Graded Project MAS6003

Registration number 160223367

April 3 25, 2017

Contents

It may be assumed that seeds from which the trees in the data-set grew were selected randomly from a seed bank and that the trees grew independently of each other.

Task 1 - Identify model

“Use the data to identify a suitable growth model for Olly Bolly Dob Dob plants aged between 3 and 25 weeks.”

We examine the data by making a line plot of height versus age for all OllyBolly trees. By inspecting the plot we can see that the growth slows down with age.

```r
g1 <- ggplot(OllyBolly, aes(x = age , y=height)) + geom_line(aes(colour=Seed)) +
  theme(legend.position="none") + labs(title="A")
```

An examination of the data also revealed that a square root transformation of the heights is required to improve homoscedasticity of the model residues. In addition I did examine if plotting the square root of height versus the logarithm of the age makes the slope constant which would simplify the model be eliminating higher order terms.

```r
g2 <- ggplot(OllyBolly, aes(x = age , y=height)) + geom_line(aes(colour=Seed))+
  scale_x_log10()+ theme(legend.position="none") + scale_y_sqrt() + labs(title="B")
```

To represent $\log(age)$, $\sqrt{age}$ and $\sqrt{height}$ I introduced a new column `logAge`, `sqrtAge` and `sqrtHeight` in the OllyBolly dataframe.

```r
OllyBolly$logAge <- log(OllyBolly$age)
OllyBolly$sqrtAge <- sqrt(OllyBolly$age)
OllyBolly$sqrtHeight <- sqrt(OllyBolly$height)
```

After those transformations, for all ages starting with 5, we have approximately an affine growth model.

$$\log(height) = \alpha + \beta\sqrt{age}$$

. The growth curve from 3 to 5 weeks seems to come from a different process. Very likely the measurement at 3 weeks is unreliable since very close to the ground. Therefore, and because we anyhow do not plan to make predictions for the early growth phase, I decided to remove the age of 3 from the dataset.

```
OllyBolly <- subset(OllyBolly, age > 4)
```

Figure 1 shows the original and the transformed growth curves.

```
multiplot(g1,g2,cols=2)
```

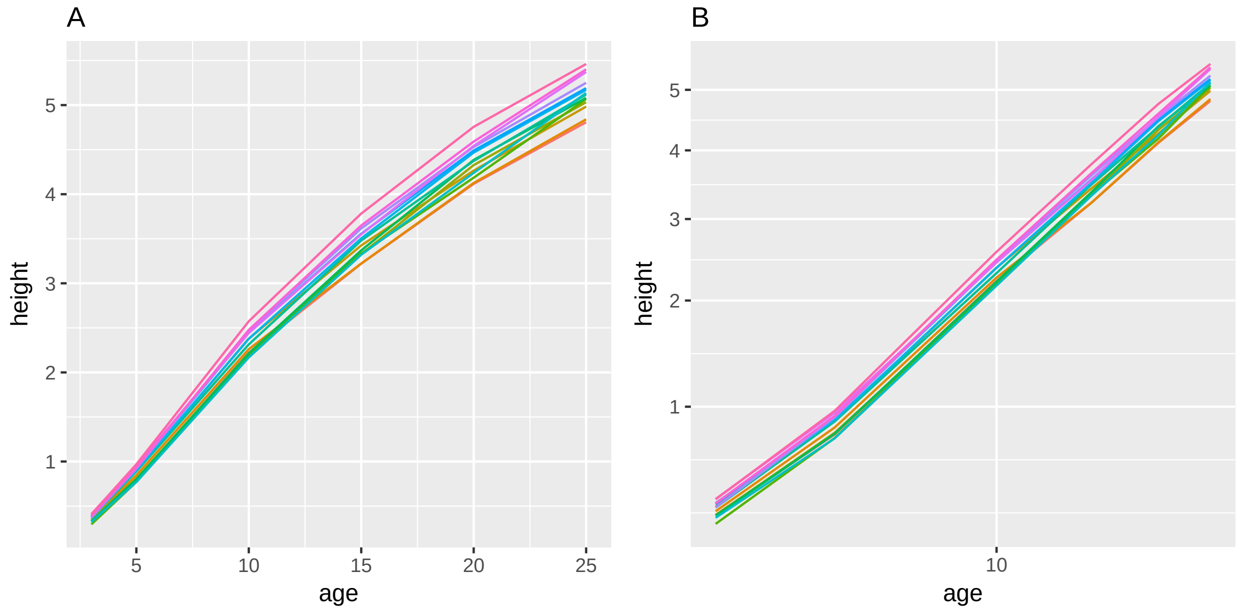

Figure 1: Tree height versus age (A), square root of height versus log10(age).

Choosing the Fixed effects model.

Since a preliminary examination of model residues for the simple affine model, discussed above, revealed that a higher order term for age needs to be included, I compared the affine model with the higher order model and tested if I can reject that they equally explain the data using ANOVA.

- $\sqrt{Height}$ depends on $\log(Age)$ only.
- $\sqrt{Height}$ depends on $\log(Age)$ and $\sqrt{age}$.

```
lme2 <- lmer(sqrtHeight ~ 1 + logAge + (1|Seed) ,
             data=OllyBolly, REML = F)
lme3 <- lmer(sqrtHeight ~ 1 + logAge + sqrtAge +  (1|Seed) ,
             data=OllyBolly, REML = F)
aovres <- anova(lme2, lme3)
obs.test.stat<- - 2*(logLik(lme2)-logLik(lme3))
1 - pchisq(obs.test.stat,1)
```

```
## 'log Lik.' 1.96995e-05 (df=4)
```

The ANOVA suggest that the more complex model very unlikely equals the simple model. In addition, I verify the obtained p-value using bootstrapping (see Appendix). Therefore, I am going to use the intercept and two variables `logAge` and `sqrtAge` to model the `logHeight`.

Examining random effects

I also examined if a more complex structure for the random effects should be used.

- Compare models with various nested uncorrelated random effects.

```
lme2R <- lmer(sqrtHeight ~ 1 + logAge + sqrtAge + (1|Seed)  ,
             data=OllyBolly, REML = F)
lme3R <- lmer(sqrtHeight ~ 1 + logAge + sqrtAge + (1|Seed) + (logAge - 1|Seed)  ,
             data=OllyBolly, REML = F)
lme4R <- lmer(sqrtHeight ~ 1 + logAge + sqrtAge + (1|Seed) + (logAge - 1|Seed) +
                (sqrtAge - 1|Seed) , data=OllyBolly, REML = F)
anovas <- anova(lme2R, lme3R, lme4R)
```

The ANOVA indicated that including $\sqrt{age}$ into the random effects does not significantly improve the model fit (results not shown). Finally, I also did examine if the correlation of the random effects should be modeled.

```
lme5R <- lmer(sqrtHeight ~ 1 + logAge + sqrtAge +  (1 + logAge|Seed) ,
             data=OllyBolly, REML = F)
anovasres <- anova(lme3R, lme5R)
obs.test.stat<- - 2*(logLik(lme3R)-logLik(lme5R))
1 - pchisq(obs.test.stat,1)
```

```
## 'log Lik.' 0.007381803 (df=6)
```

Because of the small p-value, I reject the hypothesis that model `lme5r` and `lme3r` are equal and accept that `lme5r` better explains the data.

Checking model assumptions

The model I selected is:

$$Y_{i,j} = \beta_0 + a_i + (\beta_1 + b_i)x_j + (\beta_2)x'_j + \epsilon_{ij}$$

where Y_{ij} is the height of the plant i after time j. β_0 is the population intercept, β_1 and β_2 describe the growth curve,

$$\begin{pmatrix} a_i \\ b_i \end{pmatrix} \sim N(0, \Sigma)$$

are the random effects with Σ the covariance matrix, and $\epsilon \sim N(0, \sigma^2)$ are the model residues. For the `summary` of the model see Appendix. For all further analysis I will be using a a model fit obtained using REML.

```
lmeChosen <- update(lme5R,  REML = T)
```

To check model assumptions I am using the following plots:

- I extract variances of random effects an make q-q plots for the variables Intercept and `logAge` (Panel - A and B) - looking for a straight line in both.
- Subject level residuals vs fitted values - looking for random scatter (Panel - C)

- q-q plot of subject-wise residuals to check the assumption that the errors ϵ are normally distributed (Panel - D).
- Population level residuals (Level 0) vs fitted values - looking for random scatter (Panel - E)
- Assess general fit of the model by plotting the measured tree heights vs fitted heights (Panel -F).

```r
raneffVar <- VarCorr(lmeChosen)
stdevs <-attributes(((raneffVar))$Seed)$stddev
par(mfrow=c(2,3))
randomEffects <- ranef(lmeChosen)[[1]]
qqnorm(randomEffects[,1],main=paste("A - ",names(randomEffects)[1])) #Panel - A
abline(c(0,stdevs[1]),col=2)
qqnorm(randomEffects[,2],main=paste("B - ", names(randomEffects)[2]))#Panel - B
abline(c(0,stdevs[2]),col=2)
plot(fitted(lmeChosen), resid(lmeChosen), main="C - subject level residues")# Panel - C
qqnorm(resid(lmeChosen), main="D - model residues") # Panel - D
abline(c(0,(summary(lmeChosen))$sigma),col=2)
fitted.level0<-lmeChosen@pp$X %*% fixef(lmeChosen) # Panel - E
resid.level0<-OllyBolly$sqrtHeight-fitted.level0
plot(jitter(fitted.level0),resid.level0, main="E - population level residues")
plot(fitted(lmeChosen),OllyBolly$sqrtHeight, main="F") # Panel - F
abline(0,1,col=2)
```

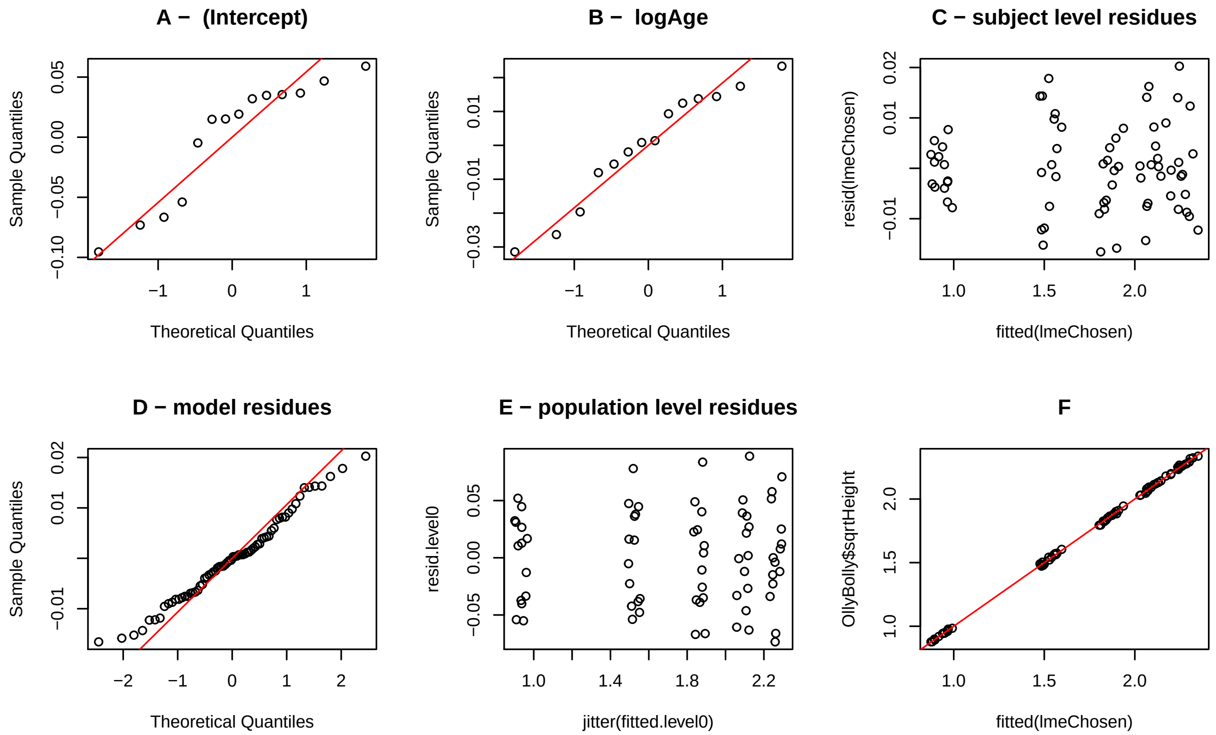

Figure 2: Diagnostic model plots. A - q-q plot of random effects for Intercept variable, B- q-q plot of random effects for logAge variable, C - subject level residues vs fitted values, D - q-q plot of subject wise residues, E - population level residuals, F - general fit of model

Task 2 - Estimate height at 25 weeks

“Estimate the height at age 25 weeks of another plant grown under similar conditions to those here, but from a different seed chosen randomly from the same seed bank.”

Using the function `simulate` we can simulate responses corresponding to the fitted model object.

```
new.height<-simulate(lmeChosen, nsim=10000)^2
new.height.25 <- new.height[OllyBolly$age==25,]
```

The expected height of the plant will be 5.1371261. Furthermore, we can also say that in 95% of cases, the height of the plant will be in the range 4.7395084, 5.546239.

Alternatively, we can use the fixed effect coefficients to predict the expected height of a tree at the age of 25.

```
fixef(lmeChosen)
```

```
## (Intercept)      logAge     sqrtAge
## -0.44103334  0.95649597 -0.07432639
```

```
(pred25 <- fixef(lmeChosen)[1] + fixef(lmeChosen)[2] *log(25) + fixef(lmeChosen)[3] * sqrt(25))^2
```

```
## (Intercept)
##    5.135556
```

Task 3 - Estimeate height at 25 given additional information about seed at age of 10 weeks.

“Furthermore, suppose that in addition, you are later told that this plant is 2.6m tall at 10 weeks of age. Provide an updated estimate of the height of this plant at age 25 weeks.”

To answer the question I am going to use simulation results from Task 2.

However, this time I selected only those simulated seeds where the height at the age of 10 is very close to 2.6 (larger than 2.59 and less 2.615).

```
new.height.10 <- new.height[OllyBolly$age==10,]
height.10_2.6 <- which(new.height.10 > (2.59) & new.height.10 < (2.615),arr.ind = T)
```

There are 1070 such seeds and their mean is 2.6016889. Only for those seeds, I examine the height at 25.

```
new.height.25 <- new.height[OllyBolly$age==25,]
height.25_height2.6At10 <- new.height.25[height.10_2.6]
```

The average height of the tree in meters after 25 weeks, for those trees that were 2.6 meter high at the age of 10 weeks, is:

```
mean(height.25_height2.6At10)
```

```
## [1] 5.522321
```

In addition, we easily can estimate the 95% confidence interval which is:

```
quantile(height.25_height2.6At10, c(0.025,0.975))
```

```
##     2.5%    97.5%
## 5.328965 5.724868
```

Figure 3 illustrates the distribution of the simulated heights at age of 25 for all trees (Panel A) and those with height 2.6 at the age of 10 weeks (Panel B).

```r
par(mfrow=c(1,2))

# Panel A
hist(unlist(new.height.25), main="A", xlab="height")
Olb25<-subset(OllyBolly, age==25)
points(Olb25$height, rep(1000, nrow(Olb25)), col=2, type="p")

# Panel B

plot(density(unlist(new.height.25)), ylim=c(0,4), main="B",xlab="height")
lines(density(height.25_height2.6At10),col=2)
legend("topleft", legend = c("all seeds", "seeds h=2.6 at 10w"), lty = 1, col=c(1,2))
```

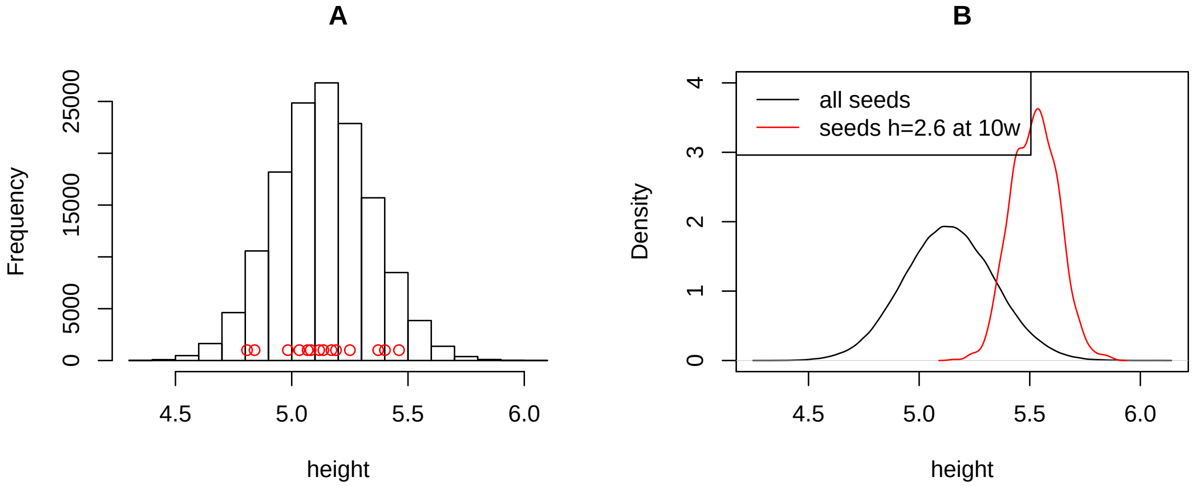

Figure 3: Panel A : Histogram of simulated heights at age 25. The red dots indicate the measured tree heights at age 25. Panel B: Density plot of heights at age 25 (black) while in red is the density plot at age 25 weeks for those seeds which had a height of 2.6 at the age of 10.

Conclusions

In the model assumption check section, it can be seen that the random effects are not ideally normally distributed (Figure 2 Panel A and B). Furthermore, although I did try to remove heteroscedasticity by square root transforming the heights the early time points exhibit less variance at the subject level as well as on the population level (Figure 2 Panel C and F). Overall, however, the model fits the data extremely well, the residues are extremely small compared with the effect size. This might indicate an overfitting. However, the fixed effect model is still rather simple, and therefore I think the model is correct and well describes the data.

Appendix

Comparison of the fixed effects model using Bootsrap.

```
N<-250
boot.test.stats<-rep(0,N)
for(i in 1:N){
  if(i %% 100 == 0){
    print(i)
    }
  new.height<-unlist(simulate(lme2))
  fm.reduced.new <-update(lme2, new.height ~ . )
  fm.full.new <- update(lme3, new.height ~ .)
  boot.test.stats[i]<- -2*(logLik(fm.reduced.new)-logLik(fm.full.new))
}
```

```
## [1] 100
## [1] 200
```

```
hist(boot.test.stats, prob=T, breaks=40, xlim = c(0, obs.test.stat+3))
curve(dchisq(x,df=1),from=0,to=40,add=T,col=3)
points(obs.test.stat,0,pch=4,col="red")
abline(v=obs.test.stat, col=2, lwd=2)
```

```
mean(boot.test.stats > obs.test.stat)
```

```
## [1] 0
```

The bootstrap analysis confirms that the more complex model significantly improves the fit with the data.

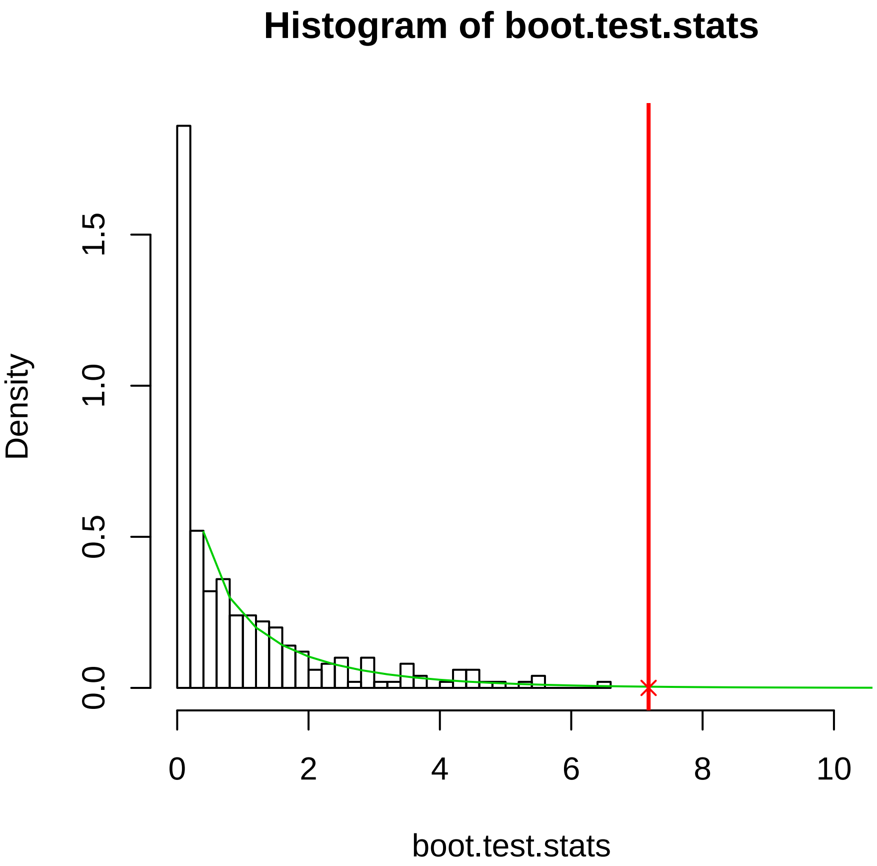

Figure 4: Hisogram of simulated test statistics. The green curve represents a chisq distribution with 1 df. The red vertical line are the observed test statistic.

Model summary

```r
summary(lmeChosen)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: sqrtHeight ~ 1 + logAge + sqrtAge + (1 + logAge | Seed)
##    Data: OllyBolly
##
## REML criterion at convergence: -333.4
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.55609 -0.61975 -0.00233  0.41004  1.90185
##
## Random effects:
##  Groups   Name        Variance  Std.Dev. Corr
##  Seed     (Intercept) 0.0029496 0.05431
##           logAge      0.0003391 0.01841  -0.70
##  Residual             0.0001135 0.01065
## Number of obs: 70, groups:  Seed, 14
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) -0.44103    0.01709  -25.81
## logAge       0.95650    0.01962   48.74
## sqrtAge     -0.07433    0.01109   -6.70
##
## Correlation of Fixed Effects:
##         (Intr) logAge
## logAge  -0.572
## sqrtAge  0.400 -0.961
```