eNote 10

Mixed Model Theory, Part II

Indhold

The previous modules have introduced a number of situations where models including random effects are very useful. In Module 4 a fairly detailed description of the mixed model theory framework was given. In this Module 10, we will elaborate one some of the theoretical issues important for the practical work with mixed models. Hopefully this will provide the reader with a better understanding of the structure and nature of these models, along with an improved ability to interpret results from these models.

A general linear mixed model can be presented in matrix notation by:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \varepsilon, \quad \text{where } \mathbf{u} \sim N(0, \mathbf{G}) \text{ and } \varepsilon \sim N(0, \mathbf{R}).$$

The vector $\mathbf{u}$ is the collection of all the random effect coefficients (just like β for the fixed effect parameters). The covariance matrix for the measurement errors $\mathbf{R} = \text{var}(\varepsilon)$ has dimension $n \times n$. In most examples $\mathbf{R} = \sigma^2\mathbf{I}$, but in some examples to be described later

in this course, it is convenient to use a different $\mathbf{R}$. The covariance matrix for the random effect coefficients $\mathbf{G} = \text{var}(\mathbf{u})$ has dimension $q \times q$, where q is the number of random effect coefficients. The structure of the $\mathbf{G}$ matrix can be very simple. If all the random effect coefficients are independent, then $\mathbf{G}$ is a diagonal matrix with diagonal elements equal to the variance of the random effect coefficients.

The covariance matrix $\mathbf{V}$ describing the covariance between any two observations in the data set, can be calculated directly from the matrix representation of the model in the following way:

$$\begin{aligned}
\mathbf{V} &= \text{var}(\mathbf{y}) = \text{var}(\mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \varepsilon) \quad [\text{from model}] \\
&= \text{var}(\mathbf{X}\beta) + \text{var}(\mathbf{Z}\mathbf{u}) + \text{var}(\varepsilon) \quad [\text{all terms are independent}] \\
&= \text{var}(\mathbf{Z}\mathbf{u}) + \text{var}(\varepsilon) \quad [\text{variance of fixed effects is zero}] \\
&= \mathbf{Z}\text{var}(\mathbf{u})\mathbf{Z}' + \text{var}(\varepsilon) \quad [\mathbf{Z} \text{ is constant}] \\
&= \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R} \quad [\text{from model}]
\end{aligned}$$

10.1 Likelihood function and parameter estimation

In Chapter 4 th following was given: The *likelihood function* L is a function of the observations and the model parameters. It returns a measure of the probability of observing a particular observation $\mathbf{y}$, given a set of model parameters β and γ. Here β is the vector of the fixed effect parameters, and γ is the vector of parameters used in the two covariance matrices $\mathbf{G}$ and $\mathbf{R}$, and hence in $\mathbf{V}$. Instead of the likelihood function L itself, it is often more convenient to work with *negative log likelihood function*, denoted ℓ. The negative log likelihood function for a mixed model is given by:

$$\begin{aligned}
\ell(\mathbf{y}, \beta, \gamma) &= \frac{1}{2}\left\{ n\log(2\pi) + \log|\mathbf{V}(\gamma)| + (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{V}(\gamma))^{-1}(\mathbf{y} - \mathbf{X}\beta) \right\} \\
&\propto \frac{1}{2}\left\{ \log|\mathbf{V}(\gamma)| + (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{V}(\gamma))^{-1}(\mathbf{y} - \mathbf{X}\beta) \right\} \qquad (10\text{-}1)
\end{aligned}$$

The symbol '$\propto$' reads 'proportional to', and is used here to indicate that only an additive constant (constant with respect to the model parameters) have been left out.

A natural and often used method for estimating model parameters, is the maximum likelihood method. The maximum likelihood method take the actual observations and chooses the parameters which make those observations most likely. In other words, the parameter estimates are found by:

$$(\hat{\beta}, \hat{\gamma}) = \underset{(\beta,\gamma)}{\text{argmin}}\, \ell(\mathbf{y}, \beta, \gamma)$$

In practice this minimum is found in three steps. 1) The estimate of the fixed effect parameters β is expressed as a function of the random effect parameters γ, as it turns out that no matter what value of the model parameters that minimize ℓ, then $\hat{\beta}(\gamma) = (\mathbf{X}'(\mathbf{V}(\gamma))^{-1}\mathbf{X})^{-1}\mathbf{X}'(\mathbf{V}(\gamma))^{-1}\mathbf{y}$. 2) The estimate of the random effect parameters is found by minimizing $\ell(\mathbf{y}, \hat{\beta}(\gamma), \gamma)$ as a function of γ. 3) The fixed effect parameters are calculated by $\hat{\beta} = \hat{\beta}(\hat{\gamma})$.

The maximum likelihood method is widely used to obtain parameter estimates in statistical models, because it has several nice properties. One nice property of the maximum likelihood estimator is "functional invariance", which means that for any function f, the maximum likelihood estimator of $f(\psi)$ is $f(\hat{\psi})$, where $\hat{\psi}$ is the maximum likelihood estimator of ψ. For mixed models however, the maximum likelihood method on average tends to underestimate the random effect parameters, or in other words the estimator is *biased* downwards.

A well known example of this bias is in a simple random sample. Consider a random sample $\mathbf{x} = (x_1, \ldots, x_N)$ from a normal distribution with mean μ and variance σ^2. The mean parameter is estimated by the average $\hat{\mu} = (1/n)\sum x_i$. The maximum likelihood estimate of the variance parameter is $\hat{\sigma}^2 = (1/n)\sum(x_i - \hat{\mu})^2$. This estimate is not often used, because it is known to be biased. Instead the unbiased estimate $\hat{\sigma}^2 = (1/(n-1))\sum(x_i - \hat{\mu})^2$ is most often used. This estimate is known as the *Restricted or Residual maximum likelihood estimate* (REML).

Now in this Chapter 10, before we elaborate on the REML approach let us detail the basic ML approach using a simple example.

|||| Example 10.1 One sample situation with NO random effects

Consider the simplest possible normal distribution setting:

$$y_i = \mu + \varepsilon_i, \quad \text{where } \varepsilon_i \sim N(0, \sigma^2).$$

The ε_i are all independent. The matrix notation for this model with $n = 6$ is:

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{pmatrix}}_{\mathbf{y}} = \underbrace{\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} \mu \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix}}_{\varepsilon},$$

where $\varepsilon \sim N(0, \sigma^2\mathbf{I})$. There is no random effects so the term with Z and $\mathbf{u}$ from the general mixed linear model expression is not there at all. Notice how the matrix representation

exactly correspond to model formulation, when the matrices are multiplied. And remember that

$$\begin{aligned}\mathbf{V} &= \text{var}(\mathbf{y}) = \sigma^2\mathbf{I} \\ &= \begin{pmatrix} \sigma^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2 \end{pmatrix}\end{aligned}$$

So this $\mathbf{V} = \mathbf{V}(\gamma) = \sigma^2\mathbf{I}$ can then be inserted into the likelihood to express the likelihood function for this particular case: (where we have only a single variance parameter γ, and only a single fixed effect parameter)

$$\begin{aligned}\ell(\mathbf{y}, \beta, \gamma) &= \frac{1}{2}\left\{n\log(2\pi) + \log|\sigma^2\mathbf{I}| + (\mathbf{y} - \mathbf{X}\beta)'(\sigma^2\mathbf{I})^{-1}(\mathbf{y} - \mathbf{X}\beta)\right\} \\ &= \frac{1}{2}\left\{n\log(2\pi) + \log(\prod_{i=1}^{n}\sigma^2) + (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)/\sigma^2\right\} \\ &= \frac{1}{2}\left\{n\log(2\pi) + n\log(\sigma^2) + (\mathbf{y} - \mu)'(\mathbf{y} - \mu)/\sigma^2\right\} \\ &= \frac{1}{2}\left\{n\log(2\pi) + n\log(\sigma^2) + \sum_{i=1}^{n}(y_i - \mu)^2/\sigma^2\right\}\end{aligned}$$

where we have used a little knowledge of matrix algebra. We see how the multivariately expressed likelihood for this simple setting boils down to a univariate normal based likelihood, which would be expressed explicitly as the product of the densities of each of the n independent normal observations:

$$L((y_1, \ldots, y_n), \mu, \sigma^2) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_i - \mu)^2}{2\sigma^2}\right)$$

And hence the negative log-likelihood becomes:

$$\begin{aligned}\ell((y_1, \ldots, y_n), \mu, \sigma^2) &= log\left[\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_i - \mu)^2}{2\sigma^2}\right)\right] \\ &= -\sum_{i=1}^{n}\left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \left(-\frac{(y_i - \mu)^2}{2\sigma^2}\right)\right]\end{aligned}$$

which can be seen to be exactly the same as above. To find the maximum likelihood estimates we find the derivatives of the negative log-likelihood function:

$$\begin{aligned}\frac{\partial\ell(\mu, \sigma^2)}{\partial\mu} &= \frac{1}{2}\left[\sum_{i=1}^{n}(-2)(y_i - \mu)/\sigma^2\right] \\ &= (n\mu - \sum_{i=1}^{n}y_i)/\sigma^2\end{aligned}$$

and

$$\frac{\partial \ell(\mu, \sigma^2)}{\partial \sigma^2} = \frac{1}{2}\left[\frac{n}{\sigma^2} - \frac{\sum_{i=1}^{n}(y_i-\mu)^2}{(\sigma^2)^2}\right]$$

And if we solve the derivatives equal to zero we get:

$$\begin{aligned}\hat{\mu} &= \bar{y} \\ \hat{\sigma} &= \frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2\end{aligned}$$

Let us see how this looks in R. First we generate some data with $n = 6$:

```
n <- 6
mu=2
sigma <- sqrt(2)
set.seed(12345)
y <- rnorm(n, mu, sigma)
mydata <- data.frame(y)
```

Then the likelihood function is defined explicitly in three different ways: (and evaluated in the maximum point)

```
## Using the mathematical expression:
minusloglik <- function(x){
  0.5*(n*log(2*pi) + n*log(x[2]) + sum((y-x[1])^2)/x[2])
}
-minusloglik(c(mean(y), (n-1)*var(y)/n))
```

[1] -9.859

```
## Using the normal density function:
minusloglik2 <- function(x) -sum(log(dnorm(y, x[1], sqrt(x[2]))))
-minusloglik2(c(mean(y), (n-1)*var(y)/n))
```

[1] -9.859

```
## Using the MULTIVARIATE normal density function:
library(mvtnorm)
minusloglik3 <- function(x) dmvnorm(y, rep(x[1], n), diag(x[2], n), log=TRUE)
-minusloglik3(c(mean(y), (n-1)*var(y)/n))
```

[1] 9.859

Let us check that we get the same from the `logLik` function:

```
lmfit <- lm(y~., data=mydata)
logLik(lmfit)
```

'log Lik.' -9.859 (df=2)

And finally, we can see that we could directly optimize the likelihood (by minimizing minus the log-likelihood) using the R-optimizer function `optim`:

```
MLoptimize <- optim(c(1,1), minusloglik)
## The results:
MLoptimize$par
```

[1] 1.887 1.566

```
## Commpare with:
c(mean(y), var(y)*(n-1)/n)
```

[1] 1.887 1.566

Let's look at the likelihood for the variance for $\mu = \bar{y}$, or rather the relative likelihood, where we divide by the maximal value:

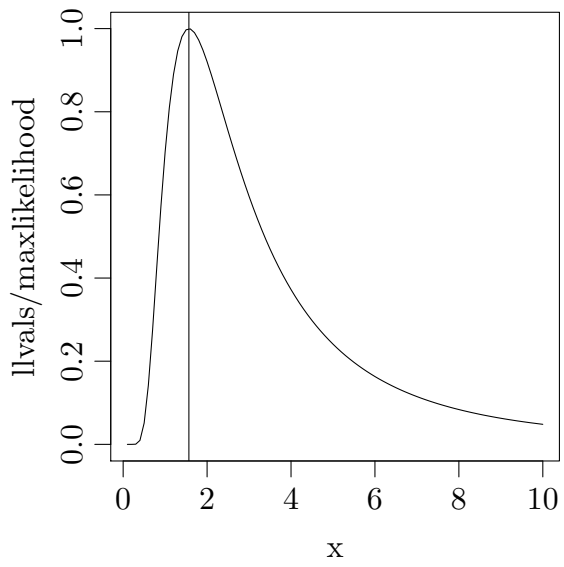
$$\frac{L(\bar{y}, \sigma^2)}{L(\bar{y}, \hat{\sigma}^2 \cdot 5/6)}$$

```
loglik2 <- function(x) prod(dnorm(y, x[1], sqrt(x[2])))

sigmalikelihood <- function(sig){loglik2(c(mean(y), sig))}

maxlikelihood <- sigmalikelihood(var(y)*(n-1)/n)

x <- seq(0.1, 10, 0.1)
llvals <- rep(0, length(x))
j <- 0
 for (i in x){
  j <- j + 1
 llvals[j] <- sigmalikelihood(i)
}
plot(x, llvals/maxlikelihood, type="l")
abline(v = var(y)*(n-1)/n)
```



The maximum point is shown.

10.2 Model testing and comparison by likelihood

We have previously used that we can investigate a hypothesis/submodel about a set of multidimensional model parameters $\theta \in \Theta$ by comparing the fit of the model under the hypothesis

$$H_0 : \theta \in \Theta_0,$$

where Θ_0 is a subspace of Θ. We do this by computing the maximum likelihood estimates and fits under both assumptions, and then finding minus twice the ratio of the two log-likelihoods:

$$G = -2 \log \frac{\max_{\text{Submodel}} L(\theta)}{\max_{\text{Model}} L(\theta)} = -2 \log \frac{\max_{\Theta_0} L(\theta)}{\max_{\Theta} L(\theta)} = 2(\ell(\hat{\theta}) - \ell(\hat{\theta}_0))$$

Under certain conditions this statistic will be approximately χ^2-distributed with degrees of freedom being the difference in model dimensions, that is, in the number of parameters in the two models, that is again, the number of parameters being "tested away".

We also saw that for testing single variance parameters equal to zero these conditions are not met, and we use either of two approaches: divide the p-value with 2 or use $1/2$ as degrees of freedom for the test. The problem is that 0 is on the boundary of the parameter space, so for testing other values than zero it would still be OK to use the χ^2_1-distribution.

It is also possible to compare different models by likelihood methods even though they are not nested within each other, as just discussed. Then typically the AIC and/or BIC values would be used:

$$\begin{aligned} \text{AIC} &= -2 \cdot \log L(\hat{\theta}) + 2 \cdot p \\ \text{BIC} &= -2 \cdot \log L(\hat{\theta}) + \log(n) p \end{aligned}$$

where n is the number of observations and p is the number of parameters in the model. One would then chose the model with the smallest AIC or BIC value. In R these can be extracted from model objects just as the likelihood value:

|||| Example 10.2 One sample situation with NO random effects

```
npar <- 2
AIC(lmfit)
```

[1] 23.72

```
-2*logLik(lmfit) + 2*npar
```

'log Lik.' 23.72 (df=2)

```
BIC(lmfit)
```

[1] 23.3

```
-2*logLik(lmfit) + log(n)*npar
```

'log Lik.' 23.3 (df=2)

Remark 10.3

For linear hypotheses within linear normal models it can be found that the G-statistic coming from comapring the likelihoods is in one-to-one correspondence with the well-known F-statistic:

$$G = h(F)$$

where h is a monotone function. This means that probability statements for G can be re-expressed to probability statements for F, and since under linear and normal assumptions the F-statistic is EXACTLY F-distributed (under the null hypothesis), whereas the G-statistic is only approximately χ^2-distributed, it is for these situations better to use the F (as we have always done anyway). But it is re-assuring to know now, that this is in fact also likelihood ratio testing.

10.3 Wald Confidence intervals

Most classically, confidence intervals for parameter estimates within the likelihood approach would be based on a quadratic approximation of the likelihood, also called the Wald interval: (also called the asymptotic normality property, a version of the Central Limit Theorem)

$$\hat{\theta}_i \pm z_{1-\alpha/2} SE_{\hat{\theta}_i}$$

where the standard error then is found from the curvature of the likelihood in the maximum point, that is, from the second derivative of the log-likelihood function, the Hessian matrix if there is more then a single model parameter, then also called the (observed) Information matrix:

$$SE_{\hat{\theta}_i} = \sqrt{(I(\hat{\theta})^{-1})_i}$$

where $I(\theta)$ is the matrix of second derivatives of the negative log-likelihood function:

$$I(\theta) = \ell''(\theta)$$

Example 10.4 One sample situation with NO random effects

Continuing the one-sample normal example from above, we can find the second derivatives. Here we give only one of the four entries of the 2×2 matrix:

$$\frac{\partial^2 \ell(\mu, \sigma^2)}{\partial \mu^2} = \frac{n}{\sigma^2}$$

It can be shown (cf. Exercise 1) that when inserting the estimated mean and variance into

$$\frac{\partial^2 \ell(\mu, \sigma^2)}{\partial \mu \partial (\sigma^2)}$$

we get a zero, so the off-diagonal element of the observed Fisher information is zero:

$$I(\hat{\theta})_{12} = 0$$

This means that the inverse of the Fisher information is just the inverse of each diagonal element, so:

$$SE_{\hat{\mu}} = \sqrt{(I(\hat{\mu}, \hat{\sigma}^2)^{-1})_{11}} = \sqrt{\frac{\hat{\sigma}^2}{n}}$$

This confirms, what we allready new about the variance of the mean!

For the variance, it can be shown that (cf. Exercise 1)

$$SE_{\hat{\sigma}^2} = \sqrt{(I(\hat{\mu}, \hat{\sigma}^2)^{-1})_{22}} = \sqrt{\frac{2(\hat{\sigma}^2)^2}{n}}$$

So the Wald confidence interval for the variance would be:

$$\hat{\sigma}^2 \pm z_{1-\alpha/2} \hat{\sigma}^2 \sqrt{\frac{2}{n}}$$

which for the example data becomes:

```
sig2hat <- var(y)*(n-1)/n
sig2hat + c(-1, 1)*qnorm(0.975)*sig2hat*sqrt(2/n)
```

```
[1] -0.206 3.338
```

The Wald confidence interval for a variance is not a good one. The sampling ditribution of the variance estimate is not nicely approximated by a normal distribution with as low an n as in this case. From basic introductory statistics classes we also know that we would usually find a confidence interval for a variance from the more proper (small sample) χ^2-distribution which one can derive from the normal assumption on the residuals in a linear normal model.

10.4 Likelihood Confidence intervals

Generally, confidence intervals should be based on the proper sampling distribution. For the mean and variance of a single normal distribution, we can find these, as already

discussed. More generally, and for us when it comes to parameters of the variance-covariance structure of a mixed model in general, where the estimates do not have analytical expressions and just appear as the maximum point of a computer based likelihood optimization, it is not possible to easily find and definitely not express analytically these small sample sampling distributions.

A very good general alternative, as it shows, is to construct a confidence interval based directly on the likelihood function. Or differently put, let us use the following classical way of thinking of the confidence interval: The interval is containing the possible values for the parameter that we accept by a (two-tailed) hypothesis test. And then let us use the likelihood ratio approach to do the testing, as shown above. A single parameter testing situation will be using the χ^2-distribution with 1 degrees of freedom and the test statistic for the hypothesis $H_0 : \theta = \theta_0$ would be:

$$-2\log\frac{L(\theta_0)}{L(\hat{\theta})}$$

as the maximum of a single value of a single parameter only can be the same value inserted

|||| Method 10.5 The likelihood confidence interval

The $(1-\alpha)100\%$ likelihood confidence interval for θ (in a one-parameter setting) is given by

$$\left\{\theta| -2\log\frac{L(\theta)}{L(\hat{\theta})} < \chi^2_{1-\alpha}(1)\right\}$$

So the 95%-confidence interval becomes:

$$\left\{\theta| -2\log\frac{L(\theta)}{L(\hat{\theta})} < 3.84\right\}$$

equivalent to:

$$\left\{\theta|\frac{L(\theta)}{L(\hat{\theta})} > \exp\left(-3.84/2\right)\right\}$$

So the 95%-confidence interval for a parameter consists of all those values with a relative likelihood value greater than 0.1466. This confidence interval will generally have much better small sample properties, than the Wald interval.

10.5 Profile likelihood

The likelihood based confidence interval approach just presented was presented as a single parameter interval in a single parameter setting. It is possible to extend this to multi-parameter confidence regions for multi-parameter settings. What is mostly needed and used in practice, though, is single-parameter confidence intervals in multi-parameter settings. For this we need the so-called *profile likelihood* principle where subsets of the parameter space, e.g. a single parameter, can be profiled by maximizing all the other parameters (given the chosen one(s)), hence expressing a profile likelihood only as a function of the chosen (single) parameter.

|||| Example 10.6 One sample situation with NO random effects

The profile likelihood function for the mean μ is:

$$L_p(\mu) = L(\mu, \hat{\sigma}^2(\mu))$$

where $\hat{\sigma}^2(\mu)$ means the functional expression where the maximum point of $L(\mu, \sigma^2)$ with respect to σ^2 for a fixed μ is given. We found this above, and can insert:

$$\begin{aligned} L_p(\mu) &= L(\mu, \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)^2}} \exp\left(-\frac{(y_i-\mu)^2}{2\frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)^2}\right) \end{aligned}$$

Now this is a single parameter expression where the likelihood confidence interval approach as given can be applied. For simple linear models this is overly complicated as we have a perfect approach based on the t-distribution. And in fact the `confint` function that for mixed models would use something similar to the above to find the profile likelihood intervals for the fixed effect parameters, will for `lm` objects do nothing else but show the t-based intervals.

Similarly, the profile likelihood for the variance σ^2 can be expressed:

$$\begin{aligned} L_p(\sigma^2) &= L(\hat{\mu}(\sigma^2), \sigma^2) \\ &= L(\bar{y}, \sigma^2) \end{aligned}$$

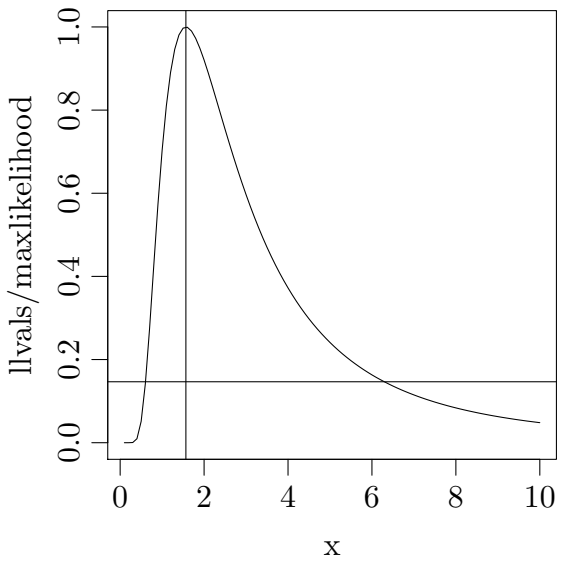
This becomes particularly simple, as the μ-estimate does not depend on the σ, so the profile likelihood is simply the two-parameter likelihood with the $\hat{\mu} = \bar{y}$ inserted for the mean μ. And, in fact we allready studied this above. Let us repeat this and show the 95%-cut off limits 0.1466:

```
lik2 <- function(x) prod(dnorm(y, x[1], sqrt(x[2])))

sigmalikelihood <- function(sig){lik2(c(mean(y), sig))}

maxlikelihood <- sigmalikelihood(var(y)*(n-1)/n)

x <- seq(0.1, 10, 0.1)
llvals <- rep(0, length(x))
j <- 0
 for (i in x){
  j <- j + 1
 llvals[j] <- sigmalikelihood(i)
}
plot(x, llvals/maxlikelihood, type="l")
abline(v = var(y)*(n-1)/n, h = exp(-3.84/2))
```



To actually get the cut-off points exactly one would have to solve for them in R:

```
CIfun <- function(sig) sigmalikelihood(sig)/maxlikelihood-exp(-3.84/2)

uniroot(CIfun, lower=0.001, upper=sig2hat)$root
```

[1] 0.604

```
uniroot(CIfun, lower=sig2hat, upper=1000)$root
```

```
[1] 6.294
```

Note how different this confidence interval is from the Wald based found above. Again, for `lm` objects, the `confint` function will not give this interval - it only provides the fixed effect intervals.

10.6 REstricted/REsidual Maximum Likelihood estimation (REML)

In Chapter 4 the following was written: The restricted (also known as residual) maximum likelihood method is a modification of the maximum likelihood method. Instead of minimizing the negative log likelihood function ℓ in step 2), the function ℓ_{re} is minimized, where ℓ_{re} is given by:

$$\frac{1}{2}\left\{\log|\mathbf{V}(\gamma)| + (\mathbf{y}-\mathbf{X}\beta)'(\mathbf{V}(\gamma))^{-1}(\mathbf{y}-\mathbf{X}\beta) + \log|\mathbf{X}'(\mathbf{V}(\gamma))^{-1}\mathbf{X}|\right\}$$

The two other steps 1) and 3) are exactly the same.

The intuition behind the raw maximum likelihood method, is that it should return the estimates that makes the actual observations most likely. The intuition behind the restricted likelihood method is almost the same, but instead of optimizing the likelihood of the observations directly, it optimizes the likelihood of the *full residuals*. The full residuals are defined as the observations minus the fixed effects part of the model. This focus on the full residuals can be theoretically justified, as it turns out, that these full residuals contain all information about the variance parameters.

This modification ensures, at least in balanced cases, that the random effect parameters are estimated without bias, and for this reason the REML estimator is generally preferred in mixed models.

Now let us as a new thing in this chapter 10 elaborate some more on how the REML likelihood appears. The REML likelihood is the likelihood function expressed for the residuals. It can argued that as only the residuals contain information about the variance parameters, we should focus on these already in the likelihood expression. This is the premise for the REML likelihod approach. The ”known variance” residuals in a mixed model are expressed as:

$$\mathbf{e} = (\mathbf{y} - \mathbf{X}\hat{\beta})$$

where

$$\hat{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

By "known variance" we mean that we use the true variance $\mathbf{V}$ in the expression. In a simple linear normal model the estimated residuals do not depend on the variance, and in these models actual residuals are independent from the mean parameter estimates. E.g. in the simplest of cases: The mean $\bar{y}$ is stochastically independent from the sample variance s^2. In the mixed model the "known variance" residuals are independent from the "known variance" fixed effect estimates, and the actual residuals are "almost" independent from the actual fixed effect estimates. If this independency property is used, we can express the likelihood function of the original data as the product of the likelihoods for residuals and parameter estimates:

$$L(\mathbf{y}, \beta, \gamma) = L(\mathbf{e}, \beta, \gamma) \cdot L(\hat{\beta}, \beta, \gamma)$$

Then isolating the residual likelihood and applying the log, we get: (remember that we use ℓ to denote the negative log-likelihood: $\ell = -\log L$)

$$-\ell(\mathbf{e}, \beta, \gamma) = -\ell(\mathbf{y}, \beta, \gamma) + \ell(\hat{\beta}, \beta, \gamma)$$

We see how the residual likelihood becomes the orginial likelihood minus the likelihood for the parameter estimates. Seen as a likelihood for the variance parameters, the orginal likelihood express the zero mean "true β" residuals multivariate normal part, whereas the second part express that the βs are actually estimated.

The (sampling) distribution of the $\hat{\beta}$ is, again expressed with the known variance is a well known thing from a general linear model point of view:

$$\hat{\beta} \sim N(\beta, (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1})$$

So expressing the multivariate normal negative log-likelihood for this distribution gives: (assume the number of parameters is p)

$$\ell(\hat{\beta}, \beta, \gamma) = \frac{1}{2}\left\{ p\log(2\pi) + \log|(\mathbf{X}'\mathbf{V}(\gamma)^{-1}\mathbf{X})^{-1}| + (\beta - \hat{\beta})'(\mathbf{X}'\mathbf{V}(\gamma)^{-1}\mathbf{X})^{-1}(\beta - \hat{\beta}) \right\}$$

The last term of this likelihood will end up having no contribution as the β will be estimated at $\hat{\beta}$ anyway (the β-estimate is unbiased) but inserting the first two, will then give the RE log-likelihood:

$$\begin{aligned}
\ell_{re}(\beta, \gamma) &= \frac{1}{2}\Big\{ n\log(2\pi) + \log|\mathbf{V}(\gamma)| + (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{V}(\gamma))^{-1}(\mathbf{y} - \mathbf{X}\beta) \\
&\quad - p\log(2\pi) - \log|(\mathbf{X}'(\mathbf{V}(\gamma))^{-1}\mathbf{X})^{-1}| \Big\} \\
&= \frac{1}{2}\Big\{ (n-p)\log(2\pi) + \log|\mathbf{V}(\gamma)| + (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{V}(\gamma))^{-1}(\mathbf{y} - \mathbf{X}\beta) \\
&\quad + \log|\mathbf{X}'(\mathbf{V}(\gamma))^{-1}\mathbf{X}| \Big\} \\
&\propto \frac{1}{2}\Big\{ \log|\mathbf{V}(\gamma)| + (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{V}(\gamma))^{-1}(\mathbf{y} - \mathbf{X}\beta) + \log|\mathbf{X}'(\mathbf{V}(\gamma))^{-1}\mathbf{X}| \Big\}
\end{aligned}$$

where we used that the matrix determinant of the inverse of a square matrix A equals the reciprocal value of the determinant, and hence changing the sign of the additional term in the likelihood.

Remark 10.7

The mixed model REML likelihood expression can also be justified by other approaches than the original residual likelihood approach. In Pawitan Y: In All Likelihood: Statistical modeling and inference using likelihood. Oxford University Press (2001) it is shown that a Baysian approach can also lead to the same likelihood expression. And also in Pawitan, one can read about the general principle of *modified profile likelihood*, and the REML likelihood function also appears as the modified profile likelihod for the variance parameters of the mixed model.

Example 10.8 One sample situation with NO random effects

Let us return the to simplest of all settings again. To express the RE-log-likelihood, we just need to find the additional term - the μ-sampling error term:

$$\begin{aligned} \log|\mathbf{X}'(\mathbf{V}(\gamma))^{-1}\mathbf{X}| &= \log|\mathbf{X}'\mathbf{X}/\sigma^2| \\ &= \log|n/\sigma^2| \\ &= \log(n) - \log(\sigma^2) \end{aligned}$$

Inserting this, will then give the Restricted negative log-likelihood function:

$$\begin{aligned} \ell_{RE}(\mathbf{y},\beta,\gamma) &= \frac{1}{2}\left\{(n-1)\log(2\pi) + n\log(\sigma^2) + \sum_{i=1}^{n}(y_i-\mu)^2/\sigma^2 + \log(n) - \log(\sigma^2)\right\} \\ &= \frac{1}{2}\left\{\log(n) + (n-1)\log(2\pi) + (n-1)\log(\sigma^2) + \sum_{i=1}^{n}(y_i-\mu)^2/\sigma^2\right\} \\ &\propto \frac{1}{2}\left\{(n-1)\log(\sigma^2) + \sum_{i=1}^{n}(y_i-\mu)^2/\sigma^2\right\} \end{aligned}$$

We see how the RE-likelihood corresponds to the original likelihood with just the single change of n into $n-1$. This has no effect on the μ-estimate but changes exactly the σ^2-estimate the well-known way, when the derivatives are found and solved:

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$$

And we can then profile this likelihood to get a confidence interval for this one:

```
REminusloglik <- function(x){
  (n-1)*log(x[2]) + sum((y-x[1])^2)/x[2]
}
optim(c(1,1), REminusloglik)$par
```

[1] 1.887 1.879

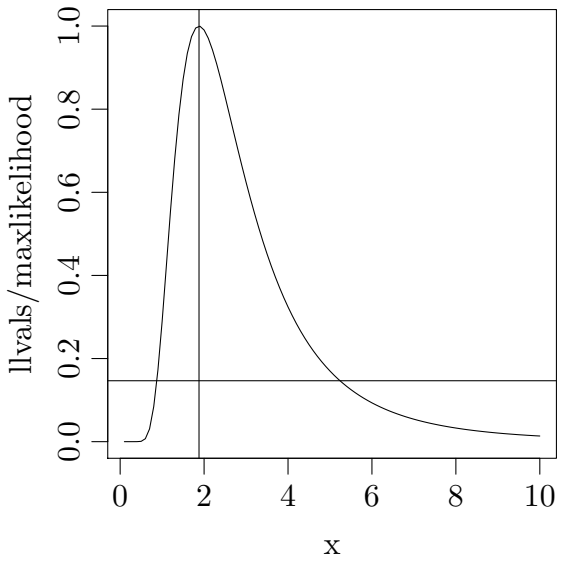
```
c(mean(y), var(y))
```

[1] 1.887 1.879

```
sigmalikelihood <- function(sig){exp(-REminusloglik(c(mean(y), sig)))}

maxlikelihood <- sigmalikelihood(var(y))

x <- seq(0.1, 10, 0.1)
llvals <- rep(0, length(x))
j <- 0
 for (i in x){
  j <- j + 1
 llvals[j] <- sigmalikelihood(i)
}
plot(x, llvals/maxlikelihood, type="l")
abline(v = var(y), h = exp(-3.84/2))
```



To actually get the cut-off points exactly one would have to solve for them in R:

```
CIfun <- function(sig) sigmalikelihood(sig)/maxlikelihood-exp(-3.84/2)

uniroot(CIfun, lower=0.001, upper=sig2hat)$root
```

[1] 0.8744

```
uniroot(CIfun, lower=sig2hat, upper=1000)$root
```

[1] 5.239

Let us just finally compare with the χ^2-based confidence interval, that would be the usual choice in a linear model:

```
(n-1)*var(y)/qchisq(0.975, n-1)
```

[1] 0.7321

```
(n-1)*var(y)/qchisq(0.025, n-1)
```

[1] 11.3

With low $n = 6$ still, the likelihood interval is considerably too low on the upper limit compared to the χ^2-based. With n a bit higher the likelihood interval becomes better.

|||| Example 10.9 One way ANOVA with random effects

Consider the one way analysis of variance model with random effect. The case with two observations withing each of three levels is used.

$$y_{ij} = \mu + b_i + \varepsilon_{ij}, \quad \text{where } b_i \sim N(0, \sigma_B^2) \text{ and } \varepsilon_{ij} \sim N(0, \sigma^2).$$

Here $i = 1, 2, 3$, $j = 1, 2$ and the random effects b_i and ε_{ij} are all independent. The matrix notation for this model is:

$$\underbrace{\begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix}}_{\mathbf{y}} = \underbrace{\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} \mu \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}}_{\mathbf{Z}} \underbrace{\begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}}_{\mathbf{u}} + \underbrace{\begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{pmatrix}}_{\varepsilon},$$

where $\mathbf{u} \sim N(0, \mathbf{G})$ and $\varepsilon \sim N(0, \sigma^2\mathbf{I})$. The covariance matrix $\mathbf{G}$ for the random effects is in this case a 3×3 diagonal matrix with diagonal elements σ_B^2. Notice how the matrix representation exactly correspond to model formulation, when the matrices are multiplied.

$$\begin{aligned}
\mathbf{V} &= \text{var}(\mathbf{y}) = \mathbf{ZGZ}' + \sigma^2\mathbf{I} \\
&= \begin{pmatrix}
\sigma^2+\sigma_B^2 & \sigma_B^2 & 0 & 0 & 0 & 0 \\
\sigma_B^2 & \sigma^2+\sigma_B^2 & 0 & 0 & 0 & 0 \\
0 & 0 & \sigma^2+\sigma_B^2 & \sigma_B^2 & 0 & 0 \\
0 & 0 & \sigma_B^2 & \sigma^2+\sigma_B^2 & 0 & 0 \\
0 & 0 & 0 & 0 & \sigma^2+\sigma_B^2 & \sigma_B^2 \\
0 & 0 & 0 & 0 & \sigma_B^2 & \sigma^2+\sigma_B^2
\end{pmatrix}
\end{aligned}$$

And it can be shown that: (If no errors were made in the derivations)

$$\begin{aligned}
\log|\mathbf{V}(\gamma)| &= 3\log(\sigma^2) + 3\log(\sigma^2+2\sigma_B^2) \\
\log|\mathbf{X}'(\mathbf{V}(\gamma))^{-1}\mathbf{X}| &= \log(n) - \log(\sigma^2+2\sigma_B^2) \\
(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{V}(\gamma))^{-1}(\mathbf{y}-\mathbf{X}\beta) &= \frac{1}{\sigma^2(\sigma^2+2\sigma_B^2)} \times \\
&\quad \left((\sigma^2+\sigma_B^2)\sum_{i=1}^{3}\sum_{j=1}^{2}(y_{ij}-\mu)^2 - 2\sigma_B^2\sum_{i=1}^{3}(y_{i1}-\mu)(y_{i2}-\mu)\right)
\end{aligned}$$

All of this comes from the fact that we can handle the 2×2-matrix

$$U = \begin{pmatrix} \sigma^2+\sigma_B^2 & \sigma_B^2 \\ \sigma_B^2 & \sigma^2+\sigma_B^2 \end{pmatrix}$$

$$\det(U) = |U| = (\sigma^2+\sigma_B^2)^2 - (\sigma_B^2)^2 = \sigma^2(\sigma^2+2\sigma_B^2)$$

$$U^{-1} = \frac{1}{\sigma^2(\sigma^2+2\sigma_B^2)}\begin{pmatrix} \sigma^2+\sigma_B^2 & -\sigma_B^2 \\ -\sigma_B^2 & \sigma^2+\sigma_B^2 \end{pmatrix}$$

So:

$$\begin{aligned}
\ell_{RE}(\mu, \sigma^2, \sigma_B^2) &\propto 3\log(\sigma^2) + 3\log(\sigma^2+2\sigma_B^2) - \log(\sigma^2+2\sigma_B^2) \\
&\quad + \frac{1}{\sigma^2(\sigma^2+2\sigma_B^2)}\left((\sigma^2+\sigma_B^2)\sum_{i=1}^{3}\sum_{j=1}^{2}(y_{ij}-\mu)^2 - 2\sigma_B^2\sum_{i=1}^{3}(y_{i1}-\mu)(y_{i2}-\mu)\right)
\end{aligned}$$

From this, one could insert $\mu = \bar{y}$, as this indeed will be the estimate here and then take the derivative with respect to the two variance parameters, equate with zero and solve. And we would get the following solution:

$$\begin{aligned}
\hat{\sigma}^2 &= MSE \\
\hat{\sigma}_B^2 &= \frac{MSE - MS_B}{2}
\end{aligned}$$

So in this case the REML approach gives the same results as the moments method based on the so-called expected mean squares (EMS) expressions. We mentioned these in the Chapter on hierarchical models in relation to a discussion of alternative ways of obtaining confidence intervals for variance parameters. For balanced settings it is possible to easily derive the EMS-expressions. And it can be done based on the factor structure diagram:

Method 10.10 Expected Mean Squares

For a random effect F, the expected mean square is

$$E(MS_F) = n_F\sigma_F^2 + \sum_{G:G>F} n_G\sigma_G^2$$

where n_F is the number of observations for each level of the factor F and the summation is over all those factors finer than F, that is, all those factors that can be "hit" by going backwards in the diagram starting from F.

Using one expression for each random effect gives a linear system of equations that can be solved for the variance components. In these balanced settings these estimates will coincide with the REML estimates. Or differently put: the system of equations coming out of the derivatives of the REML log-likelihood will in these cases be this system of EMS-values.

Example 10.11 One way ANOVA with random effects

The expected mean squares will then be:

$$E(MS_B) = 2\sigma_B^2 + \sigma^2$$

and

$$E(MSE) = \sigma^2$$

leading to the solution already expressed.

Slightly more general, if fixed effects are also present in the model, or if possibly some inbalance occur, one may still estimate variance components by a so-called moment method. In such other cases, expected mean squares are almost impossible to deduce by hand calculation. However, infact , for instance SAS PROC GLM can do this for us by means listing all the random effects in an additional RANDOM statement like:

```
proc glm;
class treat sub;
model y=treat sub;
Random sub;
```

This is not so easily available in R. But working with the REML approach would in any case be better for these cases.

10.7 Prediction of random effect coefficients

A model with random effects is typically used in situations where the subjects (or blocks) in the study are to be considered as representatives from a greater population. As such, the main interest is usually not in the actual levels of the randomly selected subjects, but in the variation among them. It is however possible to obtain predictions of the individual subject levels $\mathbf{u}$. The formula is given in matrix notation as:

$$\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta})$$

If $\mathbf{G}$ and $\mathbf{R}$ are known, $\hat{\mathbf{u}}$ can be shown to be 'the best linear unbiased predictor' (BLUP) of $\mathbf{u}$. Here, 'best' means minimum mean squared error. In real applications $\mathbf{G}$ and $\mathbf{R}$ are always estimated, but the predictor is still referred to as the BLUP.

|||| Example 10.12 One way ANOVA with random effects

Using the notation from above we can find that

$$\begin{aligned}
\hat{\mathbf{u}}_i &= (\sigma_B^2, \sigma_B^2)\, U^{-1} \begin{pmatrix} y_{i1} - \bar{y} \\ y_{i2} - \bar{y} \end{pmatrix} \\
&= (\sigma^2\sigma_B^2, \sigma^2\sigma_B^2) \begin{pmatrix} y_{i1} - \bar{y} \\ y_{i2} - \bar{y} \end{pmatrix} / \sigma^2(\sigma^2 + 2\sigma_B^2) \\
&= \frac{\bar{y}_i - \bar{y}}{\frac{\sigma^2}{2\sigma_B^2} + 1}
\end{aligned}$$

Note how these predictions are *shrinkage* versions of the fixed effect estimates: They will always be closer towards zero. The level of shrinkage depends on the ratio of the two variances: the smaller the group differences - the more shrinkage.

10.8 Exercises

|||| Exercise 1 One-sample normal model with no random effects

a) Find the second derivatives of the log-likelihood function and use that to find the two standard errors as claimed in the example of the Chapter.

|||| Exercise 2 One way ANOVA with random effects

a) Work with the second example of the chapter. Implement the REML-likelihood explicitly in R and optimize it with some toy data and compare that you get the same as you should AND the same as `lmer`.

b) Profile the likelihood with respect to each of the two variance parameters and compare with the results from `confint` applied to the `lmer` results.

c) Verify how the expression for the random effects BLUPs comes up and/or check that this is what `lmer` also finds on some toy data.