Exercises for linear mixed effect models, part 2

Timothée Bonnet

February 21, 2019

Contents

1 Uncertainty in random effects

1.1 Warming-up

* Exercise 1 Reading a summary in lme4

1. Load the dataset "thorndata.txt" to fit a linear model of "herbivory" as a function of "thorndensity". What is the estimate for the slope?

2. Add a random effect for site How does the estimate for the slope changes?

3. How much variation is explained by differences among blocks? Is there a measure of uncertainty for this estimate?

Answer of exercise 1

Part 1.

```
thorns <- read.table("thorndata.txt", header = TRUE)
lm(herbivory ~ thorndensity, data=thorns)

##
## Call:
## lm(formula = herbivory ~ thorndensity, data = thorns)
##
## Coefficients:
##  (Intercept)  thorndensity
##        3.256         0.252
```

The slope is 0.2524.

Part 2.

```
library(lme4)

## Loading required package:  Matrix

summary(lmer(herbivory ~ thorndensity + (1|site), data=thorns))

## Linear mixed model fit by REML ['lmerMod']
## Formula: herbivory ~ thorndensity + (1 | site)
##    Data: thorns
##
## REML criterion at convergence: 165.3
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -3.488 -0.606  0.109  0.523  2.873
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  site     (Intercept) 2.129    1.459
##  Residual             0.238    0.488
## Number of obs: 100, groups:  site, 5
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)     7.765      0.844    9.20
## thorndensity   -0.919      0.139   -6.63
##
## Correlation of Fixed Effects:
##             (Intr)
## thorndensty -0.632
```

The slope is now NULL.

Part 3.

```
summary(lmer(herbivory ~ thorndensity + (1|site), data=thorns))
as.numeric(VarCorr(lmer(herbivory ~ thorndensity + (1|site), data=thorns))$site)
```

The variance explained by block is (careful, if you extract the standard deviation (Std.Dev.) you need to square it to obtain a variance.). There is no measure of uncertainty for the estimate of variance in block. In the `summmary` the `Std.Dev.` is simply the square root of `Variance`.

1.2 Confidence intervals and Tests

Sometimes random effects are part of the experimental design and are in the models only to control for confounding effects. But sometimes we care about their value or their statistical significance.

* Exercise 2 CI for variance components in lme4

Use the function `confint` to estimate confidence intervals for the variance component for "site".

Answer of exercise 2

```
lmm1 <- lmer(herbivory ~ thorndensity + (1|site), data=thorns)
conf <- confint(lmm1)

## Computing profile confidence intervals ...

conf[1,]^2

##  2.5 % 97.5 %
## 0.5247 8.1932
```

* Exercise 3 Testing variance components in lme4

Use the function anova to test the statistical significance of the random effect "site" in the thorn dataset.

Answer of exercise 3

```
lm0 <- lm(herbivory ~ thorndensity , data=thorns)
lmm1 <- lmer(herbivory ~ thorndensity + (1|site), data=thorns)

anova(lm0, lmm1)#doesn't work!

## Error:  $ operator not defined for this S4 class

anova(lmm1, lm0)# mixed model must go first!

## refitting model(s) with ML (instead of REML)

## Data: thorns
## Models:
## lm0: herbivory ~ thorndensity
## lmm1: herbivory ~ thorndensity + (1 | site)
##      Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## lm0   3 224 231   -109      218
## lmm1  4 172 182    -82      164  53.4      1    2.7e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value you get is 2.6595×10^{-13}, however, be aware this is a conservative estimate!

*** Exercise 4 Does null-hypothesis testing work for random effect?

In principle we want p-values to follow a uniform distribution under the null-hypothesis; in particular the frequency of p-values below 0.05 should be 0.05 when the effect tested for is absent. Is it the case for tests of random intercepts? Simulate some data sets with a random intercept that has a variance of zero, test the significance of this random effect with anova and record the p-values. You can use the following simulation template:

```r
RandomVariance <- 0
sampsize <- 500
nbblocks <- 30

pvals <- vector(length = 1000)

for (i in 1:1000)
{
  x <- rnorm(sampsize,mean = 4, sd=0.25)
  block <- sample(x = 1:nbblocks, size = sampsize, replace = TRUE)
  blockvalues <- rnorm(n = nbblocks, mean = 0, sd = sqrt(RandomVariance))
  y <- 8 - x + blockvalues[block] + rnorm(sampsize,0,1)
  dat <- data.frame(response = y, predictor = x, block=block)

  XXX <- lm(XXX)
  XXX <- lmer(XXX)
  XXX <- anova(XXX)
  XXX
}
```

Answer of exercise 4

```r
set.seed(1234)
RandomVariance <- 0
sampsize <- 500
nbblocks <- 30

pvals <- vector(length = 1000)
altpvals <- vector(length = 1000)
for (i in 1:1000)
{
  x <- rnorm(sampsize,mean = 4, sd=0.25)
block <- sample(x = 1:nbblocks, size = sampsize, replace = TRUE)
blockvalues <- rnorm(n = nbblocks, mean = 0, sd = sqrt(RandomVariance))
y <- 8 - x + blockvalues[block] + rnorm(sampsize,0,1)
dat <- data.frame(response = y, predictor = x, block=block)
lm0 <- lm(response ~ 1 + predictor, data=dat)
lmm0 <- lmer(response ~ 1 + predictor + (1|block), data=dat )
(LRT0 <- anova(lmm0, lm0)) #mixed model must come first!
pvals[i] <- LRT0$`Pr(>Chisq)`[2] # the p-value
altpvals[i] <- 1-pchisq(LRT0$Chisq[2],0.5) # a better p-value
}
hist(pvals)
hist(altpvals)

mean(pvals<0.05)
mean(altpvals<0.05)
```

*** Exercise 5 Do kangaroos have personalities?

We want to know whether the distance at which kangaroos run away when we approach is consistent within individuals. Load the dataset "roo.csv" and fit linear mixed models of "EscapeDistance" to find out. What variables to include? Does individual repeatability explain more variance than expected by randomly grouping observations? What is the repeatability of behaviour?

Answer of exercise 5

```r
roo <- read.csv("roo.csv")
```

There is no simple answer to what model you should fit, because it depends what you think is part of intrinsic individual differences and what is not. The simplest model would be:

```
summary(lmer(EscapeDistance ~ 1 + (1|id), data=roo))

## Linear mixed model fit by REML ['lmerMod']
## Formula: EscapeDistance ~ 1 + (1 | id)
##    Data: roo
##
## REML criterion at convergence: 20185
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -4.432 -0.491  0.062  0.554  4.021
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  id       (Intercept) 59.9     7.74
##  Residual             38.3     6.19
## Number of obs: 2909, groups:  id, 826
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   31.842      0.309     103
```

That's not necessarily wrong, but be aware of what residual and id variance contain here: age, sex, date, location... should they?

The following model is probably too complex:

```
summary(lmer(EscapeDistance ~ 1 + Sex + as.factor(Age3) + (1|id) + (1|Mother), data=r

## Linear mixed model fit by REML ['lmerMod']
## Formula: EscapeDistance ~ 1 + Sex + as.factor(Age3) + (1 | id) + (1 |
##     Mother)
##    Data: roo
##
## REML criterion at convergence: 17007
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -3.942 -0.545 -0.004  0.583  6.884
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  id       (Intercept) 10.29    3.21
##  Mother   (Intercept)  2.49    1.58
##  Residual             14.62    3.82
## Number of obs: 2909, groups:  id, 826; Mother, 197
##
## Fixed effects:
##                   Estimate Std. Error t value
## (Intercept)         24.571      0.264   93.21
## SexMale              1.671      0.300    5.57
## as.factor(Age3)1    14.652      0.217   67.40
## as.factor(Age3)2    18.594      0.357   52.14
## as.factor(Age3)3    21.644      1.166   18.57
##
## Correlation of Fixed Effects:
##             (Intr) SexMal a.(A3)1 a.(A3)2
## SexMale     -0.567
## as.fct(A3)1 -0.326  0.097
## as.fct(A3)2 -0.203  0.070  0.484
## as.fct(A3)3 -0.060  0.023  0.141   0.169
```

indeed, "sex" and "Mother" (identity of an individual's mother) are part of an individual identity. Correcting for those will likely reduce the id variance. However, I think it makes sense to correct for age.

Maybe a good model would be:

```
summary(lmer(EscapeDistance ~ 1 + as.factor(Age3) + Julian + (1|Year) + (1|id) , data

## Linear mixed model fit by REML ['lmerMod']
## Formula: EscapeDistance ~ 1 + as.factor(Age3) + Julian + (1 | Year) +
##     (1 | id)
##    Data: roo
##
## REML criterion at convergence: 16982
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -3.789 -0.570  0.004  0.568  6.828
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  id       (Intercept) 13.236   3.638
##  Year     (Intercept)  0.823   0.907
##  Residual             13.985   3.740
## Number of obs: 2909, groups:  id, 826; Year, 8
##
## Fixed effects:
##                  Estimate Std. Error t value
## (Intercept)      19.28062    0.78098   24.69
## as.factor(Age3)1 16.20258    0.24313   66.64
## as.factor(Age3)2 20.43964    0.39803   51.35
## as.factor(Age3)3 23.78528    1.20335   19.77
## Julian            0.02646    0.00285    9.29
##
## Correlation of Fixed Effects:
##             (Intr) a.(A3)1 a.(A3)2 a.(A3)3
## as.fct(A3)1 -0.503
## as.fct(A3)2 -0.382  0.581
## as.fct(A3)3 -0.169  0.244   0.253
## Julian      -0.877  0.443   0.346   0.154
```

(but it is not unambiguous; you can disagree).

From there, we can test the statistical significance of repeatability as:

```
lmfull <- lmer(EscapeDistance ~ 1 + as.factor(Age3) + Julian + (1|Year) + (1|id) , da
lmreduced <- lmer(EscapeDistance ~ 1 + as.factor(Age3) + Julian + (1|Year) , data=roo
anova(lmfull, lmreduced)

## refitting model(s) with ML (instead of REML)

## Data: roo
## Models:
## lmreduced: EscapeDistance ~ 1 + as.factor(Age3) + Julian + (1 | Year)
## lmfull: EscapeDistance ~ 1 + as.factor(Age3) + Julian + (1 | Year) +
## lmfull:     (1 | id)
##           Df   AIC   BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## lmreduced  7 17793 17834  -8889    17779
## lmfull     8 16988 17036  -8486    16972   806      1     <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

and estimate repeatability:

```
as.numeric(VarCorr(lmfull)[1])/(as.numeric(VarCorr(lmfull)[1])+
                                as.numeric(VarCorr(lmfull)[2])+
                                sigma(lmfull)^2)

## [1] 0.472
```

** Exercise 6 Testing variance components in MCMCglmm

lme4 can have difficulties estimating random effect parameters when models get a bit complex. An very powerful alternative, I recommend MCMCglmm. Try and fit one of the kangaroo models. Use summary and plot to discuss the importance of the random effects.

Answer of exercise 6

```
library(MCMCglmm)
mcmod <- MCMCglmm(EscapeDistance ~ 1 + as.factor(Age3) + Julian , random =~ Year + id
summary(mcmod)
plot(mcmod)
repeatability <- mcmod$VCV[,"id"]/(mcmod$VCV[,"id"]+mcmod$VCV[,"Year"]+mcmod$VCV[,"un
plot(repeatability)
mean(repeatability); HPDinterval(repeatability)
```

2 Beyond random intercepts

So far we have considered random effects around intercepts only (that is the meaning of the "1" in the lme4 syntax (1| re)). But random effects can be around fixed effects. You may have heard of "random interactions", "random slopes", "random regressions"...

** Exercise 7 Beetles: build a model

Load the dataset "beetles.csv". It contains (fake) data from an (real) experiment on gene-by-environment interactions. The variable of interest is the mass of beetles born in two different environments, from different parents, and in different cages. Assuming that we can measure genetic varition with parent random effects, we wonder if different genomes respond differently to different environments. **Build the model corresponding to this question in lme4.**

(hints: you could start from a lm() of mass modeled by environment, then add random intercepts, and finally a little something more).

** Exercise 8 Beetles: look at the model

What are the variances related to genetic differences? How are they correlated? Does genetic variation explain a lot of the total variation we observe? Try and draw a representation of genetic variation in the two environments.

*** Exercise 9 Beetles: interpret

Interpret model outputs (use raw numbers and / or graphes) to answer the following:
Is there evidence for genetic variation? Do the two environment differ in their effects on beetles?
Is there evidence for genetic variation in the response to the environment?
Does that mean that genomes good at environment 1 are bad at environment 2?

2.1 Correlated random effects

In all of the above, we have assumed that random effect levels to be perfectly correlated (e.g., observations from the same year) or not at all correlated (e.g., observations from different years). It can be very interesting to allow for intermediate values, in particular for models of spatio-temporal autocorrelation, phylogenetics, quantitative genetics.