Model selection for Mixed Effects Models: Effects of fire on reproduction of a rare plant

In a prior demo, we demonstrated the advantages of recognizing the nature of our sampling schemes and evaluating the effects of random factors in our models. Here, we discuss how to implement model selection for models with mixed effects. We follow the procedure described in Zurr *et al.* 2009 (p: 121-122). We already provided evidence that number of reproductive structures of *Hypericum cumulicola* is significantly associated with plant height (Quintana-Ascencio *et al.* 2003). We also established that there was significant random variation on number of fruits at the population level. Now, we want to evaluate the relevance of two more fixed variables and of random variation in year, to explain variation in fecundity for this species. Number of stems complements height to characterize general plant size. We expect that plants with fewer stems will produce fewer fruits than plants with similar height but more stems. Time-since-fire (TSF) affects attributes such as nutrient and water availability, abundance of predators and competitors, potentially influencing plant resources available for reproduction. We use a model selection approach to assess the relative importance of fire and number of stems to explain variation in fruit production of *Hypericum cumulicola.*

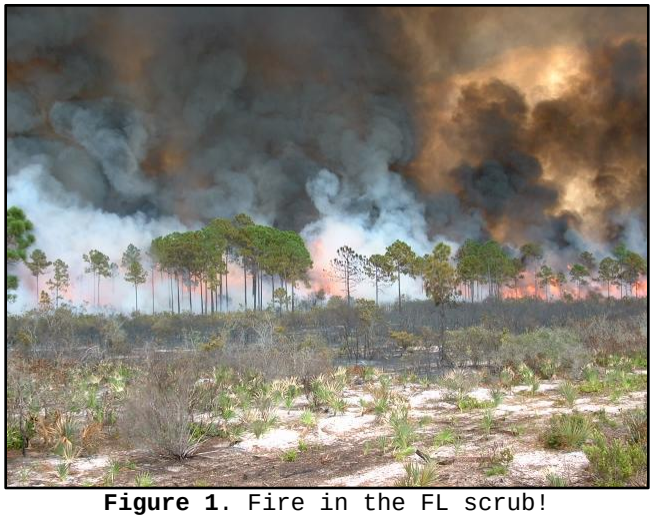

Figure 1. Fire in the FL scrub!

For this demo you will need: `Mixed model selection.R` (script), `Hypericum_data_94_07.txt` (data), and R packages: `nlme, bbmle, lme4, lattice, MuMIn.`

To run the code for Bayesian analysis (not included or commented in this document): `JAGS` version that is compatible with your R (or RStudio), `jagsUI` package, `Model_w_year.txt` (script).

We prepare the data in the same way as before, but add three new variables. Notice that we need to remove plants without data on number of stems. We transform the information based on the year of fire to build a categorical variable with three levels (recently, intermediate and long time since the last fire).

```
orig_data <- read.table("hypericum_data_94_07.txt", header=T)
dt <-subset(orig_data, !is.na(ht_init) & !is.na(st_init) & rp_init > 0 & year<1997)
yr <- unique(dt$year)
dt$lgh <- log(dt$ht_init)
dt$lfr <- log(dt$rp_init)
dt$stems <- dt$st_init
site <- unique(dt$bald)
table(dt$bald,dt$fire_year)
dt$TSF <- 1
dt$TSF[dt$fire_year <1987] <-2
dt$TSF[dt$fire_year <1973] <-3
dt$TSF <- factor(dt$TSF)
dt$fyear <- factor(dt$year)
dt$fbald <- factor(dt$bald)
I <- order(dt$lgh)
lgh <- sort(dt$lgh)
table(dt$bald,dt$TSF)
tsf <- unique(dt$TSF)
TSF <-dt$TSF
```

An initial check of the correlation between number of stems and height indicates no problems with collinearity (r = 0.339, Figure 2). We also do not find collinearity among time-since-fire and number of stems (Figure 3). We transform the number of stems to a categorical variable with eight levels (plants > 8 stems are grouped; Figure 2).

```
boxplot(dt$stems~dt$TSF)
par(mfrow=c (1,2))
plot(dt$stems,dt$ht_init,pch=16,ylab="height",
     xlab="number of stems",col="blue",cex=0.5, log="x",xlim=c(0.5,30))
summary(lm(log(dt$ht_init)~log(dt$stems)))
dt$stems[dt$stems>8] <- 8
plot(dt$stems,dt$ht_init,pch=16,ylab="height",
     xlab="number of stems",col="blue",cex=0.5, log="x",xlim=c(0.5,30))
dt$stems <- factor(dt$stems)
```

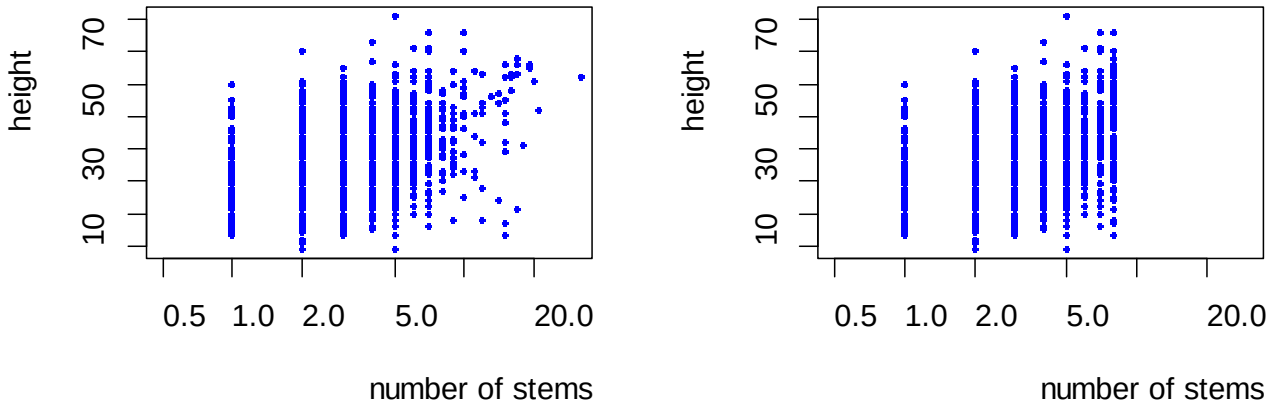

Figure 2. Plot of height as a function of number of stems (raw data on the left, as a categorical variable with eight levels on the right).

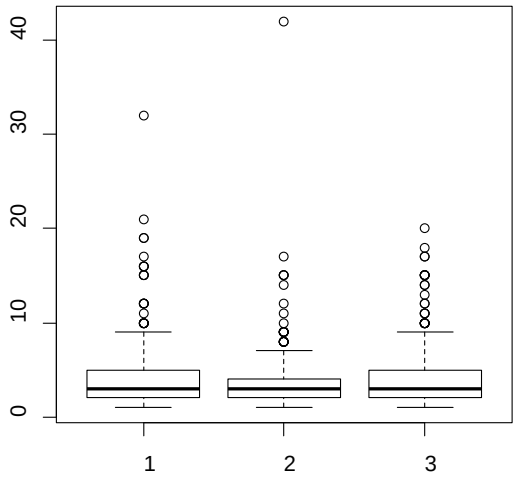

Figure 3. Plot of number of stems as a function of time-since-fire

Zuur et al. (2009) caution about the need to start with a model that includes all possible fixed effects to evaluate the best configuration for the random factors. For our data this model includes three single factors, three two-way interactions and one three-way interaction among height, number of stems and time-since-fire. We propose three options for the random configuration: (i) no random effects, (ii) random effects on the intercept given the year and population, and (iii) random effects on the intercept and the slope given the year and population. The function `lme` requires the specification of a random term. To avoid an error message we use the `gls` function for the first model. At this initial stage, we use the likelihood test with REML estimation. Because these modes take long to converge on a solution, we use the function `lmeControl` to increase the number of iterations for the last model. The AICs of these models indicate that the one with random intercept and slope is the most plausible.

```
m11 <- gls(lfr~lgh*TSF*stems,method ="REML",data=dt)
anova(m11)
M1 <- lme(lfr~lgh*TSF*stems,random=~1|year/fbald,data=dt,method ="REML")
anova(M1)
lmc <- lmeControl(niter=5200,msMaxIter=5200)
M11 <- lme(lfr~lgh*TSF*stems,random=~1 + lgh|fyear/fbald,data=dt,
       method = "REML", control=lmc)
anova(M11)

AICtab(m11,M1,M11,weights=TRUE,base = TRUE)
    AIC    df dAIC   weight
M11 3782.1 55    0.0 1
M1  3816.9 51   34.8 <0.001
m11 3995.3 49  213.2 <0.001
```

This approach warrantees that we explore the whole variation associated with the fixed factors before deciding the structure of the random factors. Arguably, fixed factors are the ones in which we are more interested. Observe that the structure of the random effects changes the inference on the fixed factors (Table 1), particularly notice the differences in “significance” for Time-since-fire.

Table 1. Analysis of variance of the models with different random structures.

Fixed effects		Assumed Independence			Random Intercept			Rand. Int. & Slope		
Source of variation	ndf	ddf	F	p	ddf	F	p	ddf	F	p
Intercept	1	1674	53327	**<0.001**	1641	3170	**<0.001**	1641	1103	**<0.001**
Height	1	1674	4127	**<0.001**	1641	4293	**<0.001**	1641	232	**<0.001**
TSF	2	1674	7.17	**0.001**	31	0.703	0.503	31	1.11	0.342
Stems	7	1674	55.45	**<0.001**	1641	60.09	**0.001**	1641	54.9	**0.001**
Height:TSF	2	1674	1.02	0.362	1641	2.302	0.100	1641	2.73	0.065
Height:stems	7	1674	2.80	**0.007**	1641	1.804	0.082	1641	1.31	0.243
Stems:TSF	14	1674	2.00	**0.014**	1641	2.811	**0.001**	1641	2.45	**0.002**
Height:Stems:TSF	14	1674	1.25	0.229	1641	1.564	0.082	1641	1.63	0.064
AIC			3995			3817			3782	

ndf =numerator degrees of freedom
dff=denominator degrees of freedom

We proceed to evaluate the optimal fixed structure of the random structure that we just found. We fit random models with the same random effects structure using ML estimation and compare their likelihood criteria (Zuur et al. 2009). We also compare them using AIC. Here we are only testing six different variants for the fixed structure that we chose based on our previous knowledge and research hypothesis, but many more are possible.

```
M11 <- lme(lfr~lgh*TSF*stems,random=~1 + lgh|fyear/fbald,data=dt,method
 ="ML",control=lmc)

M13 <- lme(lfr~lgh+TSF*stems,random=~1 + lgh|fyear/fbald,data=dt,method
 ="ML",control=lmc)

M14 <- lme(lfr~lgh+TSF+stems,random=~1 + lgh|fyear/fbald,data=dt,method
 ="ML",control=lmc)

M15 <- lme(lfr~lgh+TSF,random=~1 + lgh|fyear/fbald,data=dt,method ="ML",control=lmc)

M16 <- lme(lfr~lgh+stems,random=~1 + lgh|fyear/fbald,data=dt,method ="ML",control=lmc)

M17 <- lme(lfr~lgh*stems,random=~1 + lgh|fyear/fbald,data=dt,method ="ML",control=lmc)
```

Model M13, including additive effects of height, stems and TSF and the interaction of stems with TSF was marginally significantly different from the full model, M11 (p =0.06), and has the lowest AIC. Please note that when using the `anova` function to find differences between models, they will be tested in the order you input them, whether it is logical or not!

```
> anova(M11,M13,M14,M15,M16,M17)
    Model df      AIC      BIC    logLik   Test  L.Ratio p-value
M11     1 55 3712.116 4011.935 -1801.058
M13     2 32 3700.587 3875.026 -1818.293 1 vs 2  34.4702  0.0587
M14     3 18 3706.004 3804.126 -1835.002 2 vs 3  33.4173  0.0025
M15     4 11 4020.799 4080.762 -1999.399 3 vs 4 328.7947  <.0001
M16     5 16 3706.064 3793.284 -1837.032 4 vs 5 324.7341  <.0001
M17     6 23 3712.757 3838.136 -1833.379 5 vs 6   7.3073  0.3976

> AICtab(M11,M13,M14,M15,M16,M17,weights=TRUE,base = TRUE)
    AIC    df dAIC   weight
M13 3700.6 32    0.0 0.87976
M14 3706.0 18    5.4 0.05862
M16 3706.1 16    5.5 0.05687
M11 3712.1 55   11.5 0.00276
M17 3712.8 23   12.2 0.00200
M15 4020.8 11  320.2 < 0.001
```

Accordingly to Zuur *et al.* (2009), we should now present the summary for this model using REML. The ANOVA of this model (M13r) and the one with ML (M13) are presented for comparison. Their plots are presented below (Figures 4-6). The random effects by population and the residuals of model M13r are presented in Figures 7-8. We conclude that height and number of stems differentially affect number of fruits depending on time-since-fire, but that these effects are contingent upon variations among years and populations.

```
> anova(M13)
            numDF denDF   F-value p-value
(Intercept)     1  1664 1334.3860  <.0001
lgh             1  1664  626.1175  <.0001
TSF             2    31    1.1797  0.3208
stems           7  1664   53.5399  <.0001
TSF:stems      14  1664    2.4673  0.0019

M13r <- lme(lfr~lgh+TSF*stems,random=~1 + lgh|fyear/fbald,data=dt,method ="REML",
control=lmc)

> anova(M13r)
            numDF denDF  F-value p-value
(Intercept)     1  1664 877.9929  <.0001
lgh             1  1664 431.4187  <.0001
TSF             2    31   1.1027  0.3446
stems           7  1664  53.6127  <.0001
TSF:stems      14  1664   2.4721  0.0018
```

Below we present the statistical model for this example, where index `k` refers to individuals, index `m` to years, index `l` to populations, index `i` to the stem category, index `j` to the TSF category, β1 is the intercept, β2 is the slope for the effect of height, β3 is the coefficient for stems, β4 is the coefficient for TSF, β5 is the coefficient for the interaction between stems and TSF, α1 represents the random variation of the intercept due to year, α2 represents the random variation of the intercept due to population given the year, α3 represents the random variation of the slope due to year, α4 represents the random variation of the slope due to population given the year, and ε represents the error, or unexplained variation associated with the whole model.

$$log(reproductive\ structures)_k = (\beta_1 + a_{1m} + a_{2ml}) + (\beta_2 + a_{3m} + a_{4mi}) * log(height)_k + \beta_{3i}[stems_i] + \beta_{4j}[TSF_j] + \beta_{5ji}[TSF_j, stems_i] + \in$$

$$a_1 \sim N(0, \sigma_{a1})$$
$$a_2 \sim N(0, \sigma_{a2})$$
$$a_3 \sim N(0, \sigma_{a3})$$
$$a_4 \sim N(0, \sigma_{a4})$$
$$\epsilon \sim N(0, \sigma)$$

```
> summary(M13r)
Linear mixed-effects model fit by REML
 Data: dt
       AIC      BIC    logLik
  3764.234 3938.206 -1850.117

Random effects:
 Formula: ~1 + lgh | fyear
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 0.7130826 (Intr)
lgh         0.2473748 -1

 Formula: ~1 + lgh | fbald %in% fyear
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 1.7656596 (Intr)
lgh         0.4647955 -0.982
Residual    0.6739210

Fixed effects: lfr ~ lgh + TSF * stems
                Value      Std.Error   DF    t-value p-value
(Intercept) -7.791543[β1]  0.5668900 1664 -13.744364  0.0000
lgh          3.225019[β2]  0.1756916 1664  18.356132  0.0000
TSF2         0.017864[β3j] 0.2005082   31   0.089092  0.9296
TSF3        -0.321050[β3j] 0.2076716   31  -1.545949  0.1323
stems2       0.091776[β4i] 0.1374703 1664   0.667609  0.5045
stems3       0.211875[β4i] 0.1364376 1664   1.552908  0.1206
stems4       0.439440[β4i] 0.1429434 1664   3.074223  0.0021
stems5       0.590749[β4i] 0.1611357 1664   3.666161  0.0003
stems6       1.081824[β4i] 0.1729152 1664   6.256386  0.0000
stems7       1.296729[β4i] 0.2038312 1664   6.361779  0.0000
stems8       1.782993[β4i] 0.1831365 1664   9.735865  0.0000
TSF2:stems2  0.251562[β5ij]0.1682852 1664   1.494857  0.1351
TSF3:stems2  0.304727[β5ij]0.1779889 1664   1.712058  0.0871
TSF2:stems3  0.432558[β5ij]0.1695012 1664   2.551948  0.0108
TSF3:stems3  0.475851[β5ij]0.1768386 1664   2.690877  0.0072
TSF2:stems4  0.228175[β5ij]0.1778559 1664   1.282923  0.1997
TSF3:stems4  0.457384[β5ij]0.1868778 1664   2.447504  0.0145
TSF2:stems5  0.238509[β5ij]0.2034423 1664   1.172365  0.2412
TSF3:stems5  0.329618[β5ij]0.2036406 1664   1.618628  0.1057
TSF2:stems6  0.006405[β5ij]0.2344326 1664   0.027323  0.9782
TSF3:stems6  0.056951[β5ij]0.2271864 1664   0.250679  0.8021
TSF2:stems7 -0.072807[β5ij]0.2803267 1664  -0.259722  0.7951
TSF3:stems7 -0.124216[β5ij]0.2598069 1664  -0.478109  0.6326
TSF2:stems8 -0.381936[β5ij]0.2501994 1664  -1.526526  0.1271
TSF3:stems8 -0.383409[β5ij]0.2350758 1664  -1.631003  0.1031

Correlation:
Displays the correlations between all estimates of the parameters [deleted for space]

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-5.76848359 -0.48553033  0.08604145  0.60302384  3.06440840

Number of Observations: 1722
Number of Groups:
          fyear fbald %in% fyear
              3               36     # output was truncated
```

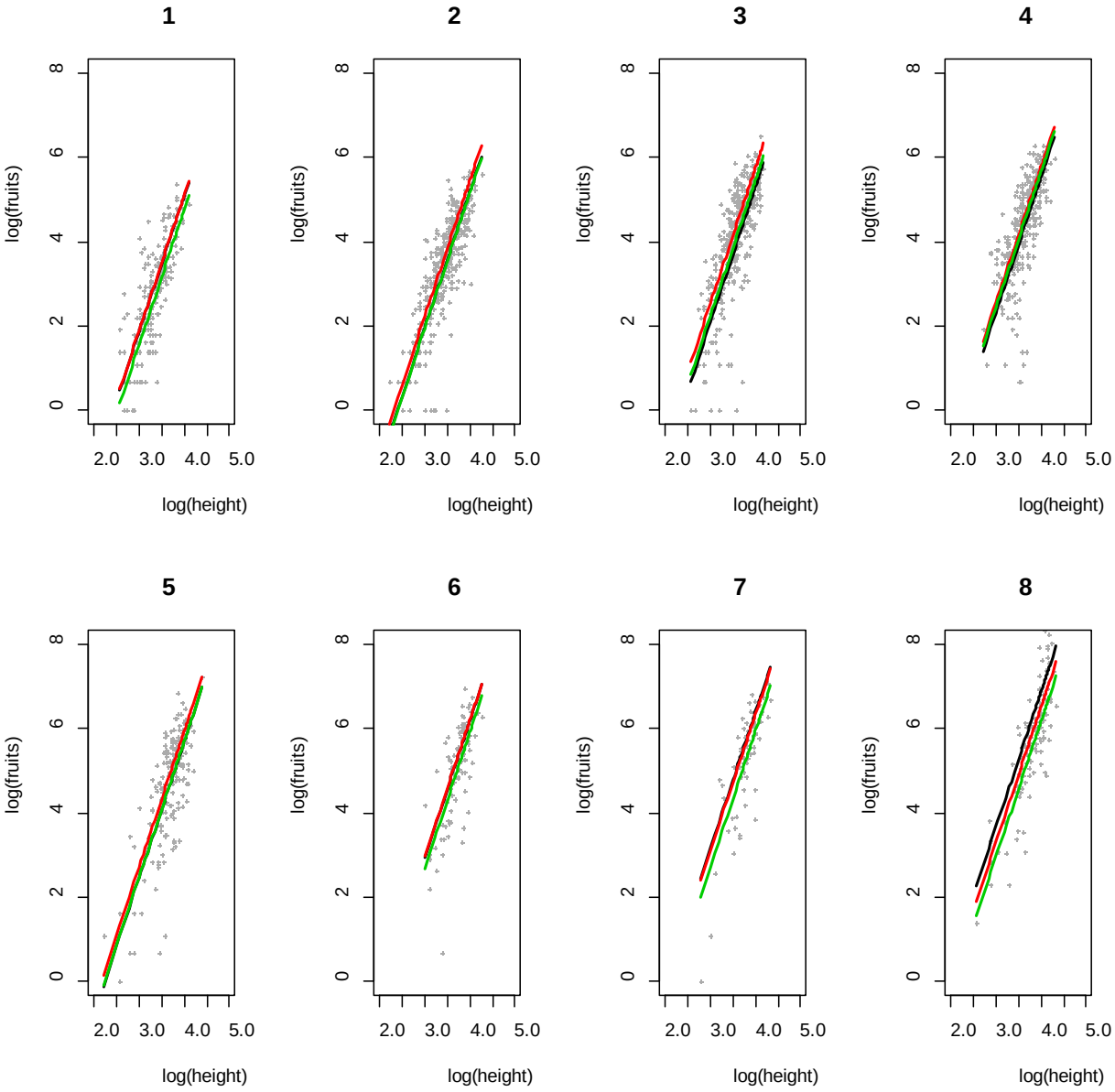

Figure 4. Plot of number of fruits as a function of height for plants with different number of stems. TSF 1 = black, TSF 2 = red, TSF 3 = green. Lines are model predictions for fixed effects of the model with random intercept and slope by year and population.

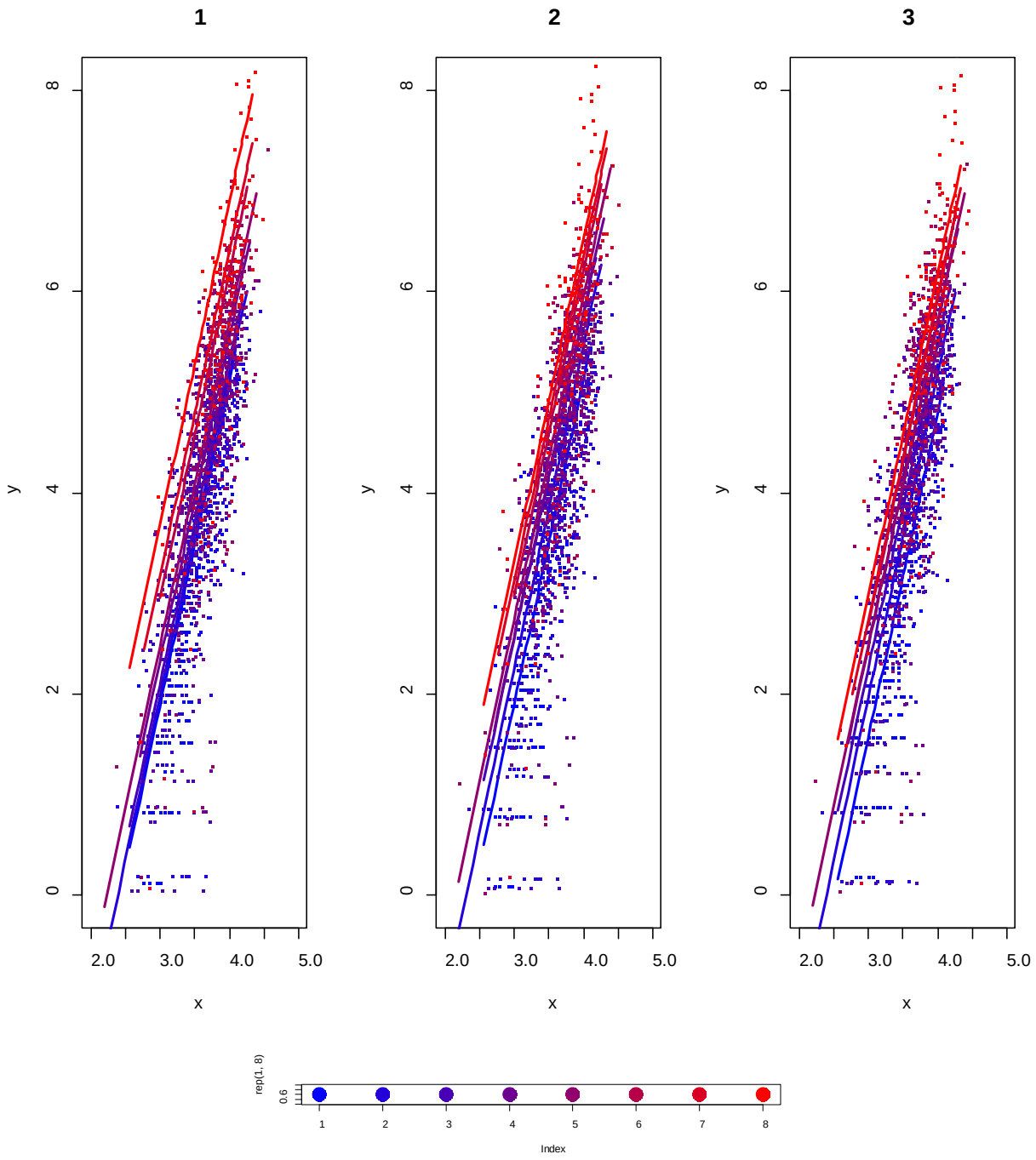

Figure 5. Plot of log number of fruits as a function of log height for plants with different number of stems (colors). Lines are model predictions for fixed effects of the model with random intercept and slope by year and population.

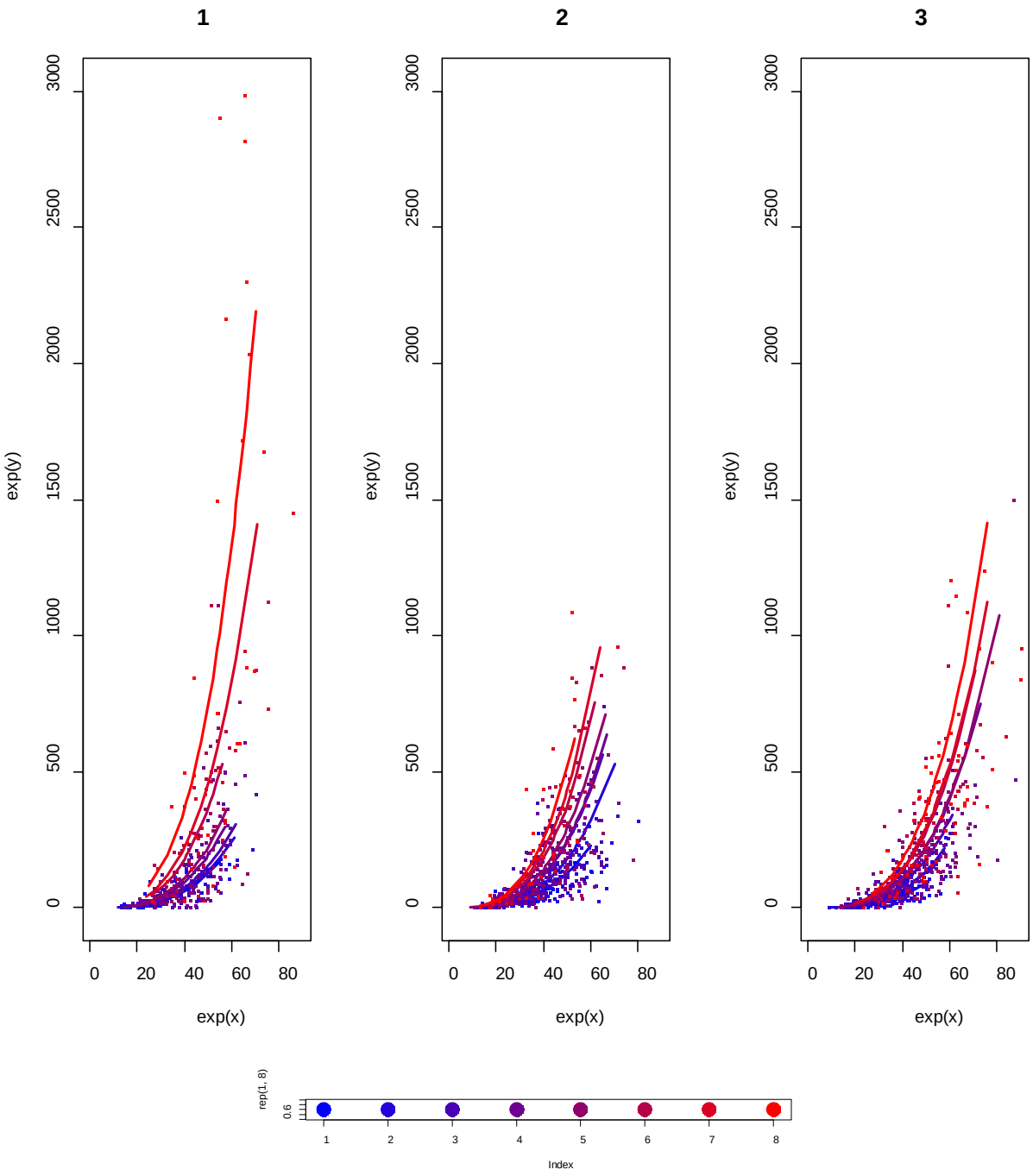

Figure 6. Plot of number of fruits as a function of height for plants with different number of stems and time-since-fire (stems in different colors, TSF in different panels). Lines are model predictions for fixed effects of the model with random intercept and slope by year and population.

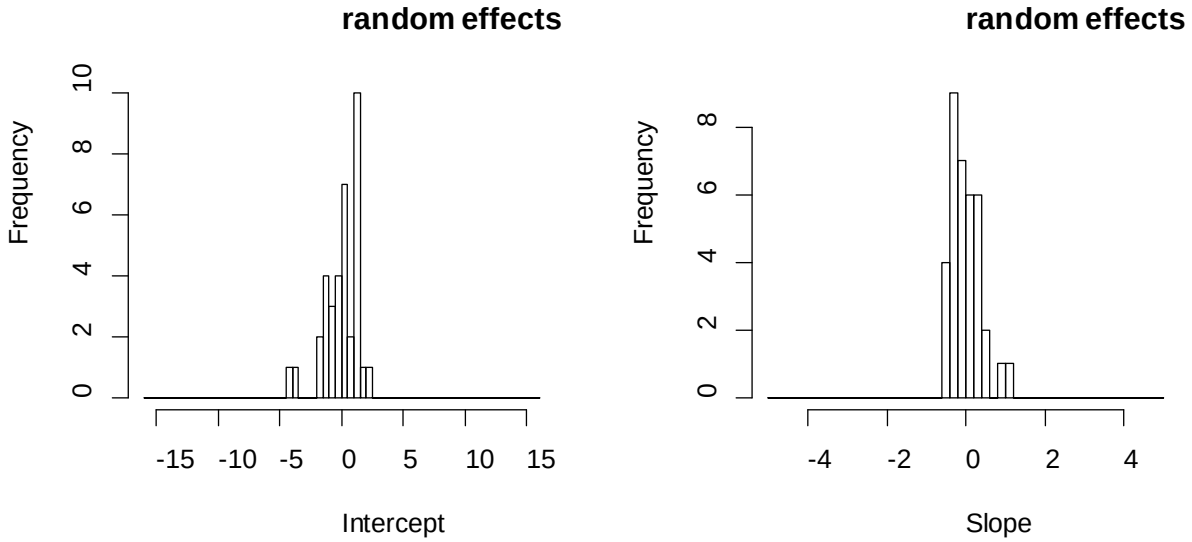

Figure 7. Random effects for population in model M13r

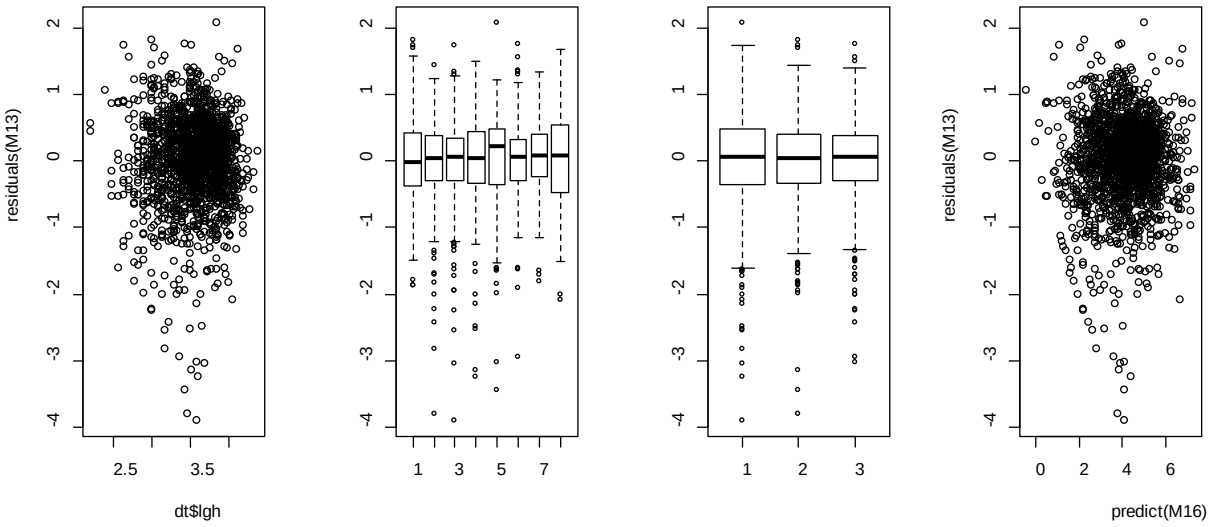

Figure 8. Residuals of model M13r

Note: See the R script for how to run the chosen model in a Bayesian framework, and the Excel file for a comparison between the output and predictions of the model with the two approaches.

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.