Introduction to Generalised Linear Models for Ecologists

Dr Niamh Mimnagh

niamh@prstats.org

https://github.com/niamhmimnagh/GLME01---Introduction-to-Generalised-Linear-Models-for-Ecologists

PR STATS

Course Outline

- A recap on the normal model
- Models for binary data
- Models for binomial data
- Models for multinomial data
- Models for count data
- Models for overdispersion data
- Models for zero-inflated data
- Bayesian models
- Models for grouped data

Why Do We Model Data?

1. To understand relationships between variables
2. To predict outcomes for new or future data
3. To test hypotheses about ecological processes
4. To simplify complex systems into interpretable components

Examples:

- Predicting species richness based on elevation, temperature or rainfall.
- Estimating plant biomass from soil nitrogen or sunlight exposure.
- Modelling bird abundance based on land-use type or proximity to water.

What is a Normal Model?

- Y_i is a response variable associated with observational or experimental unit i.
- We assume it comes from a certain probability distribution with probability mass function/probability density function f and vector of parameters θ
- In general, one of the parameters in θ is the mean of the distribution
- We also have predictors x_i that we are interested in studying
- We can link these predictors to a parameter of interest, typically the mean of the distribution

What is a Normal Model?

For the normal model we typically write for each observation:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_{2i} x_{2i} + \ \ldots + \epsilon_i$$
$$\epsilon \sim N(0, \sigma^2)$$

Each β coefficient represents the expected mean change in y for a 1-unit increase in its predictor, provided all other predictors are fixed.

We can show that, from the equation:

$$y_i = \beta_0 + \beta_1 x_{1i} + \ \ldots + \beta_p x_{pi} + \epsilon_i$$

The expected value of Y_i is:

$$E[Y_i] = \beta_0 + \beta_1 x_{1i} + \ \ldots + \beta_p x_{pi}$$

And the variance is:

$$Var(Y_i) = \sigma^2$$

Key Assumptions: Independence

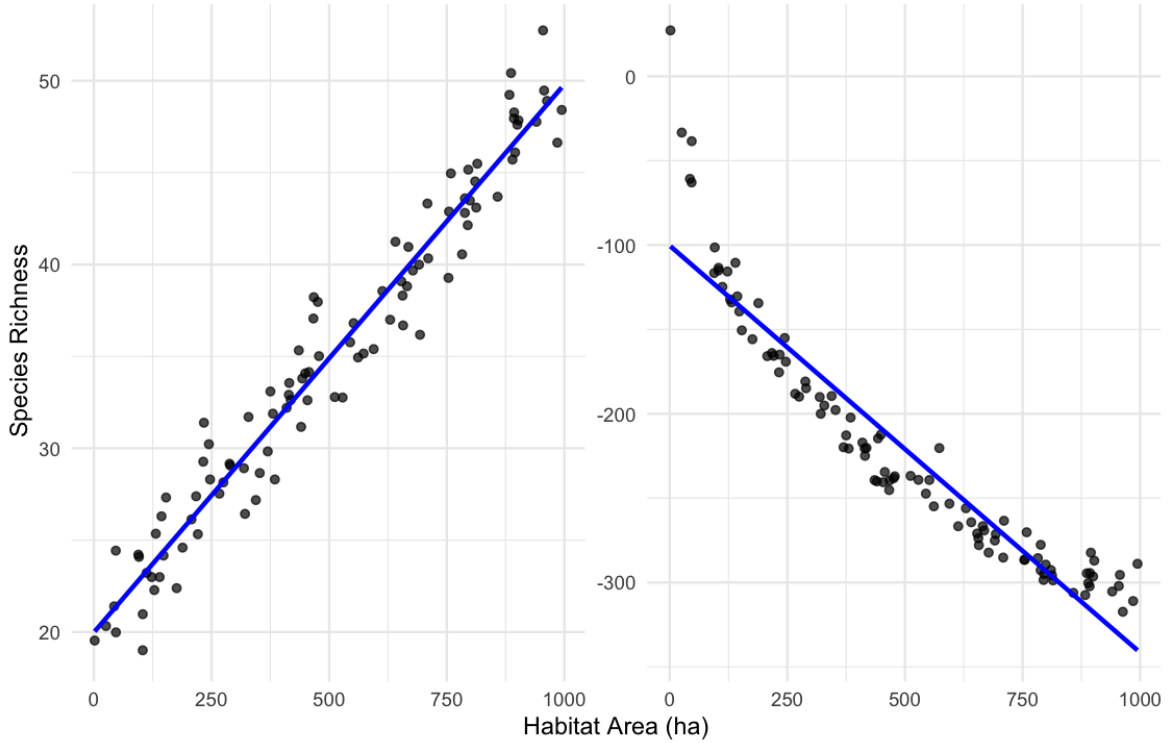
- The observations are assumed to be independent of each other. No observation should influence another.

- Independence is determined from the context of the data.

Independent	Dependent
One measurement taken per individual	Multiple measurements taken per individual
Measurements taken across random locations	Spatially correlated locations
Individuals in shared environments	Individuals in independent environments

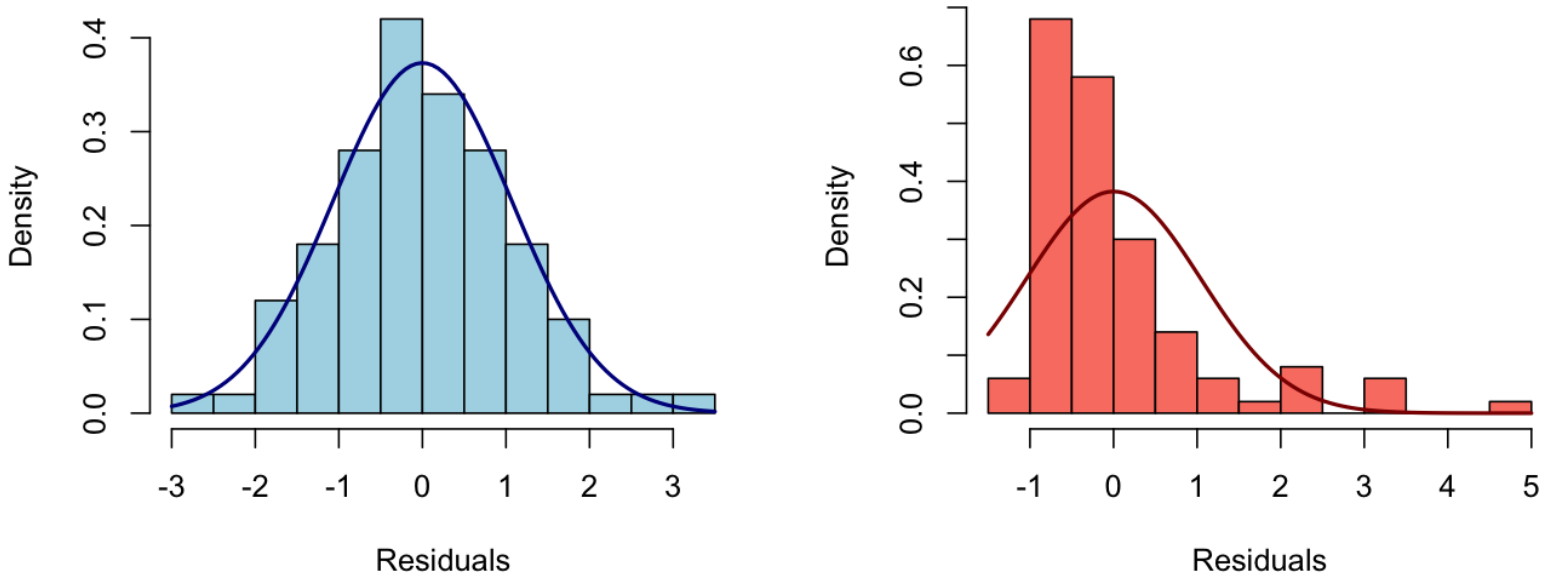
Key Assumptions: Linearity

- The relationship between the predictors and the response variable is assumed to be linear.



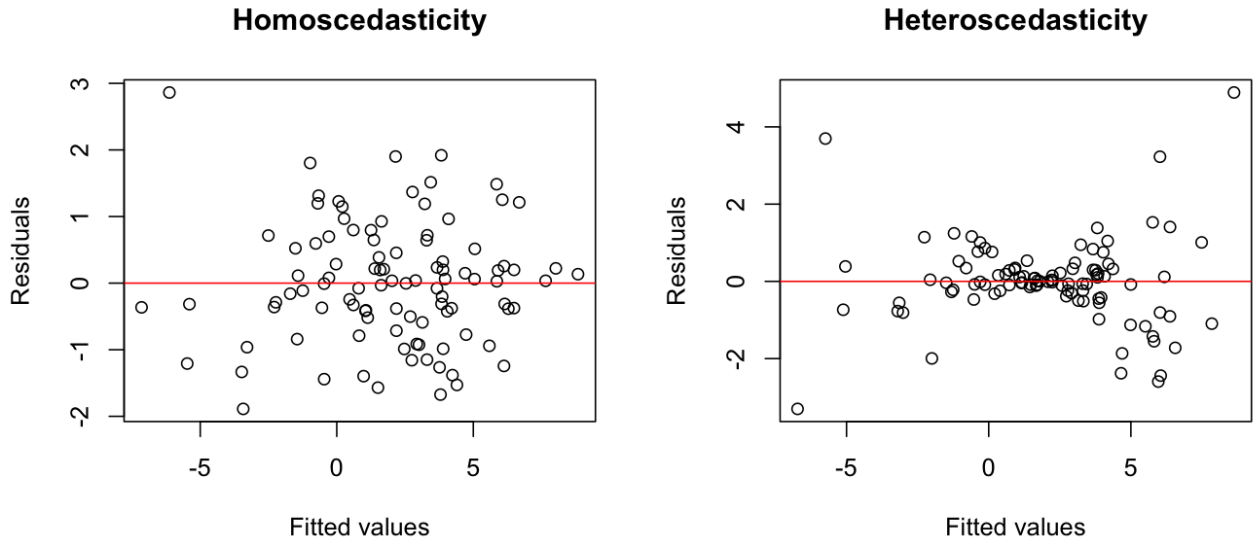
Key Assumptions: Normality of Residuals

- The residuals should be approximately normally distributed.



Key Assumptions: Homoscedasticity

- The residuals should have constant variance (the spread of the errors should be roughly equal for all predicted values).



Example: Animal Movement

- Our example investigates factors influencing average daily movement distance of tracked animals (km).
- Lets say we have 50 forest sites surveyed using GPS collars.
- Response: Average daily movement distance (km) - continuous, approximately normal.
- Predictors:
  - Habitat area (ha) - continuous
  - Canopy cover (%) - continuous

$$Movement_i = \beta_0 + \beta_1 HabitatArea_i + \beta_2 CanopyCover_i + \epsilon_i$$

Coding Demo

Categorical Predictors

- Categorical predictors may be included as factors in linear models.
- This is done using dummy coding.
- Linear models require numeric inputs
- A categorical variable with k levels cannot be directly included
- Solution: represent categories using dummy (indicator variables)
- Each dummy variable = 1 if the observation is in that category, and 0 otherwise

Dummy Coding

- Suppose habitat type has three categories: forest, grassland and wetland. What do we do with these?
- Answer: Create two dummy variables (Not three!)
- Grassland: 1 if grassland, 0 otherwise
- Wetland: 1 if wetland, 0 otherwise
- Reference level: forest (when both dummy variables are 0)

Reference Levels

- The reference category is the one with all dummies = 0
- By default (in R), it's the first level alphabetically
- You can change the reference level to:
  - Improve interpretation (e.g., compare against the most common habitat)
  - Highlight a specific category of interest

Habitat	Grassland	Wetland
Grassland	1	0
Forest	0	0
Wetland	0	1
Forest	0	0
Wetland	0	1
Grassland	1	0

Categorical Predictors

$$Movement_i = \beta_0 + \beta_1 Grassland_i + \beta_2 Wetland_i + \epsilon$$

- β_0: Mean movement in the reference habitat (forest)
- β_1: difference in movement between grassland and forest
- β_2: difference in movement between wetland and forest

Coding Demo

Interaction Terms

- Interaction occurs when the effect of one predictor depends on the level of another.
- For example, does the size of the habitat area affect richness the same way, across all habitats?
- Adding an interaction term allows different slopes for different habitats, allowing effects to vary.

Interaction Model Structure

$$Movement_i = \beta_0 + \beta_1 HabitatArea_i + \beta_2 Grassland_i + \beta_3 Wetland_i + \beta_4 (HabitatArea_i \times Grassland_i) + \beta_5 (HabitatArea_i \times Wetland_i) + \epsilon_i$$

β_0 : Baseline mean richness (forest, nitrogen $= 0$)

β_1 : Habitat Area slope for reference habitat

β_2 : Baseline difference Grassland vs Forest at Nitrogen=0

β_3 : Baseline difference Wetland vs Forest at Nitrogen=0

β_4 : Change in Habitat Area slope for Grassland vs Forest

β_5 : Change in Habitat Area slope for Wetland vs Forest

Interaction Model Structure

$$Movement_i = \beta_0 + \beta_1 HabitatArea_i + \beta_2 Grassland_i + \beta_3 Wetland_i + \beta_4(HabitatArea_i \times Grassland_i) + \beta_5(HabitatArea_i \times Wetland_i) + \varepsilon_i$$

Forest (reference): $Richness = \beta_0 + \beta_1 HabitatArea$

Grassland: $Richness = (\beta_0 + \beta_2) + (\beta_1 + \beta_4) HabitatArea$

Wetland: $Richness = (\beta_0 + \beta_3) + (\beta_1 + \beta_5) HabitatArea$

If $\beta_4 = 0$ → The effect of habitat area in the grassland is the same as the effect in the forest.

If $\beta_4 \neq 0$ → Habitat area effect on species richness is different in grassland versus forest

Why Include Interactions?

- Interaction terms are realistic: different habitats may respond differently to nutrients.
- Interaction terms can model fit when true effects are not additive.
- Caution: Interactions increase model complexity and must be justified by theory or data.

Coding Demo

Limitations of Linear Models

- Linear models are appropriate for continuous, normally distributed outcomes only.
- They are not suitable for many types of data that are commonly encountered in ecological studies, including:
  - Binary data
  - Counts
  - Proportions

Model Formulation

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_{2i} x_{2i} + \ \ldots + \epsilon_i$$
$$\epsilon \sim N(0, \sigma^2)$$

$$Y_i \sim Normal(\mu_i, \sigma^2)$$
$$\mu_i = \beta_o + \beta_1 x_i$$

What about the linear model assumptions?

They are built into this model!

Generalised Linear Models

Y can be assumed to have any distribution

So why do we use the normal distribution so often?

A History of GLMs

- Multiple linear regression: a normal model with the identity link (Legendre, Gauss, Galton, 19th Century)
- Analysis of Variance (ANOVA): a normal model with the identity link (Fisher, 1918)
- The exponential family class of distributions (Fisher, 1934)
- Probit analysis: a binomial distribution with the probit link (Bliss, 1935)
- Logistic regression: a binomial distribution wit the logit link (Berkson, 1944; Dyke and Patterson, 1952)
- Log-linear models: a Poisson distribution with the log link (Birch, 1963)
- Regression for survival data: an exponential distribution with the inverse or log links (Feigl and Zelen, 1965; Zippin and Armitage, 1966; Gasser, 1967)
- Inverse polynomials: a gamma distribution with the inverse link (Nelder, 1966)

A History of GLMs

J. R. Statist. Soc. A,
(1972), **135**, *Part* 3, p. 370

Generalized Linear Models

By J. A. NELDER and R. W. M. WEDDERBURN

Rothamsted Experimental Station, Harpenden, Herts

SUMMARY

The technique of iterative weighted linear regression can be used to obtain maximum likelihood estimates of the parameters with observations distributed according to some exponential family and systematic effects that can be made linear by a suitable transformation. A generalization of the analysis of variance is given for these models using log-likelihoods. These generalized linear models are illustrated by examples relating to four distributions; the Normal, Binomial (probit analysis, etc.), Poisson (contingency tables) and gamma (variance components).

The implications of the approach in designing statistics courses are discussed.

Keywords: ANALYSIS OF VARIANCE; CONTINGENCY TABLES; EXPONENTIAL FAMILIES; INVERSE POLYNOMIALS; LINEAR MODELS; MAXIMUM LIKELIHOOD; QUANTAL RESPONSE; REGRESSION; VARIANCE COMPONENTS; WEIGHTED LEAST SQUARES

GLMs

GLMs extend linear models to accommodate non-normal response distributions.

Structure:

1. Random component: a distribution belonging to the exponential family
2. Systematic component: a linear predictor
3. Link function: a function that links the mean to the linear predictor

GLMs

Random Component

$$Y_i \sim Normal(\mu_i, \sigma^2)$$

$$\mu_i = \beta_o + \beta_1 x_i$$

Link Function: The identity link ($g(\mu_i) = \mu_i$)

Systematic Component

GLMs

We have independent random variables $Y_i, i = 1, \ldots, n$

The linear predictor can be written as $\boldsymbol{\eta} = \boldsymbol{X\beta}$ where $\boldsymbol{X}$ is the $n \times (p+1)$ design (model) matrix and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ is the vector of model coefficients

The link function $g(.)$ relates the mean μ_i to η_i, i.e. $g(\mu_i) = \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$

Note:

$$\boldsymbol{X} = \mathbb{X}$$
$$\boldsymbol{\beta} = \vec{\beta} = \beta$$

Coding Demo

Exponential Family of Distributions

- GLMs are based on distributions from the exponential family:
  - Normal
  - Binomial
  - Poisson
- Each has a specific mean-variance relationship.
- Binary response → Binomial (e.g., species presence)
- Count data → Poisson (e.g., number of individuals)

Link Functions

- A link function connects the mean of the response to the linear predictor.
  - Identity: $g(\mu) = \mu$
  - Logit: $g(\mu) = \log(\frac{\mu}{1-\mu})$
  - Log: $g(\mu) = \log(\mu)$

Choosing the Right Model

- Pick a GLM based on:
  - Nature of response variable
  - Shape of data distribution
  - Mean-variance relationship
- Plot the data first!

Example: Presence/Absence Data

$$Presence \sim Elevation + Habitat$$

- Binary response → Binomial family
- Logit link: log-odds of presence

R code:

```
glm(Presence ~ Elevation + Habitat, family = binomial,
data = data)
```

Example: Count Data

$$Abundance \sim Rainfall + Temperature$$

- Count response → Poisson family
- Log link: log of expected count

R code:

```
glm(Abundance ~ Rainfall + Temperature, family = poisson, data = data)
```

Interpreting GLM Coefficients

- Interpretation depends on the link:
  - If we use a logit link, we interpret based on a change in the log-odds (odds ratios on the response scale).
  - If we use a log link, we interpret based on a multiplicative effect (rate ratios on the response scale).
  - If we use the identity link, we interpret based on an additive effect (direct change in the mean response).

GLMs in R

Basic syntax:

```
glm(response ~ predictors, family = ..., data = ...)
```

Common families:

- `binomial(link = 'logit')`
- `poisson(link = 'log')`
- `gaussian(link = 'identity')`
- `Gamma(link='log')`

Coding Demo