Chapter 3

Fixed Linear Effects Models

3.1 Resources

Similarly to chapter 2, this chapter on `fixed linear effects models` (FLEM) is based on the work of [Bühlmann and Mächler, 2016] and on the book [Searle, 1971].

3.2 Introduction

In chapter 2, we saw how linear regression analysis was used to describe and to quantify the relationship between a response variable and between one or more predictor variables. The type of analysis shown in chapter 2 is called "regression analysis, because the response and the predictors are all continuous variables. This means that the values of the variables in the dataset are all floating-point numbers. For datasets where predictor variables are discrete, the model is referred to as *fixed linear effects model*.

The reason why fixed linear effects models must be treated differently from regression models can best be seen by looking at an extension of our example dataset on body weight of some animals. Let us assume that besides the predictors that we have used so far, we have the breed of the animal as an additional information. Animals of different breeds have different body weights, hence we expect that the breed of the animal has an effect on its body weight. The question is how is it possible to integrate the breed of the animal into a model that describes and quantifies the different influence factors on body weight. First, we have a look at the extended dataset.

Table 3.1: Extended Dataset on Body Weight for 10 Animals

Animal	Breast Circumference	Body Weight	BCS	HEI	Breed
1	176	471	5.0	161	Angus
2	177	463	4.2	121	Angus
3	178	481	4.9	157	Simmental
4	179	470	3.0	165	Angus
5	179	496	6.8	136	Simmental
6	180	491	4.9	123	Simmental
7	181	518	4.4	163	Limousin
8	182	511	4.4	149	Limousin
9	183	510	3.5	143	Limousin
10	184	541	4.7	130	Limousin

The extension in our dataset consists of the breed for each animal. With this extension, the immediate question of how to measure "breed" arises. The breed as it is in the dataset cannot be integreated into our model. It must be converted into a numeric code. One possibility is to assign each breed to a number according to how heavy an average animal of the breed is expected to be. Because this assignment is difficult to do, as the body weight of animals within a given breed show a certain variation. For our example, the following assignment of breeds to numeric codes is assumed.

Table 3.2: Assignment of Breeds to numeric Codes

Code	Breed
1	Angus
2	Limousin
3	Simmental

For reasons of simplicity, we assume that the variable "breed" is the only predictor in a simple regression model

$$E(y_i) = b_0 + b_1 x_i \tag{3.1}$$

where $E(y_i)$ stands for the expected value of body weight (y_i) of animal i, b_0 is the intercept, x_i corresponds to the numeric code of the breed of animal i and b_1 is the regression coefficient for the breed code. The influence of the predictor variable breed code on body weight could be tested with the hypothesis $b_1 = 0$ which is done by the function `lm()` in R.

Although this analysis as described is permissible, it does come with a number of problems which show that the assumptions behind this type of model are unrealistic. This can best be shown by looking at the expected values of body weight (BW) for animals of the different breeds.

$$\begin{aligned} E(\text{BW Angus}) &= b_0 + b_1 \\ E(\text{BW Limousin}) &= b_0 + 2b_1 \\ E(\text{BW Simmental}) &= b_0 + 3b_1 \end{aligned} \tag{3.2}$$

This means, for example, that

$$\begin{aligned} E(\text{BW Limousin}) - E(\text{BW Angus}) &= E(\text{BW Simmental}) - E(\text{BW Limousin}) \\ E(\text{BW Simmental}) - E(\text{BW Angus}) &= 2\,[E(\text{BW Limousin}) - E(\text{BW Angus})] \end{aligned} \tag{3.3}$$

Depending on the data, the relations shown in (3.3) might be quite unrealistic. And even without data, only by the allocation of numerical codes to the different breed, the consequences shown in (3.3) are forced on the analysis results. The only real estimates that the analysis yields are the one of b_0 and of b_1. This will also be the case, if different numerical codes are used for the different levels of the variable.

The inherent difficulty with the analysis suggested above is the allocation of numerical codes to non-quantitative variables such as breed. Yet such varibles are of great interest in many scientific areas. Allocating numerical codes to such variables involves at least two problems.

1. Often the assignment cannot be made in a reasonable way and is thereby to a large extent an arbitrary process.
2. Making such allocations of numeric codes to different levels of a variable imposes value differences on the categories of the variable such as shown in equation (3.3).

The above state problems can best be solved by using a type of model that is often referred to as *regession on dummy* $(0, 1)$ *variables*. In the context here, we are calling these models just *fixed linear effect models*. The description of these models is deferred to a later section. We first describe an important exception in which the application of a linear regression model on discrete variables is very reasonable and has a wide range of applications.

3.3 Linear Regression Analysis for Genomic Data

The question why linear regression models can be applied to genomic data is best answered by looking at the data. In general, genomic breeding values can either be estimated using a two-step procedure or by a single step approach. At the moment, we assume that we are in the first step of the two step approach

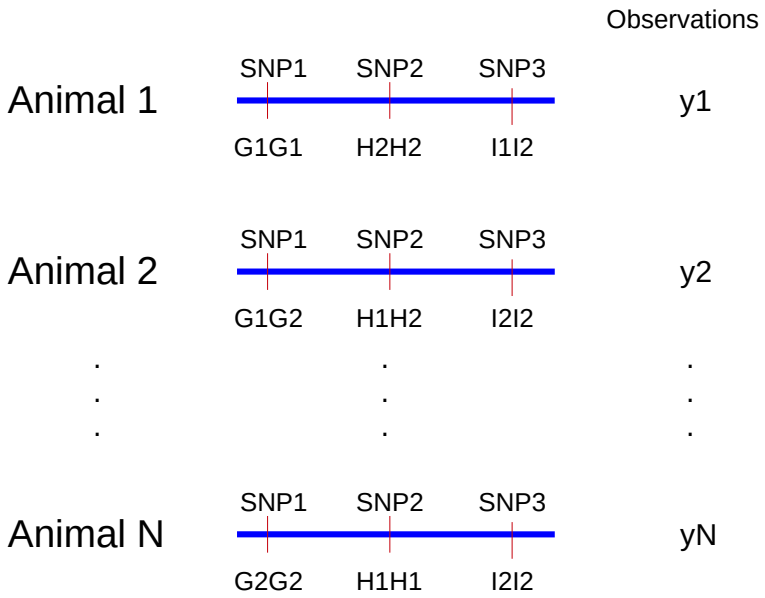

Figure 3.1: Structure of Dataset To Estimate GBV

where we estimate the marker effects (a-values) in a reference population or alternatively we have a perfect data set with all animals genotyped and with a phenotypic observation in a single step setting. Both situations are equivalent when it comes to the structure of the underlying dataset. Furthermore the same class of models can be used to analyse the type of data.

3.4 Data

As already mentioned in section 3.3, we are assuming that we have a perfect dataset for a given population of animals. That means each animal i has a phenotypic observation y_i for a given trait of interest. Furthermore, we assume to have a map of only three SNP markers. The marker loci are called G, H and I. All markers have two alleles each. Figure 3.1 tries to illustrate the structure of a dataset used to estimate genomic breeding values (GBV).

As can be seen from Figure 3.1 each of the N animals have known genotypes for all three SNP markers and they all have a phenotypic observation $y_i \quad (i = 1, \cdot, N)$. Because we are assuming each SNP marker to be bi-allelic, there are only three possible marker genotypes at every marker position. Hence marker

genotypes are discrete entities with a fixed number of levels. Due to the nature of the SNP marker genotype data, we can already say that they could be modeled as fixed effects in a fixed linear effects model. More details about the model will follow in section 3.5.

3.5 Model

The goal of our data analysis using the dataset described in section 3.4 is to come up with estimates for genomic breeding values for all animals in our dataset. The genomic breeding values will later be used to rank the animals. The ranking of the animals according to the GBV is used to select the parents of the future generation of livestock animals. It probably makes sense to distinguish between two different types of models that we have to set up. On the one side we need a model that describes the underlying genetic architecture which is present in our dataset. We will be using a so-called **genetic** model to describe this. On the other side, we have to be able to get estimates for the GBVs which requires a **statistical** model which is able to estimate unknown parameters as a function of observed data. In the end, we will realize that the two models are actually the same model but they are just different ways of looking at the same structure of the underlying phenomena. These phenomena characterize the relationship between genetic architecture of an animal and the expression of a certain phenotypic trait in that same animal.

3.5.1 Genetic Model

The availability of genomic information for all animals in the dataset makes it possible to use a polygenic model. In contrast to an infinitesimal model, a polygenic model uses a finite number of discrete loci to model the genetic part of an expressed phenotypic observation. From quantitative genetics (see e.g. [Falconer and Mackay, 1996] for a reference) we know that every phenotypic observation y can be separated into a genetic part g and an environmental part e. This leads to the very simple genetic model

$$y = g + e \tag{3.4}$$

The environmental part can be split into some fixed known systematic factors such as `herd`, `season effects`, `age` and more and into a random unknown part. The systematic factors are typically grouped into a vector of fixed effects called β. The unknown environmental random part is usually called ϵ. This allows to re-write the simple genetic model in (3.4) as

$$y = \beta + g + \epsilon \tag{3.5}$$

The genetic component g can be decomposed into contributions from the finite number of loci that are influencing the observation y. In our example dataset

(see Figure 3.1) there are three loci[1] that are assumed to have an effect on y. Ignoring any interaction effects between the three loci and thereby assuming a completely additive model, the overall genetic effect g can be decomposed into the sum of the genotypic values of each locus. Hence

$$g = \sum_{j=1}^{k} g_j \tag{3.6}$$

where for our example k is equal to three[2].

Considering all SNP loci to be purely additive which means that we are ignoring any dominance effects, the genotypic values g_j at any locus j can just take one of the three values $-a_j$, 0 or $+a_j$ where a_j corresponds to the a value from the mono-genic model (see Figure **??**). For our example dataset the genotypic value for each SNP genotype is given in the following table.

Table 3.3: Genotypic Values For All Three SNP-Loci

SNP Locus	Genotype	Genotypic Value
SNP_1	G_1G_1	a_1
SNP_1	G_1G_2	0
SNP_1	G_2G_2	$-a_1$
SNP_2	H_1H_1	a_2
SNP_2	H_1H_2	0
SNP_2	H_2H_2	$-a_2$
SNP_3	I_1I_1	a_3
SNP_3	I_1I_2	0
SNP_3	I_2I_2	$-a_3$

From the Table 3.3 we can see that always the allele with subscript 1 is taken to be that with the positive effect. Combining the information from Table 3.3 together with the decomposition of the genotypic value g in (3.6), we get

$$g = m^T \cdot a \tag{3.7}$$

where m is an indicator vector taking values of -1, 0 and 1 depending on the SNP marker genotype and a is the vector of a values for all SNP marker loci.

[1] Implicitly, we are treating the SNP-markers to be identical with the underlying QTL. But based on the fact that we have very many SNPs spread over the complete genome, there will always be SNP sufficiently close to every QTL that influences a certain trait. But in reality the unknown QTL affect the traits and not the SNPs.

[2] In reality k can be $1.5 * 10^5$ for some commercial SNP chip platforms. When working with complete genomic sequences, k can also be in the order of $3 * 10^7$.

Combining the decomposition in (3.7) together with the basic genetic model in (3.5), we get

$$y = \beta + m^T \cdot a + \epsilon \tag{3.8}$$

The result obtained in (3.5) is the fundamental decomposition of the phenotypic observation y into a genetic part represented by the SNP marker information (m) and an environmental part (β and ϵ). The a values are unknown and must be estimated. The estimates of the a values will then be used to predict the GBVs. How this estimation procedure works is described in the next section 3.5.2.

3.5.2 Statistical Model

When looking at the fundamental decomposition given in the genetic model presented in (3.8) from a statistics point of view, the model in (3.8) can be interpreted as **fixed linear effects model** (FLEM). FLEM represent a class of linear models where each model term except for the random residual term is a fixed effect. Furthermore, besides a random error term, the response is explained by a linear function of the predictor variables.

Using the decomposition given in our genetic model (see equation (3.8)) for our example dataset illustrated in Figure 3.1, every observation y_i of animal i can be written as

$$y_i = W_i \cdot \beta + M_i \cdot a + \epsilon_i \tag{3.9}$$

where

- y_i is the observation of animal i
- β is a vector of unknown systematic environmental effects
- W_i is an indicator row vector linking β to y_i
- a is a vector of unknown additive allele substitution effects (a values)
- M_i is an indicator row vector encoding the SNP genotypes of animal i and
- ϵ_i is the random unknown environmental term belonging to animal i

In the following section, we write down the definition of a FLEM and compare it to the statistical model given in (3.9).

3.6 Definition of FLEM

The multiple fixed linear effects model is defined as follows.

In a fixed linear effects model, every observation i in a dataset is characterized by a **response variable** and a set of **predictors**. Up to some random errors the response variable can be expressed as a linear function of the predictors. The

proposed linear function contains unknown parameters. The goal is to estimate both the unknown parameters and the error variance.

3.6.1 Terminology

For datasets where both the predictors and the response variables are on a continuous scale, which means that they correspond to measured quantities such as body weight, breast circumference or milk yield, the model is referred to as **multiple linear regression model**. Because the statistical model in (3.9) contains the SNP genotypes as discrete fixed effects, we are not dealing with a regression model but with a more general fixed linear effects model.

3.6.2 Model Specification

An analysis of the model given in (3.9) shows that it exactly corresponds to the definition **??**. In this equivalence, the observation y_i corresponds to the response variable. Furthermore, the unknown environmental term ϵ corresponds to the random residual part in the FLEM. Except for the random residuals the response variable y_i is a linear function of the fixed effects which corresponds to all systematic environmental effects and to all SNP genotype effects.

For the description of how to estimate the unknown parameter β and a in the model (3.9), it is useful to combine β and a into a single vector of unknown parameters and we call it b.

$$b = \begin{bmatrix} \beta \\ a \end{bmatrix} \tag{3.10}$$

Taking the equations as shown in (3.9) for all observations ($i = 1, \ldots, N$) and expressing them in matrix-vector notation, we get

$$y = Xb + \epsilon \tag{3.11}$$

where

- y is the vector of N observations
- b is the vector of all unknown fixed effects
- X is the incidence matrix linking the parameters of b to y
- ϵ is the vector of random residuals

The incidence matrix X in (3.11) can be composed from the matrices W and M by concatenating the latter two matrices, i.e.,

$$X = \begin{bmatrix} W & M \end{bmatrix} \tag{3.12}$$