z}_i^{\text{¤}\prime}\mathbf{P}^{\text{¤}}\mathbf{Z}_j^{\text{¤}}\mathbf{Z}_j^{\text{¤}\prime})\right\}_{i,j=0}^{q}\right]^{-1}\\ &= 2\left[\left\{ {}_{\mathrm{m}}\,\mathrm{tr}\left[\begin{pmatrix}\mathbf{P} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{Z}_i & \mathbf{0}\\ \mathbf{0} & \mathbf{Z}_{i_n}\end{pmatrix}\begin{pmatrix}\mathbf{Z}_i' & \mathbf{0}\\ \mathbf{0} & \mathbf{Z}_{i_n}'\end{pmatrix}\begin{pmatrix}\mathbf{P} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{Z}_j & \mathbf{0}\\ \mathbf{0} & \mathbf{Z}_{j_n}\end{pmatrix}\begin{pmatrix}\mathbf{Z}_j' & \mathbf{0}\\ \mathbf{0} & \mathbf{Z}_{j_n}'\end{pmatrix}\right]\right\}_{i,j=0}^{q}\right]^{-1}\\ &= 2\left[\left\{ {}_{\mathrm{m}}\,\mathrm{tr}\begin{pmatrix}\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j' & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\right\}_{i,j=0}^{q}\right]^{-1}\\ &= 2\left[\left\{ {}_{\mathrm{m}}\,\mathrm{tr}(\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j')\right\}_{i,j=0}^{q}\right]^{-1} \approx \mathrm{Var}(\hat{\sigma}^2_{\mathrm{REML}}).\end{aligned}$$

■

Proposition 4.1. *The model* (4.1) *is equivalent to the model*

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \begin{pmatrix} \mathbf{Z} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{u}_0 \end{pmatrix} + \mathbf{e} \,. \qquad \blacksquare$$

4.3. The accuracy of the genetic parameters estimates depending on genetic relationships

It is obvious that without genetic relationships between studied individuals we can not to estimate any genetic parameters based on measured phenotypic values. At the same time it is unclear how complicated relationships we need to get as accurate as possible estimates of genetic parameters. This question should arise at first in the phase of experiment planning. Should we cross or choose for further study possibly deeply related individuals or, in the opposite, completely unrelated individuals? In which case would we get the most accurate estimates to the parameters describing the genetic determination of observed traits?

In a mathematical sense the relationships between individuals aggregate in the covariance structure of the model. The effect of variance-covariance structure on the estimation and prediction results is not trivial (for a short discussion see, for example, Möls, 2004, p 51). In genetic studies, where a small change in genetic relationships should lead to a complicated change in variance-covariance structure, these associations are even harder to fix. In the following, a short study still waiting for further and more mathematical approach is presented.

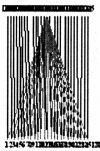
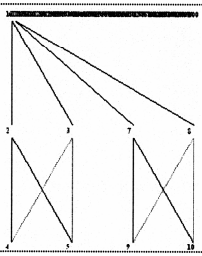
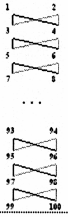
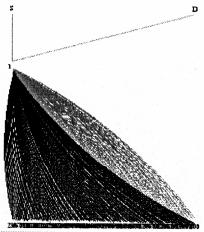
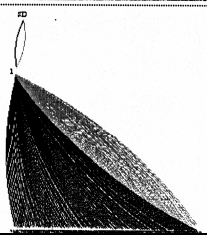
The form of the general linear mixed model (1.1) corresponding to the animal model (2.8) was assumed. For simplicity, the only fixed effect included in the model was the overall mean μ, the variance-covariance structure of random effects in the form (2.10) was assumed, the error variance $\sigma_e^2 = 1$ was used and the calculations were implemented for heritability coefficient h^2 values 0.01, 0.025 and 0.6. Corresponding values of variance component σ_a^2 are calculable by expression $\sigma_a^2 = h^2\sigma_e^2/(1-h^2)$ throughout.

Six different crossing schemes were worked out and the corresponding variance-covariance matrices were constructed (Table 4.1). The number of measured individuals was taken equal to 100.

The studied parameters were the mean square error of predictors, MSE($\hat{\mathbf{a}}$), and the asymptotic sampling variance of REML-estimator of variance component σ_a^2, Var($\hat{\sigma}^2_{a,REML}$), calculated by formulas (1.14) and (1.25), respectively.

From the results presented in Table 4.2 it is obvious that the simplest variance-covariance structure is not the best, but also not the worst. Also, the values of MSE($\hat{\mathbf{a}}$) and Var($\hat{\sigma}^2_{a,REML}$) depend in addition to variance-covariance structure on the real values of variance components or their ratios (some models changed their rankings if the values of h^2 varied) and on the number of studied individuals.

Tabel 4.1. The studied crossing or relationship schemes of 100 individuals

Model	Description	Scheme / $\mathrm{Var}(\mathbf{a}) = \sigma_a^2\mathbf{A}$ (for symmetric matrices only upper triangular block is shown)
M1	All 100 observed individuals unrelated.	$\mathrm{Var}(\mathbf{a}) = \sigma_a^2\mathbf{I}_{100}$
M2	25 individuals are half-sibs, 75 are unrelated. Parents of half-sibs are not under study.	 $\mathrm{Var}(\mathbf{a}) = \sigma_a^2\begin{pmatrix}\mathbf{G}_{hs} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_{75}\end{pmatrix}$, where $\mathbf{G}_{hs} = \begin{pmatrix}1 & 0.25 & \cdots & 0.25\\ & 1 & \cdots & 0.25\\ & & \ddots & \vdots\\ & & & 1\end{pmatrix}$
M3	25 individuals are full-sibs, 75 are unrelated. Parents of full-sibs are not under study.	 $\mathrm{Var}(\mathbf{a}) = \sigma_a^2\begin{pmatrix}\mathbf{G}_{fs} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_{75}\end{pmatrix}$, where $\mathbf{G}_{fs} = \begin{pmatrix}1 & 0.5 & \cdots & 0.5\\ & 1 & \cdots & 0.5\\ & & \ddots & \vdots\\ & & & 1\end{pmatrix}$
M4	2 families, where 2 half-sibs are crossed to produce 2 offspring (thus in both families are 5 related individuals – 1 grandparent, 2 parents and 2 offsprings); the other 90 individuals are unrelated	 $\mathrm{Var}(\mathbf{a}) = \sigma_a^2\begin{pmatrix}\mathbf{G}_f & \mathbf{0} & \mathbf{0}\\ & \mathbf{G}_f & \mathbf{0}\\ & & \mathbf{I}_{90}\end{pmatrix}$, where $\mathbf{G}_f = \begin{pmatrix}1 & 0.5 & 0.5 & 0.5 & 0.5\\ & 1 & 0.25 & 0.625 & 0.625\\ & & 1 & 0.625 & 0.625\\ & & & 1.125 & 0.625\\ & & & & 1.125\end{pmatrix}$
M5	2 unrelated individuals crossed to produce 2 offspring who are crossed to produce 2 offspring etc; 100th individual has inbreeding coefficient equal to 0.99996.	 $\mathrm{Var}(\mathbf{a}) = \sigma_a^2\begin{pmatrix}1 & 0 & 0.5 & 0.5 & 0.5 & \cdots & 0.5\\ & 1 & 0.5 & 0.5 & 0.5 & \cdots & 0.5\\ & & 1 & 0.5 & 0.75 & \cdots & 0.75\\ & & & 1 & 0.75 & \cdots & 0.75\\ & & & & 1.25 & \cdots & 1\\ & & & & & \ddots & \vdots\\ & & & & & & 1.99996\end{pmatrix}$
M6	1 offspring of unrelated individuals and her/his 99 identical clones (thus, the additive genetic relationship between all individuals equals to 1 but there is no inbreeding). Parents of the first individual are not under study, they are used just to have the correct genetic relationships.	 $\mathrm{Var}(\mathbf{a}) = \sigma_a^2\mathbf{J}_{100}$
M7	1 offspring of genetically identical (pure line) individuals and her/his 99 identical clones (thus the additive genetic relationship between individuals and the inbreeding both equal to 1). Parents of the first individual are not under study, they are used just to have the correct genetic relationships.	 $\mathrm{Var}(\mathbf{a}) = 2\sigma_a^2\mathbf{J}_{100}$

The mean square error of predictors has its smallest value in the case of Model 3, where there exist some non-null offdiagonal elements in the variance-covariance matrix (25% of observed individuals are full-sibs), but all diagonal elements were equal to σ_a^2. Weaker relationships between studied individuals (for example Model 2, where these 25% of individuals have been assumed to be half-sibs) will increase the values of MSE(**â**).

If inbreed individuals exist in the pedigree, then some diagonal elements are bigger than σ_a^2 and the values of MSE(**â**) increase. Also, the mean square error of predictors increases if the number of related individuals increased (in variance-covariance matrix there are more non-null offdiagonal elements).

The asymptotic sampling variance of REML-estimator of variance component σ_a^2 has its smallest values if the studied individuals are deeply related and the accuracy is the poorest in case of unrelated individuals.

Table 4.2. The values of MSE(**â**) and Var($\hat{\sigma}^2_{a,REML}$) in case of different variance-covariance structures and heritability coefficient values, $N = 100$, $\sigma_e^2 = 1$.

Model	MSE(**â**)			Var($\hat{\sigma}^2_{a,REML}$) x10		
	$h^2 = 0.8$	$h^2 = 0.25$	$h^2 = 0.01$	$h^2 = 0.8$	$h^2 = 0.25$	$h^2 = 0.01$
M1	83.20	25.08	1.000	5.0505	0.3591	0.2061
M2	83.46	24.52	0.998	4.4824	0.3502	0.2060
M3	81.64	23.34	0.990	3.5541	0.3372	0.2059
M4	82.51	24.78	1.004	4.2435	0.3542	0.2061
M5	249.12	34.68	1.814	0.2534	0.2131	0.2027
M6	400.00	33.33	1.010	0.2020	0.2020	0.2020
M7	801.35	66.17	2.010	0.2038	0.1995	0.2020

4.4. The dependency of the genetic parameters estimates on the genetic model choice

4.4.1. Discussion

The model choice problem is incidental to all studies where the probabilistic (usually noted as statistical) models are used. The true model, which describes the data perfectly and is never known exactly, is tried to approximate with the ideal model, which is set up based on the researcher's knowledge and experiences. In real data analysis due to the lack of data and/or computing facilities usually the simplified version of the ideal model is used.

In large population based genetic studies with a large amount of individuals the fixed environmental factors traditionally contain information about the sex, age, birth and abiding place. The form of genetic parameters estimators, hence the correctness of estimates depends first of all on the random part of the model.

The latter is defined with the genetic model, the choice of which depends on the researcher's ability to understand the genetic background of the studied trait (to set up a correct genetic model), and the possibilities to collect and analyse data based on the proposed model.

In pedigree based studies also the question about the use and exert of all available relationships between individuals can arise. Based on population genetic studies (see, for example, Kennedy and Sorensen, 1988) it is clear that the use of relationships between individuals allows to avoid the bias in estimates caused by the deviances in population from the Hardy-Weinberg law (non-random mating, migration, etc). But how many generations back we should involve in the study is unclear.

There are a few articles available using the simulations to study the effect of pedigree structures on the accuracy of estimates. Reverter and Kaiser (1996) concluded that the h^2 estimates with the smallest standard error are obtained when there is performance information on many animals closely related to foundation animals, it is better to have data about tightly related individuals from a few generations than the long term pedigree schemes. Mehrabani-Yeganeh, Gibson and Schaeffer (1999) used stochastic simulation to study the effect of using full data and pedigree structure versus more recent data and pedigree structure to obtain best linear unbiased predictors of breeding values. They concluded that using the data only from one generation would complicate the ranking of individuals by their genetic potential, especially in case of small heritabilities; but there is no big difference in selecting animals based on data from last 2 or 10 generations.

The effect of using a more or less complicated genetic model besides the statistical model depends on the real nature of the studied trait. It is logical to presume that more maternally influenced traits can only be poorly analysed with sire based models. On the other hand, if the mothers do not add something extra to the trait realized values, then the use of complicated maternal effects animal models is not necessary. Also, if in the pedigree there are many generations without data, then the use all of them is not necessary and the simpler model can be used.

Since the pedigree structure and also the real values of genetic parameters differ in different populations, then the studies dealing with the choice of genetic models should also be population based. There have been published some real data studies illustrating the effect of using different genetic models in estimating genetic parameters and in ranking animals for selection (Hagger and Schneeberger, 1995; Ferreira, MacNeil and Van Vleck, 1999; Rumph, et al, 2002).

In the following the effects of using different genetic models and different amount of pedigree information is studied in the Estonian Black Face Sheep Database example.

4.4.2. Example: lambs weaning weights analysis

The weaning weights of Estonian Black Face Sheep lambs, born in 2001–2003, and the required pedigree information were extracted from the Estonian Sheep Database. The main dataset contains observations of 1073 animals. In pedigree data, where all available ancestors were included, 2451 animals were listed (the maximum number of ancestor's generations was 12 and the number of inbreed animals was 636). As an alternative, only parents and grand-parents of animals with weaning weight values were listed producing the pedigree of 1886 animals. The pedigree structures of all available ancestors and only 2 parental generations are represented on Figure 4.1 and Figure 4.2, respectively, using the PedigreeViewer program (Kinghorn, 1999).

To study the effect of genetic model choice and the amount of pedigree information on the estimates of genetic parameters, 6 genetic models were compared, 4 of them were applied with two different amount pedigree information. Model 1 was the typical sire model discussed in Section 2.1 with only one random effect caused by sire and without supplementary relationships. Model 2 was the full-sib model with two independent random effects caused by sire and dam, respectively. Models 3, 4, 5 and 6 were the animal model, the animal model with permanent environment effects (measured as the effect of litter and defined with the dam number), the maternal effect animal model and the maternal effect animal model with permanent environment effects, respectively. In models 3, 4, 5 and 6 the reduced pedigree information was used. Models 7, 8, 9 and 10 were the same as models 3, 4, 5 and 6, respectively, but only information on all ancestors was used.

The fixed effects included in the model were weaning age, dam age, sex, weaning type and year*farm interaction. All these effects remained the same for all studied models (for additional reading about the influence of fixed effects in Estonian sheep studies see, for example, Piirsalu and Kaart, 2001).

The analyses were performed with VCE-5 software and the variance-covariance components were estimated with AIREML algorithm (Kovač and Groeneveld, 2003). The values of estimated ratios of variance components interpreted as different genetic parameters are presented in Table 4.3.

The results show, as would be predicted, that the error variance is decreasing if the complicacy of model is increasing.

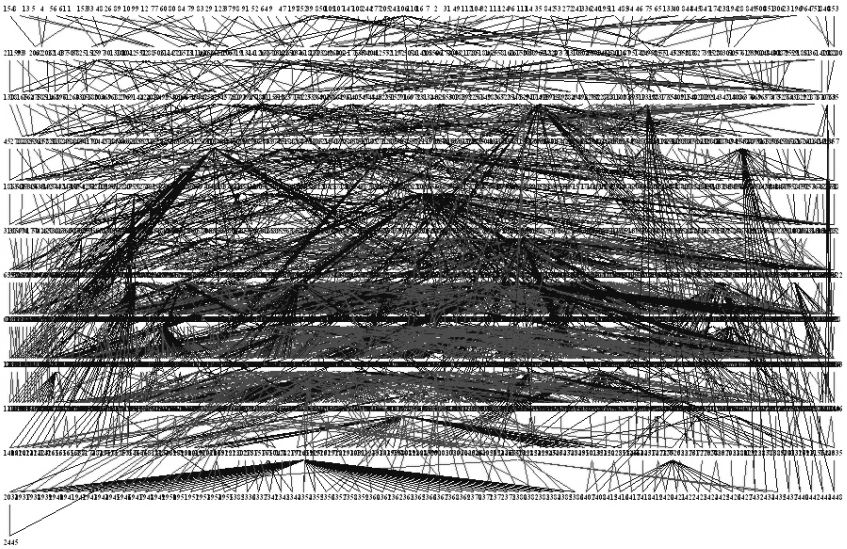


Figure 4.1. The pedigree structure of 2451 Estonian Black Face Sheep (all available ancestors includes, the number of inbreed animals was 636)

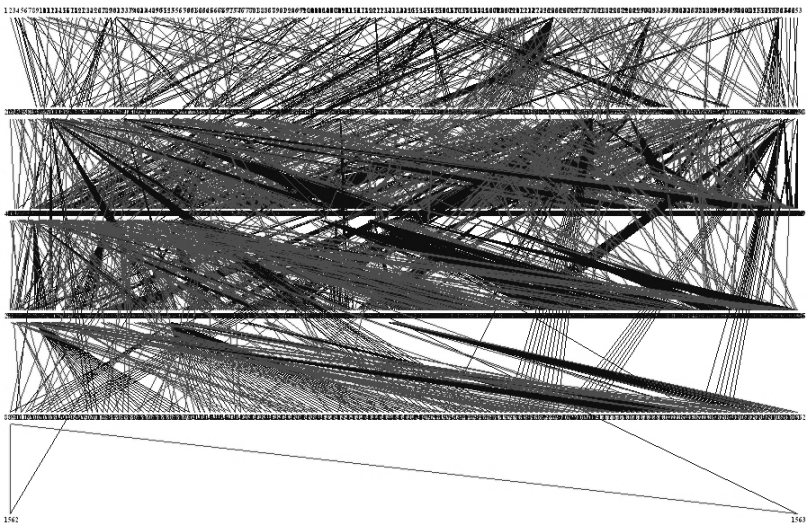


Figure 4.2. The pedigree structure of 1886 Estonian Black Face Sheep (only parents and grand-parents of animals with weaning weight values included, the number of inbreed animals was 116).

The estimates of the primary genetic parameter – the heritability coefficient – depend quite a bit on the presence of other random effects in the model.

Apparently the half- and full-sib model both overestimate the heritability coefficient, placing too much weight on the simple sire or dam effect and do not consider the non-additive genetic influences associated with these effects. Thus it seems that these simple genetic models are not suited for large population based studies and should be used only in well designed pilot studies to get a first glance on the genetic determination of an examined trait.

Comparing the pedigree based models it is obvious that including non-direct genetic or environmental effects will decrease the values of heritability coefficient as the measure of the direct additive genetic influence. It is also logical that including maternal genetic effects and permanent environment effects together will decrease both of them, as they measure quite similar effects.

The effect of using all available pedigree information compared to using only 2 parental generations is not considerable. The only bigger difference appears in the estimates of additive genetic correlation between direct and maternal effects – the mentioned parameter has much bigger estimates based on only 2 parental generations. The reason may be that in the last generations the selection of animals is implemented more correctly and in such a way that only animals with both high additive and maternal genetic effects were selected to produce offspring. But perhaps the number of studied animals was just a little too small to correctly estimate such complex genetic parameters.

Generally it seems that there is no need to use all the available pedigree information, 2 or some more generations are enough to get relatively correct estimates. Also, the right understanding of the genetic background of a studied trait and the right choice of genetic and statistical models are necessary to get proper results.

Table 4.3. The estimated genetic parameters of Estonian Black Face Sheep lambs weaning weights got with different models.

	Pedigree information	Estimated genetic parameters				
		σ_e^2	h^2	m^2	r_{am}	c^2
Model 1	sires	48.674	0.256	–	–	–
Model 2	sires+dams	44.818	0.274	–	–	–
Model 3	≤2	43.030	0.162	–	–	–
Model 4	generations	41.669	0.118	–	–	0.065
Model 5	of ancestors	41.736	0.139	0.043	0.446	–
Model 6		41.303	0.127	0.012	0.922	0.048
Model 7	≤12	42.995	0.165	–	–	–
Model 8	generations	41.619	0.129	–	–	0.066
Model 9	of ancestors	41.966	0.125	0.055	0.087	–
Model 10		41.661	0.118	0.024	0.260	0.041

BIBLIOGRAPHY

1. Ahrens, H., Pincus, R. (1981). On two measures of unbalancedness in a one-way model and their relation to efficiency. *Biometrical Journal*, **23**, 227–237.
2. Almasy, L., Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *American Journal of Human Genetics*, **62**, 1198–1211.
3. Anderson, R. L., Crump, P. P. (1967). Comparison of designs and estimation procedures for estimating parameters in a two-stage nested process. *Technometrics*, **9**, 499–516.
4. Broman, K. W. (2001). Review of statistical methods for QTL mapping in experimental crosses. *Lab Animal*, **30**, 44–52.
5. Das, K., Jiang, J., Rao, J. N. K. (2004). Mean squared error of empirical predictor. *The Annals of Statistics*, **32**, 818–840.
6. Donner, A., Koval, J., J. (1983). A note on the accuracy of Fisher's approximation to the large sample variance of an intraclass correlation. *Communications in Statistics: Computation and Simulation*, 12, 443–449.
7. Elston, R., C., Buxbaum, S., Jacobs, K., B., Olson, J., M. (2000). Haseman and Elston revisited. *Genetic Epidemiology*, **19**, 1–17.
8. Falconer, D. S., Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics. Fourth Edition*. Longman, Harlow, UK.
9. Fernando, R. L., Grossman, M. (1989). Marker assisted selection using best linear unbiased prediction. *Génétique, Sélection, Evolution*, **21**, 467–477.
10. Ferreira, G. B., MacNeil, M. D., Van Vleck, L. D. (1999). Variance components and breeding values for growth traits from different statistical models. *Journal of Animal Science*, **77**, 2641–2650.
11. Gianola, D., Hammond, K. (Eds.) (1990). *Advances in Statistical Methods for Genetic Improvement of Livestock*. Springer, Berlin, Heidelberg, New York.
12. Gill, J. L., Jensen, E. L. (1968). Probability of obtaining negative estimates of heritability. *Biometrics*, **24**, 517–526.
13. Graser, H.-U., Smith, S. P., Tier, B. (1987). A derivative-free approach for estimating variance components in animal models by restricted maximum likelihood. *Journal of Animal Science*, **64**, 1362–1370.
14. Graybill, F. A. (1983). *Matrices with Applications in Statistics, Second Edition*. Wadsworth, Belmont, California.
15. Hagger, C., Schneeberger, M. (1995). Influences of amount of pedigree information on computing time and of model assumptions on restricted maximum-likelihood estimates of population parameters in Swiss black-brown mountain sheep. *Journal of Animal Science*, **73**, 2213–2219.
16. Hammarsley, J. M. (1948). The unbiased estimate and standard error of the interclass variance. *Metron*, **15**, 189–205.
17. Hartley, H. O., Rao, J. N. K. (1967). Maximum_likelihood estimation for mixed analysis of variance model. *Biometrika*, **54**, 93–105.
18. Harville, D. A. (1997). *Matrix algebra from a statistician's perspective*. Springer-Verlag, Berlin, Heidelberg, New York.

19. Haseman, J. K., Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics*, **2**, 3–19.
20. Henderson, C. R. (1950). Estimation of genetic parameters. *Annals of Mathematical Statistics*, **21**, 309.
21. Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics*, **9**, 226–310.
22. Henderson, C. R. (1975). Best linear unbiased estimators and prediction under a selection model. *Biometrics*, **31**, 423–447.
23. Henderson, C. R. (1976). A simple method for computing the inverse of a numerator relationship matrix used in predicting of breeding values. *Biometrics*, **32**, 69–83.
24. Henderson, C. R. (1984). *Applications of Linear Models in Animal Breeding*. University of Guelph Press, Guelph, Canada.
25. Hill, W. G., Thompson, R. (1978). Probabilities of Non-Positive Definite Between-Group or Genetic Covariance Matrices. *Biometrics*, **34**, 429–439.
26. Hofer, A. (1998). Variance component estimation in animal breeding: a review. *Journal of Animal Breeding and Genetics*, **115**, 247–266.
27. Johnson, D. L., Thompson, R. (1995). Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. *Journal of Dairy Science*, **78**, 449–456.
28. Johnson, N. L., Kotz, S. (1970). *Continuous Univariate Distributions—2*. Wiley, New York.
29. Kaart, T. (1997). Probability of the estimate of heritability being negative or greater than one. In: *The 3rd Baltic Animal Breeding Conference*, Proceedings, Riia, Latvia, 1997, 57–59.
30. Kaart, T. (1998). Estimation of variance components and heritability coefficient with four different methods in case of different sire-daughter design. In: *The 4th Baltic Animal Breeding Conference*, Proceedings, Tartu, Estonia, 1998, 28–32.
31. Kaart, T. (2001). Ülevaade geneetiliste parameetrite hindamisel kasutatavatest mudelitest. *Eesti Põllumajandusülikooli Loomakasvatusinstituudi teadustöid 71*, EPMÜ Loomakasvatusinstituut, Tartu, 52–67.
32. Kaart, T. (2004). About the data designs for estimation of genetic parameters in animal breeding studies. *Acta et Commentationes Universitatis Tartuensis de Mathematica*, **8**, 113–121.
33. Kaart. T. (2005). A new approximation for the variance of the ANOVA estimate of the intraclass correlation coefficient. *Proceedings of the Estonian Academy of Sciences. Physics. Mathematics*, **54**, 243–254.
34. Kaart, T., Piirsalu, P. (2000). The complex analysis of genetic parameters in Estonian sheep breeds. In: *The 6th Baltic Animal Breeding Conference*, Proceedings, Jelgava, Latvia, 2000, 135–140.
35. Kackar, R. N., Harville, D. A. (1981). Unbiasedness of two-stage estimation and prediction procedures for mixed linear models. *Communications in Statistics: Theory and Methods*, **10**, 1249–1261.
36. Kackar, R. N., Harville, D. A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, **79**, 853–862.

37. Kakwani, N. C. (1967). The unbiasedness of Zellner's seemingly unrelated regression equations estimators. *Journal of the American Statistical Association*, **62**, 141–142.
38. Kennedy, B. W., Sorensen, D. A. (1988). Properties of mixed model methods for prediction of genetic merit. *Proceedings of the second International Conference on Quantitative genetics*. Sinauer Associates Publ. Mass., 91–104.
39. Kenward, M., G., Roger, J., H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, **53**, 983–997.
40. Khuri, A. I., Mathew, T., Sinha, B. K. (1998). *Statistical Tests for Mixed Linear Models*. Wiley, New York.
41. Kinghorn, B., Kinghorn, S. (2005). *Pedigree Viewer. Version 5.3*. http://metz.une.edu.au/~bkinghor/pedigree.htm (1.11.2005)
42. Kovač, M., Groeneveld, E. (2003). *VCE-5. User's Guide and Reference Manual*. ftp://ftp.zgr.fal.de/ (1.11.2005).
43. LaMotte, L. R. (1970). *A class of estimators of variance components*. Technical Report 10, Dept. of Statistics, University of Kentucky, Lexington.
44. Lange, K. (1997). *Mathematical and statistical methods for genetic analysis*. Springer, New York.
45. Lynch, M., Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Inc.
46. McCulloch, C. E., Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. Wiley, New York.
47. Mehrabani-Yeganeh, H., Gibson, J. P., Schaeffer, L. R. (1999). Using recent versus complete pedigree data in genetic evaluation of a closed nucleus broiler line. *Poultry Science*, **78**, 937–941.
48. Meuwissen, T. H. E., Luo, Z. (1992). Computing inbreeding coefficients in large populations. *Génétique, Sélection, Evolution*, **24**, 305–313.
49. Meuwissen, T., Goddard, M. (2000). Fine-mapping of quantitative trait loci using kinkage disequilibrium with closely linked markers. *Genetics*, **155**, 421–430.
50. Mrode, R. A. (1996). *Linear Models for the Prediction of Animal Breeding Values*. CAB International.
51. Möls, M. (2004). *Linear mixed models with equivalent predictors*. Dissertation. Tartu University Press.
52. Nahtman, T. (2004). *Permutation invariance and reparameterizations in linear models*. Dissertation. Tartu University Press.
53. Norell, L. (2001). *On the Optimum Number of Levels in the One-way Model With Random Effects*. Rapport 65, Department of Biometry and Informatics, Swedish University of Agricultural Sciences, Uppsala.
54. Norell, L. (2003). ANOVA Estimators Under Imbalance in the One-Way Random Model. *Communications in Statistics: Theory and Methods*, 32, 601–623.
55. Osborne, R., Paterson, W. S. B. (1952). On the sampling variance of heritability estimates derived from variance analysis. *Proceedings of the Royal Society of Edinburgh. Section B*, **64**, 456–461.
56. Ott, J. (1999). *Analysis of human genetic linkage, 3rd edition*. Johns Hopkins University Press, Baltimore.

57. Patterson, H. D., Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545–555.
58. Prasad, N. G. N., Rao, J. N. K. (1990). The Estimation of Mean Squared Error of Small-Area Estimators. *Journal of the American Statistical Association*, **85**, 163–171.
59. Quaas, R. L. (1976). Computing the diagonal elements of a large numerator relationship matrix. *Biometrics*, **32**, 949–953.
60. Quaas, R. L., Pollak, E. J. (1981). Modified equations for sire models with groups. *Journal of Dairy Science*, **64**, 1868–1872.
61. Rao, C. R. (1970). Estimation of heteroscedastic variances. *Journal of the American Statistical Association*, **65**, 161–170.
62. Rao, C. R. (1971). Estimation of variance and covariance components – MINQUE theory. *Journal of Multivariate Analysis*, **1**, 445–457.
63. Reents, R., Uba, M., Pedastsaar, K., Vares, T. (1996). Implementation of animal model for production traits of dairy cattle in Estonia. *Proceedings of the Open Session of the Interbull Annual Meeting, Bulletin no. 14*, Veldhoven, The Netherlands, 135–139.
64. Reverter, A., Kaiser, C. J. (1997). The role of different pedigree structures on the sampling variance of heritability estimates. *Journal of Animal Science*, **75**, 2355–2361.
65. Rumph, J. M., Koch, R. M., Gregory, K. E., Cundiff, L. V., Van Vleck, L. D. (2002). Comparison of models for estimation of genetic parameters for mature weight of Hereford cattle. *Journal of Animal Science*, **80**, 583–590.
66. SAS Institute Inc. (1999). *SAS OnlineDoc, Version 8*. Cary, NC, SAS Institute Inc.
67. Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika*, **6**, 309–316.
68. Schott, J., R. (1997). *Matrix Analysis for Statistics*. Wiley, New York.
69. Searle, S. R. (1982). *Matrix Algebra Useful for Statistics*. Wiley, New York.
70. Searle, S. R. (1987). *Linear Models for Unbalanced Data*. Wiley, New York.
71. Searle, S. R. (1998). *A Mathematical Supplement to C.R. Henderson's Book "Applications of Linear Models in Animal Breeding"*. University of Guelph Press, Guelph, Canada.
72. Searle, S. R., Casella, G., McCulloch, C. E. (1992). *Variance Components*. Wiley, New York.
73. Sham, P. (1998). *Statistics in human genetics*. Arnold, London.
74. Singh, B. (1989). A comparison of variance component estimators under unbalanced situations. *Sankhyā, Series B*, **51**, 323–330.
75. Smith, S. P., Graser, H.-U. (1986). Estimating variance components in a class of mixed models by restricted maximum likelihood. *Journal of Dairy Science*, **69**, 1156–1165.
76. Sorensen, D. A., Kennedy, B. W. (1983). The use of the relationship matrix to account for genetic drift variance in the analysis of genetic experiments. *Theoretical and Applied Genetics*, **66**, 217–220.
77. Swallow, W. H., Monahan, J. F. (1984). Monte Carlo Comparison of ANOVA, MIVQUE, REML and ML Estimators of Variance Components. *Technometrics*, **26**, 47–57.

78. Swinger, L. A., Harvey, W. R., Everson, D. O., Gregory, K. E. (1964). The variance of intraclass correlation involving groups with one observation. *Biometrics*, **20**, 818–826.
79. Visscher, P. M. (1998). On the sampling variance of intraclass correlations and genetic correlations. *Genetics*, **149**, 1605–1614.
80. Wang, S., G., Sinha, B., K., Sutradhar, B. (1992). Nonnegative estimation of variance components in unbalanced mixed linear models with two variance components. *Journal of Multivariate Analysis*, **42**, 77–101.
81. Weisstein, E. W. (2004). "Standard Error." From *MathWorld* — A Wolfram Web Resource. http://mathworld.wolfram.com/StandardError.html (1.11.2005).
82. Weller, J. I. (2001). *Quantitative Trait Loci Analysis in Animals*. CAB International.
83. Westell, R. A., Quaas, R.L., Van Vleck, L. D. (1988). Genetic groups in an animal model. *Journal of Dairy Science*, **71**, 1310–1318.
84. Willham, R. L. (1972). The role of maternal effects in animal breeding. III. Biometrical aspects of maternal effects in animals. *Journal of Animal Science*, **35**, 1288–1293.
85. Wright, S. (1922). Coefficients of inbreeding and relationship. *American Naturalist*, **56**, 330–338.
86. Zerbe, G. O., Goldgar, D. E. (1980). Comparison of intraclass correlation coefficients with the ratio of two independent F-statistics. *Communications in Statistics. Theory and Methods, Series A*, **9**, 1641–1655.

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to everyone who belived in the completion of the present thesis.

I am especially grateful to my supervisor Docent T. Möls for his longstanding supervision and guidance.

I am indebted to all my teachers of the Faculty of Mathematics and Computer Science and deeply appreciate their role in my education. I especially appreciate the people of the Institute of Mathematical Statistics, headed by Prof. T. Kollo, for their support, inspiration and critical attitude if needed.

I would like to express my gratitude to Prof. O. Saveli and Docent E. Pärna for inviting me to the Institute of Animal Husbandry of the Estonian Agricultural University over ten years ago. I am also grateful to all the researchers at the Institute of Veterinary Medicine and Animal Sciences of the Estonian Agricultural University for their multifarious scientific problems that never allowed me to become lazy.

I am grateful to PhD J. Vilo for his encouragement and patience.

I am much obliged to the Linnaeus Centre for Bioinformatics which, by the support of the European Commission program Human Research Potential & the Socio-economic Knowledge Base: Access to Research Infrastructures (project number HPRI-CT-2001-00153), covered the costs of my short research related stay in Uppsala, Sweden.

And finally, I thank my family members for their presence.

SUMMARY IN ESTONIAN

Lineaarsete segamudelite usaldusväärsus geneetilistes uuringutes

Lineaarsete segamudelite areng 20. sajandi vältel on käinud käsikäes geneetikaalaste teadmiste kasvuga. Järjest täienev ja komplitseeruv arusaamine elusorganismidel mõõdetud suuruste geneetilisest determineeritusest on nõudnud järjest täiuslikumaid mudeleid selle seotuse matemaatiliseks kirjeldamiseks. Sobivaimateks mudeliteks on osutunud lineaarsed segamudelid, sest ühelt poolt on nende abil võimalik hinnata konkreetsete andmetes fikseeritud mittegeneetiliste faktorite mõju ning teiselt poolt saab prognoosida otseselt mittevaadeldavaid geneetilisi efekte ja hinnata nende osa uuritava tunnuse varieeruvuses. Tänu otsesele majanduslikule huvile on järglastele pärandatavate geneetiliste efektide tuvastamine ja mõõtmine aastakümneid olnud loomakasvatusteaduse eesmärk, mistõttu on paljud lineaarsete segamudelite teoorias klassikaks saanud tulemused pärit just sellest valdkonnast. Näiteks ühed tuntumad 20. sajandi teise poole lineaarsete segamudelite teooria arendajad C. R. Henderson ja S. R. Searle on mõlemad aastakümneid töötanud just põllumajandusloomade aretuse vallas (Searle, 1998). Tänapäeval, millal järjest enam luuakse kõiksugu inimmeditsiinilisi andmeid, geneetikat, sugupuid jmt sisaldavaid ühtseid registreid, on lineaarsed segamudelid muutumas atraktiivseks analüüsimeetodiks ka geneetilises epidemioloogias.

Käesolevas töös on juhindutud eelkõige põllumajandusloomade aretuses kasutatavatest mudelitest, aga samas on paljud käsitletavad probleemid üldised ja võimaldavad teha katsete planeerimise ja andmete analüüsi alaseid otsuseid mistahes lineaarsete segamudelite rakendusalal.

Töö esimeses peatükis on defineeritud kasutatud maatriksite ja mudelite esitused ning toodud peamised tulemused lineaarsete segamudelite teooriast. Senise teooria üldistusena on tuletatud juhuslike efektide teist järku prognooside dispersioonimaatriksi avaldis.

Teises peatükis on peamised geneetilistes uuringutes rakendatavad mudelid esitatud paralleelselt nii geneetika kui ka matemaatilise statistika terminoloogiat ja kirjapilti kasutades. Taolise tavakäsitlusest komplekssema esituse mõte on korraga hõlmata nii teoreetilise geneetika seaduspäradel baseeruvad mudelite püstitamise põhimõtted kui ka reaalsetel andmetel põhinevatest statistilistest mudelitest hinnatavate parameetrite geneetiline sisu (Kaart, 2001).

Kolmandas peatükis on toodud töö peamised teoreetilised tulemused. Vaatluse alla on võetud lihtsaim võimalik lineaarne segamudel – ühe juhusliku faktoriga mudel – mis rakendusena geneetikas on tuntud ka poolõvede mudeli või isa mudeli (loomakasvatuses) nime all. Kirjanduse allikaile tuginedes on esitatud tulemused dispersioonikomponendi hinnangute varieeruvuse kohta, tuletatud on valemid päritavuskoefitsiendi hinnangute ja juhuslike efektide prognooside varieeruvuse kohta ning illegaalsete hinnangute tõenäosuste kohta. Nii teoreetilistele

tõestustele kui ka modelleerimiseksperimentidele tuginedes on uuritud hinnangute täpsuse sõltuvust andmete disainist ja mittetasakaalulisusest. Peatükis esitatud tulemused on osaliselt publitseeritud artiklites Kaart (2004), Kaart (2005) ja diskuteeritud konverentsi ettekannetes Kaart (1997), Kaart (1998).

Neljandas peatükis on diskuteeritud geneetiliste andmete analüüsil igapäevaste, aga üldiste lineaarsete segamudelite tavateoorias mittekäsitletavate mudelite püstituste üle. Tõestatud on tulemused juhuslike faktorite andmetes reaalsete mõõtmistulemustega mitteesindatud tasemete mõjude hindamise ja prognoosimise kohta. Väikese modelleerimiseksperimendi abil on uuritud erinevate ristamisskeemide ja põlvnemisstruktuuride mõju geneetiliste parameetrite hinnangute täpsusele. Arutletud on geneetilise mudeli valiku ning arvesse võetavate eellaspõlvkondade arvu mõju üle. Viimast diskussiooni on illustreeritud Eesti mustapealiste lammaste võõrutusmassi geneetilise hindamise näitega (pilootuuring töös esitatud näitanalüüsile on publitseeritud artiklis Kaart, Piirsalu, 2000).

CURRICULUM VITAE

Tanel Kaart

Born:	December 17, 1971 in Tartu, Estonia
Citizenship:	Estonian
Marital status:	married, 2 children
Adress:	Sõbra 40, Tartu 50106, Estonia
Contacts:	phone: +372 7 313 408
	e-mail: tanel.kaart@emu.ee

Education

1997-2002	PhD studies in Mathematical Statistics, University of Tartu
1995-1997	MSc studies in Mathematical Statistics, University of Tartu
1990-1995	BSc studies in Mathematical Statistics, University of Tartu

Professional employment

1995-2000	statistician, Institute of Animal Husbandry, Estonian Agricultural University
2000-2005	researcher, Institute of Animal Husbandry, Estonian Agricultural University
since 2003	statistician- geneticist, EGeen
since 2005	lecturer of biostatistics and population genetics, Institute of Veterinary Medicine and Animal Sciences, Estonian Agricultural University

ELULOOKIRJELDUS

Tanel Kaart

Sünniaeg ja -koht: 17. detsember 1971, Tartu
Kodakondsus: Eesti
Perekonnaseis: abielus, 2 last
Aadress: Sõbra 40, Tartu 50106
Kontaktandmed: telefon: +372 7 313 408
e-mail: tanel.kaart@emu.ee

Haridus

1997–2002	doktoriõpe, Tartu Ülikool, Matemaatikateaduskond
1995–1997	magistriõpe, Tartu Ülikool, Matemaatikateaduskond
1990–1995	bakalaureuseõpe, Tartu Ülikool, Matemaatikateaduskond

Erialane teenistuskäik

1995–2000	statistik, EPMÜ Loomakasvatusinstituut
2000–2005	teadur, EPMÜ Loomakasvatusinstituut
alates 2003	statistik-geneetik, AS EGeen
alates 2005	biostatistika ja populatsioonigeneetika lektor, EPMÜ Veterinaarmeditsiini ja loomakasvatuse instituut