Chapter 1
Models

C. R. Henderson

1984 - Guelph

This book is concerned exclusively with the analysis of data arising from an experiment or sampling scheme for which a linear model is assumed to be a suitable approximation. We should not, however, be so naive as to believe that a linear model is always correct. The important consideration is whether its use permits predictions to be accomplished accurately enough for our purposes. This chapter will deal with a general formulation that encompasses all linear models that have been used in animal breeding and related fields. Some suggestions for choosing a model will also be discussed.

All linear models can, I believe, be written as follows with proper definition of the various elements of the model. Define the observable data vector with n elements as $\mathbf{y}$. In order for the problem to be amenable to a statistical analysis from which we can draw inferences concerning the parameters of the model or can predict future observations it is necessary that the data vector be regarded legitimately as a random sample from some real or conceptual population with some known or assumed distribution. Because we seldom know what the true distribution really is, a commonly used method is to assume as an approximation to the truth that the distribution is multivariate normal. Analyses based on this approximation often have remarkable power. See, for example, Cochran (1937). The multivariate normal distribution is defined completely by its mean and by its central second moments. Consequently we write a linear model for $\mathbf{y}$ with elements in the model that determine these moments. This is

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

$\mathbf{X}$ is a known, fixed, $n \times p$ matrix with rank $= r \leq$ minimum of (n, p).

$\boldsymbol{\beta}$ is a fixed, $p \times 1$ vector generally unknown, although in selection index methodology it is assumed, probably always incorrectly, that it is known.

$\mathbf{Z}$ is a known, fixed, $n \times q$ matrix.

$\mathbf{u}$ is a random, $q \times 1$ vector with null means.

$\mathbf{e}$ is a random, $n \times 1$ vector with null means.

The variance-covariance matrix of $\mathbf{u}$ is $\mathbf{G}$, a $q \times q$ symmetric matrix that is usually non-singular. Hereafter for convenience we shall use the notation $Var(\mathbf{u})$ to mean a variance-covariance matrix of a random vector.

$Var(\mathbf{e}) = \mathbf{R}$ is an $n \times n$, symmetric, usually non-singular matrix. $Cov(\mathbf{u}, \mathbf{e}') = \mathbf{0}$, that is, all elements of the covariance matrix for $\mathbf{u}$ with $\mathbf{e}$ are zero in most but not all applications.

It must be understood that we have hypothesized a population of $\mathbf{u}$ vectors from which a random sample of one has been drawn into the sample associated with the data vector, $\mathbf{y}$, and similarly a population of $\mathbf{e}$ vectors is assumed, and a sample vector has been drawn with the first element of the sample vector being associated with the first element of $\mathbf{y}$, etc.

Generally we do not know the values of the individual elements of $\mathbf{G}$ and $\mathbf{R}$. We usually are willing, however, to make assumptions about the pattern of these values. For example, it is often assumed that all the diagonal elements of $\mathbf{R}$ are equal and that all off-diagonal elements are zero. That is, the elements of $\mathbf{e}$ have equal variances and are mutually uncorrelated. Given some assumed pattern of values of $\mathbf{G}$ and $\mathbf{R}$, it is then possible to estimate these matrices assuming a suitable design (values of $\mathbf{X}$ and $\mathbf{Z}$) and a suitable sampling scheme, that is, guarantee that the data vector arose in accordance with $\mathbf{u}$ and $\mathbf{e}$ being random vectors from their respective populations. With the model just described

$$\begin{aligned} E(\mathbf{y}) &= \text{mean of } \mathbf{y} = \mathbf{X}\boldsymbol{\beta}. \\ Var(\mathbf{y}) &= \mathbf{ZGZ}' + \mathbf{R}. \end{aligned}$$

We shall now present a few examples of well known models and show how these can be formulated by the general model described above. The important advantage to having one model that includes all cases is that we can thereby present in a condensed manner the basic methods for estimation, computing sampling variances, testing hypotheses, and prediction.

1 Simple Regression Model

The simple regression model can be written as follows,

$$y_i \;=\; \mu \;+\; x_i\,\alpha \;+\; e_i.$$

This is a scalar model, y_i being the i^{th} of n observations. The fixed elements of the model are μ and α, the latter representing the regression coefficient. The concomitant variable associated with the i^{th} observation is x_i, regarded as fixed and measured without error.

Note that in conceptual repeated sampling the values of x_i remain constant from one sample to another, but in each sample a new set of e_i is taken, and consequently the values of y_i change. Now relative to our general model,

$$\begin{aligned} \mathbf{y}' &= (y_1 \quad y_2 \quad \ldots \quad y_n), \\ \boldsymbol{\beta}' &= (\mu \quad \alpha), \\ \mathbf{X}' &= \begin{bmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & \ldots & x_n \end{bmatrix}, \text{ and} \\ \mathbf{e}' &= (e_1 \quad e_2 \quad \ldots \quad e_n) \end{aligned}$$

Zu does not exist in the model. Usually **R** is assumed to be $\mathbf{I}\sigma_e^2$ in regression models.

2 One Way Random Model

Suppose we have a random sample of unrelated sires from some population of sires and that these are mated to a sample of unrelated dams with one progeny per dam. The resulting progeny are reared in a common environment, and one record is observed on each. An appropriate model would seem to be

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

y_{ij} being the observation on the j^{th} progeny of the i^{th} sire.

Suppose that there are 3 sires with progeny numbers 3, 2, l respectively. Then **y** is a vector with 6 elements.

$$\begin{aligned} \mathbf{y}' &= (y_{11} \quad y_{12} \quad y_{13} \quad y_{21} \quad y_{22} \quad y_{31}), \\ \mathbf{x}' &= (1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1), \\ \mathbf{u}' &= (s_1 \quad s_2 \quad s_3), \text{ and} \\ \mathbf{e}' &= (e_{11} \quad e_{12} \quad e_{13} \quad e_{21} \quad e_{22} \quad e_{23}), \\ Var(\mathbf{u}) &= \mathbf{I}\sigma_s^2, \\ Var(\mathbf{e}) &= \mathbf{I}\sigma_e^2, \end{aligned}$$

where these two identity matrices are of order 3 and 6, respectively.

$$Cov(\mathbf{u}, \mathbf{e}') = \mathbf{0}.$$

Suppose next that the sires in the sample are related, for example, sires 2 and 3 are half-sib progeny of sire l, and all 3 are non-inbred. Then under an additive genetic model

$$Var(\mathbf{u}) = \begin{bmatrix} 1 & 1/2 & 1/2 \\ 1/2 & 1 & 1/4 \\ 1/2 & 1/4 & 1 \end{bmatrix} \sigma_s^2.$$

What if the mates are related? Suppose that the numerator relationship matrix, $\mathbf{A}_m$, for the 6 mates is

$$\begin{pmatrix} 1 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1 & 1/4 & 0 & 0 \\ 1/2 & 0 & 1/4 & 1 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 1 & 1/4 \\ 0 & 1/2 & 0 & 0 & 1/4 & 1 \end{pmatrix}.$$

Suppose further that we invoke an additive genetic model with $h^2 = 1/4$. Then

$$Var(\mathbf{e}) = \begin{pmatrix} 1 & 0 & 1/30 & 1/30 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1/30 & 1/30 \\ 1/30 & 0 & 1 & 1/60 & 0 & 0 \\ 1/30 & 0 & 1/60 & 1 & 0 & 0 \\ 0 & 1/30 & 0 & 0 & 1 & 1/60 \\ 0 & 1/30 & 0 & 0 & 1/60 & 1 \end{pmatrix} \sigma_e^2.$$

This result is based on $\sigma_s^2 = \sigma_y^2/16$, $\sigma_e^2 = 15\ \sigma_y^2/16$, and leads to

$$Var(\mathbf{y}) = (.25\ \mathbf{A}_p + .75\ \mathbf{I})\ \sigma_y^2,$$

where $\mathbf{A}_p$ is the relationship matrix for the 6 progeny.

3 Two Trait Additive Genetic Model

Suppose that we have a random sample of 5 related animals with measurements on 2 correlated traits. We assume an additive genetic model. Let $\mathbf{A}$ be the numerator relationship matrix of the 5 animals. Let

$$\begin{pmatrix} g_{11} & g_{12} \\ g_{12} & g_{22} \end{pmatrix}$$

be the genetic variance-covariance matrix and

$$\begin{pmatrix} r_{11} & r_{12} \\ r_{12} & r_{22} \end{pmatrix}$$

be the environmental variance-covariance matrix. Then h^2 for trait 1 is $g_{11}/(g_{11}+r_{11})$, and the genetic correlation between the two traits is $g_{12}/(g_{11}\ g_{22})^{1/2}$. Order the 10 observations, animals within traits. That is, the first 5 elements of $\mathbf{y}$ are the observations on trait 1. Suppose that traits 1 and 2 have common means μ_1, μ_2 respectively. Then

$$\mathbf{X}' = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{pmatrix},$$

and

$$\boldsymbol{\beta}' = (\mu_1 \quad \mu_2).$$

The first 5 elements of $\mathbf{u}$ are breeding values for trait 1 and the last 5 are breeding values for trait 2. Similarly the errors are partitioned into subvectors with 5 elements each. Then $\mathbf{Z} = \mathbf{I}$ and

$$\mathbf{G} = Var(\mathbf{u}) = \begin{pmatrix} \mathbf{A}\ g_{11} & \mathbf{A}\ g_{12} \\ \mathbf{A}\ g_{12} & \mathbf{A}\ g_{22} \end{pmatrix},$$

$$\mathbf{R} = \text{Var}\ (\mathbf{e}) = \begin{pmatrix} \mathbf{I}\ r_{11} & \mathbf{I}\ r_{12} \\ \mathbf{I}\ r_{12} & \mathbf{I}\ r_{22} \end{pmatrix},$$

where each $\mathbf{I}$ has order, 5.

4 Two Way Mixed Model

Suppose that we have a random sample of 3 unrelated sires and that they are mated to unrelated dams. One progeny of each mating is obtained, and the resulting progeny are assigned at random to two different treatments. The table of subclass numbers is

	Treatments	
Sires	1	2
1	2	1
2	0	2
3	3	0

Ordering the data by treatments within sires,

$$\mathbf{y}' = \begin{pmatrix} y_{111} & y_{112} & y_{121} & y_{221} & y_{222} & y_{311} & y_{312} & y_{313} \end{pmatrix}.$$

Treatments are regarded as fixed, and variances of sires and errors are considered to be unaffected by treatments. Then

$$\mathbf{u}' = \begin{pmatrix} s_1 & s_2 & s_3 & st_{11} & st_{12} & st_{22} & st_{31} \end{pmatrix}.$$

$$\mathbf{Z} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

$$Var(\mathbf{s}) = \mathbf{I}_3\ \sigma_s^2, \quad Var(\mathbf{st}) = \mathbf{I}_4\ \sigma_{st}^2, \quad Var(\mathbf{e}) = \mathbf{I}_8\ \sigma_e^2.$$

$$Cov(\mathbf{s}, \quad (\mathbf{st}')) = \mathbf{0}.$$

This is certainly not the only linear model that could be invoked for this design. For example, one might want to assume that sire and error variances are related to treatments.

5 Equivalent Models

It was stated above that a linear model must describe the mean and the variance-covariance matrix of $\mathbf{y}$. Given these two, an infinity of models can be written all of which yield the same first and second moments. These models are called linear equivalent models.

Let one model be $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$ with $Var(\mathbf{u}) = \mathbf{G}$, $Var(\mathbf{e}) = \mathbf{R}$. Let a second model be $\mathbf{y} = \mathbf{X}_*\boldsymbol{\beta}_* + \mathbf{Z}_*\mathbf{u}_* + \mathbf{e}_*$, with $Var(\mathbf{u}_*) = \mathbf{G}_*$, $Var(\mathbf{e}_*) = \mathbf{R}_*$. Then the means of $\mathbf{y}$ under these 2 models are $\mathbf{X}\boldsymbol{\beta}$ and $\mathbf{X}_*\boldsymbol{\beta}_*$ respectively. $Var(\mathbf{y})$ under the 2 models is

$$\mathbf{ZGZ}' + \mathbf{R} \text{ and } \mathbf{Z}_*\mathbf{G}_*\mathbf{Z}_*' + \mathbf{R}_*.$$

Consequently we state that these 2 models are linearly equivalent if and only if

$$\mathbf{X}\boldsymbol{\beta} = \mathbf{X}_*\boldsymbol{\beta}_* \text{ and } \mathbf{ZGZ}' + \mathbf{R} = \mathbf{Z}_*\mathbf{G}_*\mathbf{Z}_*^{'} + \mathbf{R}_*.$$

To illustrate, $\mathbf{X}\boldsymbol{\beta} = \mathbf{X}_*\ \boldsymbol{\beta}_*$ suppose we have a treatment design with 3 treatments and 2 observations on each. Suppose we write a model

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

then

$$\mathbf{X}\boldsymbol{\beta} \quad = \quad \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ t_1 \\ t_2 \\ t_3 \end{pmatrix}.$$

An alternative model is

$$y_{ij} = \alpha_i + e_{ij},$$

then

$$\mathbf{X}_*\boldsymbol{\beta}_* \quad = \quad \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}.$$

Then if we define $\alpha_i \ = \ \mu \ + \ t_i$, it is seen that $E(\mathbf{y})$ is the same in the two models. To illustrate with two models that give the same $Var(\mathbf{y})$ consider a repeated lactation model. Suppose we have 3 unrelated, random sample cows with 3, 2, 1 lactation records, respectively. Invoking a simple repeatability model, that is, the correlation between any pair of records on the same animal is r, one model ignoring the fixed effects is

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

$$Var(\mathbf{c}) \ = \ Var \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} \quad = \quad \begin{pmatrix} r & 0 & 0 \\ 0 & r & 0 \\ 0 & 0 & r \end{pmatrix} \sigma_y^2.$$

$$Var(\mathbf{e}) \;=\; \mathbf{I}_6\;(1-r)\;\sigma_y^2.$$

An alternative for the random part of the model is

$$y_{ij} \;=\; e_{ij},$$

where **Zu** does not exist.

$$Var(\boldsymbol{\epsilon}) \;=\; \mathbf{R} \;=\; \begin{pmatrix} 1 & r & r & 0 & 0 & 0 \\ r & 1 & r & 0 & 0 & 0 \\ r & r & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & r & 0 \\ 0 & 0 & 0 & r & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \sigma_y^2.$$

Relating the 2 models,

$$\begin{aligned} \sigma_\epsilon^2 &= \sigma_c^2 \;+\; \sigma_e^2. \\ Cov(\epsilon_{ij},\; \epsilon_{ij},) &= \sigma_c^2 \;\text{ for }\; j \;\neq j'. \end{aligned}$$

We shall see that some models are much easier computationally than others. Also the parameters of one model can always be written as linear functions of the parameters of any equivalent model. Consequently linear and quadratic estimates under one model can be converted by these same linear functions to estimates for an equivalent model.

6 Subclass Means Model

With some models it is convenient to write them as models for the "smallest" subclass mean. By "smallest" we imply a subclass identified by all of the subscripts in the model except for the individual observations. For this model to apply, the variance-covariance matrix of elements of **e** pertaining to observations in the same smallest subclass must have the form

$$\begin{pmatrix} v & & c \\ & \ddots & \\ c & & v \end{pmatrix},$$

no covariates exist, and the covariances between elements of **e** in different subclasses must be zero. Then the model can be written

$$\bar{\mathbf{y}} = \bar{\mathbf{X}}\boldsymbol{\beta} + \bar{\mathbf{Z}}\mathbf{u} + \boldsymbol{\epsilon}.$$

$\bar{\mathbf{y}}$ is the vector of "smallest" subclass means. $\bar{\mathbf{X}}$ and $\bar{\mathbf{Z}}$ relate these means to elements of $\boldsymbol{\beta}$ and $\mathbf{u}$. The error vector, $\boldsymbol{\epsilon}$, is the mean of elements of $\mathbf{e}$ in the same subclass. Its variance-covariance matrix is diagonal with the i^{th} diagonal element being

$$\left(\frac{v}{n_i} + \frac{n_i - 1}{n_i} c \right) \sigma_e^2,$$

where n_i is the number of observations in the i^{th} subclass.

7 Determining Possible Elements In The Model

Henderson(1959) described in detail an algorithm for determining the potential lines of an ANOVA table and correspondingly the elements of a linear model. First, the experiment is described in terms of two types of factors, namely main factors and nested factors. By a main factor is meant a classification, the "levels" of which are identified by a single subscript. By a nesting factor is meant one whose levels are not completely identified except by specifying a main factor or a combination of main factors within which the nesting factor is nested. Identify each of the main factors by a single unique letter, for example, B for breeds and T for treatments. Identify nesting factors by a letter followed by a colon and then the letter or letters describing the main factor or factors within which it is nested. For example, if sires are nested within breeds, this would be described as S:B. On the other hand, if a different set of sires is used for each breed by treatment combination, sires would be identified as S:BT. To determine potential 2 factor interactions combine the letters to the left of the colon (for a main factor a colon is implied with no letters following). Then combine the letters without repetition to the right of the colon. If no letter appears on both the right and left of the colon this is a valid 2 factor interaction. For example, factors are A,B,C:B. Two way combinations are AB, AC:B, BC:B. The third does not qualify since B appears to the left and right of the colon. AC:B means A by C interaction nested within B. Three factor and higher interactions are determined by taking all possible trios and carrying out the above procedure. For example, factors are (A, D, B:D, C:D). Two factor possibilities are (AD, AB:D, AC:D, DB:D, DC:D, BC:D). The 4^{th} and 5^{th} are not valid. Three factor possibilities are (ADB:D, ADC:D, ABC:D, DBC:D). None of these is valid except ABC:D. The four factor possibility is ADBC:D, and this is not valid.

Having written the main factors and interactions one uses each of these as a subvector of either $\boldsymbol{\beta}$ or $\mathbf{u}$. The next question is how to determine which. First consider main factors and nesting factors. If the levels of the factor in the experiment can be regarded as a

random sample from some population of levels, the levels would be a subvector of $\mathbf{u}$. With respect to interactions, if one or more letters to the left of the colon represent a factor in $\mathbf{u}$, the interaction levels are subvectors of $\mathbf{u}$. Thus interaction of fixed by random factors is regarded as random, as is the nesting of random within fixed. As a final step we decide the variance-covariance matrix of each subvector of $\mathbf{u}$, the covariance between subvectors of $\mathbf{u}$, and the variance- covariance matrix of $(\mathbf{u}, \mathbf{e})$. These last decisions are based on knowledge of the biology and the sampling scheme that produced the data vector.

It seems to me that modelling is the most important and most difficult aspect of linear models applications. Given the model everything else is essentially computational.