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 Beef Cattle 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

Table 3.1: Extended Dataset on Body Weight for Beef Cattle Animals

Animal	Breast Circumference	Body Weight	BCS	HEI	Breed
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 Section 3.4. 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 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 using a marker-effect model. 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 this type of data.

3.3.1 Data

As already mentioned in Section 3.3, we are assuming that each animal i has a phenotypic observation y_i for a given trait of interest. Furthermore, every animal has a genotype consisting 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 such a dataset used to estimate marker effects for the three SNP.

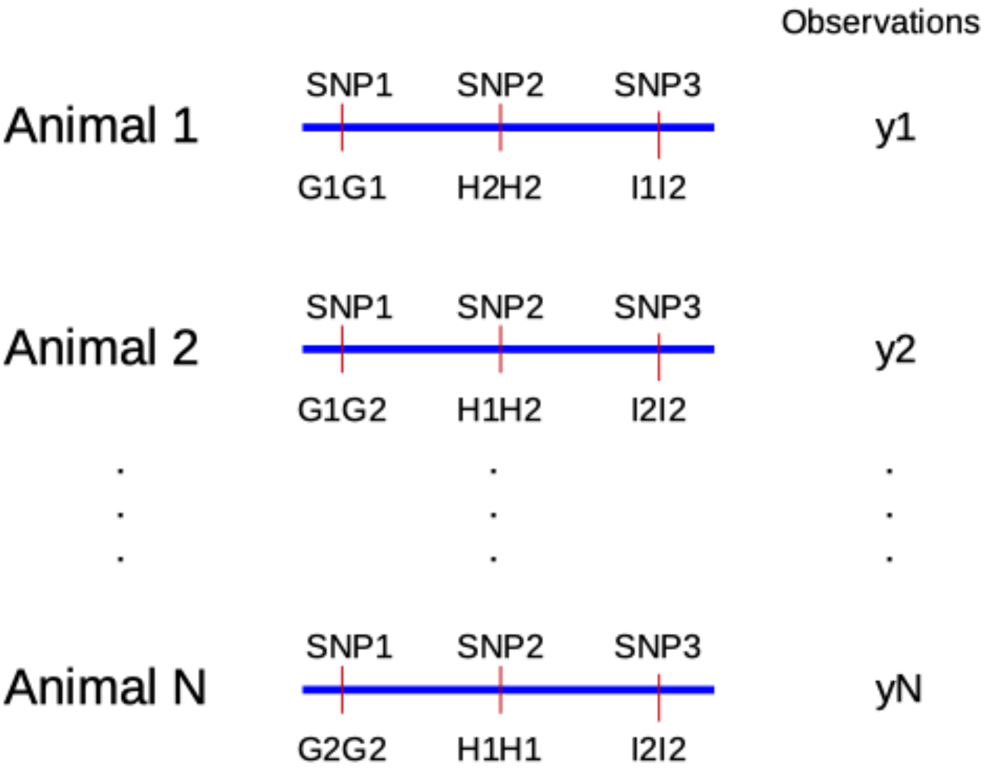

Figure 3.1: Structure of Dataset To Estimate GBV

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 $y_i \quad (i = 1, \dots, N)$. 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. Hence, in principle the marker genotypes occur in discrete levels such as the breed of an animal from dataset shown in Table 3.1. Because we are interested in the maker-effect at each locus and the relationships shown in equation 3.3 which are imposed by the use of a linear regression model on the discrete genotype variables, contain the marker effects, the regression model can be used for the analysis of genomic data. More details about the model will follow in section Section 3.3.2.

3.3.2 Model

The goal of our data analysis using the dataset described in section Section 3.3.1 is to come up with estimates for maker effects at each SNP locus. The marker effects can be used to predict 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 seams reasonable to distinguish between two different types of models. On the one hand we need a model that describes the underlying genetic architecture of the observed phenotypic values in our dataset. We are using a so-called **genetic** model to describe the relationship between genetic background and expressed phenotype of interest. On the other hand, we have to be able to get estimates for marker effects and the GBVs which requires a **statistical** model. Only with the latter we are going to be 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.

3.3.3 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. These effects are currently not of interest and are ignored for the moment. To allow for more flexibility, we include a general intercept term μ into the model. The unknown environmental random part is usually called ϵ. This allows to re-write the simple genetic model in Equation 3.4 as

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

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. 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 Equation 3.6, we get

[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 $3 * 10^7$.

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

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. Combining the decomposition in Equation 3.7 together with the basic genetic model in Equation 3.5, we get

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

The result obtained in Equation 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 (μ 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 Section 3.3.4.

3.3.4 Statistical Model

When looking at the fundamental decomposition given in the genetic model presented in Equation 3.8 from a statistics point of view, the model in Equation 3.8 corresponds to a linear model. In a linear model, the response is explained by a linear function of the predictor variables plus a random error term.

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

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

where

- y_i is the observation of animal i
- μ is a general intercept term
- 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

3.3.5 Genomic Regression Analysis

Although, the predictor variables in the model shown in Equation 3.9 are discrete genotypes which can take only three states, namely the three genotypes of a biallelic locus, it is still possible to model such genomic data with a regression model. The reason for this is that the chosen encoding of the three genotypes into values -1, 0 and 1 is biologically meaningful. This can be seen by looking at expectations of different phenotypic values. For reasons of simplicity, we assume that the phenotypic value y is only affected by a single bi-allelic locus G. Furthermore, locus G has a purely additive effect on the observed phenotypic values. Hence the genotypic values of the three genotypes G_1G_1, G_1G_2 and G_2G_2 at locus G are a_G, 0 and $-a_G$, respectively. Hence for three animals with three different genotypes, the model Equation 3.9 can be written as

$$\left.\begin{array}{l}\text{Animal i with genotype } G_1G_1 \\ \text{Animal j with genotype } G_1G_2 \\ \text{Animal k with genotype } G_2G_2\end{array}\right\} \begin{array}{l} y_i = \mu + 1 * a_G + \epsilon_i \\ y_j = \mu + 0 * a_G + \epsilon_j \\ y_k = \mu + (-1) * a_G + \epsilon_k \end{array}$$

From this we can see that the expected values of the phenotypic values can be written as

$$\begin{aligned} E(y_i) &= \mu + a_G \\ E(y_j) &= \mu \\ E(y_k) &= \mu - a_G \end{aligned}$$

The differences between the expectations of the phenotypic values of animals with different genotypes can now be written as

$$E(y_i) - E(y_j) = E(y_j) - E(y_k) = a_G$$

This difference corresponds to the allele substitution effect a_G at locus G. Hence the chosen encoding of the genotypes G_1G_1, G_1G_2 and G_2G_2 as 1, 0 and -1 has an internal biological meaning and the regression coefficient of the observed phenotypic values on the encoded genotypes provides the allele substitution effect.

3.4 Regression On Dummy Variables

In general, both the response variable and the predictor variables of a regression model 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 predictor variable `Breed` in the extended dataset given in

Table 3.1 is a discrete variable. That means, observations of such a variable can only take a certain number of values. These values are determined by the nature of the variable. For our example with the breeds of animals, the observed values can only come from the existing breeds of that species from which the observations were generated.

The discussion of regression on dummy variables is fascilitated by the notioon of **factors** and **levels**. This terminology is adapted from the literature of experimental design. In the study of the influence of an animals breed on its body weight, we are interested in the extent to which each breed is associated to the body weight. Thus we want to see whether a group of animals from a particular breed show specific values for their body weights and whether these values are different from the body weights of animals from a different breed.

The problem of discrete variables not being measureable is acknowledged by the introduction of the terms "factor" and "levels". Hence a discrete variable is referred to as a "factor". The possible values that a factor can take are called "levels". The concept of levels enables us to quantify differences between the effects that different levels of a factor have on a certain response variable. Translating the concept of levels and factors to our extended dataset (Table 3.1) means that the breed of an animal is a "factor" and the different breeds are correspond to the different levels of the factor "breed".

3.4.1 Model

The goal of the model that we are going to develop is to quantify the effect of each level of the factor "breed" on the response variable "body weight". In a first step, all other variables with a potential influence on body weight are ignored. Hence, we are just looking at the possible effect of the breed on body weight. This is done by setting up a regression on three independent variables x_1, x_2 and x_3

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + e_i \tag{3.10}$$

In this context y_i is the body weight of animal i and b_0 and e_i are the intercept and the random error term which were already found in the regression analysis of Chapter 2. Corresponding to the independent variables x_1, x_2 and x_3 are the regression coefficents b_1, b_2 and b_3, respectively. Depending on the definition of the independent variables x, the regression coefficients b will turn out to be terms that lead to estimates of the differences of the effects of the different levels on the response variable.

For the definition of the independent variables x, it is important to note that each animal can only have one breed[3] associated to it. Each level of the factor "breed" is assigned to one of the indendent variables x_1, x_2 or x_3. This assignment is completely arbitrary. The assignment given in Table 3.4 is proposed.

[3] At this point, we assume that all animals are pure-bred. Alternatively, we would interpret crosses as further distinct levels of the factor "breed".

Table 3.4: Assignment of Breeds to Independen Variables

Breed	Independent Variable
Angus	x_1
Limousin	x_2
Simmental	x_3

For a given animal i that is in breed j, the independent variable assigned to breed j is 1 and all other independent variables are set to 0. This means for animal 1 from breed Angus, the variable x_1 is set to 1 and all other variables are set to 0.

For our example shown in Table 3.1 when only looking at body weight as response and breed as a factor, y_{ij} stands for the j^{th} animal with breed-level i. Then with $e_{ij} = y_{ij} - E(y_{ij})$, the model is the same as in Chapter 2, except for the two subscripts and for the ordering the observations according to the levels of the breed factor.

$$
\begin{aligned}
y_{11} &= b_0 + b_1 * 1 + b_2 * 0 + b_3 * 0 + e_{11} \\
y_{12} &= b_0 + b_1 * 1 + b_2 * 0 + b_3 * 0 + e_{12} \\
\cdots &= \cdots \\
y_{33} &= b_0 + b_1 * 0 + b_2 * 0 + b_3 * 1 + e_{33}
\end{aligned}
\tag{3.11}
$$

The system of equations shown in 3.11 can be converted into matrix-vector notation which turns the model in the familiar form

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \tag{3.12}$$

where $\mathbf{y}$ and $\mathbf{e}$ are both vectors of the same length as there are observations in the dataset and are defined the same way as in the regression in Chapter 2. The vector $\mathbf{b}$ contains the intercept as the first component and regression coefficients for each level of the factor "breed" in the model. The matrix $\mathbf{X}$ is called "design matrix" and contains zeros and ones that link the regression coefficients of the appropriate level to the observations.

Analogously to the regression model in Chapter 2 the properties of the components in vector $\mathbf{e}$ of random residuals are such that $E(\mathbf{e}) = \mathbf{0}$ and $var(\mathbf{e}) = I\sigma^2$. Applying the least squares procedure to Equation 3.12 yields the same normal equations

$$\mathbf{X}^T\mathbf{X}\mathbf{b}^{(0)} = \mathbf{X}^T\mathbf{y} \tag{3.13}$$

Due to the definition of the matrix $\mathbf{X}$, it does not have full column rank. Thus the models as shown in Equation 3.12 that contains factors is also referred to as "models not of full rank".

An important consequence of the rank deficiency of the matrix $\mathbf{X}$ is that the inverse $(\mathbf{X}^T\mathbf{X})^{-1}$ of $(\mathbf{X}^T\mathbf{X})$ does not exist. However the use of a generalized inverse of $(\mathbf{X}^T\mathbf{X})$ solutions to the normal equation Equation 3.13 can be found.

3.4.2 Parameter Estimation In Models Not Of Full Rank

The goal of model Equation 3.12 is to get an estimate for the unknown parameters in vector $\mathbf{b}$.

The normal equations in Equation 3.13 are written with the symbol $\mathbf{b}^{(0)}$ to denote that the equations do not have a single solution $\mathbf{b}^{(0)}$ in the sense that we were able to compute them in the case of the regression model. In the case where X^TX is singular, there are infinitely many solutions $\mathbf{b}^{(0)}$. These solutions can be expressed as

$$\mathbf{b}^{(0)} = (\mathbf{X}^T\mathbf{X})^{-}\mathbf{X}^T\mathbf{y} \tag{3.14}$$

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

3.4.3 Generalized Inverse Matrices

A generalized inverse matrix $\mathbf{G}$ of a given matrix $\mathbf{A}$ is defined as the matrix that satisfies the equation $\mathbf{AGA} = \mathbf{A}$. The matrix $\mathbf{G}$ is not unique. Applying the concept of a generalized inverse to a system of equations $\mathbf{Ax} = \mathbf{y}$, it can be shown that $\mathbf{x} = \mathbf{Gy}$ is a solution, if $\mathbf{G}$ is a generalized inverse of $\mathbf{A}$. Because $\mathbf{G}$ is not unique, there are infinitely many solutions corresponding to $\tilde{\mathbf{x}} = \mathbf{Gy} + (\mathbf{GA} - \mathbf{I})\mathbf{z}$ where $\mathbf{z}$ can be an arbitrary vector of consistent length. Applying these statements concerning generalized inverses and solutions to systems of equations to Equation 3.14, it means that $\mathbf{b}^{(0)}$ is not a unique solution to Equation 3.13 because the generalized inverse $(\mathbf{X}^T\mathbf{X})^{-}$ is not unique. As a consequence of that non-uniqueness, the solution $\mathbf{b}^{(0)}$ is not suitable as an estimate of the unknown parameter vector $\mathbf{b}$.

3.4.4 Estimable Functions

The numeric solution of the analysis of the example dataset given in Table 3.1 is the topic of an exercise. When developing that solution, we will see that some linear functions of $\mathbf{b}^{(0)}$ can be found which do not depend on the choice of the generalized inverse $(\mathbf{X}^T\mathbf{X})^{-}$. Such functions are called **estimable functions** and can be used as estimates for the respective functions of the unknown parameter vector $\mathbf{b}$. 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 the breed of the animal as an independent factor. The dataset for that example is given in Table 3.5.

Table 3.5: Example Showing Estimable Functions

Animal	Breed	Observation
1	Angus	16
2	Angus	10
3	Angus	19
4	Simmental	11
5	Simmental	13
6	Limousin	27

As shown before, we want to estimate the effect of the breed on the observation. This can be done with the following fixed effects model.

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

with

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

The vector $\mathbf{b}$ of unknown parameters consist of the intercept μ which was previously called b_0 and the three breed 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 3.6.

Table 3.6: Solution of Normal Equations

Elements of Solution	b_1^0	b_2^0	b_3^0	b_4^0
μ^0	14	15.5	15.25	1519.5
α_1^0	1	-0.5	-0.25	-1504.5
α_2^0	-2	-3.5	-3.25	-1507.5
α_3^0	13	11.5	11.75	-1492.5

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

This problem can be addressed, if we are not considering the single elements of a solution vector $\mathbf{b}^0$, but linear functions of these elements. Examples of such linear functions are shown in Table 3.7.

Table 3.7: 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 values of the expressions shown in Table 3.7 are invariant to whatever solution b^0 is selected. Because this invariance statement is true for all solutions $\mathbf{b}^0$, these functions are of special interest which corresponds to

- $\alpha_1^0 - \alpha_2^0$: estimate of the difference between breed effects for Angus and Simmental
- $\mu^0 + \alpha_1^0$: estimate of the general mean plus the breed effect of Angus
- $\mu^0 + 1/2(\alpha_2^0 + \alpha_3^0)$: estimate of the general mean plus mean effect of breeds Simmental and Limousin

3.4.4.1 Definition of Estimable Functions

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

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 $\mathbf{q}^T\mathbf{b}$ is estimable, if

$$\mathbf{q}^T\mathbf{b} = \mathbf{t}^T E(\mathbf{y})$$

for some vector $\mathbf{t}$. That means, if there exists a vector $\mathbf{t}$, such that $\mathbf{t}^T E(\mathbf{y}) = \mathbf{q}^T\mathbf{b}$, then $\mathbf{q}^T\mathbf{b}$ is said to be estimable. For our example shown in Table Table 3.5, the expected value of the observations of all animals with breed Angus is obtained by

$$E(y_{1j}) = \mu + \alpha_1$$

with $\mathbf{t}^T = \left[\begin{array}{cccccc} 1 & 1 & 1 & 0 & 0 & 0 \end{array}\right]$ and $\mathbf{q}^T = \left[\begin{array}{cccc} 1 & 1 & 0 & 0 \end{array}\right]$

3.4.5 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).

3.4.5.1 Form of Estimable Function

If $\mathbf{q}^T\mathbf{b}$ is estimable, then $\mathbf{q}^T\mathbf{b} = \mathbf{t}^T E(\mathbf{y})$ for some $\mathbf{t}$. By definition $E(\mathbf{y}) = \mathbf{X}\mathbf{b}$ and therefore, $\mathbf{q}^T\mathbf{b} = \mathbf{t}^T\mathbf{X}\mathbf{b}$. Because estimability is not a concept that depends on $\mathbf{b}$, this result is true for all values of $\mathbf{b}$. Therefore

$$\mathbf{q}^t = \mathbf{t}^T\mathbf{X}$$

for some vector $\mathbf{t}$.

3.4.5.2 Invariance to Solutions $\mathbf{b}^0$

If $\mathbf{q}^T\mathbf{b}$ is estimable, the linear function $\mathbf{q}^T\mathbf{b}^0$ is invariance to whatever solution of the normal equation

$$\mathbf{X}^T\mathbf{X}\mathbf{b}^0 = \mathbf{X}^T\mathbf{y}$$

is used for $\mathbf{b}^0$. This is because

$$\mathbf{q}^T\mathbf{b}^0 = \mathbf{t}^T\mathbf{X}\mathbf{b}^0 = \mathbf{t}^T\mathbf{X}\mathbf{G}\mathbf{X}^T\mathbf{y}$$

where $\mathbf{G}$ is a generalized inverse of $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}\mathbf{G}\mathbf{X}^T$ is invariant to $\mathbf{G}$ which means that it is the same for any choice of $\mathbf{G}$. This can be seen by the definition of G which has to satisfy that

$$\mathbf{X}^T\mathbf{X}\mathbf{G}\mathbf{X}^T\mathbf{X} = \mathbf{X}^T\mathbf{X}$$

in order to be a generalised inverse of $\mathbf{X}^T\mathbf{X}$. Because $\mathbf{X}^T$ is not a null matrix, it follows that $\mathbf{X}\mathbf{G}\mathbf{X}^T\mathbf{X} = \mathbf{X}$. For any other generalised inverse matrix $\mathbf{F}$ of $\mathbf{X}^T\mathbf{X}$, we can write analogeously to above with the generalised inverse $\mathbf{G}$ that $\mathbf{X}^T\mathbf{X}\mathbf{F}\mathbf{X}^T\mathbf{X} = \mathbf{X}^T\mathbf{X}$ which implies that $\mathbf{X}\mathbf{F}\mathbf{X}^T = \mathbf{X}\mathbf{G}\mathbf{X}^T$. This can be shown to be true for any generalised inverse of $\mathbf{X}^T\mathbf{X}$

3.4.5.3 Testing for Estimability

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

$$\mathbf{q}^T\mathbf{H} = \mathbf{q}^T$$

with $\mathbf{H} = \mathbf{G}\mathbf{X}^T\mathbf{X}$. This is proved by the fact that if $\mathbf{q}^T\mathbf{b}$ is estimable, then $\mathbf{q}^T = \mathbf{t}^T\mathbf{X}$ and $\mathbf{q}^T\mathbf{H} = \mathbf{t}^T\mathbf{X}\mathbf{H} = \mathbf{t}^T\mathbf{X}\mathbf{G}\mathbf{X}^T\mathbf{X} = \mathbf{t}^T\mathbf{X} = \mathbf{q}^T$.

3.4.5.4 BLUE of Estimable Function

BLUE stands for Best Linear Unbiased Estimation. The BLUE of the estimable function $\mathbf{q}^T\mathbf{b}$ is $\mathbf{q}^T\mathbf{b}^0$ that is

$$\widehat{\mathbf{q}^T\mathbf{b}} = \mathbf{q}^T\mathbf{b}^0 \tag{3.15}$$

where here the "hat" stands for "BLUE of". For a proof of Equation 3.15, it has to be shown that properties of BLUE hold. The **linearity** holds because $\mathbf{q}^T\mathbf{b}^0$ is a linear function of the observations, because $\mathbf{q}^T\mathbf{b}^0 = \mathbf{q}^T\mathbf{G}\mathbf{X}^T\mathbf{y}$. **Unbiasedness** is checked by inspecting $E(\mathbf{q}^T\mathbf{b}^0)$

$$\begin{aligned}
E(\mathbf{q}^T\mathbf{b}^0) &= \mathbf{q}^T E(\mathbf{b}^0) \\
&= \mathbf{q}^T E(\mathbf{G}\mathbf{X}^T\mathbf{y}) \\
&= \mathbf{q}^T\mathbf{G}\mathbf{X}^T E(\mathbf{y}) \\
&= \mathbf{q}^T\mathbf{G}\mathbf{X}^T\mathbf{X}\mathbf{b} \\
&= \mathbf{q}^T\mathbf{H}\mathbf{b} \\
&= \mathbf{t}^T\mathbf{X}\mathbf{H}\mathbf{b} \\
&= \mathbf{t}^T\mathbf{X}\mathbf{b} \\
&= \mathbf{q}^T\mathbf{b}
\end{aligned}$$

using $\mathbf{X} = \mathbf{X}\mathbf{H} = \mathbf{X}\mathbf{G}\mathbf{X}^T\mathbf{X}$

To show that $\mathbf{q}^T\mathbf{b}^0$ is the best estimator among all linear estimators, we need to show that it has minimum variance. The variance of $\mathbf{q}^T\mathbf{b}^0$ is

$$\begin{aligned}
var(\mathbf{q}^T\mathbf{b}^0) &= \mathbf{q}^T \cdot var(\mathbf{b}^0) \cdot \mathbf{q} \\
&= \mathbf{q}^T \cdot var(\mathbf{G}\mathbf{X}^T\mathbf{y}) \cdot \mathbf{q} \\
&= \mathbf{q}^T\mathbf{G}\mathbf{X}^T \cdot var(\mathbf{y})\mathbf{X}\mathbf{G}^T\mathbf{q} \\
&= \mathbf{q}^T\mathbf{G}\mathbf{X}^T\mathbf{X}\mathbf{G}^T\mathbf{q}\sigma^2 \\
&= \mathbf{q}^T\mathbf{G}\mathbf{X}^T\mathbf{X}\mathbf{G}^T\mathbf{X}^T\mathbf{t}\sigma^2 \\
&= \mathbf{q}^T\mathbf{G}\mathbf{X}^T\mathbf{t}\sigma^2 \\
&= \mathbf{q}^T\mathbf{G}\mathbf{q}\sigma^2 \qquad (3.16)
\end{aligned}$$

with $var(y) = \mathbf{I}\sigma^2$. Suppose $\mathbf{k}^T\mathbf{y}$ is some other linear unbiased estimator of $\mathbf{q}^T\mathbf{b}$ different from $\mathbf{q}^T\mathbf{b}^0$. Because $\mathbf{k}^T\mathbf{y}$ is unbiased $E(\mathbf{k}^T\mathbf{y}) = \mathbf{q}^T\mathbf{b}$ and so $\mathbf{k}^T\mathbf{X} = \mathbf{q}^T$. Therefore,

$$cov(\mathbf{q}^T\mathbf{b}^0, \mathbf{k}^T\mathbf{y}) = cov(\mathbf{q}^T\mathbf{G}\mathbf{X}^T\mathbf{y}, \mathbf{k}^T\mathbf{y}) = \mathbf{q}^T\mathbf{G}\mathbf{X}^T\mathbf{k}\ \sigma^2 = \mathbf{q}^T\mathbf{G}\mathbf{q}\ \sigma^2$$

Looking at the variance of the difference between $\mathbf{q}^T\mathbf{b}^0$ and $\mathbf{k}^T\mathbf{y}$

$$\begin{aligned}
var(\mathbf{q}^T\mathbf{b}^0 - \mathbf{k}^T\mathbf{y}) &= var(\mathbf{q}^T\mathbf{b}^0) + var(\mathbf{k}^T\mathbf{y}) - 2cov(\mathbf{q}^T\mathbf{b}^0, \mathbf{k}^T\mathbf{y}) \\
&= var(\mathbf{k}^T\mathbf{y}) - \mathbf{q}^T\mathbf{G}\mathbf{q}\ \sigma^2 \\
&= var(\mathbf{k}^T\mathbf{y}) - var(\mathbf{q}^T\mathbf{b}^0) > 0
\end{aligned}$$

because a variance has to be positive. Hence $var(\mathbf{k}^T\mathbf{y}) > var(\mathbf{q}^T\mathbf{b}^0)$ which can be shown for any linear unbiased estimator $\mathbf{k}^T\mathbf{y}$ and hence $\mathbf{q}^T\mathbf{b}^0$ is “best”.

3.5 Contrasts

Contrasts are linear combinations of parameters. In R, contrasts are used to determine which estimable functions are used to produce results of a linear model analysis that are shown to a user. Furthermore, the user has the option to choose among different contrasts which are already available by default. It is also possible for the user to create custom made contrasts. This section introduces the basic idea of contrasts and how they are used in R.

Let us go back to our example datasets containing body weight and breed of different animals shown in Table Table 3.8.

```

Attaching package: 'dplyr'


The following objects are masked from 'package:stats':

    filter, lag


The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

Table 3.8: Body Weight and Breed of Beef Cattle Animals

Animal	Body Weight	Breed
1	471	Angus
2	463	Angus
3	481	Simmental
4	470	Angus
5	496	Simmental
6	491	Simmental
7	518	Limousin
8	511	Limousin
9	510	Limousin
10	541	Limousin

3.5.1 Contrasts in R

The contrasts used in R can be seen from the function `contrasts()`. For our example dataset with body weight and breed of animals, we get

```
(mat_ctr <- contrasts(as.factor(tbl_flem_bw_breed$Breed)))
```

```
          Limousin Simmental
Angus            0         0
Limousin         1         0
Simmental        0         1
```

The information in the above shown contrasts matrix reflects the model terms in the columns of the matrix. Hence from the above matrix it can be seen that there are two terms associated with breeds in any linear model that considers breed as a factor. These two terms are Limousin and Simmental. The rows of the above shown contrasts matrix reflect the encoding of the different levels in the dataset. All animals of breed `Angus` are encoded with both zeroes for the two model terms. `Limousin` animals receive a code of 1 for the first model term and a code of 0 for the second term. Animals of breed `Simmental` receive a 0 for the first term and a 1 for the second term. The above contrasts matrix does not show the intercept. The intercept term is implicitly coded as 1 for all animals.

3.5.2 Model Matrix

The assignment of codes to the different data records can also be seen in the model matrix. In R the model matrix is obtained as a result of the function `model.matrix()`. The model matrix that goes together with the above shown contrasts for the factor `Breed` in our dataset is shown below.

```
lm_bw_br <- lm(`Body Weight` ~ Breed, data = tbl_flem_bw_breed)
(mat_X <- model.matrix(lm_bw_br))
```

```
  (Intercept) BreedLimousin BreedSimmental
1           1             0              0
2           1             0              0
3           1             0              1
4           1             0              0
5           1             0              1
6           1             0              1
7           1             1              0
```

```
8             1             1              0
9             1             1              0
10            1             1              0
attr(,"assign")
[1] 0 1 1
attr(,"contrasts")
attr(,"contrasts")$Breed
[1] "contr.treatment"
```

From the above shown model matrix, it can be seen that the encoding contained in the contrasts matrix is applied to the data records.

3.5.3 Estimable Functions

The type of estimable functions that are used in a given linear model analysis can be found by first extending the contrasts matrix by a column of all ones, reflecting the encoding of the intercept term.

```
mat_ctr_ext <- cbind(matrix(c(rep(1, nrow(mat_ctr))), ncol = 1), mat_ctr)
colnames(mat_ctr_ext)[1] <- colnames(mat_X)[1]
mat_ctr_ext
```

```
          (Intercept) Limousin Simmental
Angus               1        0         0
Limousin            1        1         0
Simmental           1        0         1
```

The matrix of estimable functions is obtained by computing the inverse of the extended contrasts matrix

```
(mat_estf <- solve(mat_ctr_ext))
```

```
            Angus Limousin Simmental
(Intercept)     1        0         0
Limousin       -1        1         0
Simmental      -1        0         1
```

Each row of the matrix of estimable functions corresponds to a model term. Each column can be seen as one component of the solution to the least squares normal equation. The estimate of the intercept term corresponds to the solution for the first breed level in the normal equations.

The estimate for the model term `Limousin` corresponds to the difference beween the solution for the second breed level minus the solution of the first breed level. The estimate of the effect of the term `Simmental` is the difference between the last solution and the first breed level.

3.5.4 Validation

The results on the investigated connection between contrasts and estimable functions is validated with our example dataset. For this validation, we first need a set of solutions to the least squares normal equations. As the first step, we set up the design matrix $\mathbf{X}$ and use it to compute the crossproduct $\mathbf{X}^T\mathbf{X}$

```
mat_X <- model.matrix(lm(`Body Weight` ~ 0 + Breed, data = tbl_flem_bw_breed))
mat_X <- cbind(matrix(1, nrow = nrow(tbl_flem_bw_breed), ncol = 1), mat_X)
dimnames(mat_X) <- NULL
mat_xtx <- crossprod(mat_X)
mat_xtx
```

```
     [,1] [,2] [,3] [,4]
[1,]   10    3    4    3
[2,]    3    3    0    0
[3,]    4    0    4    0
[4,]    3    0    0    3
```

The generalized inverse $(\mathbf{X}^T\mathbf{X})^-$ provided by the function `MASS::ginv()` of package `MASS` is used to come up with a solution to the least squares normal equation

$$\mathbf{X}^T\mathbf{X}\mathbf{b}^0 = \mathbf{X}^T\mathbf{y}$$

A solution for $\mathbf{b}^0$ is

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

For our dataset we get

```
vec_y <- tbl_flem_bw_breed$`Body Weight`
mat_xty <- crossprod(mat_X, vec_y)
mat_xtx_ginv <- MASS::ginv(mat_xtx)
mat_b0 <- crossprod(mat_xtx_ginv,mat_xty)
mat_b0
```

```
          [,1]
[1,] 369.33333
[2,]  98.66667
[3,] 150.66667
[4,] 120.00000
```

These solutions are used to construct the effect results computed by the function `lm()` in R. The summary table looks as follows

```
lm_bw_br <- lm(`Body Weight` ~ Breed, data = tbl_flem_bw_breed)
(smry_lm_bw_br <- summary(lm_bw_br))
```

```

Call:
lm(formula = `Body Weight` ~ Breed, data = tbl_flem_bw_breed)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.0000  -7.5000  -0.1667   2.7500  21.0000 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     468.000      6.097  76.758 1.68e-11 ***
BreedLimousin    52.000      8.066   6.447 0.000351 ***
BreedSimmental   21.333      8.623   2.474 0.042575 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.56 on 7 degrees of freedom
Multiple R-squared:  0.8597,	Adjusted R-squared:  0.8196 
F-statistic: 21.44 on 2 and 7 DF,  p-value: 0.001035
```

From the matrix of estimable functions

```
mat_estf
```

```
            Angus Limousin Simmental
(Intercept)     1        0         0
Limousin       -1        1         0
Simmental      -1        0         1
```

we can see that the intercept estimate corresponds to the mean body weight of all Angus animals. Which is

```
library(dplyr)
mean((tbl_flem_bw_breed %>% filter(Breed == "Angus"))$`Body Weight`)
```

```
[1] 468
```

The estimate for the effect BreedLimousin is the difference between the third and the second component in the solution vector $\mathbf{b}^0$

```
mat_b0[3] - mat_b0[2]
```

```
[1] 52
```

Similarly, the estimate for effect BreedSimmental is the difference between the last component of the solution vector and the second component of the solution vector.

```
mat_b0[4] - mat_b0[2]
```

```
[1] 21.33333
```