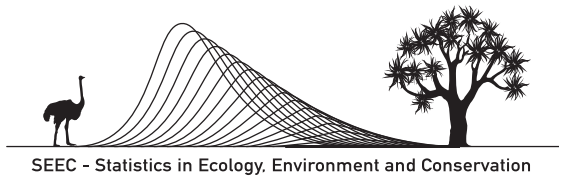
SEEC stats toolbox seminar series:

Generalized Linear Mixed Models

Mzabalazo Z. Ngwenya

Centre for Statistics in Ecology, Environment and Conservation (SEEC)



Department of Statistical Sciences
University of Cape Town

"GLMMs are surprisingly challenging to use even for statisticians ..... We strongly recommend that researchers proceed with caution by making sure they have a good understanding of the basics of linear and generalized mixed models before taking the plunge into GLMMs"

Bolker *et al* (2008)

Outline

1. Introduction
   - Why generalized linear models
   - Why mixed models
2. Generalized Linear Mixed Models
   - Estimation of parameters
   - Inference (Model selection and Hypothesis testing)
   - Mulitmodel inference
   - Checking assumptions

Table 1. Data set for illustration

Variable	Description
Site	Study site
Type	Method of seed dispersal
Species	Type of Acacia species
Seedbank size (count)	Response
Stand age	Explanatory variable

Strydom M., Veldtman R., Ngwenya M.Z., Esler K.J. (2017). Invasive Australian Acacia seed banks: Size and relationship with stem diameter in the presence of gall-forming biological control agents. *PLoS ONE*, **12** (8)

Introduction

1.1 Why generalized linear models

- Most of the data in ecology, environment and conservation studies is non-normal; e.g. count data, binary data, proportions
  - Count data examples:
    number of individuals of a certain species in an area, clutch sizes of birds
  - Binary data examples:
    presence-absence of a species in a locale, infection status of individuals with regards to a certain disease
  - Proportional data examples:
    sex ratios, infection rates, mortality rates within a group or area

- To overcome non-normality of data and analyse these data with linear models
  - Apply transformations
  - Use non parametric tests
  - Rely on the robustness of classical ANOVA
- However GLMs are a more suitable tool for such type of analysis
- To use GLMs all one has to do is
  1. Specify distribution of your data
  2. Specify link function

Link function:

The function that describes the relationship between the mean of the response and a linear combination of the covariates

Table 2. Distributions and associated link functions for the various types of data commonly encountered in ecology

Data	Distribution	Link
Count	Poisson	Log
Binary	Bernoulli	Logit
Proportions	Binomial	Logit

`glm`{**stats**} and `glmer`{**lme4**} - will fit these models for you

- It may happen that your data may show more variation than what would be expected under the distributions shown in Table 2 - Overdispersion
- In such situations one can use the quasi-Poisson or negative binomial distribution to model count data instead of the Poisson distribution
- Similarly one can use the quasi-Binomial distribution for proportions
- For GLMs overdispersion can be tested for by applying the `dispersion.test{`**AER**`}` function on a fitted model - values greater than 7.5 indicate overdispersion

1.2 Why mixed models

- Most environmental and ecological studies are observational and include
  - Natural blocking; species, sites
  - Repeated observation of the same subjects over time
  - Samples of observational units from a larger population
- Want a way to model this variability (random variation)
- Modeling random variation allows one to extrapolate results to individuals and populations beyond the study sample; make inferences about the general population
- If random variation not accounted for all inferences are limited to study sample
- Therefore need to model this "random effect"

Random effect:

Grouping variable for which we are trying to control for e.g. site, biome, observation time

- Random effects are formed from categorical variables whose levels are sampled from a larger population
- Interest is not on the effect of the random variable on the response. Instead interest is in the variation exhibited by each level of the random effects

Fixed effect:

Factors whose levels are experimentally determined or in which were are interested in determining the specific effects of each level

- These are variables we expect to have an effect on the response. Interest is on these effects: differences among levels/treatments and interactions

Note:

It is common to have situations where strictly speaking a variable could be classified as a fixed or random effect. Eventual assignment of variables will thus depend on the context of the study, research questions to be answered and/or experimental design employed.

Variable	Description
Site	Study site
Type	Method of seed dispersal
Species	Type of Acacia species
Seedbank size (count)	Response
Stand age	Explanatory variable

- Types of random effects
  - Block random effects:
    These are effects that apply equally - usually natural groupings e.g. species, site

    `(1|site)`
  - Nested random effects:
    Appears in situations where we have multiple random effects that follow some kind of hierarchy e.g. species within genus

    `(1|type/species)` or `(1|type) + (1|type:species)`
  - Crossed random effects:
    Arise when there are multiple random effects that affect our sample units independently, e.g. time and block

    `(1|site) + (1|type)`

Generalized Linear Mixed Models

Combine generalized linear models and linear mixed models to form a very powerful tool

To use one has to specify

1. Distribution of data
2. Link function
3. Structure of the random effects

2.1 Estimation of parameters

1. Fixed effect parameters - Effects of covariates
2. Random effects variance - Variation across groups

Approaches to estimation

- Maximum Likelihood (ML) and Restricted Maximum Likelihood: `glmmML`{**MASS**}
- Pseudo/penalized quasilikelihood (PQL): `glmmPQL`{**MASS**}
- Laplace approximation: `glmer`{**lme4**}, `glmmML`{**MASS**}, `glmmadmb`{**glmmADMB**}
- Gauss-Hermite quadrature (GHQ): `glmer`{**lme4**}
- Markov Chain Monte Carlo (MCMC): `glmer`{**lme4**}, `MCMCglmm`{**MCMCglmm**}

- ML and REML:
  - ML tends to underestimate random effect standard deviations except in very large data
  - REML is better at obtaining unbiased random effect standard deviations
  - Slow and sometimes computationally infeasible when there are a large number of random effects.
- PQL:
  - Will yield biased parameter estimates of fixed effects if the random effects are large. This is especially true for Binary data
  - Performs poorly for Poisson data when the mean number of counts per treatment combination is less than 5
  - Poor performer for Binomial data where the expected number of success and failure are both less than 5

- Laplace approximation
  - Better accuracy than PQL
  - Computes actual likelihood and hence allows for likelihood based inference
- GHQ
  - More accurate than Laplace
  - Slow and speed decreases further with increasing number of random effects
  - Hence not feasible for analysis with more than 2 or 3 random factors

- MCMC
  - Gives comparable results to likelihood methods when data sets are highly informative and a vague prior is used
  - Extends easily to multiple random effects - need large data sets to do this though
  - MCMC involves difficult technical details making their correct and effective use potentially difficult

Take home:

Laplace approximation with REML estimation will be best for most problems.

2.2 Inference - Model selection and hypothesis testing

Inference includes

- Inspecting parameter estimates and their confidence intervals
- Testing (biological) hypothesis
- Determining best model and evaluating goodness of fit of models

These processes and objectives are not mutually exclusive

Model selection:

Compare fits of candidate models to find best model. We seek to balance goodness of fit and model complexity

Model selection can be implemented via one of two ways

1. Traditional null hypothesis testing approach
   - Likelihood Ratio Test - `lrtest`{**lmtest**}; `anova`{**stats**}
2. Information theoretic approach
   - Akaike Information Criterion - `AIC`{**stats**}; `AICc`{**AICcmodavg**} (AIC_c for small samples; QAIC and quasi-AIC for overdispersed data)
   - Bayesian Information Criterion - `BIC`{**stats**}
   - Deviance Information Criterion - `DIC`{**AICcmodavg**}

2.2.1 Null hypothesis testing approach

- Simple nested models are tested against more complex models
- This approach can lead to suboptimal models - model selected as "best" depends on the order of testing
- Unreliable for small to moderate samples
- LRT gives no indication for the relative support of competing models
- In LRTs multiple pairwise comparisons are performed which increases possibility of type-I-error

2.2.2 Information theoretic approach

- All IT methods have a term which penalizes complex models - hence the do well in balancing goodness of fit and model complexity
- Allow for simultaneous comparisons of multiple competing models which are nested or non-nested
- Competing models can be ranked which is useful where ere there is more than on plausible hypothesis
- Provide a basis for averaging parameter estimates and predictions across various models which has the following benefits
    1. Provide more accurate estimates of parameters and predictions
    2. Enable construction of confidence intervals that correctly account for model uncertainty

2.3 Multimodal Inference

- Well supported in R - **AICcmodavg** and **MuIn**; (we use the later package here)
- Model averaging for inference involves 3 main steps

1. Generating model set:

   Use domain knowledge to form models which represent hypothesis of interest/possible scenarios - *model.set*

2. Identifying set of models with good support

   `top.models<-get.model(model.set, cumsum(weight)` $\leq$ `0.95)`

   The above command will form the 95% confidence set of model.

3. Model averaging

   `avg.model <- model.avg(top.models)`

```
Confidence set for the best model
Method:  raw sum of model probabilities

95% confidence set:
       K    AICc Delta_AICc AICcWt
mod 3 10 4111.21       0.00   0.52
mod 2 11 4112.83       1.63   0.23
mod 1 13 4113.48       2.28   0.17
mod 8 12 4114.86       3.65   0.08

Model probabilities sum to 1
```

- Once you have obtained averaged model you can then inspect the model in the usual manner; `summary()`, `confint()`
- Similarly prediction can be made in the usual manner using the `predict()` function

2.4 Checking assumptions

- To check for overdispersion

  `dispersion_glmer`{**blemco**}

  values greater than 1.4 indicate overdispersion

- Other diagnostics
  - Plots of the residuals like in linear model are the most prevalent (graphical) diagnostic tool
  - Pearson and deviance residuals are commonly used; it is advised that one should stick with deviance residuals

Bibliography

Bolker B.M., Brooks M.E., Clark C.J., Geange S.W., Poulsen J.R., Stevens H.H. and White J.S. (2008).
Generalized linear mixed models: a practical guide for ecology and evolution.
Trends in Ecology and Evolution, **24**, 127-134.

Burnham K.P. and Anderson D.R. (2011)
AIC model selection and multimodal inference in behavioral ecology: some background, observations and comparisons.
Behavioral Ecology and Sociobiology, **65**, 23-35.

Faraway J.J (2016).
Extending the Linear Model with R: generalized linear, mixed effects and nonparametric regression. Chapman and Hall/CRC, New York.

Gruber C.E. Nakagawa S., Laws R.J. and Jamieson I.G. (2011).
Multimodal inference in ecology and evolution: challages and solutions.
Journal of Evolutionary Biology, **24**, 699-711.

Johnson J.B. and Omland K.S. (2004).
Model selection in ecology and evolution.
Trends in Ecology and Evolution, **19**, 101-108.

Zuur A.F., Ieno E.N., Walker H.J., Saveliev A.A. and Smith G.M (2009)
Mixed Effects Models and Extensions in Ecology with R. Springer, New York.