4 Mixed Linear Effects Models

Mixed linear effects models are a very useful tool in the analysis of data with some dependencies. In all statistical analyses that we have seen so far the assumption of independence between observations was central. One way of expressing this independence assumption is via the variance-covariance matrix ($var(\mathbf{e})$) of the vector ($\mathbf{e}$) of residuals. In mathematical terms this can be written as

$$var(\mathbf{e}) = \mathbf{I} * \sigma_e^2 \tag{4.1}$$

which means that the variance-covariance matrix ($var(\mathbf{e})$) is proportional to the identity matrix $\mathbf{I}$ with the variance component σ_e^2 as proportionality factor.

In what follows, the models that account for different dependency structures are described.

4.1 Repeated Observations

It is quite common to have repeated observations of the same traits or characteristics from a group of animals. Observing the same characteristic of the same animal multiple times is expected to yield a more accurate description of any relationship between different traits such as body weight and breast circumference. If we apply that line of thought to the example data used in Chapter 2, we would have repeated measurements of breast circumference and body weight of the same animals. Such a dataset is shown in Table 4.1 for a selected number of animals.

Table 4.1: Repeated Observations for Body Weight and Breast Circumference

Animal	Breast Circumference	Body Weight
2	177.0000	463.0000
2	177.3129	468.8940
2	177.3292	467.8753
5	179.0000	496.0000
5	178.6501	495.0033
5	178.7485	493.6563
7	181.0000	518.0000

Table 4.1: Repeated Observations for Body Weight and Breast Circumference

Animal	Breast Circumference	Body Weight
7	180.9819	509.3221
7	181.1467	506.5958
10	184.0000	541.0000
10	184.5957	547.3609
10	183.1749	533.9288

In Table 4.1, the column entitiled `Animal` is no longer a running counter which enumerates the observation records. In this repeated observation dataset, the column `Animal` denotes for which animal the measurements was observed. The association between observations and animals is shown in Figure 4.1.

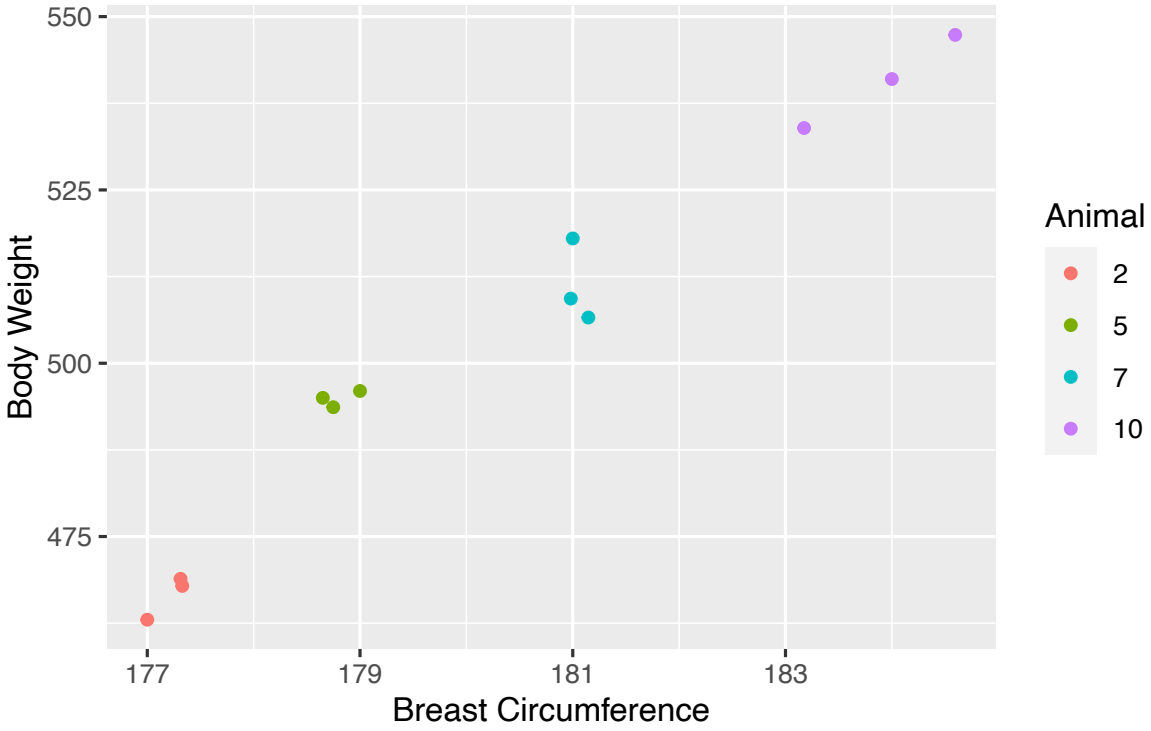

Figure 4.1: Repeated Observations of Breast Circumference and Body Weight

The color codes in Figure 4.1 identify the observations for the same animal. This shows that observations for the same animal tend to be grouped together. This grouping has to be considered in the statistical analysis of such a dataset.

4.1.1 Statistical Analysis

In principle, the dataset shown in Table 4.1 can be analysed with a linear regression model. But from the plot (Figure 4.2) of the residuals versus the fitted values, it becomes clear that the residuals are grouped according to the animals from which the measurement was taken. Due to the small size of the dataset, the grouping effect according to the animal does not show up as clearly as intended. But never the less, this grouping indicates that the assumption of independent residuals is violated.

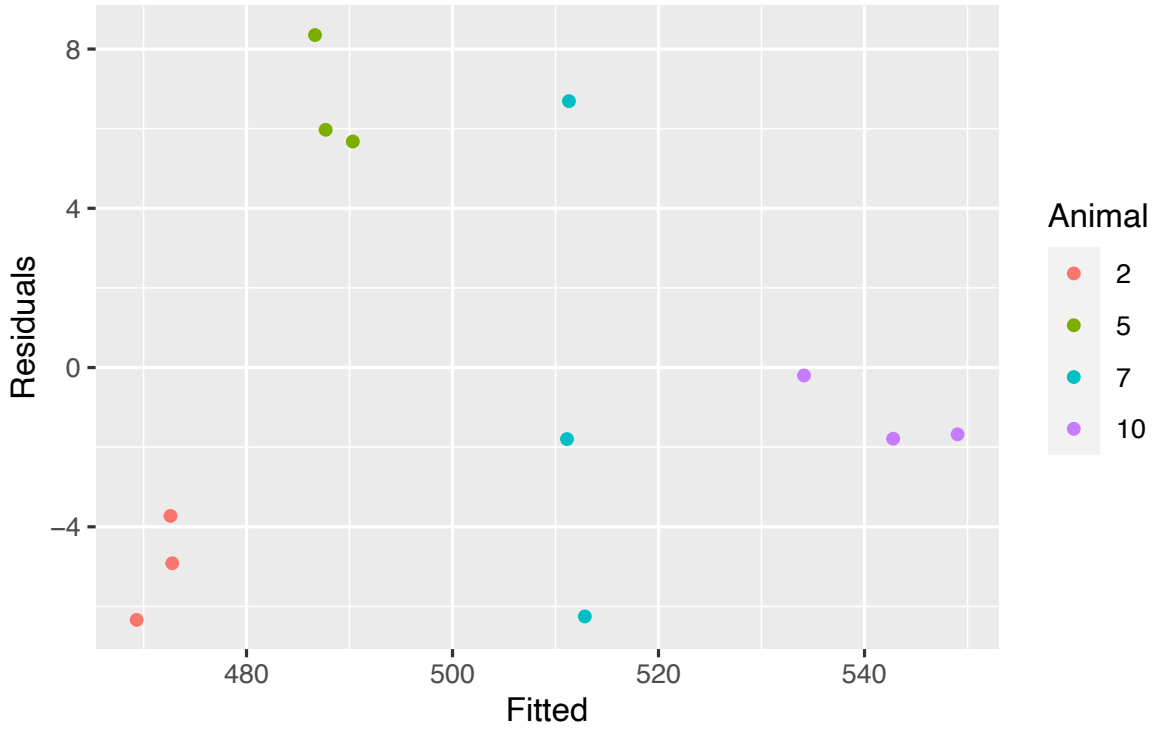

Figure 4.2: Residuals vs. Fitted Values Plot for Linear Regression Model of Repeated Observation Data

4.1.2 Analysis of Variance

Traditionally repeated measurement data have been analyzed using a statistical technique referred to as analysis of variance (ANOVA). ANOVA is a general method that has been used for a long time to assess the variability of different factors in a dataset. This is done by constructing a specific type of table (ANOVA-table) which presents the essential features of a given dataset (see (Shayle R. Searle, Casella, and McCulloch 1992) for more details). The structure and the properties of an ANOVA-table can best be demonstrated by an analysis of

a dataset that shows the influence of the factor breed on body weight of animals shown in Table 4.2.

Table 4.2: Body Weight and Breed for 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

A one-factor analysis of variance of the data shown in Table 4.2 can answer the question, whether the factor `Breed` has an influence on the response variable `Body Weight`. An ANOVA in R can be constructed by the function `aov()` as follows.

```
aov_bw_breed <- aov(`Body Weight` ~ Breed, data = tbl_bw_breed)
(smry_aov_bw_breed <- summary(aov_bw_breed))
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
Breed        2   4783  2391.5   21.44 0.00103 **
Residuals    7    781   111.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The result of the one-way ANOVA of `Body Weight` ond `Breed` shows that it is very unlikely that `Breed` does not have any influence on `Body Weight`. The presented test-statistic from an F-Test is the same that is also shown by the summary results of a result from the `lm()` function. The ANOVA table which is presented by the `summary()` function applied to the `aov`-object contains also an estimate ($\widehat{\sigma_e^2}$) of the residual variance component (σ_e^2). The estimate corresponds to the mean sum of squares for the component `Residuals`. For our dataset the estimate is 111.5. Taking the square root of this value results in the `Residual standard error` shown in the summary output of an `lm()`-analysis.

Extending the dataset shown in Table 4.2 to multiple observations for a selected number of animals results in the dataset given in Table Table 4.3.

Table 4.3: Repeated Observations of Body Weight and Breed for Beef Cattle Animals

Animal	Body Weight	Breed
2	463.0000	Angus
2	468.8940	Angus
2	467.8753	Angus
5	496.0000	Simmental
5	495.0033	Simmental
5	493.6563	Simmental
7	518.0000	Limousin
7	509.3221	Limousin
7	506.5958	Limousin
10	541.0000	Limousin
10	547.3609	Limousin
10	533.9288	Limousin

Applying an ANOVA on the dataset given in Table 4.3 allows to check whether there is variation between measurements of the same animal.

```
aov_bw_breed_rep <- aov(`Body Weight` ~ Breed + Error(Animal), data = tbl_rep_obs_breed)
summary(aov_bw_breed_rep)
```

```
Error: Animal
          Df Sum Sq Mean Sq F value Pr(>F)
Breed      2   7356    3678   2.826  0.388
Residuals  1   1302    1302

Error: Within
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  8  183.8   22.98
```

The above ANOVA results show that taking into account the repeated measurement structure of the data greatly reduces the mean squared residuals. On the other hand due to the low number of animals in the dataset, the null-hypothesis of the factor `Breed` having no effect on `Body Weight` could not be rejected.

While ANOVA is a widely used method and the above results show that we were able to correctly separate the variation between breeds and within a series of observation for the same animal, it has a major disadvantage. ANOVA cannot handle so-called **unbalanced** data very well. Unbalanced data means that the number of observations per factor level or per animal is

not the same. Because the problem of unbalanced data occurs quite frequently even in planned experiments, ANOVA is not used that often nowadays. The problems of unbalanced data can be addressed by a different class of models, called the mixed linear effects models.