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

In the last demo we discussed how to implement model selection for linear mixed models. Here, we discuss model selection for mixed effects models with binary responses (GLMM) by combining procedures described in Crawley (2007) and Zuur *et al*. (2009). These approaches are still being developed so examples are scarce and documentation is limited. There are several procedures in R to complete this work, and their results are not always consistent (Zuur *et al.* 2009 p: 323-325). However, the results for our data were commensurate. We encourage you to continue to review future improvements as they become available.

We evaluate the relevance of three fixed variables to explain survival variation in this species (Quintana-Ascencio *et al.* 2003). We decided not to include number of reproductive structures in this model because of the high co-linearity between this variable and plant height (see previous demo). We evaluate the effect of height (cm), number of stems and time-since-fire (TSF) on *Hypericum cumulicola* survival, taking into account the random effects of population and year. We use a model selection approach to assess the relative importance of the fixed and random factors.

Figure 1. Dying *Hypericum cumulicola*

For this demo you will need: `GLMM.R` (script), `hypericum_data_94_07.txt` (data), and R packages: `nlme, bbmle, lme4, lattice, optimx.`

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, `GLMM_wBayes.R` (script), `Model_w_year binary intercept.R` (script).

We prepare the data as we have before but to include survival, the variable *fate* needs to be reorganized into *surv* to convert “rip” to zeros (dead) and everything else to ones (alive). We also scaled height (lgh) to facilitate convergence of the models.

```
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$bald)
dt$surv <-1
dt$surv[dt$fate =="rip"] <- 0
table(dt$surv,dt$fate)
I <- order(dt$lgh)
lgh <- sort(dt$lgh)
table(dt$bald,dt$TSF)
tsf <- unique(dt$TSF)
tsf <- sort(tsf)
TSF <-dt$TS
dt$stems[dt$stems>8] <- 8
dt$stems <- factor(dt$stems)
dt$lghc <- scale(dt$lgh)
```

We check for co-linearity and find a significant association among time-since-fire and plant height (Figure 2). There is evidence that the average height is higher in long-unburned populations than in populations more recently burned and with intermediate time-since-fire, but there is enough variation in height to proceed with our analysis.

```
pairs(subset(dt,select=c(ht_init, rp_init,stems)))
boxplot(dt$lgh~dt$TSF)
summary(lm(dt$lgh~factor(dt$TSF)))
```

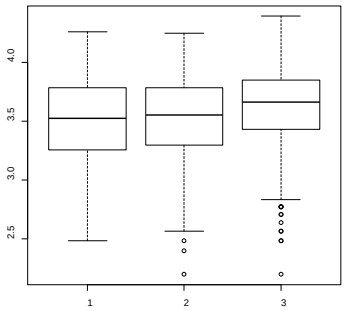

Figure 2. Plot of height (cm) as a function of time-since-fire

Once again following Zuur *et al.* (2009), to evaluate the best configuration for the random factors we use a saturated model for the fixed effects (height * stems * TSF). We propose three options for the random configuration: (i) no random effects, (ii) random intercept and (iii) random intercept and slope. We use year and population again, but in this case we assumed them as independent because models with population nested in years did not converge. We used the function *glmer* to specify the binomial family, but since this function requires the specification of a random term we are forced to use *glm* for the non-random model.

Additionally, we have the concern that some of the models did not reach convergence so we use an optimization procedure (*optimx + method = "nlminb"*) to address this issue. These facts prevent us from using REML for the comparison of the three models, so tentatively we compare the AICs of models with ML. This comparison indicates that the model with random effects only on the intercept is the more plausible model.

```
require(optimx)
m1 <- glm(surv~lghc*TSF*stems,data=dt,family =binomial)
m2 <- glmer(surv~lghc*TSF*stems + (1|fbald) + (1|year),data=dt,family =binomial)
m2_nlminb <- update(m2,control=glmerControl(optimizer="optimx",
             optCtrl=list(method="nlminb")))
m3 <- glmer(surv~lghc*TSF*stems + (lghc|fbald)+ (lghc|year),data=dt,family =binomial)
m3_nlminb <- update(m3,control=glmerControl(optimizer="optimx",
             optCtrl=list(method="nlminb")))

AICtab(m1,m2_nlminb,m3_nlminb,weights=TRUE,base=TRUE))

        AIC    dAIC   df weight
m2_nlminb 2136.4    0.0 50 0.962
m3_nlminb 2142.9    6.5 54 0.038
m1        2230.6   94.2 48 <0.001
```

We proceed to evaluate fixed effects, using the random structure that we just found was most plausible (Zuur *et al.* 2009). We only use the optimization procedure, on those models that did not reached convergence with the regular call. The time it takes for the procedure to finish and whether it converges or not may vary depending on your computer.

```
M11 <- glmer(surv~lghc*TSF*stems + (1|fbald)+(1|year),data=dt,family=binomial)
M11_nlminb <- update(M11,control=glmerControl(optimizer="optimx",
    optCtrl=list(method="nlminb")))
M13 <- glmer(surv~lghc+TSF*stems + (1|fbald)+(1|year),data=dt,family =binomial)
M13_nlminb <- update(M13,control=glmerControl(optimizer="optimx",
    optCtrl=list(method="nlminb")))
M14 <- glmer(surv~lghc+TSF+stems + (1|fbald)+(1|year),data=dt,family =binomial)
M15 <- glmer(surv~lghc+TSF + (1|fbald)+(1|year),data=dt,family =binomial)
M16 <- glmer(surv~lghc+stems + (1|fbald)+(1|year),data=dt,family =binomial)
M17 <- glmer(surv~lghc*stems + (1|fbald)+(1|year),data=dt,family =binomial)
M18 <- glmer(surv~lghc*stems + TSF + (1|fbald)+(1|year),data=dt,family =binomial)
M19 <- glmer(surv~lghc*TSF + stems + (1|fbald)+(1|year),data=dt,family =binomial)
M20 <- glmer(surv~lghc*TSF + stems*TSF + (1|fbald)+(1|year),data=dt,family =binomial)
M20_nlminb <- update(M20,control=glmerControl(optimizer="optimx",
       optCtrl=list(method="nlminb")))

AICtab(M11_nlminb,M13_nlminb,M14,M15,M16,M17,M18,M19,M20_nlminb,weights=TRUE,base=TRUE)

           AIC    dAIC   df weight
M19        2105.3    0.0 15 0.395
M13_nlminb 2105.5    0.2 27 0.363
M20_nlminb 2106.8    1.5 29 0.185
M14        2109.9    4.6 13 0.040
M18        2111.7    6.4 20 0.016
M16        2117.3   12.0 11 <0.001
M17        2119.6   14.3 18 <0.001
M15        2131.8   26.5 6  <0.001
M11_nlminb 2136.4   31.1 50 <0.001
```

There are three models providing significant information (M19, M13 and M20). We chose M20, which includes the interactive effects of height and stems with TSF because it integrates the

information of the other two (retains the two important interactions without adding the three-way interaction). The formula, summary and plots of model M20 are presented below (Figure 3). We conclude that increasing height and time-since-fire tends to decrease survival when compared to recently burned populations. The effect of number of stems differentially affects survival depending of time-since-fire. There is considerable random variation by population and by year. Figure 4 shows the residuals for model M20.

```
summary(M20_nlminb)
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: surv ~ lghc * TSF + stems * TSF + (1 | fbald) + (1 | year)
   Data: dt
Control: glmerControl(optimizer = "optimx", optCtrl = list(method = "nlminb"))

     AIC      BIC   logLik deviance df.resid
  2106.8   2264.9  -1024.4   2048.8     1693

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.7381 -0.9305  0.2943  0.8173  2.3938

Random effects:
 Groups Name        Variance Std.Dev.
 fbald  (Intercept) 0.4711   0.6864
 year   (Intercept) 0.1145   0.3383
Number of obs: 1722, groups:  fbald, 14; year, 3

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.03730[β1]   0.66261   4.584 4.57e-06 ***
lghc        -0.36543[β2]   0.17306  -2.112 0.034722 *
TSF2        -2.76851[β4j]  0.73647  -3.759 0.000170 ***
TSF3        -2.46280[β4j]  0.76598  -3.215 0.001303 **
stems2      -0.49468[β3i]  0.58361  -0.848 0.396642
stems3      -0.83734[β3i]  0.54440  -1.538 0.124026
stems4      -1.16465[β3i]  0.56812  -2.050 0.040365 *
stems5      -0.86474[β3i]  0.65839  -1.313 0.189045
stems6      -1.46078[β3i]  0.66424  -2.199 0.027865 *
stems7      -1.79443[β3i]  0.79418  -2.259 0.023854 *
stems8      -2.98295[β3i]  0.74799  -3.988 6.66e-05 ***
lghc:TSF2    0.31356[β5j]  0.19485   1.609 0.107568
lghc:TSF3    0.26729[β5j]  0.19928   1.341 0.179832
TSF2:stems2  0.49834[β6ij] 0.65362   0.762 0.445806
TSF3:stems2  0.09303[β6ij] 0.67799   0.137 0.890856
TSF2:stems3  0.50193[β6ij] 0.62403   0.804 0.421205
TSF3:stems3  0.95240[β6ij] 0.64696   1.472 0.140989
TSF2:stems4  0.98897[β6ij] 0.65239   1.516 0.129537
TSF3:stems4  1.21458[β6ij] 0.68034   1.785 0.074222 .
TSF2:stems5  0.38880[β6ij] 0.76060   0.511 0.609229
TSF3:stems5  0.08767[β6ij] 0.76204   0.115 0.908412
TSF2:stems6  0.98568[β6ij] 0.82339   1.197 0.231265
TSF3:stems6  0.99879[β6ij] 0.80643   1.239 0.215519
TSF2:stems7  0.60739[β6ij] 1.00739   0.603 0.546554
TSF3:stems7  0.97139[β6ij] 0.93495   1.039 0.298817
TSF2:stems8  3.07249[β6ij] 0.90574   3.392 0.000693 ***
TSF3:stems8  1.49963[β6ij] 0.88518   1.694 0.090237 .
```

Below we present the logistic regression 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 height and TSF, β6 is the coefficient for the interaction between stems and TSF, α1 represents the random variation of the intercept due to population, and α2 represents the random variation of the intercept due to year.

$$Survival_k \sim Bin(\pi_k)$$

$$Logit(\pi)_k = (\beta_1 + a_{1l} + a_{2m}) + \beta_2 * height_k + \beta_{3i}[stems_i] + \beta_{4j}[TSF_j] + \beta_{5[TSFj]} * height_k + \beta_{6ji}[TSF_j, stems_i]$$

$$a_1 \sim N(0, \sigma_{a1})$$

$$a_2 \sim N(0, \sigma_{a2})$$

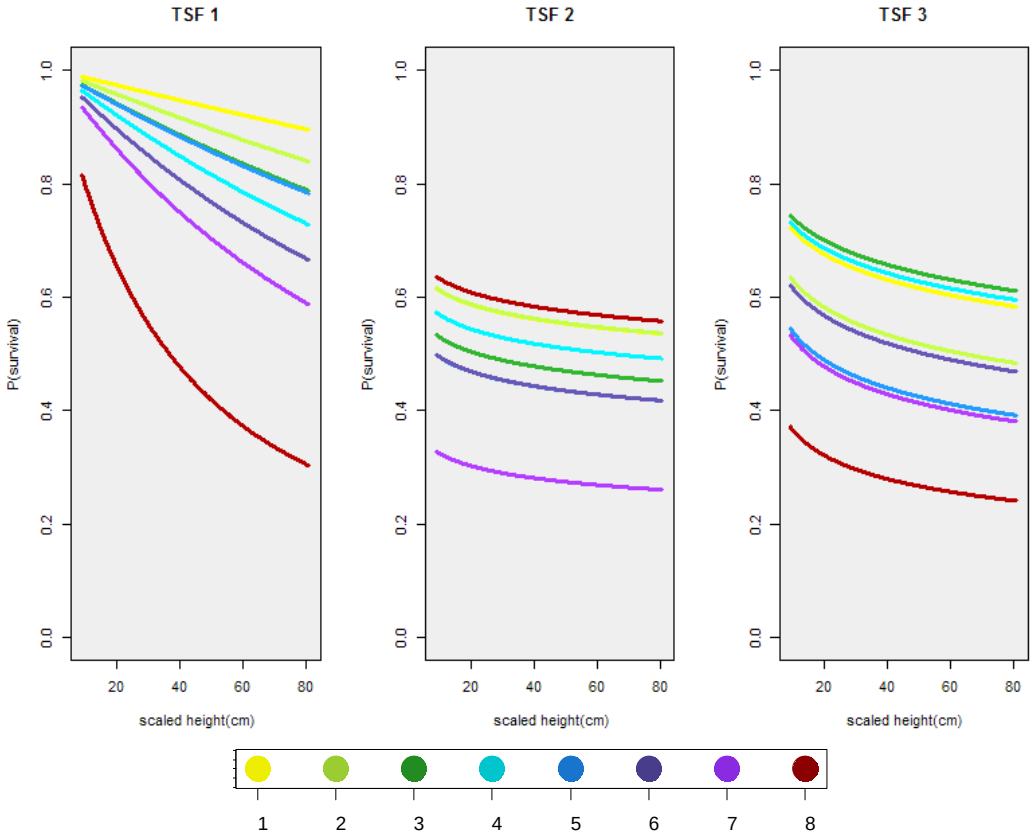

Figure 3. Plots of survival as a function of height for plants with different number of stems and time-since-fire (x=height, y=survival, stems in different colors)

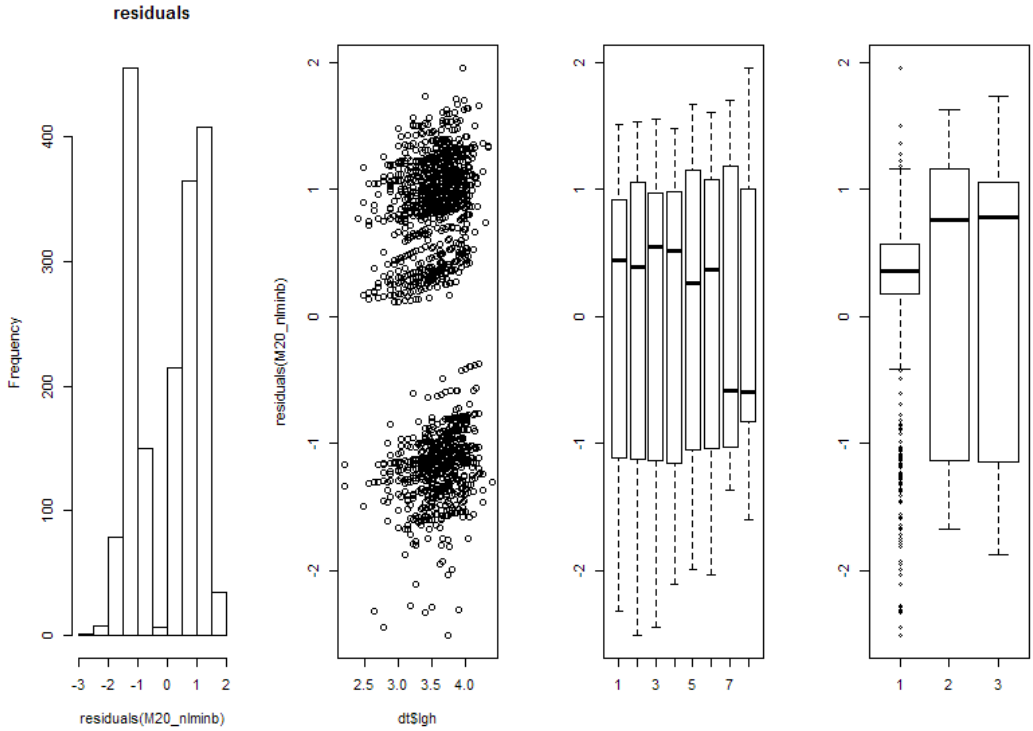

Figure 4. Residuals of model M20

Note: See an associated 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

Crawley, M. J. 2006. The R Book. Wiley.

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.