# Chapter 2

# Fixed Linear Effects Models

## 2.1 Other Resources

This chapter is based on the work of [Buehlmann and Maechler, 2014]. Apart from 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` (FLEM) 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 1.4) or by a single step approach (see section 1.5). At the moment, we assume that we are in the first step of the two step approach where we estimate the $a$ effects 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 and with respect to the proposed model to analyse the data.

## 2.3 Data

As already mentioned in section 2.2, we are assuming that we have a perfect data set 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 just have a map of three SNP markers. The marker loci are called $G$, $H$ and

$I$. Each of the markers has just two alleles. Figure 2.1 tries to illustrate the structure of a dataset used to estimate GBV.

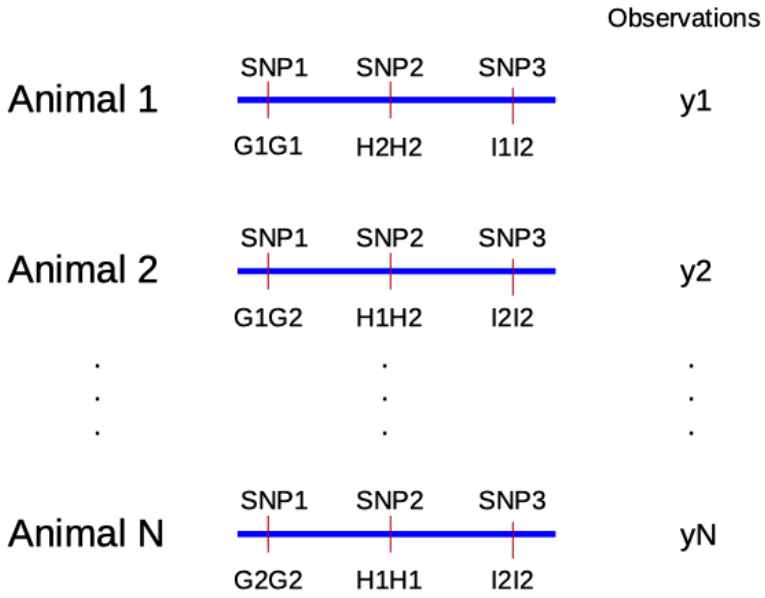


Figure 2.1: Structure of Dataset To Estimate GBV

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 $y_i \quad (i = 1, \cdot, N)$. 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 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 at some point being 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 underlying phenomena.

### 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 $\beta$. The unknown environmental random part is usually called $\epsilon$. 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 loci[1] that are assumed to have an effect on $y$. Ignoring any interaction effects between the three loci, we can decompose the overall genetic effect $g$ into the some of the genotypic values of each locus. Hence

$$g = \sum_{j=1}^{k} g_j \tag{2.3}$$

[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.

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 1.4). 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

$$g = M \cdot a \tag{2.4}$$

where $M$ is an indicator matrix taking values of $-1$, 0 and 1 depending on the SNP marker genotype and $a$ is a vector of $a$ values. Combining the decomposition in (2.4) together with the basic genetic model in (2.2), we get

$$y = \beta + M \cdot a + \epsilon \tag{2.5}$$

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 ($\beta$ and $\epsilon$). The $a$ values are unknown an 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]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 $3 * 10^7$.

### 2.4.2 Statistical Model

When looking at the fundamental decomposition given in the statistical 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.

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

$$y_i = W_i \cdot \beta + M_i \cdot a + \epsilon_i \tag{2.6}$$

where

- $y_i$ is the observation of animal $i$
- $\beta$ is a vector of unknown systematic environmental effects
- $W_i$ is an indicator row vector linking $\beta$ 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
- $\epsilon_i$ 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 $\epsilon$ 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 $\beta$ and $a$ in the model (2.6), it is useful to combine $\beta$ and $a$ into a single vector of unknown parameters and we call it $b$.

$$b = \begin{bmatrix} \beta \\ a \end{bmatrix} \tag{2.7}$$

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

$$y = Xb + \epsilon \tag{2.8}$$

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 $y$
- $\epsilon$ 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.,

$$X = \begin{bmatrix} W & M \end{bmatrix} \tag{2.9}$$

## 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 data set 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

$$y = x * b + \epsilon \tag{2.10}$$

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 $\epsilon$ 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.

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

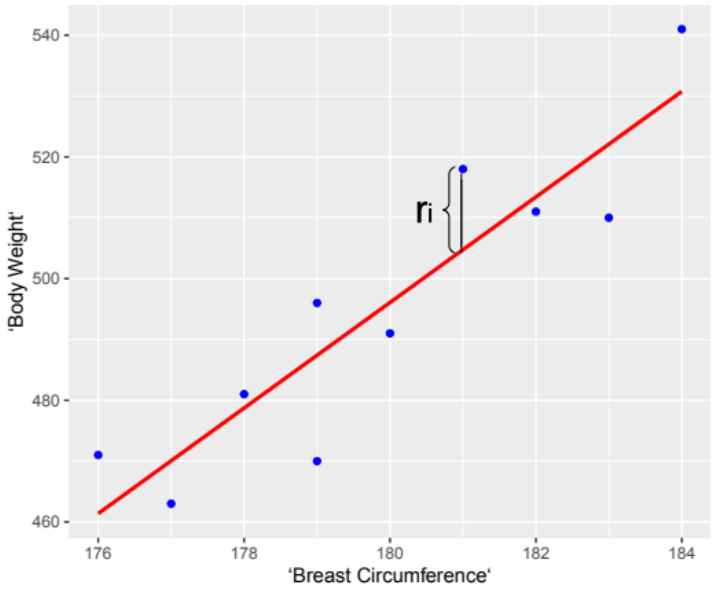


Figure 2.2: Regression of Body Weight On Breast Circumference

$$r_i = y_i - x_i * \hat{b} \tag{2.11}$$

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

$$\hat{b}_{LS} = argmin_b ||y - Xb||^2 \tag{2.12}$$

where $||.||$ 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 as follows

$$LS = ||y-Xb||^2 = (y-Xb)^T(y-Xb) = y^Ty-y^TXb-b^TX^Ty+b^TX^TXb \tag{2.13}$$

The derivative of $LS$ with respect to $b$ is

$$\frac{\partial LS}{\partial b} = -y^TX - y^TX + 2 * b^TX^TX \tag{2.14}$$

The minimum is found by setting $\frac{\partial LS}{\partial b}$ to 0.

$$\frac{\partial LS}{\partial b} = -y^TX - y^TX + 2 * \hat{b}^TX^TX = 0 \tag{2.15}$$

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

$$X^TX\hat{b} = X^Ty \tag{2.16}$$

For a regression model, we know that $X$ has full column rank[3]. That means we can solve the normal equations (2.16) explicitly for $\hat{b}$.

[3]In 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.

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

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 $\epsilon$. Because $\epsilon$ 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 $\sigma^2$ 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 $\sigma^2$. But the estimate of $\sigma^2$ based on the residuals is often declared to be the `least squares estimate` of $\sigma^2$. The residuals $r_i$ as defined in (2.11) are estimates of the error terms $\epsilon_i$. As a matter of fact the residuals can be used to estimate $\sigma^2$. This estimate is given by

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

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

## 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 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

$$y_i = b_0 + x_i * b_1 + \epsilon_i \tag{2.19}$$

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

$$y_i = b_0 + b_1 * x_i + b_2 * x_i^2 + \cdots + b_k * x_i^k + \epsilon_i \tag{2.20}$$

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

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

## 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 $\widehat{y_s}$ for an animal $s$ is computed based on the measured breast circumference $x_s$ of animal $s$ as follows

$$\widehat{y_s} = \hat{b} * x_s \tag{2.22}$$

It has to be noted that the prediction $\widehat{y_s}$ 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.