Why Linear Mixed Effects Models? Reproductive output of a rare plant: random effects of populations

In demography is habitual to estimate reproductive output of individuals as a component together with other vital rates in population models (Quintana-Ascencio et al. 2003). We could use *Hypericum cumulicola* data to evaluate a linear regression model to predict number of reproductive structures of individuals with different heights as if individuals from the same population were independent. However, we recognize that plants in the same population are likely to be more similar to each other (and consequently less independent) than those in other populations. These random effects should be considered to avoid pseudo-replication and provide generality to our results. Here, we introduce a method to incorporate hierarchical random effects in our models.

Figure 1. *Hypericum cumulicola in the scrub*

We prepare the data

```
orig_data <- read.table("hypericum_data_94_07.txt", header=T)
dt <- subset(orig_data, !is.na(ht_init) & rp_init > 0 & year < 1997)
yr <- unique(dt$year)
height <- dt$ht_init
fruits <- dt$rp_init
id <- dt$tag
year <- dt$year
```

We call the libraries of two packages that we will need during the analysis

```
library(nlme)
library(bbmle)
```

A table is created to deposit the coefficients of the effects of the three models that we will evaluate.

```
table_coef <- array(0,c(3,2))
colnames(table_coef) <- c("intercept","slope")
rownames(table_coef) <- c("no mixed","random slope","random intercept & slope")
```

We estimate the coefficients of the model assuming complete independence of the data and plot the model (in blue) in Figure 2.

Complete independence

$$log(reproductive\ structures)_k = \beta_1 + \beta_2 * log(height)_k \qquad \epsilon \sim N(0, \sigma)$$

```
m1 <- lm(lfr~lgh,data=dt)
summary(m1)
table_coef[1,] <- m1$coefficients

Beta_1 <-Beta_2 <- array(0,c(1,length(site)))
colnames(Beta_1)=site
colnames(Beta_2)=site


par(mfrow=c (1,1))

plot(dt$lgh,dt$lfr,pch=16,ylab="log(fruits)",
     xlab="log(height)",col="grey",cex=0.5,ylim=c(0,8),main="Individual populations")

for (j in 1:length(site)){
MU <- lm(lfr~lgh,subset=(bald==site[j]),data=dt)
Mi <- summary(MU)
x1 <- dt$lgh[dt$bald==site[j]]
K <-  order(x1)
lines(sort(x1),predict(MU)[K],col="red",lwd=1.1)
Beta_1[j] <- Mi$coefficients[1,1]
Beta_2[j] <- Mi$coefficients[2,1]}
I <- order(dt$lgh)
lines(sort(dt$lgh),predict(m1)[I],col="blue",lwd=3)

Beta_1
Beta_2

mB1 <- mean(rowSums(Beta_1)/14)
mB2 <- mean(rowSums(Beta_2)/14)

hist(mB1-Beta_1,5,main="Intercept")
hist(mB2-Beta_2,5,main="Slope")
```

The output of the general model under the assumption of complete independence should be familiar. This model explains approximately 66 % of the variance.

```
Call:
lm(formula = lfr ~ lgh, data = dt)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2555 -0.4869  0.0289  0.5258  4.8065

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.35037    0.17916  -41.03   <2e-16 ***
```

```
lgh          3.23867    0.05043   64.22   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.8164 on 2099 degrees of freedom
Multiple R-squared: 0.6627,     Adjusted R-squared: 0.6625
F-statistic:  4124 on 1 and 2099 DF,  p-value: < 2.2e-16
```

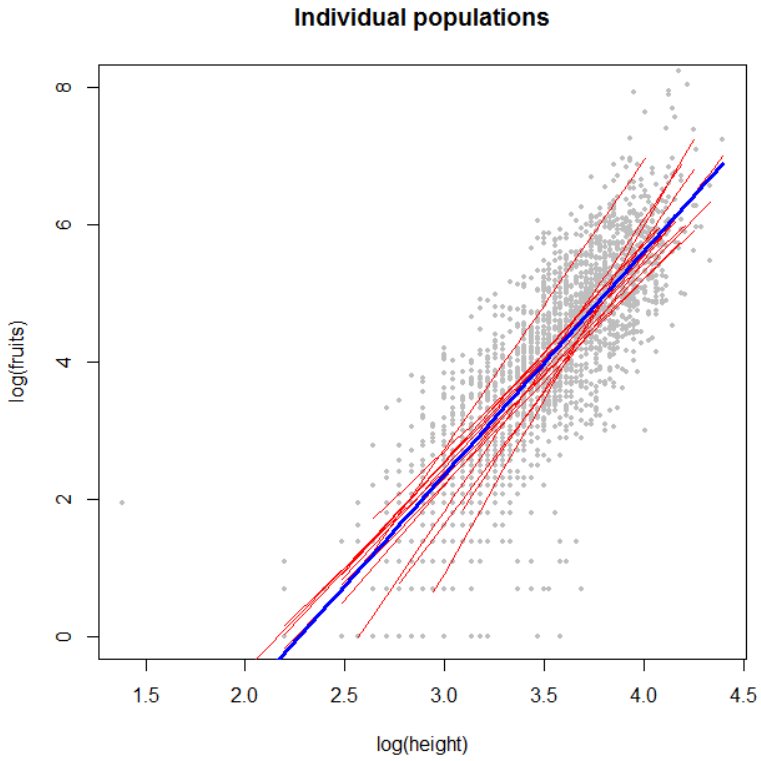


Figure 2. *Plot of log(height) vs log(fruits), data as points in grey, overall model assuming independence in blue and population specific models in red.*

We now estimate the coefficients of the model of each population. We plot the models for each of the 14 populations studied (in red) in Figure 2. We notice that there is variation for the models for each of these conditions. We could model the log (number of fruits) as a linear function of the log (height) using these specific intercepts (*Beta 1*) and slopes (*Beta 2*) for each population, but this will significantly reduce the degrees of freedom and limit the generality of our interpretation. A histogram of deviations of the coefficients of these models from those of the model with complete independence are in Figure 3. The coefficients per model by population are listed in Table 1, first the “intercept” (`Beta_1`) and then the “slope” (`Beta_2`).

Population specific

$$log(reproductive\ structures)_{ik} = \beta_{1_i} + \beta_{2_i} * log(height)_{ik} \qquad \epsilon_i \sim N(0, \sigma)$$

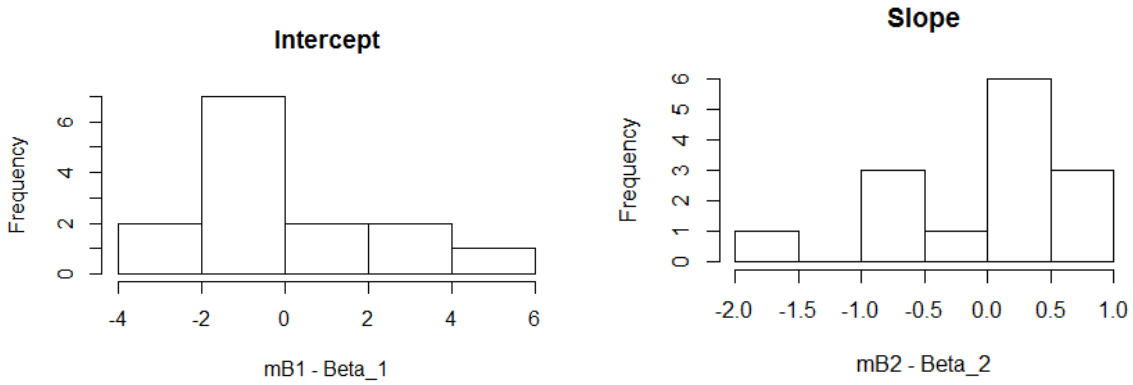


Figure 3. *Histograms of deviations of coefficients of models by population from those of the model with complete independence*

Table 1. *Coefficients per population and year. We use the ones with data grouped by population (all years).*

Beta1														
Yr/pop	**1**	**29**	**32**	**42**	**50**	**57**	**59**	**62**	**67**	**87**	**88**	**91**	**93**	**103**
1994	-10.6	-3.9	-13.7	-8.1	-6.4	6.1	-8.2	-10.3	-7.3	-8.2	-11.1	-6.3	-6.5	-6.0
1995	-8.1	-10.4	-10.4	-7.0	-10.3	-19.7	-6.3	-6.8	-6.8	-6.8	-10.3	-5.6	-7.1	-8.2
1996	-10.1	-8.0	-8.0	-6.4	-10.7	-18.8	-7.8	-5.1	-4.2	-10.2	-9.4	-4.3	-7.3	-3.7
All yrs	**-9.84**	**-6.53**	**-11.3**	**-7.02**	**-10.8**	**-14.2**	**-7.0**	**-6.9**	**-6.7**	**-7.8**	**-9.8**	**-5.5**	**-6.7**	**-6.0**

Beta2														
Yr/pop	**1**	**29**	**32**	**42**	**50**	**57**	**59**	**62**	**67**	**87**	**88**	**91**	**93**	**103**
1994	4.05	2.29	4.95	3.36	2.72	-0.94	3.55	4.20	3.14	3.49	4.63	3.01	3.01	2.82
1995	3.39	4.06	4.04	3.22	4.12	6.53	3.00	3.12	3.05	3.07	4.33	2.75	3.08	3.39
1996	3.85	3.43	3.32	2.94	4.17	6.18	3.36	2.62	2.26	3.80	4.01	2.35	3.14	2.16
All yrs	**3.83**	**3.02**	**4.26**	**3.16**	**4.23**	**5.05**	**3.19**	**3.15**	**2.97**	**3.33**	**4.18**	**2.72**	**3.01**	**2.80**

If we are interested in more general inference (and better use of the data), we could instead estimate the variation around the intercept (or both the intercept and slope) and assume that they are normally distributed. For more details on these assumptions see Zuur et al. 2009. We apply the function `lme` to obtain this model. We convert the variables population (bald) as factor.

```
dt$fbald <- factor(dt$bald)
```

We start assuming random intercepts by population but a common slope. In this case we ignore year effects. We specify the random component: `random=~1|fbald`. It represents that our data were nested within populations.

$$log(reproductive\ structures)_k = \beta_1 + \alpha_{1_i} + \beta_2 * log(height)_k$$

$$\alpha_1 \sim N(0, \sigma_1)$$
$$\epsilon \sim N(0, \sigma)$$

```
M1 <- lme(lfr~lgh,random=~1|fbald,data=dt)
summary(M1)
table_coef[2,] <- M1$coefficients$fixed
```

The output is presented below. It includes the model AIC and BIC. The residual variance is $\sigma^2 = 0.78^2 = 0.61$, the variance for the random effect due to population within year equal to $0.26^2=0.07$. The fixed effect intercept was 7.80 (s.e. = 0.19) and the fixed slope 3.36 (0.05).

```
Linear mixed-effects model fit by REML
 Data: dt
       AIC      BIC    logLik
  4984.305 5006.902 -2488.152

Random effects:
 Formula: ~1 | fbald
        (Intercept)  Residual
StdDev:   0.2580349 0.7823796

Fixed effects: lfr ~ lgh
                Value  Std.Error   DF   t-value p-value
(Intercept) -7.799101 0.19345517 2086 -40.31477       0
lgh          3.363026 0.05086119 2086  66.12165       0
 Correlation:
    (Intr)
lgh -0.93

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-5.4382227 -0.5645786  0.0408648  0.6112008  6.4232962

Number of Observations: 2101
Number of Groups: 14
```

We plot this model (Figure 4). The line in blue is the model obtained with the fixed components. The lines in red represent the variation estimated by population as their displacement from the population curve. The random intercept models are curves that shift by a factor that is normally distributed with a given variance. If the variance is large the shift is greater. The `fitted` command takes an argument from the function `lme`. The level = 0 takes the fitted values for fix effects, the level =1 takes those of population.

```
dt$fbald <- factor(dt$bald)
M1 <- lme(lfr~lgh,random=~1|fbald,data=dt,method ="REML")
summary(M1)
table_coef[2,] <- M1$coefficients$fixed
cr <- ranef(M1)

F0 <-fitted(M1,level=0)
F1 <-fitted(M1,level=1)
lgh <- sort(dt$lgh)

plot(lgh,predict(m1)[I],lwd=1,type="l",ylab="log(fruits)",xlab="log(height)",ylim=c(0,
8),main="Random Intercept", col="black")

points(dt$lgh,dt$lfr,pch=16,ylab="log(fruits)",
     xlab="log(height)",col="grey",cex=0.5,ylim=c(0,8))
```

```
for (j in 1:length(site)){
x1 <- dt$lgh[ dt$bald==site[j]]
y1 <- F1[dt$bald==site[j]]
K <-  order(x1)
lines(sort(x1),y1[K],col="red")
                        }
lines(lgh,F0[I],lwd=3,type="l",col="blue" )
```

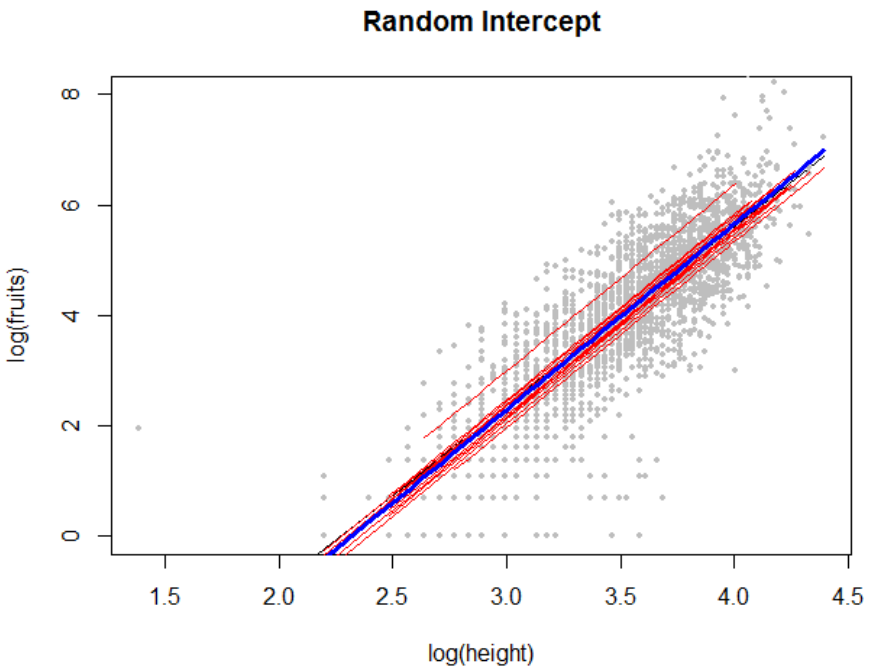


Figure 4. *Plot of log (height) vs log (fruits), data as points in grey, predicted fix effects of a model with random intercepts in blue and adding random intercept effects by population in red.*

Nakagawa and Schielzeth (2013) provide an equation describing the proportion of variance for mixed models. The variance explained by the fixed factor alone, which is called the marginal R^2 is:

$$R^2_{\mathrm{GLMM}(m)} = \frac{\sigma_f^2}{\sigma_f^2 + \sum_{l=1}^{u} \sigma_l^2 + \sigma_e^2 + \sigma_d^2}$$

The fixed effects variance is in the numerator and in the case of the current model is 67.7 %.

The proportion of the variance explained by both the fixed and random effects is called conditional R^2 and is described by

$$R^2_{\mathrm{GLMM}(c)} = \frac{\sigma_f^2 + \sum_{l=1}^{u} \sigma_l^2}{\sigma_f^2 + \sum_{l=1}^{u} \sigma_l^2 + \sigma_e^2 + \sigma_d^2}$$

In this equation, the numerator contains both the variance of the fixed, as well as the sum of the random variance components for each level. In this model is equal to 70.6 %.

We can now try a model that estimates random intercepts and slopes. This is specified in the model as `random=~1 + lgh|fbald`. In the output the estimated value $2.2^2 = 4.86$ indicates the variance in intercepts and $0.61^2 = 0.37$ that of the slopes. The negative correlation (-0.99) between random intercepts and random slopes indicates that populations with high intercepts tend to have lower slopes.

$$log(rep\ struct)_k = \beta_1 + \alpha_{1_i} + [\beta_2 + \partial_{2_i}] * log(height)_k$$

$$\alpha_1 \sim N(0, \sigma_1)$$

$$\partial_1 \sim N(0, \sigma_2)$$

$$\epsilon \sim N(0, \sigma)$$

```
M11 <- lme(lfr~lgh,random=~1 + lgh|fbald,data=dt,method ="ML")
summary(M11)
table_coef[3,] <- M11$coefficients$fixed
cr <- ranef(M11)

F0 <-fitted(M11,level=0)
F1 <-fitted(M11,level=1)
lfrs <- sort(dt$lgh)

plot(lfrs,predict(m1)[I],lwd=2,type="l",ylab="log(fruits)",xlab="log(height)",ylim=c(0
,8)
     ,main="Random Intercept & Slope",col="green")

points(dt$lgh,dt$lfr,pch=16,ylab="log(fruits)",
     xlab="log(height)",col="grey",cex=0.5,ylim=c(0,8))

for (j in 1:length(site)){
x1 <- dt$lgh[dt$bald==site[j]]
y1 <- F1[dt$bald==site[j]]
K <-  order(x1)
lines(sort(x1),y1[K],col="red")
                          }
lines(lgh,F0[I],lwd=3,type="l",col="blue")
```

The output is presented below.

```
Linear mixed-effects model fit by maximum likelihood
```

```
 Data: dt
     AIC      BIC  logLik
  4891.4 4925.301 -2439.7

Random effects:
 Formula: ~1 + lgh | fbald
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 2.2037935 (Intr)
lgh         0.6108079 -0.99
Residual    0.7584316

Fixed effects: lfr ~ lgh
                Value Std.Error   DF   t-value p-value
(Intercept) -8.200610 0.6175165 2086 -13.27999       0
lgh          3.469046 0.1712917 2086  20.25228       0
 Correlation:
    (Intr)
lgh -0.991

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-5.67788940 -0.56574901  0.03189972  0.62002756  5.75040517

Number of Observations: 2101
Number of Groups: 14
```

The variance explained is similar to the previous model. For fixed effects is equal to 67.8 % and for random effects is 74.0 %

We plot the fix effects model (in blue), those for the estimated shifts by population (in red) and that of the model under the assumption of complete independence (in green) in Figure 5.

Table 2. *Comparison of the coefficients of the population level for the three approaches (only fixed effects for mixed models), and their AICs. The model with random intercept and random slope was the most informative.*

	Intercept	s.e	slope	s.e	AIC
no mixed	-7.35	0.18	3.24	0.05	5113.8
random slope	-7.80	0.19	3.36	0.05	4984.3
random intercept & slope	-8.20	0.62	3.47	0.17	**4891.4**

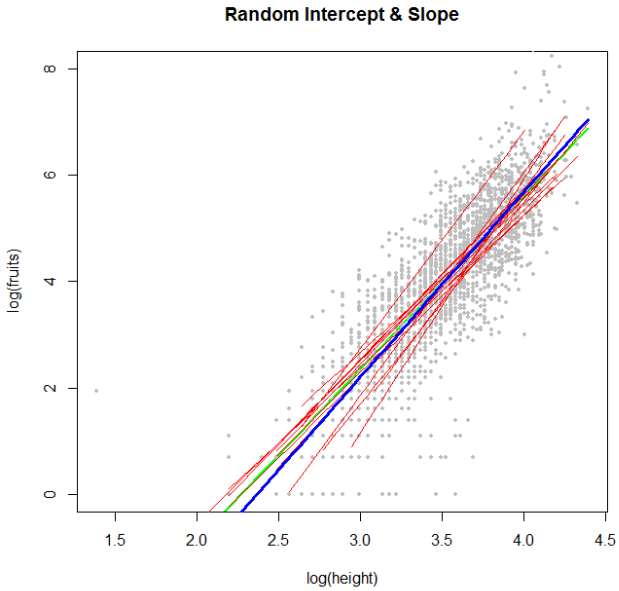


Figure 5. *Plot of log (height) vs log (fruits), data as points in grey, predicted fix effects of a model with random intercepts and slopes in blue and adding random intercept effects by population in red. The model assuming complete independence is in green.*

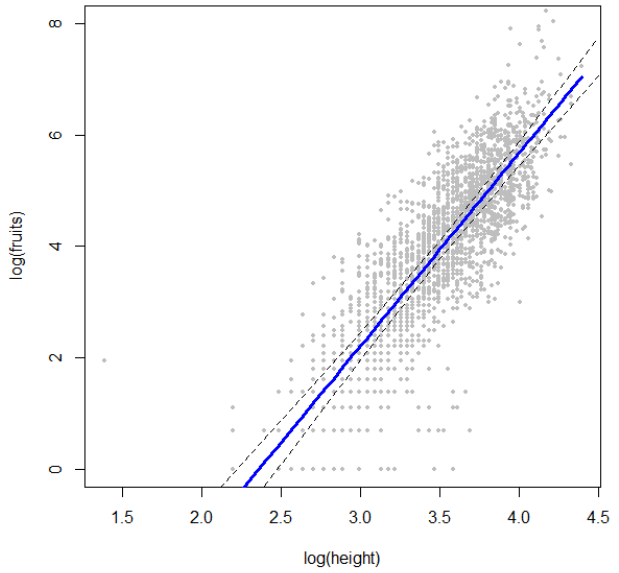


Figure 6. *Plot of log (height) vs log (fruits), data as points in grey, predicted fix effects of a model with random intercepts and slopes in blue (continuous line) and its confidence intervals (discontinuous line).*

The last model is the most plausible and to validate this model we plot its residuals to the fitted values and the observed values (Figure 7).

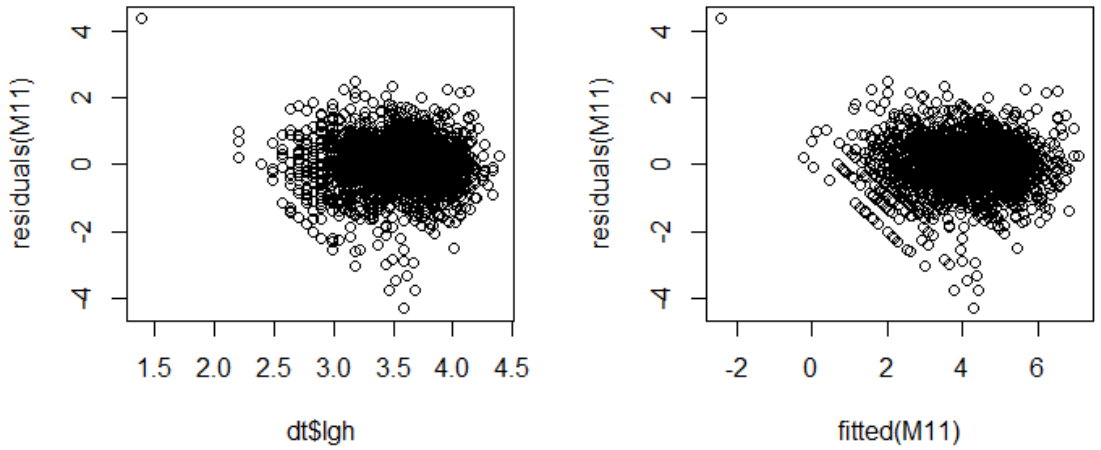


Figure 7. *Plot of residuals of the mix effects model with random intercept and slope to the fitted values and the observed values (Figure 6).*

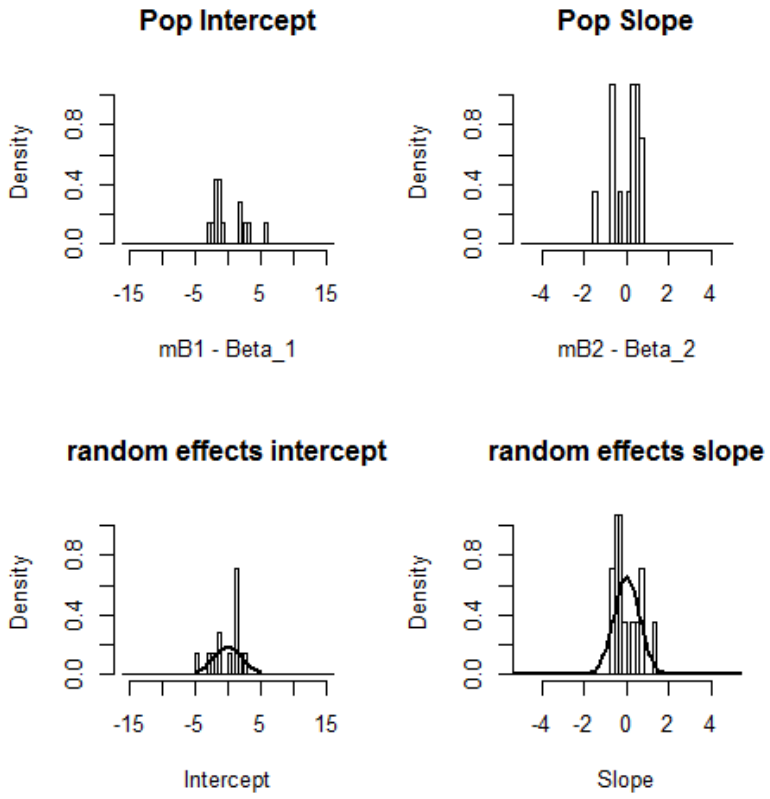


Figure 8. *Spread of the deviations of the coefficients under complete independence and after the mixed model with random slope and intercept by population.*

Note:

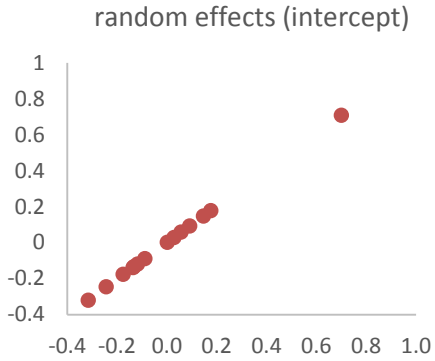
The restricted maximum likelihood estimation method (REML) is the default in function `lme`. This procedure "corrects the degrees of freedom" because the parameters in the model are not independently estimated under maximum likelihood (ML) (Zurr et al. 2009). If the number of fixed covariates is small compared to the number of observations their differences are minor. The Tables below present their differences for the models that we evaluated above. AIC and BIC based on REML are not comparable with AIC and BIC obtained by ML because for REML n* = n-p (Zuur et al. 2009).

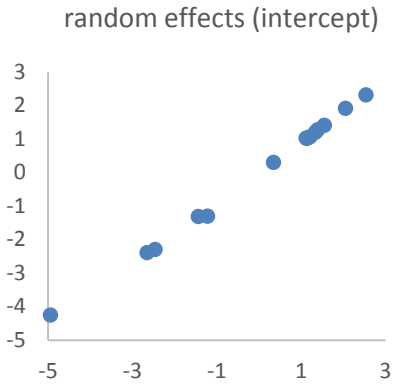
Component	**REML**	**ML**
S dev random Intercept: Population	0.258	0.247
Random: Residual	0.782	0.782
Fixed: Intercept	-7.80 (0.193)	-7.80 (0.190)
Fixed: Slope	3.36 (0.051)	3.36 (0.051)
Correlation	-0.93	-0.93
AIC	4984.3	4976.7
BIC	5006.9	4999.3
Component	**REML**	**ML**
Variance random Intercept: Population	2.30	2.20
Variance random Slope: Population	0.638	0.61
Random: Residual	0.758	0.758
Fixed: Intercept	-8.21 (0.64)	-8.20 (0.62)
Fixed: Slope	3.47 (0.18)	3.47(0.18)
Correlation	-0.99	-0.99
AIC	4896.1	4891.4
BIC	4930.0	4925.3

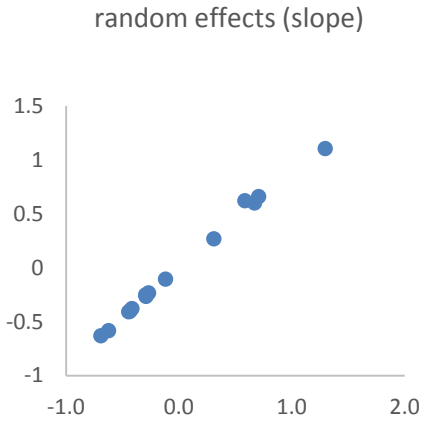
Summary of Frequentist and Bayesian approaches (diffuse priors)

	Frequentist		*Bayesian*		*Frequentist*		*Bayesian*	
	B1	s.e	B1	s.e	B2	s.e	B2	s.e
`slope`	-7.80	0.19	-7.80	0.20	3.36	0.05	3.36	0.05
`intercept & slope`	-8.20	0.62	-8.18	0.66	3.47	0.17	3.46	0.18
`Random residual`					0.758		0.759	0.012

Plots of random effects per population (x axis = frequentist, y axis = Bayesian)







Quintana-Ascencio, P. F., E. S. Menges, and C. Weekley. 2003. A fire-explicit population viability analysis of *Hypericum cumulicola* in Florida rosemary scrub. Conservation Biology, 17: 433-449.

Zuur, A.F., E.N. Ieno, N.J. Walker, A. Savaliev, G.M. Smith. 2009. Mixed effects models and extensions in Ecology with R. Springer.

Nakagawa, S., and H. Schielzeth. 2013. A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods in Ecology and Evolution 4(2): 133-142.