Dependent data: Mixed Models

Jan van den Broek

2020

1 Introduction

1.1 Dependent data

From 79 cows the protein content in the milk was measured in the several weeks following calving. Consider two of those measurements for cow i: the protein content from the first week y_{i1} and the protein content of the second week y_{i2}. These measurements are continuous variables for which the normal distribution can be used. The models to be considered are linear models.

But something special is going on here: two observations are taken from the same cow which is the sampling unit. These observations are likely to be dependent. To illustrate, suppose that, from a cow, you were given the first observation, which was a high number, and were asked to predict the observation from the second week. Because the first observation is high you might anticipate that the observation in the second week is also relatively high. The same argument can be used if the first observation is low. So the first observation contains information about the observation in the second week, since both are from the same cow. Now suppose you were given the first observation of cow number 1 and were asked to predict the second observation of cow number 2. This is quite impossible since the observations from cow number 1 contains no information about the observations from cow number 2. Another way to think about this, is in terms of the probability distribution. If there was one observation for each cow then we could take the normal distribution for this. In that case each sampling unit gives one observation and the probability distribution is about 1 observation. But in the case that for each sampling unit two observations are obtained, the probability distribution must deal with two observations (variables) at the same time. This probability distribution is called a bi-variate probability distribution.

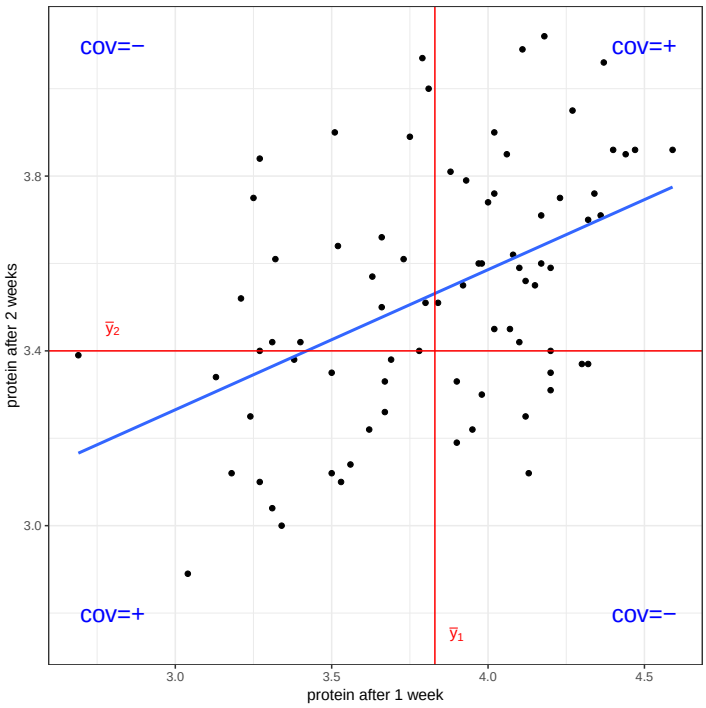

Figure 1: Relation between the protein level in week one and in week three.

1.2 Covariance and correlation

So, in this case there are two variables which are both continuous: protein level on week 1 (y_1) and week 2 (y_2) after calving. We would like to have a measure for the dependence between these two variables. Since these variables are continuous and the normal distribution is used, we can look at the linear dependence of these variable. A measure for the linear dependence between two variables is the so-called covariance:

$$cov(y_1, y_2) = \frac{1}{n-1} \sum_{i=1}^{n} (y_{i1} - \bar{y}_1)(y_{i2} - \bar{y}_2)$$

This formula looks similar to the formula of the variance. If, for example, the y_1 variable is replaced by the y_2 then the formula gives the variance of y_2. Figure 1, showing the scatter plot of y_1 and y_2, illustrates how the covariance measures the linear dependence between variables y_1 and y_2. The means of both variables are given by red lines in the plot. These lines divide the plot in 4 parts. For the upper right part the y_1-coordinates (on the horizontal axis) are larger than their mean so $(y_{i1} - \bar{y}_1)$ is positive. The y_2-coordinates (on the vertical axis) are larger than their mean so $(y_{i2} - \bar{y}_2)$ is also positive

and thus $(y_{i1} - \bar{y}_1)(y_{i2} - \bar{y}_2)$ is also positive. The points in the upper right part contribute a positive number to the covariance. For the lower right part the y_1-coordinates (on the horizontal axis) are larger than their mean so $(y_{i1} - \bar{y}_1)$ is positive. The y_2-coordinates (on the vertical axis) are lower than their mean so $(y_{i2} - \bar{y}_2)$ is negative and thus $(y_{i1} - \bar{y}_1)(y_{i2} - \bar{y}_2)$ is negative. The points in the lower right part contribute a negative number to the covariance. With the same arguments the points in the upper left part contribute a negative number and the points in the lower left part contribute a positive number to the covariance because in that case both parts are negative. If the positive contributions are larger than the negative contributions, the covariance is positive. This is the case for a rising elongated scatter plot. The more elongation the higher the covariance is. For a decreasing elongated cloud there are more negative contributions than positive and the covariance will be negative. If the number of positive contributions is approximately equal to the negative ones, the covariance is approximately zero.

The size of the covariance has no meaning since it depends on the scale on which the variables are measured. If for instance a variable is measured in kg and that is changed to gram, then the covariance will change, it is multiplied by 1000. This is a drawback of the covariance. In order to get a measure that is scale independent, one can divide the covariance by the standard deviations of y_1 and y_2:

$$r = \frac{cov(y_1, y_2)}{s_{y_1} s_{y_2}}$$

which is called the correlation coefficient. This correlation coefficient measures the linear dependence between 2 variables. The correlation is 1 if all points are on an increasing straight line, it is -1 if all points are on a decreasing straight line and the correlation is zero if the best fitting straight line is horizontal (e.g. if there is no elongation in the cloud).

The correlation between y_1 and y_2 is calculated as 0.46. This is not very high but different from zero which can be seen from the scatter plot. There are more points that contribute a positive value as compared to points that contribute a negative value.

1.3 Comparing t-statistics

In order to see how the dependence in the data is influencing the analysis of the data, compare the independent analysis case to the dependent one. Suppose there are two columns of data. The independent analysis treats these columns as independent, which means that it is assumed that y_1 was measured from different cows as y_2. Variable y_1 is measured from cows from

group number 1 and y_2 from cows in group number 2. In that case a two sample t-statistic could be calculated:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where s_1^2 and s_2^2 are the variances in the two samples. Since it is often assumed that the population variances in both groups are the same, this one variance is estimated as a weighted mean of s_1^2 and s_2^2 denoted with s^2. Both variances in the denominator are then replaced by s^2. The outcome of this t-statistic is 5.4 based on $78 + 78 - 2 = 154$ degrees of freedom (One cow had missing values and was deleted from the analysis). ($\bar{y}_1 = 3.83, \bar{y}_2 = 3.53, sd_1 = .401, sd_2 = .284$).

If the data is considered dependent, then the paired t-statistic is calculated. This is done by first reducing the two observations per cow to one: for each cow the values for y_2 and y_1 are subtracted to get the difference: $d_i = y_{1i} - y_{2i}$. The mean of this difference is calculated:

$$\begin{aligned}
\bar{d} &= \frac{1}{n}\sum_{i=1}^{n} d_i \\
&= \frac{1}{n}\sum_{i=1}^{n}(y_{1i} - y_{2i}) \\
&= \frac{1}{n}\sum_{i=1}^{n} y_{1i} - \frac{1}{n}\sum_{i=1}^{n} y_{2i} \\
&= \bar{y}_1 - \bar{y}_2
\end{aligned}$$

and the paired t-statistic then is:

$$t = \frac{\bar{d}}{\sqrt{\frac{s_d^2}{n}}} = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_d^2}{n}}}$$

where s_d^2 is the variance of the differences d_i.

Comparing these t-statistics one can see that the difference is in the denominator, it is in the way the variances are calculated, thus also giving different degrees of freedom. In general the formula for calculating the variance of the difference of two dependent variables is:

$$var(y_1 - y_2) = var(y_1) + var(y_2) - 2 \cdot cov(y_1, y_2)$$

The variances of the differences $d_i = y_{1i} - y_{2i}$ can be estimated by using the formula above or by calculating the variance of the difference s_d^2 which is 0.37

giving a t-statistic of 7.13 based on 77 degrees of freedom. If the variables are independent then $cov(y_1, y_2) = 0$ and to calculate the variance of the difference,$var(y_1 - y_2)$, the separate variances of the variables can just be added up as is done with the independent t-statistic.

So the influence of the dependencies in the data on the data-analysis is shown in the variance: the means and the difference in means are calculated the same in the independent case as in the dependent one but the difference is in the variance.

1.4 Conclusion

- Observations on the same sampling unit are very likely dependent.
- This (linear) dependence in the case of continuous data can be measured with the correlation coefficient.
- The dependence in the data is influencing the data-analysis through the variance. With dependent data the variances are different as compared to the independent case and thus are the degrees of freedom also different.

2 Random effects

2.1 The concept

If the data are dependent, the model for the data should take that dependence into account. In order to illustrate how this is achieved, consider a scatter plot of two variables, y_1 and y_2 as is shown in part A of figure 3. Just for illustration these two variables can be anything and have mean zero. Taking zero means makes no difference since one might think of that as two variables, from which the mean is subtracted (centered variables). The red cloud is a scatter plot of these two variables. As can be seen they are independent. The cloud is not stretched in any way. Now suppose a number is added to both y_1 and y_2. Take as an example that 5 is added to both the variables. This means that the y_1-coordinates are shifted 5 to the right and the y_2-coordinates are shifted 5 upwards. The whole cloud is shifted to the right upper corner. The resulting y_1-coordinates and the y_2-coordinates have the number 5 in common. So now there are two separate clouds. Suppose another number is added, let's say -3 to the two variables, then the y_1-coordinates are shifted 3 to the left and the y_2-coordinates are shifted 3 downwards. Then

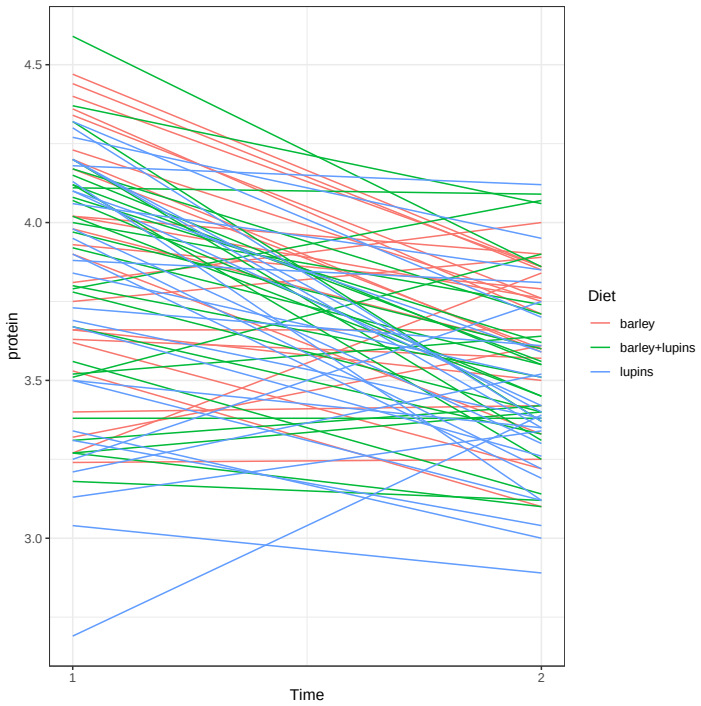

Figure 2: Protein content in the milk of cows in the 2 weeks following calving

there are 3 separate clouds. If there is yet another number added to the two variables then there are 4 separate clouds. The coordinates of these shifted clouds have a number in common, one cloud the number 5, an other cloud the number -3 etc. This is the case for the linear models where there are different groups of observations. All observations within a group have their group mean in common. In that case one can calculate a group effect for every group. As can be seen from the individual clouds, adding a number to each of two independent variables does not change their independence. Or to put it differently, if two (shifted) variables have a number in common, then that does not change their independence. This case that variables can have numbers in common, resulting in different clouds, is called fixed effects.

> If two variables y_1 and y_2 share a fixed effect then that does not change their independence.

Now suppose we add a random number to the two variables. So for instance a random number is drawn from a normal distribution with a certain variance and then this number is added to both the y_1-coordinate and the y_2-coordinate.

What does it mean when the outcome of a variable is considered to be

random? To get an idea of what random means, think of throwing a die. Suppose a die is going to be thrown and one wonders about the possible outcomes. Random means that there can be a 1 with probability $\frac{1}{6}$, there can be a 2 with probability $\frac{1}{6}$,there can be a 3 with probability $\frac{1}{6}$, $\cdots$, and there can be a 6 with probability $\frac{1}{6}$.

> Considering the outcome of a variable as random means that all possible outcomes with their attached probabilities are taken into account.

Adding a random number to both variables means we have to take all possible outcomes of that variable with their attached probabilities into account and average over it. So if a random number from a continuous distribution is added to y_1 and y_2, then there are not 1, 2 or 3 numbers added but infinitely many, and we have to take all of them into account. As a consequence there are infinitely many clouds which we all have to take into account. This means that we have to consider the whole resulting cloud and not the separate clouds. The whole resulting cloud is a stretched one which means that there is correlation. Figure 3 part C and D show 100 of these possible clouds.

If the same random number is added to two variable, we have to take all possible values of that random number with their attached probability into account, which means we have to take all possible clouds into account and thus have to consider the resulting stretched cloud which implies correlation between the two variables. Or differently, if two variables have a random effect in common, they will be correlated.

> If two variables y_1 and y_2 share a random effect then they will be correlated, because then all possible values of the random effect have to be taken into account resulting in a stretched cloud.

Figure 3 part C shows 100 possible clouds were the random affects are drawings from a normal distribution with mean 0 and standard deviation 3 whereas part D shows the case where the random effects are drawings from a normal distribution with a standard deviation of 1. If the standard deviation of the distribution of the random effects is reasonably small as in part D, then these random effects will be concentrated around zero resulting in a condensed cloud. In that case the correlation between the resulting variables will not be large. If, however the standard deviation of the distribution of the random effects is reasonably large as in part C then these random effects will be mostly around zero but large positive or large negative values are also possible resulting in a stretched cloud. In that case the correlation of the resulting variables will be large. So the level of the correlation is determined

by the standard deviation (or the variance) of the distribution of the random effects. So as said in section 1.4, the dependence in the data is influencing the model through the variance. With random effects the parameter of interest is the variance between the groups (cows) and not the individual group (cow) effects.

The variance of the resulting variable, the variable with the random effects added, will be the variance the original variable had plus the variance of the random effect. So suppose the variables $y_1$and y_2 have variance σ^2 and the random effects b come from a normal distribution, independent of the y's, with variance σ_b^2 then the variance of the resulting variables $y_1 + b$ and $y_2 + b$ will be $\sigma_b^2 + \sigma^2$. So the variance of the resulting variables will be larger than the original ones. The size of the variance of the random effects determines the stretching of the resulting cloud and thus the size of the correlation between the resulting variables.

This is how the dependence of the data is modeled, by using random effects and the variance of these random effects determine the size of the correlation.

2.2 The concept in a model: shared random effects model

Consider the cow example again, where for each cow the protein content in the milk was measured in the weeks following calving. There are two observations for each cow. A linear model could be used for this situation, treating the cow numbers as groups and thus having two observations within each group. A linear model for this situation might be $y_{ij} = \mu_i + \epsilon_{ij}$, where μ_i is the mean of cow number i and ϵ_{ij} is the residual, the difference between observation y_{ij} and the group mean. Another formulation of this model is $y_{ij} = \mu + (\mu_i - \mu) + \epsilon_{ij}$, where μ is the overall mean and $(\mu_i - \mu)$ is the i th group (cow) effect. Using different notation this model can be written as: $y_{ij} = \beta_0 + b_{0i} + \epsilon ij$, where β_0 is the overall mean, b_{0i} is the effect of group number i and $\beta_0 + b_{0i} = \mu + (\mu_i - \mu) = \mu_i$, the mean of group i. This group effect depends on the overall mean and the group mean, so is the same for each observation y_{ij} with each group. Or, to put it differently, all observations within a group share the same group effect. If this group effect is treated as fixed and we thus have 79 different cow groups, then this does not change the independence. This is just a linear model for independent data. Because in this case only the 79 groups are taken into account conclusions are concerned with these 79 groups (cows) only. So for each cow its effect can be calculated.

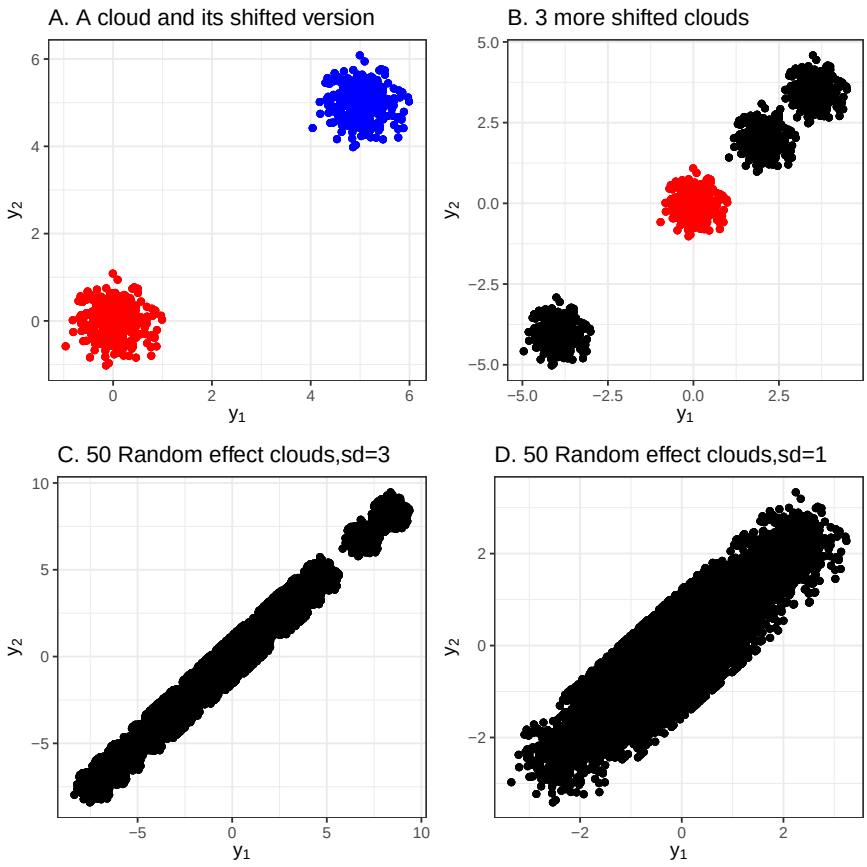

Figure 3: Fixed and random effects

If however, the group effects b_{0i} are random drawings from a normal distribution with mean zero and variance σ_b^2, then we have to take all possible values (infinitely many) of that random effect with their attached probabilities into account and as a result we have to consider one big stretched cloud, instead of many separate clouds in the fixed effect case. This means that the resulting variables are correlated. The random effects have mean zero. This means that most of the randomly drawn effects will be around zero. The variance determines the range of possible random effects. If the variance is large then a stretched cloud is obtained, and there will be a high correlation, if the variance is low, the cloud will be more condensed and the correlation will be low.

The random effect b_{0i} depends on i and thus on the cow. This means that there is one random effect for all the observations on each cow, these random effects coming from the same normal distribution. This is also called a shared random effects model.

Because now the cow effects are considered as random drawings from a normal distribution, one can consider the cows as sampled from a large population of cows. So the conclusions can be generalized to the population.

3 More than two observations on each sampling unit

Data consisting of two measurements of the same variable on each sampling unit is called paired data. If a variable is measured more often one speaks of repeated measures. One might for instance measure the concentration of something on different places of the body. If the variable is measured at different time points then the data is called longitudinal.

Let's extend the discussion by looking at more observation per sampling unit. From every cow 4 observations at different time points are considered. This is thus an example of longitudinal data. The data is illustrated in figure 4. Time is put on the x-axis and protein level on the y-axis. For every cow a line is drawn. From this figure one can see the variation between cows and also that on average there is a decline in protein level but this decline also seems to be different between cows. The cows are divided in 3 groups. Each group is given a different diet. The colours of the lines show these groups.

Since the cow is the observational unit, these 4 observations are dependent with the same arguments as in the case with two observations per cow: observations on the same cow (= observational or sampling unit) are dependent whereas observations from different cows are independent. And again,

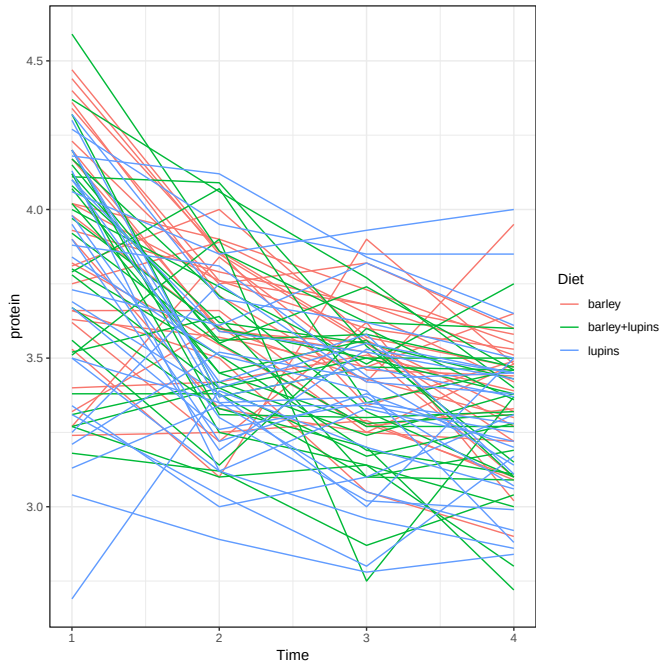

Figure 4: Protein content in the milk of cows in the 4 weeks following calving

if one observes dependent data, this dependency must be incorporated in the model since otherwise wrong conclusions might be drawn from the data. Putting random effects in the model is doing just that. Let's look at a linear model for the case where cow is the grouping variable and there are 4 observations for each cow. With this linear model one can write the 4 observations for cow number i as:

$$y_{i1} = \beta_0 + b_{0i} + \epsilon_{i1} \quad y_{i2} = \beta_0 + b_{0i} + \epsilon_{i2}$$
$$y_{i3} = \beta_0 + b_{0i} + \epsilon_{i3} \quad y_{i4} = \beta_0 + b_{0i} + \epsilon_{i4}$$

The anova model is $y_{ij} = \beta_0 + b_{0i} + \epsilon_{ij}$, where y_{ij} is the protein level of cow i at week j, β_0 is the overall mean, b_{0i} is the effect of cow i and ϵ_{ij} is the residual of observation y_{ij}. This is the linear model for the one way anova case. In this case the b_{0i} are fixed numbers. It represents the deviation for cow i from the overall mean. To put it differently, the mean for cow i is $\beta_0 + b_{0i}$, just as in the case of two observations above. Because the observations for cow i have a fixed effect in common they are modeled as independent. Since in that case the cow effects are fixed, conclusions from the study are limited to the cows in the study.

When the b_{0i} are treated as being drawn from a $N(0, \sigma^2_{int})$-distribution, this is they are random numbers, then the observations share a random effect

and as a consequence, by considering all possible outcomes of the random effect, the observations on the same cow are dependent. The variance between the cows is σ^2_{int} where int stands for intercept. The correlation between any two observations on the same cow is the same, because the drawings are from the same distribution, with the same variance, so the stretching is the same. This correlation structure is called the **exchangeable correlation structure** or **compound symmetry**. Since the cow effects are random, one can regard the cows in the study as a random sample from a population of cows. This means that the conclusion from this study can be generalized to the population of cows from which the cows in the study were sampled.

The fixed effect part of this model, β_0, is the mean protein level of an average cow. An average cow is a cow with random effect zero.

4 Random intercept model

Let's extend the model with the random effects above to include a linear time effect:

$$y_{ij} = \beta_0 + b_{0i} + \beta_1 Time + \epsilon_{ij}$$

Just as in a regression model, time is treated as a continuous variable and β_1 is the regression coefficient. In this model there is only one regression coefficient so the decline is modeled the same for every cow. The intercepts, however, are different for every cow. This is illustrated in figure 5. Recall the model for the analysis of covariance : $y_{ij} = \alpha + (\alpha_i - \alpha) + \beta Time + \epsilon_{ij}$. The general intercept is α and $(\alpha_i - \alpha)$ is the effect of group number i and $\alpha + (\alpha_i - \alpha) = \alpha$ is the mean of group number i. Now replace α by β_0 and $(\alpha_i - \alpha)$ by b_{0i} and the model above is obtained where the cows are the groups. In this model the intercept for cow i is $\beta_0 + b_{0i}$ where β_0 is the general mean of the intercepts and b_{0i} is the deviation from the general intercept for cow i. So this model has different intercepts for the cows and for every cow the same slope. So for every cow different lines are fitted but the lines are parallel. If the b_{0i} are treated as random, then the observations on the same cow have a random number in common and are thus correlated. The correlation is the same for every pair of observations from the same cow (exchangeable correlation structure). This model is called a random intercept model. As can be seen from the equation above, the model contains a random part and a fixed part:

Random part $b_{0i} + \epsilon_{ij}$, the random cow intercepts and the random residuals.

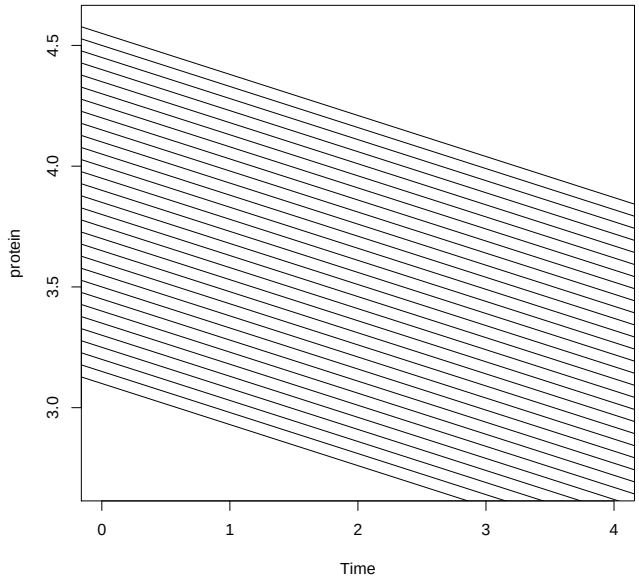

Figure 5: Different intercepts, same slope

fixed part $\beta_0 + \beta_1 Time$, the general intercept and the regression on time.

One recognizes the fixed part as an ordinary regression model. This fixed part is a description for the average cow, for a cow with random effect zero. This fixed part can, as with ordinary linear models, be modified and extended:

The 79 cows were divided in three groups, and each group got a different diet. One can then add a grouping variable to the linear part of the model. There are now several different models possible for the fixed part:

1. The model contains an intercept and time as a continuous variable. So the fixed part is a regression model.

2. One can add the grouping variable diet to the regression model above. The fixed part is then an analysis of covariance model.

3. To the analysis of covariance model the interaction between time and diet can be added. The fixed part is then a model with different intercepts and different slopes.

4. Instead of treating time as a continuous variable, one can treat time as a grouping variable. One then has two grouping variables in the model:

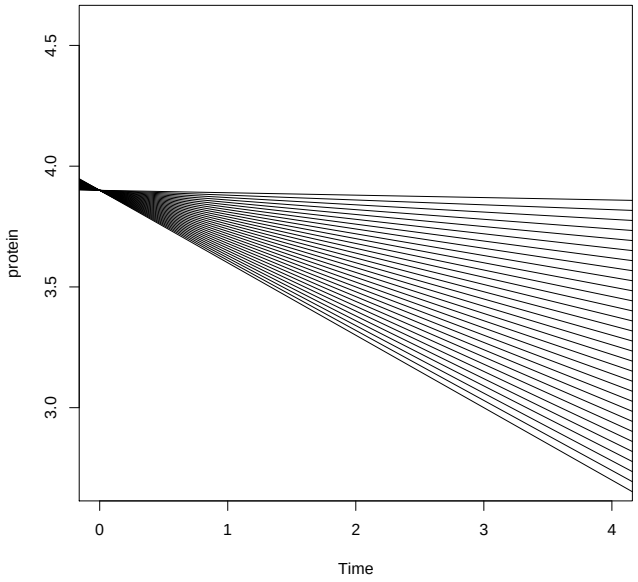

Figure 6: The same intercepts, different slopes

diet and time. This is a two-way anova model or a two way factorial model.

5. In the model with the two grouping variables one can add the interaction between diet and time. This is a full two way factorial model.

5 Random coefficients model

Instead of taking different intercepts for every cow, one could take the slopes different per cow:

$$y_{ij} = \beta_0 + (\beta_1 + b_{1i})Time + \epsilon_{ij} = \beta_0 + \beta_1 Time + b_{1i}Time + \epsilon_{ij}$$

Now there is only one intercept but every cow has her own slope: $\beta_1 + b_{1i}$. This is illustrated in figure 6. There, β_1 is the overall slope and b_{1i} is the deviation from the overall slope for cow i. The 4 observations for cow i can now be described as:

$$y_{i1} = \beta_0 + (\beta_1 + b_{1i})Time + \epsilon_{i1} \quad y_{i2} = \beta_0 + (\beta_1 + b_{1i})Time + \epsilon_{i2}$$
$$y_{i3} = \beta_0 + (\beta_1 + b_{1i})Time + \epsilon_{i3} \quad y_{i4} = \beta_0 + (\beta_1 + b_{1i})Time + \epsilon_{i4}$$

But the random effects b_{1i} do not get the same weight, as can be seen from the model. How heavily b_{1i} is weighted depends on time. Suppose the time points are 1, 2, 3 and 4. Then:

y_{i1} depends on $1 \times b_{1i}$
y_{i2} depends on $2 \times b_{1i}$
y_{i3} depends on $3 \times b_{1i}$
y_{i4} depends on $4 \times b_{1i}$

The b_{1i} can be treated as random numbers from a $N(0, \sigma^2_{slope})$-distribution, where σ^2_{slope} is the variance between the slopes of the cows. In that case the observations from the same cow have a random component in common and thus they are correlated. The weight of the random b_{1i} on the observation gets larger as the time progresses. This means that the observations on the same cow are correlated (they have a random effect in common) but this correlation depends on time. So, with this model the correlation between y_{i1} and y_{i2} and the correlation between y_{i1} and y_{i3} are different because the times at which y_{i2} and y_{i3} are measured are different. Usually the correlation between two observations decreases if they are further apart in time. So with the random part one models the correlation structure: if only random intercepts are taken the correlation between the observations is constant (exchangeable), with random coefficients the correlations depend on time.

Here the model also consists of a random and a fixed part:

Random part $b_{1i}Time + \epsilon_{ij}$, the random cow slopes and the random residuals.

Fixed part $\beta_0 + \beta_1 Time$, the general intercept and the regression on time.

The fixed part models the pattern in the data: e.g. is there an increase or decrease over time. So:

> **In a mixed effect model the random part models the correlation structure and the fixed part models the patterns in the data.**

6 Model with random intercepts and random coefficients

A model with a different intercept and different slope for each cow is:

$$y_{ij} = \beta_0 + b_{0i} + (\beta_1 + b_{1i})Time + \epsilon_{ij} = (\beta_0 + b_{0i}) + \beta_1 Time + b_{1i}Time + \epsilon_{ij}$$

Now every cow has her own intercept $\beta_0 + b_{0i}$ and her own slope: $\beta_1 + b_{1i}$. The 4 observations for cow i can now be described as:

$$
\begin{aligned}
y_{i1} &= \beta_0 + b_{0i} + (\beta_1 + b_{1i})Time + \epsilon_{i1} \\
y_{i2} &= \beta_0 + b_{0i} + (\beta_1 + b_{1i})Time + \epsilon_{i2} \\
y_{i3} &= \beta_0 + b_{0i} + (\beta_1 + b_{1i})Time + \epsilon_{i3} \\
y_{i4} &= \beta_0 + b_{0i} + (\beta_1 + b_{1i})Time + \epsilon_{i4}
\end{aligned}
$$

The b_{0i} are treated as a draw from a $N(0, \sigma^2_{int})$-distribution,where σ^2_{int} is the variance between the intercepts per cow. The b_{1i} can be treated as a random number from a $N(0, \sigma^2_{slope})$-distribution, where σ^2_{slope} is the variance between the slopes of the cows. In that case the observations from the same cow have two random components in common, one of which is weighted by time. The observations from the same cow are thus correlated and the correlation structure again depends on time. The correlation structure in this model is more complex as compared to the random coefficient model. Again the model consists of a random and a fixed part:

Random part $b_{0i} + b_{1i}Time + \epsilon_{ij}$, the random cow intercepts and slopes and the random residuals.

fixed part $\beta_0 + \beta_1 Time$, the general intercept and the regression on time.

The fixed part in this model is again just a regression model. Also here a linear model with diet can be used. And time can be taken as a grouping variable and interactions can be added, just as with the other models.

7 Fitting the models

In order to see how the likelihood in the case of random effects is determined, recall averaging. To average, multiply each observation with one over the number of observations and sum up. Instead of one over the number of observations one can use the probability of that particular outcome. The result of this is also called the expected value. So multiply an observation with its probability and then sum up all these products. If the observations are not discrete but continuous then one has to integrate instead of summing up.

This is used in order to obtain the likelihood in the continuous random effect case. First one assumes that the observations given the random effects (this is the fixed effect case; different groups for each cow), $P(y_{ij}|b_{0i})$, are normally distributed with mean μ and variance σ^2. This is multiplied with the probability of the random effects, $P(b_{0i})$ giving $P(y_{ij}|b_{0i})P(b_{0i})$. One can recognize the formula $P(a|b) * P(b) = P(a \ \& \ b)$. So $P(y_{ij}|b_{0i})P(b_{0i}$ is the joint distribution of the observations and the random effects. Now this joint distribution is integrated over all possible values of b_{0i} to get the distribution

of the observations and thus the likelihood. What is important here, is that with this method all possible values of the random effects with their attached probabilities are taken into account.

In this way one can write down the likelihood for each model discussed above, and maximize this likelihood with respect to all the parameters in the model in order to obtain the maximum likelihood estimates. If these estimates are plugged into the likelihood then the maximum value of the likelihood is obtained for the model used. One can then fit another model (by leaving out one of the terms) and obtain the maximum of the likelihood for that model. Then these two models can be compared using the AIC or the likelihood ratio test. This is just the standard likelihood procedure. This procedure works fine for the fixed effect part. So if a model is used with for instance diet and time in the fixed part, then one can fit a model with only time in it and compare the maximum of the likelihoods under both models with the AIC or the likelihood ratio test.

For the random effects part the story is a little different. Using the standard likelihood procedure here gives estimates for the different variance components that can be biased. For this reason the likelihood procedure is modified. The likelihood is transformed in such a way that the transformed likelihood does no longer depend on the fixed part of the model. Attention is restricted to the random effects part. For that reason one calls this the restricted likelihood. This restricted likelihood is maximized to obtain the maximum likelihood estimates, now called restricted maximum likelihood estimates. This procedure is known as restricted maximum likelihood estimation: REML. Because the transformation used in REML depends on the fixed part of the model, a change in this fixed part gives another transformation and thus also a different random part. If one then compares the models not only the change in the fixed part is measured but also a change in the random part. For this reason REML is not appropriate for testing the fixed part of the model.

A reasonable procedure is to first fit the complete model and test the random part using REML. After that, refit this model using maximum likelihood and test the fixed part. This procedure is illustrated below.

A model for the protein data is fitted with the grouping variables Diet and Time and their interaction as fixed effects part. The random part contains random cow intercepts and random time slopes. The variables in the data frame are Cow, Diet and Time. To fit a model with fixed and random effects, the function lmer() from the lme4 library can be used. This library has to be loaded first:

```
library(lme4)
```

```
# first get the data
mlk <- nlme::Milk[nlme::Milk$Time<5,]
fit <- lmer(protein~factor(Time)+factor(Diet)+
       factor(Diet):factor(Time)+(1+Time|Cow),data = mlk)
```

or without loading the lme4 library:

```
# first get the data
mlk <- nlme::Milk[nlme::Milk$Time<5,]
fit <- lme4::lmer(protein~factor(Time)+factor(Diet)+
       factor(Diet):factor(Time)+(1+Time|Cow),data = mlk)
```

The fixed part of the model is stated in the usual way. The random part of the model is the part between brackets: (1+Time|Cow). The part 1+Time shows the random part of the model. The 1 gives the random intercepts and the Time the random slopes. The |Cow part indicates that the random effects are per cow, so cow is treated as factor. So the 1 in this model represents the random cow intercepts. This is a model with random (cow) intercepts and random (cow) slopes. Note that the 1 in (1+Time|Cow) can be left out so (Time|Cow) represents the same model.

```
summary(fit, correlation = FALSE)

Linear mixed model fit by REML ['lmerMod']
Formula: protein ~ factor(Time) + factor(Diet) +
    factor(Diet):factor(Time) + (1 + Time | Cow)
   Data: mlk

REML criterion at convergence: 90.6

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.67360 -0.53186  0.03346  0.46926  2.39156

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Cow      (Intercept) 0.134877 0.36726
          Time        0.007439 0.08625  -0.87
 Residual             0.039697 0.19924
Number of obs: 315, groups:  Cow, 79

Fixed effects:
```

```
                                          Estimate Std. Error t value
(Intercept)                                3.886800   0.071183  54.603
factor(Time)2                             -0.247943   0.059672  -4.155
factor(Time)3                             -0.388800   0.066076  -5.884
factor(Time)4                             -0.510400   0.076511  -6.671
factor(Diet)barley+lupins                 -0.025689   0.098786  -0.260
factor(Diet)lupins                        -0.128652   0.098786  -1.302
factor(Time)2:factor(Diet)barley+lupins -0.073169   0.082321  -0.889
factor(Time)3:factor(Diet)barley+lupins -0.126756   0.091699  -1.382
factor(Time)4:factor(Diet)barley+lupins -0.072933   0.106180  -0.687
factor(Time)2:factor(Diet)lupins          -0.082428   0.082321  -1.001
factor(Time)3:factor(Diet)lupins           0.003615   0.091699   0.039
factor(Time)4:factor(Diet)lupins           0.046326   0.106180   0.436
```

From the part Random effects it can be seen that the standard deviation between the intercepts is 0.367 (variance is 0.135) and between the slopes 0.086. The correlation between these two estimates is -0.87. This means that the two random effects are not modeled as independent. The random intercepts and the random slope have a bi-variate distribution instead of each having an independent normal distribution. So this is a model with correlated random intercept and random slope. This is default in the lme4 library. To specify a model in which the random intercept and slope are not correlated one can use (1+Time||Cow).

Sometimes one finds that the correlation between the estimate for the random intercept and slope is -1. This means that they cannot be distinguished. In that case the model cannot be estimated. A remedy for this is to change the time scale in the random effects, by dividing it by an appropriate number. If that does not work one can try to fit a model with only random slopes or only random intercepts.

To examine the covariances and the variances between the protein values at the four different time points:

```
Marginal variance covariance matrix
         1        2        3        4
1 0.126680 0.066751 0.046521 0.026291
2 0.066751 0.093657 0.041170 0.028379
3 0.046521 0.041170 0.075515 0.030467
4 0.026291 0.028379 0.030467 0.072252
  Standard Deviations: 0.35592 0.30603 0.2748 0.2688
```

So the variances on the 4 time points are $.127, .094, .076$ and $.072$. The standard deviations (square roots) are at the bottom of the output. The

covariance between the protein values at times 1 and 2 is 0.0667, between times 1 and 3 0.0465, etc. From this the correlations between the protein values at different time points can be calculated. For instance the correlation between the protein values at time 1 and time 2 is the covariance between the protein values at time 1 and time 2 (0.066751) divided by the standard deviation of the protein values at time 1 (0.35592) times the standard deviation of the protein values at time 2 (0.30603), so $\frac{0.066751}{0.35592 \cdot 0.30603} = 0.6128$. One can do this for all possible combinations of time-points and obtain the following correlations:

```
          1         2         3         4
1 1.0000000 0.6128287 0.4756462 0.2748118
2 0.6128287 1.0000000 0.4895457 0.3449902
3 0.4756462 0.4895457 1.0000000 0.4124714
4 0.2748118 0.3449902 0.4124714 1.0000000
```

As can be seen the correlation between the protein values at time 1 and time 2 is 0.61, between times 1 and 3 it is 0.48 and between times 1 and 4 0.27. This illustrates something seen more often in time series: the further away the time points are, the lower their correlation. Values on time points close together have a higher correlation than values on time points further away. Note that in this model the correlation between values at times 2 and 3 for instance are not the same as those at times 1 and 2.

The model is fitted using the REML procedure by default (REML=TRUE). This is used to determine what model for the random part is best according to the data. Next two other models are fitted with the same fixed effects but with different random effects. First a model with random intercepts only

```
fit2 <- lmer(protein~factor(Time)+factor(Diet)+
        factor(Diet):factor(Time)+(1|Cow),data = mlk)
```

```
Model with random slope only
fit3 <- lmer(protein~factor(Time)+factor(Diet)+
        factor(Diet):factor(Time)+(-1+Time|Cow),data = mlk)
```

-1 means "no intercept". Another way to do this is with (0+Time—Cow). In order to see which of these models fits the data best let's compare the AIC's

```
AIC(fit,fit2,fit3)
     df      AIC
fit  16 122.5821
fit2 14 138.3831
fit3 14 182.8353
```

So clearly the model with the random intercepts and slopes fits the data best as compared to the other models. Now the fixed effects can be tested, beginning with the interaction term. But first the model has to be fitted again using the maximum likelihood method instead of REML. After that a drop1() command is used to see if the interaction is needed:

```
fitml <- lmer(protein~factor(Time)+factor(Diet)+
         factor(Diet):factor(Time)+(1+Time|Cow),
         REML=FALSE, data = mlk)
 drop1(fitml)
Single term deletions

Model:
protein ~ factor(Time) + factor(Diet) +
    factor(Diet):factor(Time) + (1 + Time | Cow)
                           Df    AIC
<none>                        73.170
factor(Time):factor(Diet)  6 66.792
```

The model where the interaction is deleted has the lowest AIC so we can remove the interaction. Let's fit that model:

```
fitml2 <- lme4::lmer(protein~factor(Time)+factor(Diet)+
          (1+Time|Cow), REML=FALSE, data = mlk)
drop1(fitml2)
Single term deletions

Model:
protein ~ factor(Time) + factor(Diet) + (1 + Time | Cow)
             Df     AIC
<none>           66.792
factor(Time)  3 166.985
factor(Diet)  2  67.739
```

There is not much difference between the first and third AIC's so we prefer the simpler model, that is the model with Diet deleted, so only Time in it. There is not enough evidence from the data that the diets are different w.r.t. the protein level. One can use confint() to obtain profile likelihood confidence intervals.

8 Logistic regression with random effects

Suppose that, in the cow example, at every time point it was measured whether or not the protein level was high. One then measures binary data. This data should be modeled with logistic regression. To deal with the dependencies in the data one can add a random intercept and random slopes to the linear part of the model, just as with an ordinary linear model. That is, with a logistic regression model the dependencies are modeled on the logit scale. The logistic model becomes:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + b_{0i} + (\beta_1 + b_{1i})Time$$

β_0 are the general log-odds at Time 0, and b_{0i} are the deviations from these general log-odds for every cow. These b_{0i} are taken as draws from a normal distribution. In this way the data are modeled as being dependent, but on the logit scale. β_1 is the general coefficient for time and the b_{1i} are the deviations from this general time effect, now taken as random draws. This model cannot be fitted with lmer. Instead the glmer function from the lme4 library can be used. To fit for example a logistic regression model with random intercepts, and with diet and time as grouping variables plus their interaction for the fixed part, one uses:

```
mlk$highprotein <- 1*(mlk$protein>3.5) #arbitrary choice
fith <- glmer(highprotein~factor(Time)+factor(Diet)+
        factor(Diet):factor(Time)+(1|Cow),family=binomial,
        data=mlk)

Generalized linear mixed model fit by maximum likelihood
(Laplace Approximation) ['glmerMod']
 Family: binomial  ( logit )
Formula: highprotein ~ factor(Time) + factor(Diet) +
         factor(Diet):factor(Time) +      (1 | Cow)
   Data: mlk

     AIC      BIC   logLik deviance df.resid
   339.8    388.6   -156.9    313.8      302

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.7141 -0.4008 -0.0865  0.3727  2.8154
```

```
Random effects:
 Groups Name        Variance Std.Dev.
 Cow    (Intercept) 4.261    2.064
Number of obs: 315, groups:  Cow, 79

Fixed effects:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                                2.65965    0.83435   3.188  0.00143 **
factor(Time)2                             -0.90172    0.89179  -1.011  0.31195
factor(Time)3                             -2.22688    0.87507  -2.545  0.01093 *
factor(Time)4                             -4.21687    0.99792  -4.226 2.38e-05 ***
factor(Diet)barley+lupins                 -0.18307    1.11129  -0.165  0.86915
factor(Diet)lupins                        -1.38568    1.05503  -1.313  0.18905
factor(Time)2:factor(Diet)barley+lupins -1.47117    1.23299  -1.193  0.23280
factor(Time)3:factor(Diet)barley+lupins -1.76819    1.26317  -1.400  0.16157
factor(Time)4:factor(Diet)barley+lupins -2.19465    1.50138  -1.462  0.14381
factor(Time)2:factor(Diet)lupins          -1.25576    1.18476  -1.060  0.28918
factor(Time)3:factor(Diet)lupins          -0.21632    1.16387  -0.186  0.85255
factor(Time)4:factor(Diet)lupins          -0.02988    1.32995  -0.022  0.98208
```

The random effects part shows that the variance between the cows on a log odds scale is 4.261. The fixed effect part show the log odds ratio's with their standard errors. The random effects show the variance between the cows on a log-odds scale. One can use drop1() to see which effects are needed in the model

```
Single term deletions

Model:
highprotein ~ factor(Time) + factor(Diet) + factor(Diet):factor(Time) +
    (1 | Cow)
                          Df    AIC
<none>                       339.83
factor(Time):factor(Diet)  6 333.06
```

So there is no evidence in the data that the interaction is needed and thus it can be left out. The drop1() can then be used on the main effects model and if no terms can be left out confint() can be used to obtain profile likelihood confidence intervals.

```
Generalized linear mixed model fit by maximum likelihood
(Laplace Approximation) ['glmerMod']
```

```
 Family: binomial  ( logit )
Formula: highprotein ~ factor(Time) + factor(Diet) + (1 | Cow)
   Data: mlk

     AIC      BIC   logLik deviance df.resid
   333.1    359.3   -159.5    319.1      308

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.08546 -0.39535 -0.07314  0.36723  2.11468

Random effects:
 Groups Name        Variance Std.Dev.
 Cow    (Intercept) 3.896    1.974
Number of obs: 315, groups:  Cow, 79

Fixed effects:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                 3.2437     0.6831   4.749 2.05e-06 ***
factor(Time)2              -1.8296     0.4980  -3.674 0.000239 ***
factor(Time)3              -2.8661     0.5493  -5.217 1.81e-07 ***
factor(Time)4              -4.8572     0.7241  -6.708 1.97e-11 ***
factor(Diet)barley+lupins  -1.4739     0.7031  -2.096 0.036041 *
factor(Diet)lupins         -1.8513     0.7185  -2.577 0.009979 **
---

confint(fith2)
Computing profile confidence intervals ...
                               2.5 %     97.5 %
.sig01                      1.315483  2.8527667
(Intercept)                 2.019117  4.7470100
factor(Time)2              -2.866508 -0.8975357
factor(Time)3              -4.031490 -1.8584034
factor(Time)4              -6.424776 -3.5601676
factor(Diet)barley+lupins -2.986824 -0.1267651
factor(Diet)lupins        -3.409608 -0.4896184
```

So for instance the log odds ratio for lupins vs barley is -1.85, with a confidence interval of $(-3.41, -0.49)$. The OR $= exp(-1.85) = 0.157$

Exercises

Exercises part 1

Summary of R-commands: To get the data

```
data(Milk,package="nlme")
milk <- Milk[Milk$Time<3,]
```

or

```
milk <- nlme::Milk[nlme::Milk$Time<3,]
```

For the t test use the function t.test() and use ?t.test for help.
To make the plot with two time points:

```
library(ggplot2)
ggplot(milk,aes(x=Time,y=protein,group=Cow))+
  geom_line(aes(color=Diet))+
  scale_x_continuous(breaks=c(1,2))+
  theme_bw()
```

To fit the model with random intercepts

```
library(lme4)
fit <- lmer(protein~factor(Time)+factor(Diet)+
       factor(Diet):factor(Time)+(1|Cow),data = milk)
```

In the second part of dependent data this will be explained in more detail.

1. From the protein example use the data from time points 1 and 3.
   (a) Calculate the outcome for the t test for two samples and calculate the standard error used in this t test.
   (b) Do the same for the paired t test.
   (c) Compare the two standard errors. Why is the standard error for the paired test smaller?
   (d) Fit a random intercept model for this data with Time as a fixed effect.
   (e) Compare the 3 analyses. Compare their estimates and the outcomes of the t-statistic.

2. Investigators at the University of North Carolina Dental School followed the growth of 27 children (16 males, 11 females) from age 8 until age 14. Every two years they measured the distance between the pituitary and the pterygomaxillary fissure, two points that are easily identified on x-ray exposures of the side of the head.
The data is in the nlme library. The variables are:
distance: a numeric vector of distances from the pituitary to the pterygomaxillary fissure (mm). These distances are measured on x-ray images of the skull.
age: a numeric vector of ages of the subject (yr).
Subject: a factor indicating the subject on which the measurement was made.
Sex: a factor with levels Male and Female.
To get the data: (Od $<-$ nlme::Orthodont)

   (a) Fit a linear mixed model with exchangeable correlation structure and with age, sex and their interaction as fixed effects. (Use REML=FALSE, see ?lmer for details)

   (b) Check the residuals of the model.

   (c) Use drop1() to see which terms are needed in the model using AIC.

3. In the data set grouseticks in the lme4 library the number of ticks on the heads of red grouse chicks sampled in the field is recorded. Some chicks came from the same broods (nests). Variables in the data set are:
INDEX: (factor) chick number (observation level)
TICKS: number of ticks sampled
BROOD: (factor) brood number
HEIGHT: height above sea level (meters)
YEAR: year
LOCATION: (factor) geographic location code
cHEIGHT: centered height, derived from HEIGHT

   (a) Fit a model for TICKS with a random BROOD effect and with HEIGHT and YEAR as fixed effects.

   (b) Test, with likelihood ratio tests, which variables need to be in the model.

   (c) Criticize this model.

Exercises part 2

Summary of R-commands:
To make the plot:

```
library(ggplot2)
ggplot(milk,aes(x=Time,y=protein,group=Cow))+
  geom_line(aes(color=Diet))+
  theme_bw()
```

To fit a model with fixed and random effects with lmer():

```
library(lme4)
fit <- lmer(protein~factor(Time)+factor(Diet)+
factor(Diet):factor(Time)+(1+Time|Cow),data = milk)
```

Add REML=FALSE for maximum likelihood estimates. To see the estimates and their standard errors:

```
summary(cow.fit)
```

1. In the protein example in the handout there were 4 time points. In the actual data set there were many more time points. To select the first 7 time points from the data:

   ```
   milk<-nlme::Milk[nlme::Milk$Time<8,]
   ```

   (a) Fit the model with diet and time as grouping variables (factors), and their interaction. For the random part take random intercepts and random slopes.

   (b) See, by fitting different models, which random effects are needed in the model. (Just like in the text.)

   (c) See, by fitting different models, what the fixed part of the model should be.

   (d) Discuss the difference with the analysis from the text and explain this.

2. In the nlme library there is a data set on the body weights of rats, measured over 64 days. The body weights of the rats (in grams) are measured on day 1 and every seven days thereafter until day64, with an extra measurement on day 44. The experiment started several weeks before "day 1". There are three groups of rats, each on a different diet. The data set is named BodyWeight. In this data set Diet is a factor, the other variables are numeric covariates.

(a) Get the BodyWeight data set, get information about this dataset (?nlme::BodyWeight) and make a plot like figure 1 of the handout.

(b) Which random effects are needed in the model? Take for the fixed part a model with Diet and Time as grouping variables, and their interaction.

(c) Check which fixed parts are needed in the model.

(d) Describe the analysis you did, and the conclusions that can be drawn.

3. Read in the data file osteochon.csv.

(a) Fit a logistic regression model with independent variables father, ground and height. (Remember to use factor() for father and ground.)

(b) Discuss the estimates and standard errors for the fathers.

(c) Now fit a logistic regression model with random father effects. Center the variable height by subtracting the mean.

(d) Compare the AIC's of both models and discuss which model fits best. Also compare the estimates of height.

(e) Discuss the effect of using random father effects.

4. **Extra exercise**
The deviance residuals for the model in exercise 3 in Exercises part 1 are between -3 and 5.6. This is reasonably large so one might wonder whether the Poisson model fits the data well. It seems there is more variance in the data then the Poisson distribution can account for. Perhaps another distribution gives a better fit. One way to get another distribution which can handle larger variance, is to use the Poisson distribution and take as the mean of this distribution a random variable with a gamma distribution. If one averages over all possible random effects one obtains the so-called negative binomial distribution. Note that a random draw is taken from a distribution so here each observation has a random effect which is different from taking a random effect for a group of observations as in the shared random effect models from the text. Instead of using the gamma distribution for the mean one could also use a log-normal distribution for the mean (and thus a normal distribution on the log-scale). This model is called a Poisson-lognormal model. This can be fitted by using a random effect for the individual observations:

```
lme4::glmer(TICKS~YEAR+cHEIGHT+(1|INDEX),
data = lme4::grouseticks, family="poisson")
```

To fit a Poisson-lognormal model with random brood effects:

```
lme4::glmer(TICKS~YEAR+cHEIGHT+(1|BROOD)+(1|INDEX),
data = lme4::grouseticks, family="poisson")
```

This model has random effects on the observation level to model a larger variance, **and** shared random effects to model the dependence between observations within a brood.

- Fit this model and compare it with the model from exercise 3 (Exercises part 1) using AIC.
- Check the residuals.