### 5.1.4 Linear Mixed Effects Models

Linear mixed effects models or just *mixed models* are a combination or a merger of fixed linear effects models and random models. That means mixed models contain both fixed effects and random effects. A first example of a dataset which can be modelled with a mixed model is shown in Table 5.3. In this dataset, the factor `Breed` is regarded as a fixed effect whereas the influence of the animal on a single measurement is considered as a random effect. Hence, body weight $y_{ijk}$ which corresponds to repeated observation $k$ of animal $j$ from breed $i$ can be written as

$$y_{ijk} = b_0 + b_i + \alpha_j + e_{ijk} \tag{5.9}$$

where $b_0$ is the intercept, $b_i$ is the fixed effect of breed $i$, $\alpha_j$ is the random effect of animal $j$ and $e_{ijk}$ is the random residual. In matrix-vector notation equation (5.9) takes the form

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\alpha + \mathbf{e} \tag{5.10}$$

where $\mathbf{y}$ is the vector of length $n$ containing responses, $\mathbf{b}$ the vector of length $p$ with covariates, $\alpha$ the vector of length $q$ with random effects related to the repeated observations for an animal and $\mathbf{e}$ is the vector of length $n$ with random residuals. The matrices $\mathbf{X}$ and $\mathbf{Z}$ relate the different effects to the observations.

From equation (5.9), we cannot really tell any difference to a fixed linear effects model. Only with the specification of the distributional properties of all the model components, it becomes clear that equation (5.9) specifies a mixed model. The mixed model is characterized by two random variables

1. a $q$-dimensional vector of random effects represented by the random variable $\alpha^*$
2. a $n$-dimensional vector of responses represented by the random variable $y^*$

The datasets to be analysed with mixed models contain observations, denoted by the vector $\mathbf{y}$. Values $\alpha$ of $\alpha^*$ are not observed and hence are unknown. When specifying the distributional properties of a mixed model, the unconditional distribution of $\alpha^*$ and the conditional distribution of $(y^*|\alpha^*)$ are given. The description of these distributions involve the form of the distribution and the values of the distributional parameters. The observations of the responses and of the covariates are used to estimate these parameters. The unconditional distribution of $\alpha^*$ and the conditional distribution of $(y^*|\alpha^*)$ are both assumed to be multivariate normal distributions.

$$
\begin{aligned}
(y^*|\alpha^*) &\sim \mathcal{N}(\mathbf{Xb} + \mathbf{Z}\alpha, \sigma^2 * I) \\
\alpha^* &\sim \mathcal{N}(\mathbf{0}, \Sigma)
\end{aligned} \tag{5.11}
$$

The analysis of the dataset shown in Table 5.3 can be analysed using the function `lmer()` of package `lme4`.

```
mlem__rep_obs_breed <- lme4::lmer(`Body Weight` ~ Breed + (1|Animal),
                                  data = tbl_rep_obs_breed)
summary(mlem__rep_obs_breed)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: `Body Weight` ~ Breed + (1 | Animal)
##    Data: tbl_rep_obs_breed
##
## REML criterion at convergence: 61.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.3714 -0.5383  0.0640  0.3213  1.4305
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Animal   (Intercept) 426.21   20.645
##  Residual              22.98    4.794
## Number of obs: 12, groups:  Animal, 4
##
## Fixed effects:
##                Estimate Std. Error t value
## (Intercept)      466.59      20.83  22.400
## BreedLimousin     59.44      25.51   2.330
## BreedSimmental    28.30      29.46   0.961
##
## Correlation of Fixed Effects:
##             (Intr) BrdLms
## BreedLimosn -0.816
## BreedSmmntl -0.707  0.577
```

The variance components obtained by `lme4::lmer()` are the same as what we have seen before as results of ANOVA. This is because, we are looking at balanced data.

In livestock breeding, linear mixed effects models are of interest when it comes to the evaluation of the genetic potential of selection candidates. From quantitative genetics, we know that parents with a superior genetic potential produce offspring which are better on average compared to the mean performance of animals from the same generation. The genetic potential of an animal is quantified

by a concept which is referred to as *breeding value*. In what follows, we describe how breeding values for animals can be predicted using mixed models.

## 5.2 Sire Model

In a first application of mixed models for predicting breeding values, observation of daughter performance records were used to predict breeding values for sires. Assuming that sires are unrelated, i.e. they do not share any common ancestors, sire breeding values can be predicted similarly to the analysis of the repeated observations dataset. A possible mixed model for such an analysis might look as follows

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zs} + \mathbf{e} \tag{5.12}$$

where $\mathbf{y}$ is the vector of length $n$ with responses, $\mathbf{b}$ is the vector of length $p$ with fixed effects, $\mathbf{s}$ is the vector of length $q$ with sire breeding values and $\mathbf{e}$ is the vector of length $n$ with random residuals. Matrices $\mathbf{X}$ and $\mathbf{Z}$ are design matrices which relate the observations to the respective effects.

A dataset that can be used to be analysed with a model such as shown in equation (5.12) is given in by the `milk` dataset of the package `pedigreemm`. The first six lines of this dataset are shown in Table 5.5.

Table 5.5: First six lines of milk dataset from package pedigreemm

| id | lact | herd | sire | dim | milk | fat | prot | scs |
|---|---|---|---|---|---|---|---|---|
| 6489 | 1 | 89 | 2 | 286 | 18420 | 784 | 607 | 2.01 |
| 6489 | 2 | 89 | 2 | 305 | 21592 | 954 | 644 | 1.64 |
| 6489 | 3 | 89 | 2 | 203 | 15834 | 779 | 490 | 1.86 |
| 6490 | 1 | 89 | 2 | 281 | 20683 | 785 | 610 | 2.71 |
| 6490 | 2 | 89 | 2 | 277 | 20050 | 841 | 582 | 2.83 |
| 6490 | 3 | 89 | 2 | 289 | 21891 | 884 | 650 | 2.63 |

In Table 5.5 the columns `milk`, `fat`, `prot` and `scs` stand for milk yield, fat yield, protein yield and somatic cell score for a given lactiation of a cow, respectively and can be selected as response variables. The columns `lact`, `herd` and `dim` are lactation, herd number and days in milk, respectively and these columns can be used as fixed effects or covariates. The sire column is used for the random sire breeding value effect in the model.

As already stated, if we assume that these sires are unrelated, the dataset can be analysed as shown above using the function `lme4::lmer`.