

What about non-independent data? Reproductive output of a rare plant with random effects of populations (mixed models)

In demography it is habitual to estimate reproductive output of individuals together with other vital rates for population models (Quintana-Ascencio et al. 2003). We could use *Hypericum cumulicola* data to evaluate a generalized linear model to predict the number of reproductive structures of individuals with different heights, assuming that individuals from the same population were independent from each other. However, we can also 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. Studying *Hypericum cumulicola* in the FL scrub

For this demo you will need to download and install:

The scripts LMM.R, LMM_rand_intercept.R and LMM_rand_intercept&slope.R

The data file Hypericum_data_94_07.txt

The following R packages: bbmle, ggplot2, jagsUI, lattice, lme4, MuMIn, nlme and rjags

A JAGS version that is compatible with your R (or RStudio) and jagsUI and rjags packages.

As in previous occasions we start by preparing the data (excluding non-reproductive individuals and log transforming both variables)...

```
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$year)
dt$lgh <- log(dt$ht_init)
dt$lfr <- log(dt$rp_init)
site <- unique(dt$bald)
```

We create a table to deposit the coefficients of the effects of the three models that we will evaluate with a frequentist approach.

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

First we estimate the coefficients of the model assuming **complete independence** of the data (this appears as a blue line in Figure 2). The following equation represents the model, where k is each individual.

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

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

The output of the general model under the assumption of complete independence should be familiar. This model explains approximately 66 % of the variance and shows a positive and statistically significant effect of $\log(\text{height})$ on $\log(\text{number of fruits})$.

```
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
```

But remember that the data comes from 14 distinct populations of the plant, so next we estimate the **population specific** coefficients of the model (these appear as red lines in Figure 2). There are some variations for the model in each population. We could predict the $\log(\text{number of fruits})$ as a linear function of the $\log(\text{height})$ using the specific intercepts (β_1) and slopes (β_2) for each population, but this would significantly reduce the degrees of freedom of our analysis, and limit the generality of our interpretation. The following equation represents the model, where k is each individual and i is each population.

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

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

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)
```

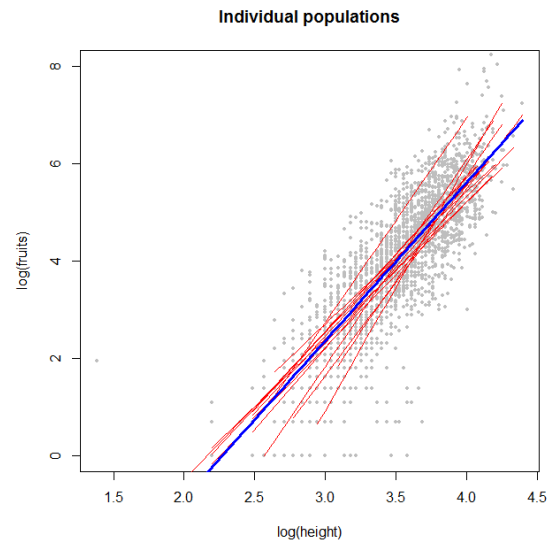


Figure 2. Plot of log(height) vs log(fruits), data as points in grey, model assuming complete independence in blue and population specific models in red.

Histograms of the deviations between the coefficients of the population specific models and the model with complete independence are shown in Figure 3. The coefficients are listed in Table 1, first the “intercept” (Beta_1) and then the “slope” (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")
```

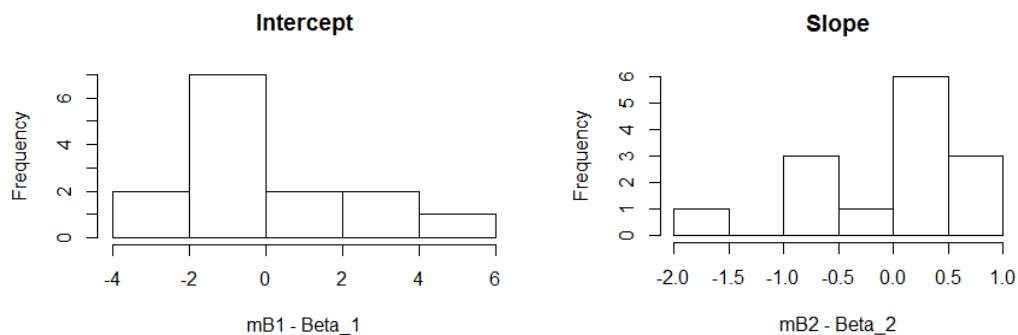


Figure 3. Histograms of deviations of coefficients of models by population from those of the model with complete independence

Table 1. Coefficients per population.

Pop	1	29	32	42	50	57	59	62	67	87	88	91	93	103
Beta1	-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	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

However, 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 a normally distributed random variable. For more details on these assumptions see Zuur *et al.* 2009. We use the function `lme` to obtain this model after converting the population variable (`bald`) into a factor. Our second model for the whole dataset assumes random intercepts for each population, but a common slope. We specify the random component as `random=~1|fbald`. This specification represents that our individual data was nested within populations but that we expect the variation among populations to be random.

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

$$\alpha_{1i} \sim N(0, \sigma_i)$$

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

```
dt$fbald <- factor(dt$bald)
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 fixed effect intercept was **-7.80** (SE = 0.19) and the fixed slope **3.36** (SE = 0.05).

```
Linear mixed-effects model fit by maximum likelihood
Data: dt
      AIC      BIC    logLik
4976.699 4999.299 -2484.349

Random effects:
Formula: ~1 | fbald
      (Intercept)  Residual
StdDev:   0.2478152  0.7821963

Fixed effects: lfr ~ lgh
              Value Std.Error   DF   t-value p-value
(Intercept) -7.796553 0.19250598 2086 -40.50032    0
lgh          3.362330 0.05086058 2086  66.10875    0
Correlation:
      (Intr)
lgh -0.934

Standardized Within-Group Residuals:
      Min      Q1      Med      Q3      Max
-5.43943318 -0.56617883  0.04190261  0.61074623  6.42325663

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

We plot this model in Figure 4. The blue line 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 fixed effects, the level = 1 takes those the random one (population).

```
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")
}
```

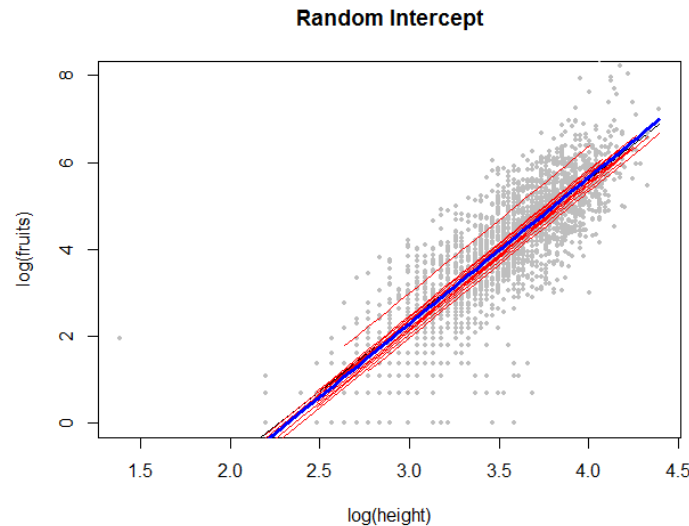


Figure 4. Plot of log (height) vs log (fruits), data as points in grey, predicted fixed effects in blue and random intercepts by population in red.

We can now try a model that estimates random intercepts and slopes. This is specified in the model as `random=~1 + lgh|fbald`.

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

$$\alpha_1 \sim N(0, \sigma_{\alpha i})$$

$$\partial_1 \sim N(0, \sigma_{\partial i})$$

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

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

In the output presented below, the estimated value of $2.2^2 = 4.86$ indicates the random variance in the intercepts and $0.61^2 = 0.37$ the one in the slopes. The negative correlation (-0.99) between random intercepts and random slopes indicates that populations with high intercepts tend to have lower slopes.

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

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

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

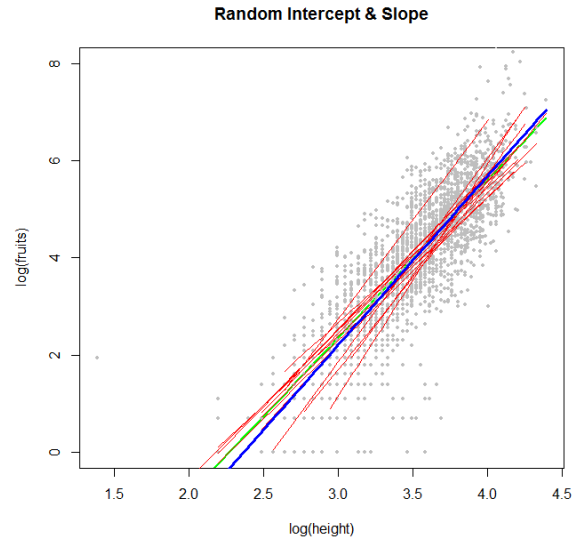


Figure 5. Plot of log (height) vs log (fruits), data as points in grey, predicted fixed effects in blue and random effects on intercepts and slopes by population in red. The model assuming complete independence is in green.

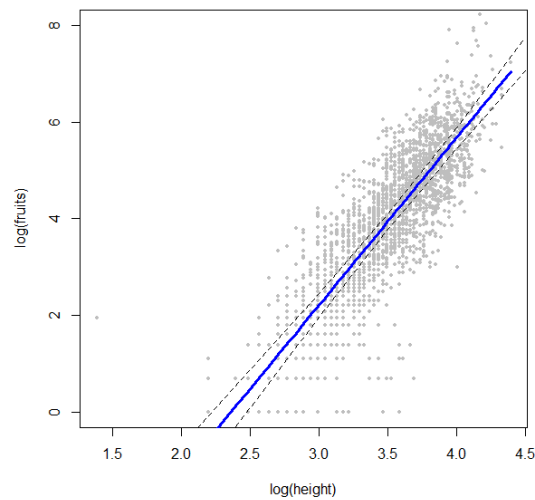


Figure 6. Plot of log (height) vs log (fruits), data as points in grey, predicted fix effects of a model with random intercepts and slopes in blue (continuous line) and its confidence intervals (discontinuous line).

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.

Model	Intercept	SE	slope	SE	AIC
no mixed	-7.35	0.18	3.24	0.05	5113.8
random intercept	-7.80	0.19	3.36	0.05	4984.3
random intercept & slope	-8.20	0.62	3.47	0.17	4891.4

The last model is the most plausible and to validate this model we plot its residuals in relation to the observed and fitted values (Figure 7). See R script for all the last part of the code.

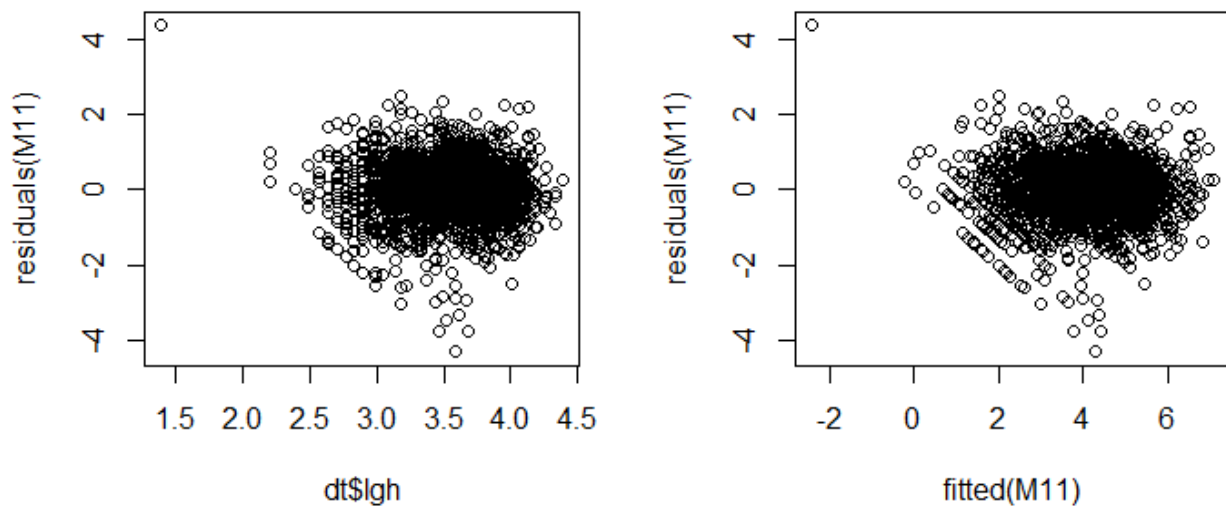


Figure 7. Residual plots for the mixed effects model with random intercept and slope (fitted and observed values).

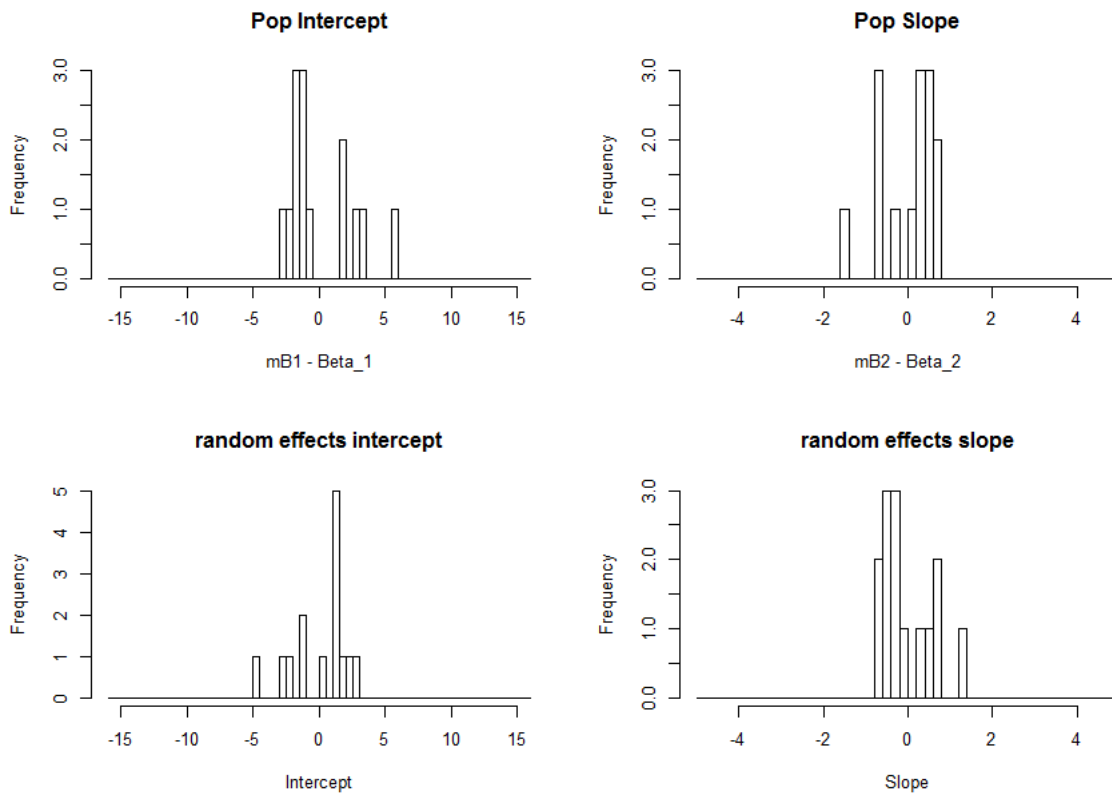


Figure 8. Spread of the deviations of the coefficients under complete independence and after the mixed model with random slope and intercept by population.

Note: The restricted maximum likelihood estimation method (REML) is the default method in the 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. Table 3 presents 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).

Table 3. Comparison of REML and ML methods for M1 and M11 models.

M1 - Component	REML	ML
<i>Random effects</i>		
StdDev Intercept	0.258	0.247
StdDev Residual	0.782	0.782
<i>Fixed effects</i>		
Intercept	-7.80 (0.193)	-7.80 (0.190)
Slope	3.36 (0.051)	3.36 (0.051)
Correlation	-0.93	-0.93
AIC	4984.3	4976.7
BIC	5006.9	4999.3
M11 - Component	REML	ML
<i>Random effects</i>		
StdDev Intercept	2.30	2.20
StdDev Slope	0.638	0.610
StdDev Residual	0.758	0.758
<i>Fixed effects</i>		
Intercept	-8.21 (0.64)	-8.20 (0.62)
Slope	3.47 (0.18)	3.47(0.18)
Correlation	-0.99	-0.99
AIC	4896.1	4891.4
BIC	4930.0	4925.3

Next we do the analysis of the mixed models from a Bayesian approach; using uninformative priors (see code in the R script). As we can see from Table 4 and Figure 9, the estimates and standard error of the coefficients (β_1 – intercept and β_2 – slope) from the two models are almost identical with the two statistical approaches.

Table 3. Summary of results with frequentist and Bayesian approaches

Model	Frequentist		Bayesian		Frequentist		Bayesian	
	β_1	SE	β_1	SE	β_2	SE	β_2	SE
Random intercept	-7.80	0.19	-7.80	0.20	3.36	0.05	3.36	0.05
Random intercept & slope	-8.20	0.62	-8.18	0.66	3.47	0.17	3.46	0.18

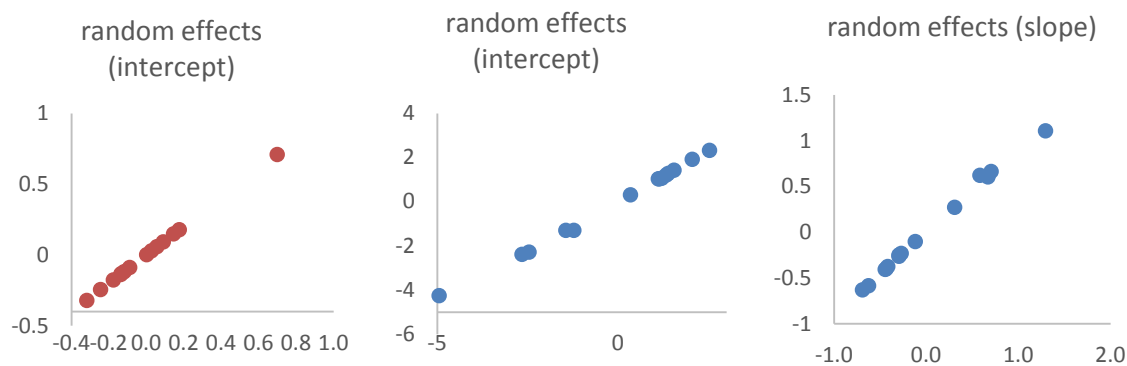


Figure 9. Plots of random effects per population (x = frequentist, y = Bayesian).

References

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.