This material has been published in CAB Reviews, issue 7, No. 016, 2012, doi: 10.1079/PAVSNNR20127019, the only accredited archive of the content that has been certified and accepted after peer review.

The electronic version of this article is the definitive one. It is located here: http://www.cabi.org/cabreviews

Potential of GLMM in modelling invasive spread: Supplement C

Modelling enemy release during range expansion of *Ilex aquifolium* with a two component GLMM

J. Thiele1* and B. Markussen2

Address: [1] Institute of Landscape Ecology, University of Muenster, Robert-Koch-Str. 28, 48149 Muenster, Germany. [2] Department of Basic Sciences and Environment, University of Copenhagen, Thorvaldsensvej 40, 1871 Frederiksberg C, Denmark.
*Correspondence: J. Thiele. Email: jan.thiele@uni-muenster.de

Introduction

In this worked example, we present a two component Generalized Linear Mixed Model (GLMM) modelling a counting response with exceedingly many zeros. The emphasis will be on model construction and validation. To our knowledge only ADMB (if the likelihood functions are coded directly) and the GLIMMIX procedure in SAS allows for a flexible modelling of two-component GLMMs, so the statistical analysis will be done in SAS V9.2. However, the analysis did not proceed as smoothly as we hoped, and we will comment on some of the problems we encountered.

Data example

The evergreen tree *Ilex aquifolium* reaches its north-eastern distribution border in Denmark, where the natural population is present in Jutland and Funen but not on the eastern islands. However, recent field observations have revealed presence of *Ilex aquifolium* in most eastern parts of Denmark. This range expansion seems to be related to climate change and land use. Skou et al. (2011) studied the interplay between the range expansion and the insect herbivore *Phytomyza ilicis* (Diptera: Agromycidae) in a transplant experiment, in particular the hypothesis of enemy release during range expansion of *Ilex aquifolium.* In that study several response variables were collected and analysed, but in this example we will only discuss one of the counting responses, which required special attention in the statistical analysis due to exceedingly many zeros. The dataset that we analyse here contains the variables shown in Table 1.

The transplants were of the genotype *Blue Angel* or *Madame Briot* and were planted at 1 and 10 meter distances from a host tree in the studied populations with *natural* genotype. Thus, the variables *genotype* and *distance* are connected in the sense that *genotype=natural* implies *distance=0* and vice versa. The continuous variable *popstandard* is a proxy for the size of the 18 studied populations, which are labelled by the variable *location*. The populations are classified by the variable *class* according to being a natural population in Jutland (*naturalJ*), an escape population in Jutland (*escapeJ*), or an escape population on the eastern island Zealand (*escapeZ*). Since the

proxy *popstandard* is constant within each population, it is necessary to use *location* as a random effect, if we want to study the effect of the population size. From each population several leaves from one host tree and up to 20 transplants were collected, and the number of feeding scars from the insect herbivore was counted. The individual trees within the populations are identified by the variable *label*, which is a random effect nested in the random effect *location*. In total 3503 leaves from 306 trees from 18 locations were investigated, out of which 420 leaves from 50 trees from 17 locations had feeding scars. The response variable *scars* contains the counts from the individual leaves. The marginal distribution of *scars* has exceedingly many zeros (Fig. 1), and hence it is not possible to model this neither by a data transformation nor by a counting distribution including a dispersion parameter like the negative binomial.

Table 1 Names and characteristics of the variables in the dataset.

Variable	Levels	Effect
genotype	3 (natural, Blue Angel, Madame Briot)	Fixed
distance	3 (0, 1, 10)	Fixed
class	3 (naturalJ, escapeJ, escapeZ)	Fixed
popstandard	Continuous and positive	Fixed
location	18 (Bjerringbro, …, Stenholt)	Random
label	Up to 21 trees within each *location*	Nested random
scars	Counts (0, 1, 2, …)	Response

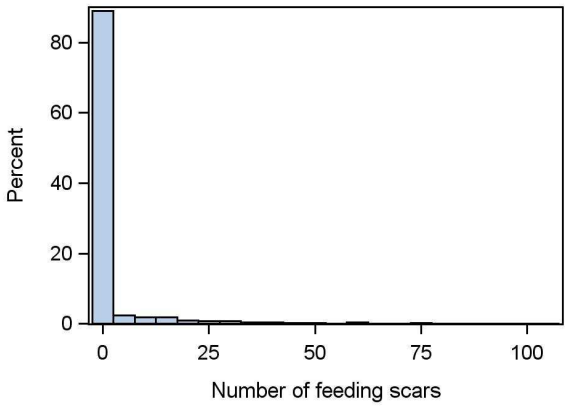

Figure 1 Histogram of the response variable *scars*.

Two component GLMM

In Skou et al. (2011), the number of feeding scars was modelled by a two component GLMM. The first component is a binomial GLMM that models the number of leaves with a strictly positive number of feeding scars against the number of leaves without feeding scars. The second component is a Linear Mixed Model (LMM) that models the logarithm of the number of feeding scars if scars were present. These two GLMMs make sense by themselves, one should just keep in mind that the normal component should be interpreted conditionally on the binomial component. However, if the link functions for the two components are chosen such that they have the same interpretation, then it makes sense to test the hypothesis that the two components share

the same parameters. If this hypothesis is not rejected, then the data is summarized by a single set of parameters and the associated interpretation. In our case, we will use a log link for the binomial GLMM in conjunction with the identity link for the LMM on the log-transformed positive counts. In the affirmative case, the interpretation will be given in terms of joint ratios of the risk of finding feeding scars and of the ratio between the numbers of scars when present.

To specify the two components we make two new datasets. For the binomial component the variable *y* counts the number of leaves with *scars>0* and the variable *total* counts the total number of leaves investigated for each tree. For the normal component we only include the observations with *scars>0* for which we define *y=log(scars)*. The initial models in both components are given by an analysis of covariance design including the factorial effects of *genotype*, *distance*, *class*, *genotype*distance* and the associated interactions with the continuous covariate *popstandard*.

Validation of binomial component

The Pearson chi-square statistic for overdispersion in the conditional distribution given the predicted random effects has a ratio of 0.35 to the degrees of freedom. Thus, there is no indication of non-modelled overdispersion. In order to validate the log link and linearity against the continuous covariate *popstandard* we investigate the cumulative residuals as proposed by Lin et al. (2002). This method is not implemented in PROC GLIMMIX, but it is available in PROC GENMOD. However, since PROC GENMOD didn't converge when the interaction *genotype*distance*popstandard* was included in the model, we were forced to exclude this term in the validation step. Figure 2 shows the observed cumulative residuals. If the model is valid, then the observed cumulative residuals should be similar to the simulations from the model.

Although the model fit could be better we will not invalidate the model. The associated Kolmogorov-Smirnov goodness-of-fit test based on 1000 simulations give p=0.062 for the log link and p=0.140 for linearity.

To investigate the distribution of the random effects we fit the GLMM with PROC GLIMMIX using the Laplace approximation. Doing this we get predictions for the random effects in terms of maximum a posteriori estimates. These estimates are displayed in Fig. 3. Although we do not expect these plots to have precise diagnostic power we see that the random effects of *location* are close to a normal distribution.

Validation of the normal component

Figure 4 shows diagnostic plots of the normal component of the model. Neither of these plots gives raise to concern. In the residual plots the levels of the variable *distance* have been coded (blue circles=0, green crosses=1, red pluses=10), and we see that there is no indication of variance heterogeneity, say, against this variable.

The normal quantile plot in Fig. 4 provides an example, where the normal distribution is rejected by the standard goodness-of-fit tests (Shapiro-Wilks p=0.0061, Kolmogorov-Smirnov p<0.0100, Cramer-von-Mises p=0.0060, Anderson-Darling p<0.0050) despite our clear decision not to invalidate the model.

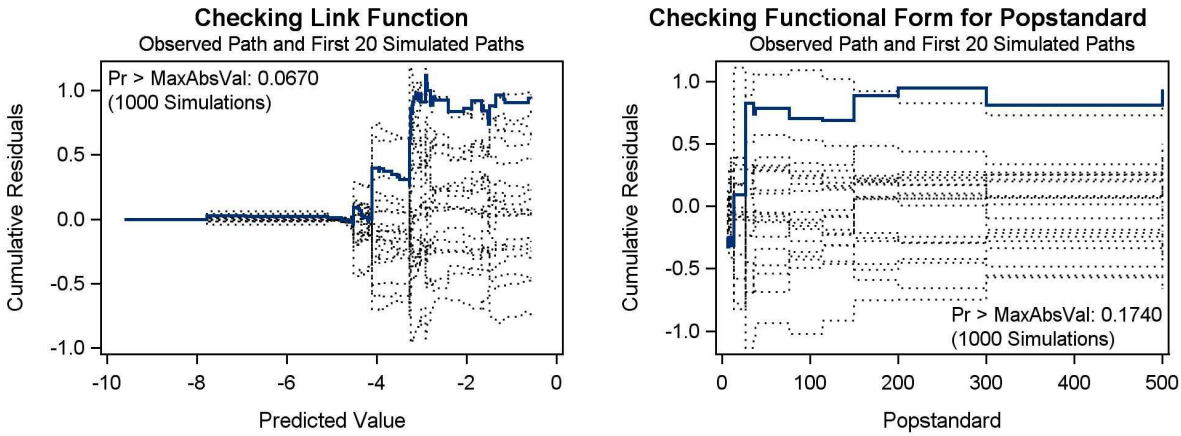

Figure 2 Cumulative residuals against the linear predictor and against the continuous covariate *popstandard* together with 20 simulations from the model.

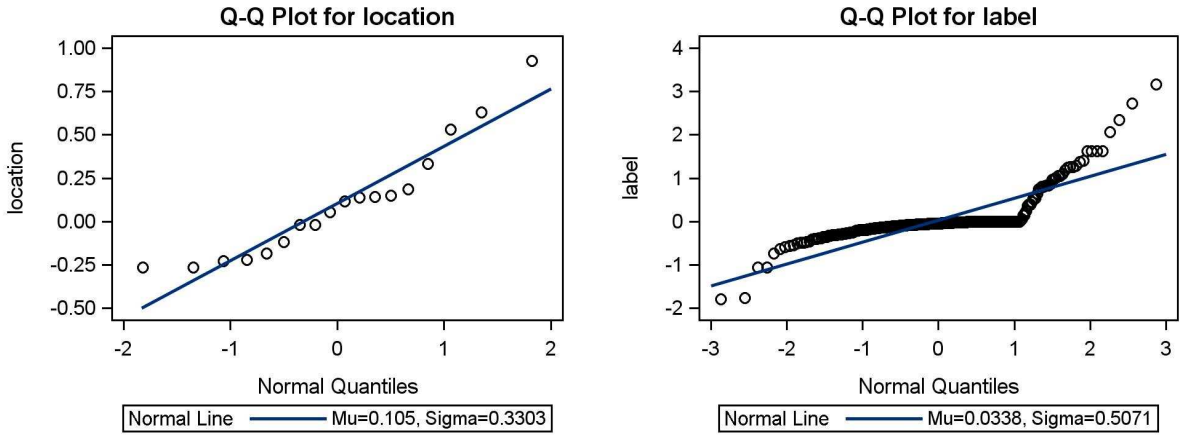

Figure 3 Normal quantile plots of a posteriori estimates of the levels of the random effects of location and label.

For the normal component the variance component for *location* estimates to zero. Since there is no random effect of *location*, we only display the normal quantile plot for the Best Linear Unbiased Predictions (BLUP) for *label* (Fig. 5).

There is a single outlier in this plot, namely the host tree from the population in *Hornbysand* situated on Zealand. From this tree, 116 leaves were investigated and 38 leaves had feeding scars. If this tree is removed from the dataset, then the variance components for the normal component both estimate to zero. In this worked example, we have chosen to remove this particular tree from the statistical analysis and proceed with a normal component without random effects.

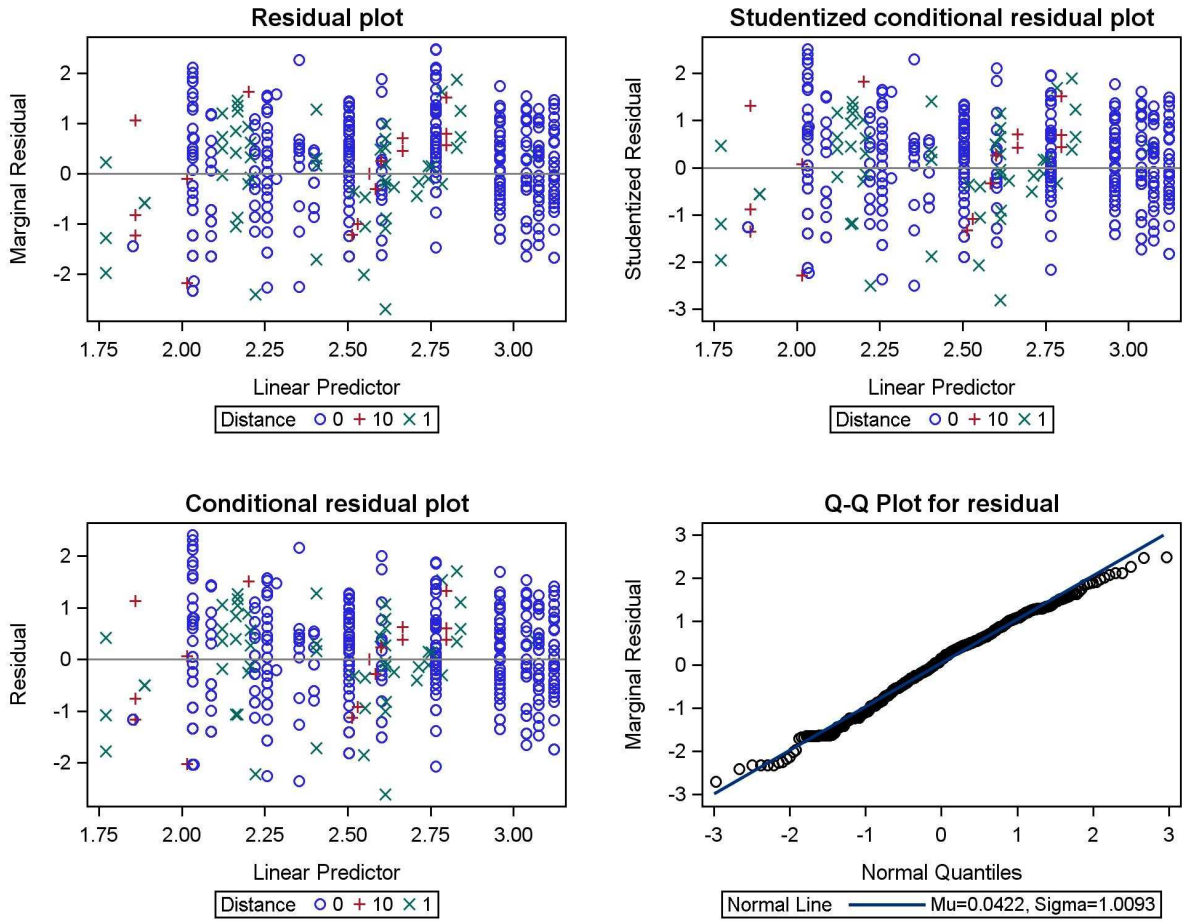

Figure 4 Some diagnostic plots of the normal model component: marginal residual plot, conditional residual plot, studentized conditional residual plot, normal quantile plot of conditional residuals.

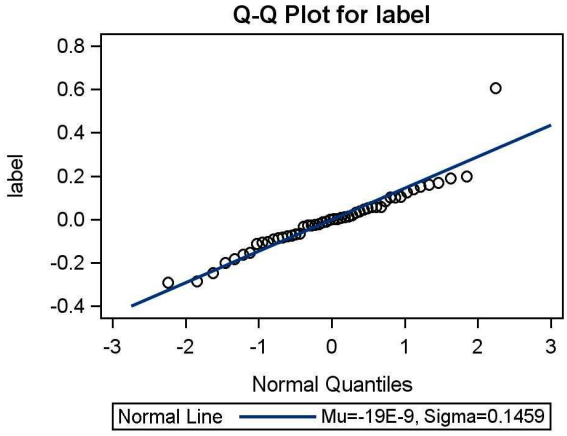

Figure 5 Normal quantile plot of Best Linear Unbiased Predictions for the random effect of *label*.

Building the two component model

To combine the binomial and the normal components into a joint model, we stack the two datasets on top of each other and adjoin four new variables *(response, link, binomial, normal)* that take the values *(binomial, log, 1, 0)* for the binomial component and *(normal, identity, 0, 1)* for the normal component (Table 2).

Table 2 Additional variables for the two component model.

Variable	Levels	Effect
response	2 (binomial, normal)	Fixed + Selection of distribution
link	2 (log, identity)	Selection of link function
binomial	Continuous (0 or 1)	Random
normal	Continuous (0 or 1)	Random

We have also added a new variable called *gd*, which simply is the concatenation of the variables *genotype* and *distance*. This variable is encoding the interaction *genotype*distance*, and it has 5 levels. Below we show a print out of the data from Stokkebro. Here 8 of the 47 leaves from the host plant had feeding scars (see observation 29), and none of the leaves from the 12 transplants had feeding scars (see observation 30 to 41). The logarithms of the 8 strictly positive counts of feeding scars are listed in observation 21 to 28.

Obs	Label	genotype	Location	class	Distance	Poopstand	Feed	gd	response	link	total	normal	binomial	y
21	1	natural	Stokkebr	escapeJ	0	76	14	natural*0	normal	identity	1	1	0	2.63906
22	1	natural	Stokkebr	escapeJ	0	76	10	natural*0	normal	identity	1	1	0	2.30259
23	1	natural	Stokkebr	escapeJ	0	76	6	natural*0	normal	identity	1	1	0	1.79176
24	1	natural	Stokkebr	escapeJ	0	76	5	natural*0	normal	identity	1	1	0	1.60944
25	1	natural	Stokkebr	escapeJ	0	76	19	natural*0	normal	identity	1	1	0	2.94444
26	1	natural	Stokkebr	escapeJ	0	76	6	natural*0	normal	identity	1	1	0	1.79176
27	1	natural	Stokkebr	escapeJ	0	76	10	natural*0	normal	identity	1	1	0	2.30259
28	1	natural	Stokkebr	escapeJ	0	76	18	natural*0	normal	identity	1	1	0	2.89037
29	1	natural	Stokkebr	escapeJ	0	76	0	natural*0	binomial	log	47	0	1	8.00000
30	2	mb	Stokkebr	escapeJ	10	76	0	mb*10	binomial	log	9	0	1	0.00000
31	3	mb	Stokkebr	escapeJ	10	76	0	mb*10	binomial	log	6	0	1	0.00000
32	4	mb	Stokkebr	escapeJ	1	76	0	mb*1	binomial	log	4	0	1	0.00000
33	5	ba	Stokkebr	escapeJ	1	76	0	ba*1	binomial	log	4	0	1	0.00000
34	6	ba	Stokkebr	escapeJ	10	76	0	ba*10	binomial	log	10	0	1	0.00000
35	7	ba	Stokkebr	escapeJ	10	76	0	ba*10	binomial	log	4	0	1	0.00000
36	8	ba	Stokkebr	escapeJ	1	76	0	ba*1	binomial	log	6	0	1	0.00000
37	9	ba	Stokkebr	escapeJ	1	76	0	ba*1	binomial	log	7	0	1	0.00000
38	10	ba	Stokkebr	escapeJ	1	76	0	ba*1	binomial	log	4	0	1	0.00000
39	11	ba	Stokkebr	escapeJ	10	76	0	ba*10	binomial	log	8	0	1	0.00000
40	12	ba	Stokkebr	escapeJ	10	76	0	ba*10	binomial	log	2	0	1	0.00000
41	13	ba	Stokkebr	escapeJ	10	76	0	ba*10	binomial	log	7	0	1	0.00000

The fixed effects from the two components are used in the joint model together with their interaction with *response* and together with the main effect of *response*. Doing this, the fixed effects vary freely in the two components in the initial model. In order to have separate variance components on the binomial and the normal responses, we use the dummy variables *binomial* and *normal*. For the binomial component we use *label* nested in *location*. For the normal component we have no random effects since their variance components were estimated at zero. PROC GLIMMIX allows us to select the response distribution and the link

function separately for each observation. This is done using the variables *response* and *link*, respectively. The syntax looks as follows:

```
proc glimmix data=scars method=laplace;
  class genotype distance class location label response;
  model y/total = response
        class genotype distance genotype*distance
        popstandard class*popstandard genotype*popstandard
        distance*popstandard genotype*distance*popstandard
        /**/
        response*class response*genotype
        response*distance response*genotype*distance
        response*popstandard response*class*popstandard
        response*genotype*popstandard response*distance*popstandard
        response*genotype*distance*popstandard
      / solution dist=byobs(response) link=byobs(link);
  random binomial / subject=location nofullz;
  random binomial / subject=label*location nofullz;
run;
```

If the initial model also included random effects for the normal component, then we would add corresponding RANDOM statements replacing the dummy variable *binomial* by the dummy variable *normal.*

Model reduction

The statistical analysis continues by backward reduction of the fixed effects. PROC GLIMMIX provides Wald F-tests for the fixed effects, and comparing the fitted likelihood we may compute likelihood ratio tests manually. All of this is quite time consuming in SAS since the backward model reduction is not automated. Furthermore, we also encountered severe convergence problems. These problems were partly overcome tweaking the numerical optimization via the NLOPTIONS statement, and by using the full-rank coding *gd* of the interaction *genotype*distance* (Cheng et al. 2010). However, it was not possible to restart PROC GLIMMIX at the variance components found in the previous iteration. Doing this in a PARMS statement almost consistently resulted in the error message:

```
ERROR: Values given in PARMS statement are not feasible.
```

Mathematically, this does not make sense since it always should be possible to restart a numerical optimization at the present estimate. The impossibility to restart the numerical optimization is most unfortunate since it makes the comparison of the likelihoods prone to instabilities due to the numerical optimization. The steps in the model reduction are summarized below:

Step	Reduction	p(Wald F)	-2logL	df	test df	LR test	p(LR)
1	Full model		481.52	27	.	.	.
2	response*class*popstandard	0.1948	1484.96	25	2	3.44619	0.17851
3	response*gd*popstandard	0.2852	1492.03	22	3	7.07161	0.06965
4	response*popstandard	0.0783	1493.74	20	2	1.70535	0.42627
5	response*class	0.2120	1497.48	19	1	3.74561	0.05295
6	response*genotype*distance	??	1500.44	18	1	2.95072	0.08584
7	response*genotype	0.7698	1500.18	17	1	-0.25571	1.00000
8	genotype*distance*popstandard	??	1501.55	16	1	1.36975	0.24186
9	genotype*popstandard	0.3759	1503.34	15	1	1.78804	0.18116

The fixed effects in the final model are listed below:

Effect	Num Df	Den Df	Wald F	p(Wald F)
response	1	370	471.49	<0.0001
class	2	370	4.53	0.0114
genotype	1	370	14.28	0.0002
distance	1	370	16.15	<0.0001
*genotype*distance*	1	370	4.64	0.0318
*response*distance*	2	370	18.83	<0.0001
popstandard	1	370	36.98	<0.0001
*class*popstandard*	2	370	14.23	<0.0001
*distance*popstandard*	2	370	3.77	<0.0001

Since the scales of the binomial and the normal components are incomparable (probabilities against log counts), we a priori expected that the main effect of *response* would be significant. Beside this, it is interesting that the variable *response* only appears in the interaction with *distance.* This means that except for the main effect of *distance* all the other effects may be assumed to have the same influence on the ratio of the probabilities of finding feeding scars and of numbers of feeding scars when these are non-zero.

Variance components

The final model contains variance components on the binomial part that models the log-probability for having some feeding scars. The estimates for the variance components are:

Effect	Variance estimate	Standard error
location	0.4670	0.2327
label	0.8527	0.2287

In particular, we see that the variation between trees is almost twice as big as the variation between locations.

Model predictions

As an example of the model predictions, the following table displays the estimated ratios between the three population types. These ratios are found by exponentiation of the pairwise differences of the least squares means.

Corrected for population size	Comparison	Estimated ratio	95% confidence interval
Yes	escapeJ vs. escapeZ	0.41	0.24 ; 0.69
	escapeJ vs. naturalJ	1.34	1.01 ; 1.77
	escapeZ vs naturalJ	3.26	1.93 ; 5.50
No	escapeJ vs. escapeZ	1.47	1.14 ; 1.90
	escapeJ vs. naturalJ	0.95	0.74 ; 1.21
	escapeZ vs naturalJ	0.64	0.50 ; 0.84

The ratios are markedly different depending on whether the estimates are corrected for the population sizes or not. This is due to a large difference between the population sizes in the three population types.

Class	N	Mean population size	Min ; Max
escapeZ	6	32	5.00 ; 114
escapeJ	5	61	5.00 ; 200
naturalJ	6	245	6.25 ; 500

If we ignore the random effects in the binomial component, then it is also possible to compute estimates for mean number of feeding scars. Recall that the strictly positive counts are modelled by a normal distribution on the logarithmic scale. If this distribution has mean μ and variance σ^2, then the properties of the log-normal distribution gives that mean of the strictly positive counts equals $exp(\mu+\sigma^2/2)$. Furthermore, suppose that the logarithm of the probability for strictly positive counts equals ξ. Then the logarithm of the mean of X, where X is the number of feeding scars, equals

$$log(mean(X)) = log(P(X>0)*mean(X|X>0)) = log(P(X>0))+log(mean(X|X>0)) = \xi+\mu+\sigma^2/2.$$

In this equation, ξ and μ are the linear predictors in the binomial and the normal components, respectively, and σ^2 is the residual variance in the normal component. Estimates and standard error for $\xi+\mu$ and for $\sigma^2/2$ are easily assessable. If we ignore the correlation between the two estimates, then Pythagoras theorem may be invoked to compute the standard error for $\xi+\mu+\sigma^2/2$. As an example, we find the following estimates and confidence intervals for the mean number of feeding scars in a transplant tree at *distance=1m* in a population with *popstandard=50*:

Class	**Mean number of feeding scars**	**95% confidence interval**
escapeZ	0.51	0.25 ; 1.04
escapeJ	0.56	0.26 ; 1.20
naturalJ	0.47	0.21 ; 1.05

These means and the ratios reported above all show that the numbers of feeding scars are not markedly different in the three population types. Thus, there is no support in favour of the hypothesis of enemy release. This conclusion is consistent with the findings of Skou et al. (2011), who as mentioned above also analysed several other responses.

Acknowledgement

Many thanks go to Anne-Marie T. Skou, Lene Sigsgaard and Johannes Kollmann for providing the data set.

References

Cheng J, Edwards LJ, Maldonado-Molina MM, Komro KA, Muller KE. Real longitudinal data analysis for real people: Building a good enough mixed model. Statistics In Medicine 2010;29:504–520.

Lin EY, Wei LJ, Ying Z. Model-checking techniques based on cumulative residuals. Biometrics 2002;58:1–12.

Skou AMT, Markussen B, Sigsgaard L, Kollmann J. No evidence for enemy release during range expansion of an evergreen tree in northern Europe. Environmental Entomology 2011;40:1183–1191.