- ϵ is the vector of random residuals

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

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

3.5 Regression On Dummy Variables

In a regression model 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 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".

3.5.1 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

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + e_i \tag{3.13}$$

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 defintion 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.

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.

$$\begin{aligned} y_{11} &= b_0 + b_1 * 1 + b_2 * 0 + b_3 * 0 + e_{11} \\ y_{12} &= b_0 + b_1 * 1 + b_2 * 0 + b_3 * 0 + e_{12} \\ \cdots &= \cdots \\ y_{33} &= b_0 + b_1 * 0 + b_2 * 0 + b_3 * 1 + e_{33} \end{aligned} \tag{3.14}$$

The system of equations shown in (3.14) can be converted into matrix-vector notation which turns the model in the familiar form

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \tag{3.15}$$

[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".

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 (3.15) yields the same normal equations

$$\mathbf{X}^T\mathbf{X}\mathbf{b}^{(0)} = \mathbf{X}^T\mathbf{y} \tag{3.16}$$

Due to the definition of the matrix $\mathbf{X}$, it does not have full column rank. Thus the models as shown in (3.15) 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 (3.16) can be found.

3.5.2 Parameter Estimation In Models Not Of Full Rank

The goal of model (3.15) is to get an estimate for the unknown parameters in vector $\mathbf{b}$.

The normal equations in (3.16) 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

$$\mathbf{b}^{(0)} = (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{y} \tag{3.17}$$

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

3.5.3 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 (3.17), it means that $\mathbf{b}^{(0)}$ is not a unique solution to (3.16) because

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

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 $\mathbf{b}$.

3.5.4 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.

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

Table 3.6: 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 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

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

The above equations have infinitely many solutions. Four of them are shown below in Table 3.6.

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.

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

3.5.4.1 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.

… {.definition #defestfun name="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 $\mathbf{q}^T\mathbf{b}$ 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 3.5, the expected value of the observations of all animals with breed Angus is obtained by

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

with $\mathbf{t}^T = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}$ and $\mathbf{q}^T = \begin{bmatrix} 1 & 1 & 0 & 0 \end{bmatrix}$

3.5.5 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].

- *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 $\mathbf{t}$.

- *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 $\mathbf{G}$.

3.5.5.1 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 $\mathbf{H} = \mathbf{G}\mathbf{X}^T\mathbf{X}$. A proof of that can be found in [Searle, 1971].