

Chapter 2

Fixed Linear Effects Models

2.1 Resources

This chapter on `fixed linear effects models` (FLEM) is based on the work of [Buehlmann and Maechler, 2014]. Besides that, there are many other resources for the topic of `Multiple Linear Regressions`. An interesting online book is [Lilja, 2016].

2.2 Motivation

Why is the topic of fixed linear effects models important for the analysis of genomic data? This question is best answered when looking at the data. In chapter 1, we saw that genomic breeding values can either be estimated using a two-step procedure (see section 1.9) or by a single step approach (see section 1.10). At the moment, we assume that we are in the first step of the two step approach 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 data.

2.3 Data

As already mentioned in section 2.2, 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 2.1 tries to illustrate the structure of a dataset used to estimate genomic breeding values (GBV).

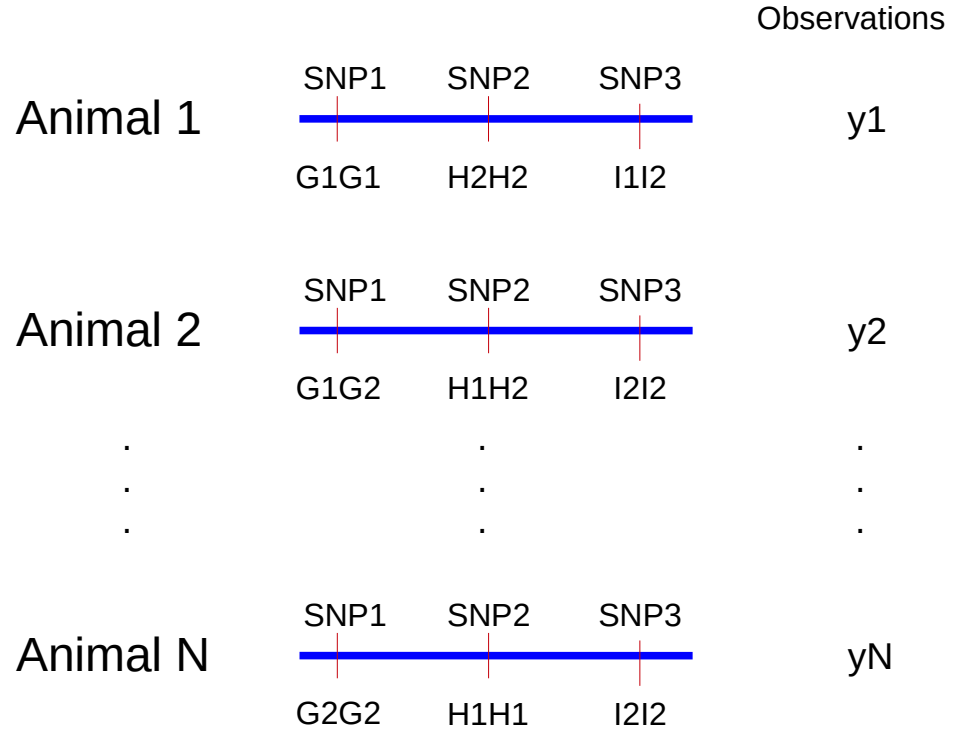


Figure 2.1: Structure of Dataset To Estimate GBV

As can be seen from Figure 2.1 each of the N animals have known genotypes for all three SNP markers and they all have a phenotypic observation y_i ($i = 1, \dots, N$). 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 2.4.

2.4 Model

The goal of our data analysis using the dataset described in section 2.3 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.

2.4.1 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

$$y = g + e \tag{2.1}$$

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 (2.1) as

$$y = \beta + g + \epsilon \tag{2.2}$$

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 2.1) there are three loci¹ that are assumed to have an effect on y .

¹Implicitly, 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

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

$$g = \sum_{j=1}^k g_j \quad (2.3)$$

where for our example k is equal to three².

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 1.5). For our example dataset the genotypic value for each SNP genotype is given in the following table.

Table 2.1: 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 2.1 we can see that always the allele with subscript 1 is taken to be that with the positive effect. Combining the information from Table 2.1 together with the decomposition of the genotypic value g in (2.3), we get

$$g = m^T \cdot a \quad (2.4)$$

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. Combining the decomposition in (2.4) together with the basic genetic model in (2.2), we get

$$y = \beta + m^T \cdot a + \epsilon \quad (2.5)$$

²the unknown QTL affect the traits and not the SNPs.

²In 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 $3 * 10^7$.

The result obtained in (2.2) 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 2.4.2.

2.4.2 Statistical Model

When looking at the fundamental decomposition given in the genetic model presented in (2.5) from a statistics point of view, the model in (2.5) 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 (2.5)) for our example dataset illustrated in Figure 2.1, every observation y_i of animal i can be written as

$$y_i = W_i \cdot \beta + M_i \cdot a + \epsilon_i \quad (2.6)$$

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 (2.6).

2.5 Definition of FLEM

The multiple fixed linear effects model is defined as follows.

Definition 2.1 (Fixed Linear Effects Model). 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.

2.5.1 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 (2.6) 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.

2.5.2 Model Specification

An analysis of the model given in (2.6) shows that it exactly corresponds to the definition 2.1. 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 (2.6), it is useful to combine β and a into a single vector of unknown parameters and we call it b .

$$b = \begin{bmatrix} \beta \\ a \end{bmatrix} \quad (2.7)$$

Taking the equations as shown in (2.6) for all observations ($i = 1, \dots, N$) and expressing them in matrix-vector notation, we get

$$y = Xb + \epsilon \quad (2.8)$$

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 (2.8) can be composed from the matrices W and M by concatenating the latter two matrices, i.e.,

$$X = [W \quad M] \quad (2.9)$$

2.6 Parameter Estimation Using Least Squares

The method of parameter estimation is explained using the simpler case of a regression model. That means both the predictors and the response variables are on a continuous scale. As a further simplification, we assume that there is only one predictor variable and one response variable. The predictor variable is called x and the response variable is called y . The model is still the same as shown in (2.8). The matrix X has just one column with the measured values of the predictor variable and b is just a scalar unknown parameter. The vector y contains the observed values for the response values.

The goal of the analysis of the simple dataset is to find an estimate of the scalar b such that the linear combination of X and b best explains the values in y . How we can find such an estimation procedure that allows us to calculate an estimate of b is explained using a small example dataset in the following subsection.

2.6.1 An Example Dataset

A widely use example dataset for such a simple regression analysis in animal breeding consists of measurements of **body weight** (BW) and **breast circumference** (BC) for a given group of animals.

Table 2.2: Dataset for Regression of Body Weight on Breast Circumference for ten Animals

Animal	Breast Circumference	Body Weight
1	176	471
2	177	463
3	178	481
4	179	470
5	179	496
6	180	491
7	181	518
8	182	511
9	183	510
10	184	541

The dataset shown above is taken from Table 9.1 in [Essl, 1987]. One of the possible reasons for fitting a regression from BW on BC is that the latter is easier to measure. The measured values of BC can be used to predict BW once we have determined the regression coefficient. For this prediction, we use BW as response variable y and BC as predictor variable x . This leads to the regression model

$$y = x * b + \epsilon \quad (2.10)$$

where y is the vector of body weights and x is the vector of breast circumferences. b is a scalar value which is unknown and ϵ is the vector of random unknown error terms. The goal is to determine b such that the predictor variable best explains the response variable. How b is determined is explained with the following plot.

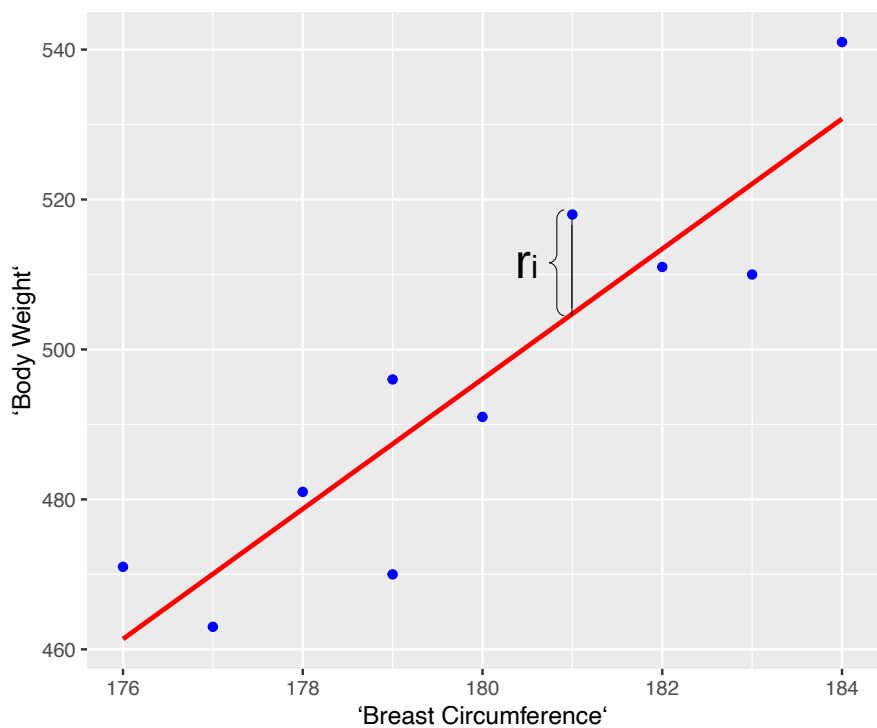


Figure 2.2: Regression of Body Weight On Breast Circumference

In Figure 2.2 the blue points correspond to the data points given by the dataset shown in Table 2.2. The red line corresponds to the regression line defined by the unknown regression parameter b . The distance between the data points to the projection in the direction of the y -axis corresponds to the residual r . For a given data point i , the residual r_i is computed as

$$r_i = y_i - x_i * \hat{b} \quad (2.11)$$

where \hat{b} denotes a concrete estimated value of b . For a different choice of a value of \hat{b} , different values for the residuals r_i can be computed. Our goal is to find the value of \hat{b} that results in the smallest residuals r_i . In order to avoid cancellation of positive and negative values of the residuals, the r_i values are squared and added. This sum of the squared residuals is used as a measure of how good a given regression line determined by \hat{b} fits a given set of data points. Because we want to have a good fit this means that the sum of the squared residuals should be as small as possible.

The method that determines \hat{b} such that the sum of the squared residuals is minimal is called **Least Squares**. In a general formula with more than one predictor variables we can write the least squares estimate \hat{b}_{LS} as

$$\hat{b}_{LS} = \operatorname{argmin}_b \|y - Xb\|^2 \quad (2.12)$$

where $\|\cdot\|$ denotes the Euclidean norm. The estimate \hat{b}_{LS} can be found by finding the minimum of $\|y - Xb\|^2$. The minimum of $\|y - Xb\|^2$ is found by first taking the derivative with respect to b and the setting that derivative to 0. The derivative of $\|y - Xb\|^2$ with respect to b can be computed by starting with the derivation of the least squares criterium LS as follows

$$LS = \|y - Xb\|^2 = (y - Xb)^T (y - Xb) = y^T y - y^T Xb - b^T X^T y + b^T X^T Xb \quad (2.13)$$

The derivative of LS with respect to b corresponds to the gradient vector of LS with respect to b which contains the derivative of LS with respect to every component b_i of the vector b . This gradient vector is

$$\frac{\partial LS}{\partial b} = -X^T y - X^T y + 2 * X^T Xb \quad (2.14)$$

The minimum is found by setting $\frac{\partial LS}{\partial b}$ to 0 which is satisfied by a specific vector called \hat{b} .

$$\frac{\partial LS}{\partial b} = -X^T y - X^T y + 2 * X^T X\hat{b} = 0 \quad (2.15)$$

From equation (2.15), we get the so-called least squares **Normal Equations** for \hat{b} .

$$X^T X\hat{b} = X^T y \quad (2.16)$$

For a regression model, we know that X has full column rank³. That means we can solve the normal equations (2.16) explicitly for \hat{b} .

$$\hat{b} = (X^T X)^{-1} X^T y \quad (2.17)$$

Equation (2.17) presents a solution to the estimation problem of the unknown parameter b in the regression problem. There is one additional unknown parameter that we have not mentioned so far. The regression model contains the random error terms ϵ . Because ϵ is random, we have to specify the expected value and the variance. The error terms are deviations of the predicted values from the observed data points. Hence the expected values $E[\epsilon]$ must be 0. The variance σ^2 of the error terms is an additional unknown parameter that has to be estimated from the data. One way of estimating the error variance from the data is shown in subsection 2.6.2.

2.6.2 Variance of Errors

The least squares procedure itself does not yield an estimate of the error variance σ^2 . But the estimate of σ^2 based on the residuals is often declared to be the **least squares estimate** of σ^2 . The residuals r_i as defined in (2.11) are estimates of the error terms ϵ_i . As a matter of fact the residuals can be used to estimate σ^2 . This estimate is given by

$$\widehat{\sigma^2} = \frac{1}{n-p} \sum_{i=1}^n r_i^2 \quad (2.18)$$

The factor $(n-p)^{-1}$ in (2.18) is used, because it leads the estimate $\widehat{\sigma^2}$ to be unbiased, which means $E[\widehat{\sigma^2}] = \sigma^2$.

2.7 Different Types of Linear Regressions

2.7.1 Regression Through The Origin

The regression model as it was proposed in (2.10) for the dataset of body weight and breast circumference defines a line in the $x - y$ -plane. This line shown in Figure 2.2. What is not shown in the plot, but what becomes clear from the model is that the regression line goes through the origin of the coordinate system. Mathematically the origin is given by $x = 0$ and $y = 0$. In this regression model, the origin is the fixed point which is on the regression line. The fixed point together with the estimated regression coefficient \hat{b} uniquely define

³In a regression model, all values in the matrix X are real values. Hence no column of X will be a linear combination of any other columns and therefore X has full column rank.

the regression line. From a geometrical point of view the estimated regression coefficient defines the slope of the regression line.

2.7.2 Regression With Intercept

Depending on the data analysed with a regression model, it does not make sense to force the regression line to run through the origin. This can be avoided by including an additional fixed term in the regression model. This term is called the **intercept**. A regression model with an intercept can be written as

$$y_i = b_0 + x_i * b_1 + \epsilon_i \quad (2.19)$$

The term b_0 corresponds to the value of the response variable y when the value of the predictor x is 0. Then the fixed point of the regression line is no longer the origin, but the point $x = 0$ and $y = \widehat{b}_0$. The slope of the regression line is determined by \widehat{b}_1 . In matrix-vector notation the intercept b_0 is added to the vector of unknown parameters b and the design-matrix X has to be augmented by a column of all ones on the left.

2.7.3 Regression With Transformed Predictor Variables

Regression models can also contain different transformations of the predictor variables. As an example, we can include any higher order polynomial functions of predictor variables such as

$$y_i = b_0 + b_1 * x_i + b_2 * x_i^2 + \dots + b_k * x_i^k + \epsilon_i \quad (2.20)$$

Although the model (2.20) contains non-linear functions of the predictors x_i , the function is still linear in the unknown parameters b_j ($j = 1, \dots, k$) and hence the model (2.20) is still a linear regression model.

Transformations of the predictor variables are not restricted to polynomial functions. Many different kinds of transformations are possible. An example is shown in the following equation

$$y_i = b_0 + b_1 * \log(x_i) + b_2 * \sin(\pi x_i) + \epsilon_i \quad (2.21)$$

2.8 Predictions

One goal of estimating the regression coefficient was that we want to be able to predict the response based on concrete values of the predictor variables. For our example with the body weight and the breast circumference, this means that

we want to measure the breast circumference of an animal for which we do not know the body weight. Then based on the estimated regression coefficient, we want to be able to predict the body weight of that animal.

The computation of the regression coefficient for the dataset shown in Table 2.2 will be the topic of an exercise. But let us assume that we have computed the value of \hat{b} , then the predicted value of the body weight \hat{y}_s for an animal s is computed based on the measured breast circumference x_s of animal s as follows

$$\hat{y}_s = \hat{b} * x_s \quad (2.22)$$

It has to be noted that the prediction \hat{y}_s is only valid, if the measured value x_s is close to the measured predictors that were used to estimate \hat{b} . For our example with body weight and breast circumference, we could not use the same regression line to predict the body weight for calves, if \hat{b} was estimated with data of adult bulls.

2.9 Regression On Dummy Variables

In a regression model (such as shown in (2.10)) both the response variable and the predictor variables are continuous variables. Examples of such variables are **body weight** and **breast circumference** which are both measured and the measurements are expressed as real numbers. In contrast to such a regression model, the statistical model shown in (2.6) has a continuous response, but the predictor variables are discrete variables. The predictor variables are assumed to be genotypes of a certain set of SNP genotypes and hence these genotypes can only have a fixed number of states. Under the assumption of bi-allelic Loci, a SNP locus can have just three genotypes and hence the predictor variable that is used to represent any given SNP-locus can only take three discrete states.

Figure 2.3 shows the difference between a regression model as the one of **body weight** on **breast circumference** and a fixed linear effects model where one locus has an effect on a quantitative trait. In the left diagram of Figure 2.3 the red line denotes the regression line. This line is meaningful because on the x-axis and on the y-axis every single point of the red line would be valid observations. On the x-axis of the diagram on the righthand side, only three values are possible. In the diagram they are shown as Genotypes G_1G_1 , G_1G_2 and G_2G_2 . We will see very soon that in our statistical model, they will be encoded by 1, 0 and -1 . The response variable in the diagram on the right of Figure 2.3 is a continuous random variable, similarly to the regression model shown in the left diagram. This combination of continuous response variable on a discrete type of variable lead to the term **regression on dummy variables** because the predictor variables are not continuous but just discrete levels of a certain factor. In this lecture, we are using **fixed linear effects model** rather than regression on dummy variables for the same type of model. The

term of fixed linear effects model was used, because in the next chapter in Genomic BLUP we are going to introduce mixed linear effects model which are an extension of the fixed linear effects model used in this chapter.

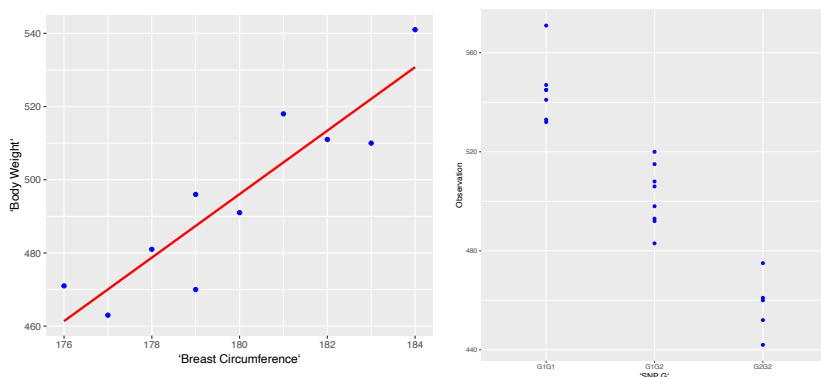


Figure 2.3: Comparison Between Regression Model And Fixed Linear Effects Model With An SNP-Locus As A Discrete Predictor Variables

2.9.1 Fixed Linear Effects Model For SNP Data

We are using genetic data and assume that the SNP genotypes have an effect on a quantitative trait. Our goal is to predict genomic breeding values based on the information from the SNP genotypes for the quantitative traits. We have seen that under some simplifying assumptions of additivity of the genetic effects, the genomic breeding values depend on the absolute value of the genotypic values (a values) of the homozygous SNP genotypes. Hence all we need to know from our analysis of the data under a fixed linear effects model are the a values for each SNP locus. The decomposition of the phenotypic observation shown in 2.4.1 under the assumed genetic model tells us that the phenotypic observation can be explained as a linear function of the genotypic values of the SNP genotypes plus a random error term. The fact that our genetic model is a fixed linear effects model that uses phenotypic observations as response and SNP loci as predictors allows us to set up the following model for an example data set shown in the following subsection.

2.9.2 Example Data Set With SNP Loci And A Phenotypic Observation

We are using the dataset shown in Table 2.3 as an example on how to use a fixed linear effects model to estimate the genotypic value of the SNP genotypes.

Table 2.3: Animals With Two SNP Loci Affecting A Quantitative Trait

Animal	SNP G	Genotypic Value G	SNP H	Genotypic Value H	Observation
1	G_1G_1	a_G	H_1H_2	0	510
2	G_1G_2	0	H_1H_1	a_H	528
3	G_1G_2	0	H_1H_1	a_H	505
4	G_1G_1	a_G	H_2H_2	$-a_H$	539
5	G_1G_1	a_G	H_1H_1	a_H	530
6	G_1G_2	0	H_1H_2	0	489
7	G_1G_2	0	H_2H_2	$-a_H$	486
8	G_2G_2	$-a_G$	H_1H_1	a_H	485
9	G_1G_2	0	H_2H_2	$-a_H$	478
10	G_2G_2	$-a_G$	H_1H_2	0	479
11	G_1G_1	a_G	H_1H_2	0	520
12	G_1G_1	a_G	H_1H_1	a_H	521
13	G_2G_2	$-a_G$	H_1H_2	0	473
14	G_2G_2	$-a_G$	H_1H_2	0	457
15	G_1G_2	0	H_1H_1	a_H	497
16	G_1G_2	0	H_1H_2	0	516
17	G_1G_1	a_G	H_1H_2	0	524
18	G_1G_1	a_G	H_1H_2	0	502
19	G_1G_1	a_G	H_2H_2	$-a_H$	508
20	G_1G_2	0	H_1H_2	0	506

Instead of fitting individual effects for the different SNP genotypes to explain the response variable, we are directly including the genotypic values a_G and a_H into the fixed effects linear model. How the genotypic values are related to the SNP genotypes is also shown in Table 2.3. For all animals in Table 2.3, we can write the model equations in matrix-vector notation as

$$y = Xb + \epsilon \quad (2.23)$$

where y is the vector of observations, b is a vector of genotypic values plus an intercept, X is a design matrix linking the elements in b to y and ϵ is a vector of random errors. Writing out the matrices and vectors leads to

$$\begin{bmatrix} 510 \\ 528 \\ 505 \\ 539 \\ 530 \\ 489 \\ 486 \\ 485 \\ 478 \\ 479 \\ 520 \\ 521 \\ 473 \\ 457 \\ 497 \\ 516 \\ 524 \\ 502 \\ 508 \\ 506 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & -1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & -1 \\ 1 & -1 & 1 \\ 1 & 0 & -1 \\ 1 & -1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & -1 & 0 \\ 1 & -1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & -1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} b_0 \\ a_G \\ a_H \end{bmatrix} + \epsilon \quad (2.24)$$

2.9.3 Parameter Estimation In A Fixed Linear Effects Model

The goal for model (2.23) is to get an estimate for the unknown parameters b_0 , a_G and a_H . In section 2.6 we saw how unknown parameters can be estimated for a regression model using least squares. When applying the least squares method, we did not make any assumptions about the predictor variables. The minimization of the sum of the squared residuals can also be applied for the fixed linear effects model. This minimization leads to the same normal equations

$$X^T X b^{(0)} = X^T y \quad (2.25)$$

So far everything was identical to the case of the regression model. But when trying to find a solution for (2.25) we have to account for the different nature of the design matrix X . In the regression model this matrix X contains real numbers. In our example of a fixed linear effects model, the matrix X just contains just the three number -1 , 0 and 1^4 . The fact that the matrix X contains only a few discrete values makes it very likely that X does not have full column rank. That means it is very likely that some columns of X can be expressed as linear combinations of other columns. This linear dependence of the columns of X causes the matrix $X^T X$ to be singular and hence the inverse of $X^T X$ cannot be computed. Whenever the matrix $X^T X$ is singular, the solution given in (2.17) cannot be computed.

The normal equations in (2.25) are written with the symbol $b^{(0)}$ to denote that the equations do not have a single solution $b^{(0)}$ in the sense that we were able to compute them in the case of the regression model. In the case where $X^T X$ is singular, there are infinitely many solutions $b^{(0)}$. These solutions can be expressed as

$$b^{(0)} = (X^T X)^- X^T y \quad (2.26)$$

where $(X^T X)^-$ stands for a **generalized inverse** of the matrix $X^T X$.

2.9.4 Generalized Inverse Matrices

A generalized inverse matrix G of a given matrix A is defined as the matrix that satisfies the equation $AGA = A$. The matrix G is not unique. Applying the concept of a generalized inverse to a system of equations $Ax = y$, it can be shown that $x = Gy$ is a solution, if G is a generalized inverse of A . Because G is not unique, there are infinitely many solutions corresponding to $\tilde{x} = Gy + (GA - I)z$ where z can be an arbitrary vector of consistent length. Applying these statements concerning generalized inverses and solutions to systems of equations to (2.26), it means that $b^{(0)}$ is not a unique solution to (2.25) because the generalized inverse $(X^T X)^-$ is not unique. As a consequence of that non-uniqueness, the solution $b^{(0)}$ cannot be used as an estimate of the unknown parameter vector b .

2.9.5 Estimable Functions

The numeric solution of the analysis of the example dataset given in Table 2.3 is the topic of an exercise. When developing that solution, we will see that some linear functions of $b^{(0)}$ can be found which do not depend on the choice of the generalized inverse $(X^T X)^-$. Such functions are called **estimable functions**

⁴In most other fixed linear effects models, the design matrix contains just 0 and 1.

Table 2.4: Example Showing Estimable Functions

Animal	SNP G	SNP H	SNP I	Observation
6	G_1G_1	H_1H_2	I_1I_2	16
6	G_1G_1	H_1H_2	I_1I_2	10
6	G_1G_1	H_1H_2	I_1I_2	19
6	G_1G_2	H_1H_1	I_1I_2	11
6	G_1G_2	H_1H_1	I_1I_2	13
6	G_1G_2	H_1H_2	I_1I_1	27

and can be used as estimates for the respective functions of the unknown parameter vector b . The idea of estimable functions can be demonstrated with the following example.

Let us assume that we have a small data set of 6 animals with observations in a particular traits and given genotypes at three loci. The dataset for that example is given in Table 2.4.

As shown before, we want to estimate the marker effects at the three loci G , H and I . This can be done with the following fixed effects model.

$$y = Xb + e$$

with

$$y = \begin{bmatrix} 16 \\ 10 \\ 19 \\ 11 \\ 13 \\ 27 \end{bmatrix}, X = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \text{ and } b = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix}$$

The vector b of unknown parameters consist of the general mean μ and the three marker effects α_1 , α_2 and α_3 . Based on the above information, the normal equations can be written as

$$\begin{bmatrix} 6 & 3 & 2 & 1 \\ 3 & 3 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu^0 \\ \alpha_1^0 \\ \alpha_2^0 \\ \alpha_3^0 \end{bmatrix} = \begin{bmatrix} 96 \\ 45 \\ 24 \\ 27 \end{bmatrix}$$

The above equations have infinitely many solutions. Four of them are shown below in Table 2.5.

Table 2.5: Solution of Normal Equations

Elements of Solution	b_1^0	b_2^0	b_3^0	b_4^0
μ^0	16	14	27	-2982
α_1^0	-1	1	-12	2997
α_2^0	-4	-2	-15	2994
α_3^0	11	13	0	3009

Table 2.6: Estimates of Estimable Functions

Linear Function	b_1^0	b_2^0	b_3^0	b_4^0
$\alpha_1^0 - \alpha_2^0$	3.0	3.0	3.0	3.0
$\mu^0 + \alpha_1^0$	15.0	15.0	15.0	15.0
$\mu^0 + 1/2(\alpha_2^0 + \alpha_3^0)$	19.5	19.5	19.5	19.5

The differences between the same elements in the four numerical solutions make it clear why no solution b^0 can be used as estimates for the unknown parameters in b .

This problem can be addressed, if we are not considering the single elements of a solution vector b^0 , but linear functions of these elements. Examples of such linear functions are shown in Table 2.6.

The values of the expressions shown in Table 2.6 are invariant to whatever solution b^0 is selected. Because this invariance statement is true for all solutions b^0 , these functions are of special interest which corresponds to

- $\alpha_1^0 - \alpha_2^0$: estimate of the difference between marker effect of locus G and locus H
- $\mu^0 + \alpha_1^0$: estimate of the general mean plus the marker effect of locus G
- $\mu^0 + 1/2(\alpha_2^0 + \alpha_3^0)$: estimate of the general mean plus mean effect of loci H and I

2.9.5.1 Definition of Estimable Functions

In summary the underlying idea of estimable functions are that they are linear functions of the parameters b that do not depend on the numerical solutions b^0 of the normal equations. Because estimable functions are functions of the parameters b , they can be expressed as $q^T b$ where q^T is a row vector. In a more formal way estimable functions can be described by the following definition.

Definition 2.2 (Estimable Function). A (linear) function of the parameters b is defined as **estimable**, if it is identically equal to some linear function of the expected value of the vector of observations y .

This means the linear function $q^T b$ is estimable, if

$$q^T b = t^T E(y)$$

for some vector t . That means, if there exists a vector t , such that $t^T E(y) = q^T b$, then $q^T b$ is said to be estimable. For our example shown in Table 2.4, the expected value of the observations of all animals with a genotype $G_1 G_1$ is obtained by

$$E(y_{G_1 G_1}) = \mu + \alpha_1$$

with $t^T = [1 \ 1 \ 1 \ 0 \ 0 \ 0]$ and $q^T = [1 \ 1 \ 0 \ 0]$

2.9.6 Properties of Estimable Functions

Among the many properties we are here just listing the ones that are considered important. The complete list of properties can be found in [Searle, 1971].

- *Form of estimable function.* If $q^T b$ is estimable, then $q^T b = t^T E(y)$ for some t . By definition $E(y) = Xb$ and therefore, $q^T b = t^T Xb$. Because estimability is not a concept that depends on b , this result is true for all values of b . Therefore

$$q^t = t^T X$$

for some vector t .

- *Invariance to solutions b^0 .* If $q^T b$ is estimable, the linear function $q^T b^0$ is invariance to whatever solution of the normal equation

$$X^T X b^0 = X^T y$$

is used for b^0 . This is because

$$q^T b^0 = t^T X b^0 = t^T X G X^T y$$

where G is a generalized inverse of $X^T X$ and $X G X^T$ is invariant to G which means that it is the same for any choice of G .

2.9.6.1 Testing for Estimability

A given function $q^T b$ is estimable, if some vector t can be found, such that $t^T X = q^T$. For a known value of q , it might not be easy to find a vector t satisfying $t^T X = q^T$. Alternatively to finding a vector t , estimability of $q^T b$ can also be investigated by seeing whether q has the property that

$$q^T H = q^T$$

with $H = GX^T X$. A proof of that can be found in [Searle, 1971].