# A gentle introduction to generalized linear models

## Analysis of Ecological and Environmental Data

## QERM 514

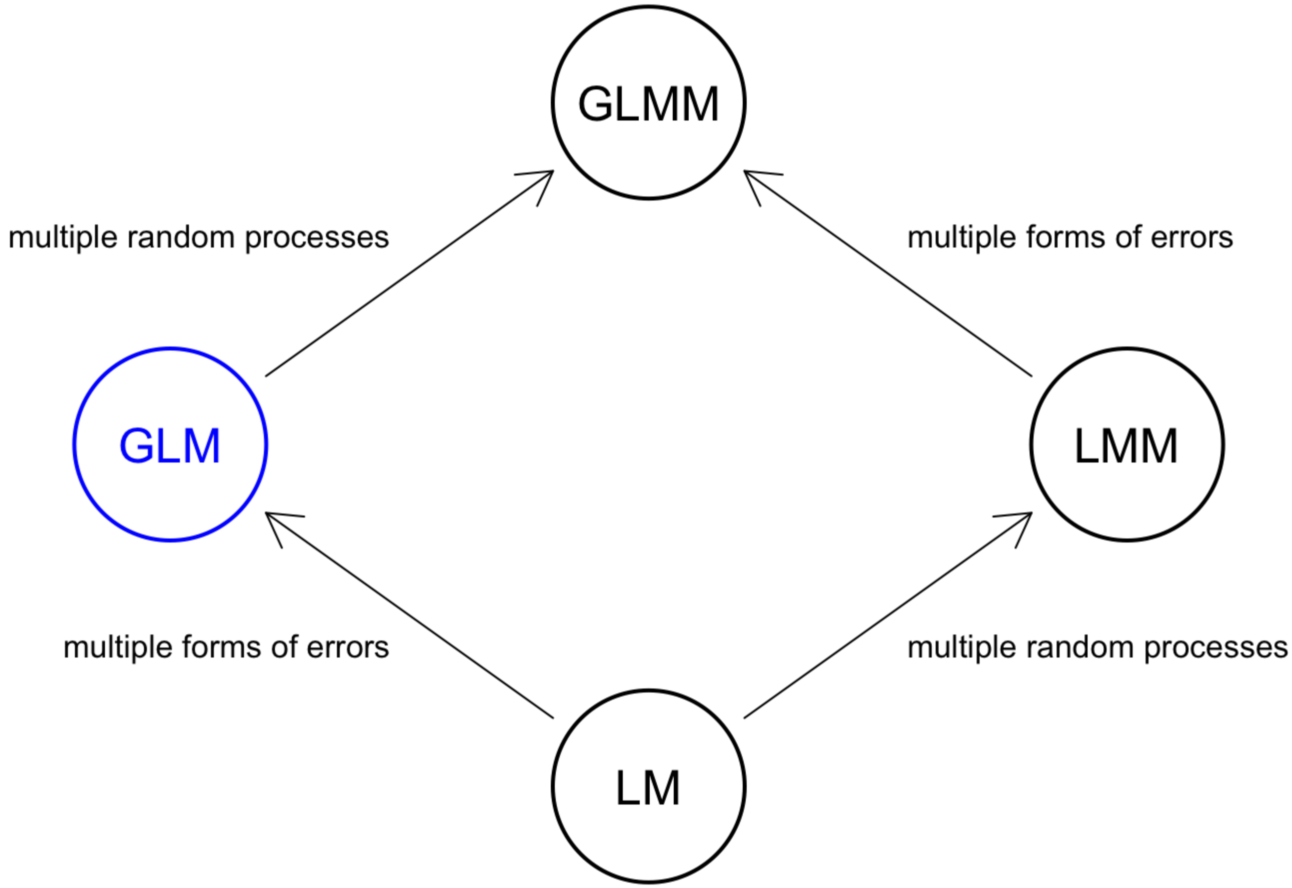
Mark Scheuerell

8 May 2020

# Goals for today

- Understand the 3 elements of a generalized linear model
- Understand how to identify the proper distribution for a generalized linear model
- Understand the concept of a link function

# Forms of linear models



# Ecological data

## At the individual level

1 Detection → presence/absence

2+ Detections → survival, movement

# Ecological data

At the individual level

1 Detection → presence/absence

2+ Detections → survival, movement

1 Measurement → fecundity, age, size

2+ Measurements → growth

# Ecological data

At the population level

Detections → presence/absence

Counts → density or survival/movement

# Data types

Discrete values

Sex

Age

Fecundity

Counts/Census

Survival (individual)

# Discrete data

Given the prevalence of discrete data in ecology (and elsewhere), we seek a means for modeling them

# Generalized linear models (GLMs)

GLMs were developed by Nelder & Wedderburn in the 1970s

They include (as special cases):

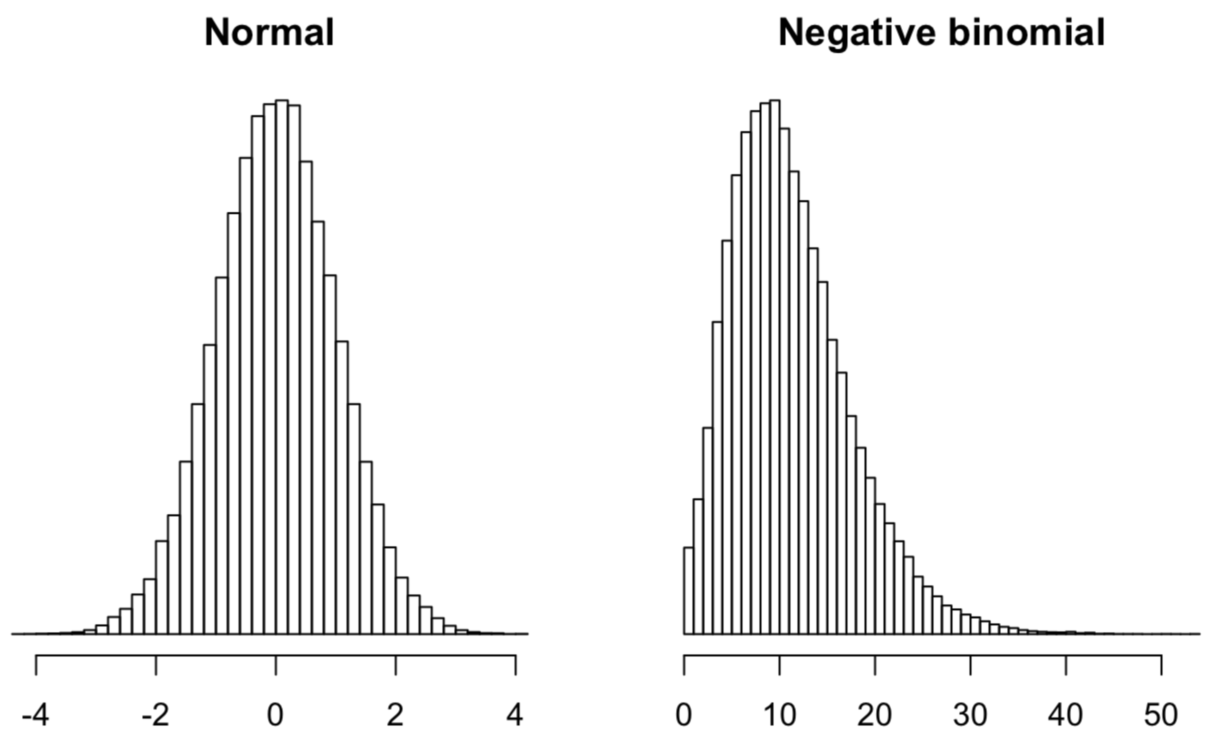
- linear regression
- ANOVA
- logit models
- log-linear models
- multinomial models

# Generalized linear models (GLMs)

In particular, GLMs can explicitly model discrete data as outcomes

# A very important question

What is the distributional form of the random process(es) in my data?



# Distribution for discrete counts

The Poisson distribution is perhaps the best known

It gives the probability of a given number of events occurring in a fixed interval of time or space

# Poisson distribution

## Examples

- the number of Prussian soldiers killed by horse kicks per year from 1868 - 1931
- the number of new COVID-19 infections per day in the US
- the number of email messages I receive per week from students in QERM 514

# Poisson distribution

It's unique in that it has one parameter $\lambda$ to describe both the mean *and* variance

$$y_i \sim \text{Poisson}(\lambda)$$

$$\text{Mean}(y) = \text{Var}(y) = \lambda$$

# Distribution for the ratio of counts

Ratios (fractions) are also very important in ecology

They convey proportions such as

- survivors / tagged individuals
- infected / susceptible
- student emails / total emails

# Distribution for the ratio of counts

The simplest ratio has as denominator of 1 & and numerator of either 0 or 1

For an individual, this can represent

- present (1/1) or absent (0/1)
- alive (1/1) or dead (0/1)
- mature (1/1) or immature (0/1)

# Bernoulli distribution

The Bernoulli distribution describes the probability of a single “event” $y_i$ occurring

$$y_i \sim \text{Bernoulli}(p)$$

where

$$\text{Mean}(y) = p \quad \text{Var}(y) = p(1 - p)$$

# Binomial distribution

The binomial distribution is closely related to the Bernoulli

It describes the number of $k$ “successes” in a sequence of $n$ independent Bernoulli “trials”

For example, the number of heads in 4 coin tosses

# Binomial distribution

For a population, these could be

- $k$ survivors out of $n$ tagged individuals
- $k$ infected individuals out of $n$ susceptible individuals
- $k$ counts of allele A in $n$ total chromosomes

# Generalized linear models (GLMs)

*Three important components*

1. Distribution of the data

Are they counts, proportions?

# Generalized linear models (GLMs)

*Three important components*

1. Distribution of the data

2. Link function $g$

Specifies the relationship between the linear predictor $\eta = \mathbf{X}\boldsymbol{\beta}$ and the mean $\mu$ of the distribution

$$g(\mu) = \eta$$

# Generalized linear models (GLMs)

*Three important components*

1. Distribution of the data

2. Link function $g$

3. Linear predictor $\eta$

$$\eta = \mathbf{X}\boldsymbol{\beta}$$

# Common link functions

| Distribution | Link function | Mean function |
| --- | --- | --- |
| Identity | $1(\mu) = \mathbf{X}\boldsymbol{\beta}$ | $\mu = \mathbf{X}\boldsymbol{\beta}$ |
| Log | $\log(\mu) = \mathbf{X}\boldsymbol{\beta}$ | $\mu = \exp(\mathbf{X}\boldsymbol{\beta})$ |
| Logit | $\log\left(\frac{\mu}{1-\mu}\right) = \mathbf{X}\boldsymbol{\beta}$ | $\mu = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})}$ |

# Canonical links

Where did we find these link functions?

For the exponential family of distributions (eg, Normal, Gamma, Poisson) we can write out the distribution of $y$ as

$$f(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} - c(y,\phi)\right)$$

$\theta$ is the *conanical* parameter of interest

$\phi$ is a scale (variance) parameter

# Exponential family

$$f(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} - c(y,\phi)\right)$$

We seek some *canonical* function $g$ that connects $\eta$, $\mu$, and $\theta$ such that

$$g(\mu) = \eta$$
$$\eta \equiv \theta$$

# Normal distribution

$$f(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} - c(y,\phi)\right)$$

$$\Downarrow$$

$$f(y;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}}\exp\left(\frac{(y-\mu)^2}{2\sigma^2}\right)$$

with $\theta = \mu$ and $\phi = \sigma^2$

$a(\phi) = \phi \qquad b(\theta) = \frac{\theta^2}{2} \qquad c(y,\phi) = -\frac{\frac{y^2}{\phi} + \log(2\pi\phi)}{2}$

# Normal distribution

$$f(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} - c(y,\phi)\right)$$

$$\Downarrow$$

$$f(y;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}}\exp\left(\frac{(y-\mu)^2}{2\sigma^2}\right)$$

with $\theta = 1(\mu)$ and $\phi = \sigma^2$

$a(\phi) = \phi \qquad b(\theta) = \frac{\theta^2}{2} \qquad c(y,\phi) = -\frac{\frac{y^2}{\phi} + \log(2\pi\phi)}{2}$

# Poisson distribution

$$f(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} - c(y,\phi)\right)$$

$$\Downarrow$$

$$f(y;\mu) = \frac{\exp(-\mu)\mu^y}{y!}$$

with $\theta = \log(\mu)$ and $\phi = 1$

$a(\phi) = 1 \qquad b(\theta) = \exp(\theta) \qquad c(y,\phi) = -\log(y!)$

# Binomial distribution

For the binomial distribution there are several possible link functions

- logit
- probit
- complimentary log-log

# Generalized linear models (GLMs)

The word *generalized* means these models are broadly applicable

For example, GLMs include linear regression models

# Writing an LM as a GLM

$$y_i = \alpha + \beta x_i + \epsilon_i$$
$$\epsilon_i \sim \mathrm{N}(0, \sigma^2)$$

# Writing an LM as a GLM

$$y_i = \alpha + \beta x_i + \epsilon_i$$
$$\epsilon_i \sim \mathrm{N}(0, \sigma^2)$$
$$\Downarrow$$
$$y_i = \mu_i + \epsilon_i$$
$$\mu_i = \alpha + \beta x_i$$
$$\epsilon_i \sim \mathrm{N}(0, \sigma^2)$$

# Writing an LM as a GLM

$$
\begin{aligned}
y_i &= \mu_i + \epsilon_i \\
\mu_i &= \alpha + \beta x_i \\
\epsilon_i &\sim \mathrm{N}(0, \sigma^2) \\
&\Downarrow \\
y_i &= \epsilon_i \\
\mu_i &= \alpha + \beta x_i \\
\epsilon_i &\sim \mathrm{N}(\mu_i, \sigma^2)
\end{aligned}
$$

# Writing an LM as a GLM

$$
\begin{aligned}
y_i &= \epsilon_i \\
\mu_i &= \alpha + \beta x_i \\
\epsilon_i &\sim \mathrm{N}(\mu_i, \sigma^2) \\
&\Downarrow \\
y_i &\sim \mathrm{N}(\mu_i, \sigma^2) \\
\mu_i &= \alpha + \beta x_i
\end{aligned}
$$

# Writing an LM as a GLM

$$y_i \sim \mathrm{N}(\mu_i, \sigma^2)$$
$$\mu_i = \alpha + \beta x_i$$
$$\Downarrow$$
$$y_i \sim \mathrm{N}(\mu_i, \sigma^2)$$
$$1(\mu_i) = \mu_i$$
$$\mu_i = \alpha + \beta x_i$$

# Writing an LM as a GLM

$$y_i \sim \mathrm{N}(\mu_i, \sigma^2)$$
$$1(\mu_i) = \mu_i$$
$$\mu_i = \alpha + \beta x_i$$
$$\Downarrow$$

$$\text{data distribution: } \; y_i \sim \mathrm{N}(\mu_i, \sigma^2)$$

$$\text{link function: } \; 1(\mu_i) = \mu_i$$

$$\text{linear predictor: } \; \mu_i = \alpha + \beta x_i$$

# Example of a GLM

Log-density of live trees per unit area $y_i$ as a function of fire intensity $F_i$

$$\text{data distribution: } \; y_i \sim \mathrm{N}(\mu_i, \sigma^2)$$

$$\text{link function: } \; 1(\mu_i) = \mu_i$$

$$\text{linear predictor: } \; \mu_i = \alpha + \beta F_i$$

# Rethinking density

We have been considering (log) density itself as a response

$$\text{Density}_i = f(\text{Count}_i, \text{Area}_i)$$
$$\Downarrow$$
$$\text{Density}_i = \frac{\text{Count}_i}{\text{Area}_i}$$

# Rethinking density

We have been considering (log) density itself as a response

$$\text{Density}_i = f(\text{Count}_i, \text{Area}_i)$$
$$\Downarrow$$
$$\text{Density}_i = \frac{\text{Count}_i}{\text{Area}_i}$$

With GLMs, we can shift our focus to

$$\text{Count}_i = f(\text{Area}_i)$$

# Example of a GLM

Counts of live trees $y_i$ as a function of area surveyed $A_i$ and fire intensity $F_i$

$$
\begin{aligned}
\text{data distribution:} \quad & y_i \sim \text{Poisson}(\lambda_i) \\
\text{link function:} \quad & \log(\lambda_i) = \mu_i \\
\text{linear predictor:} \quad & \mu_i = \alpha + \beta_1 A_i + \beta_2 F_i
\end{aligned}
$$

# Example of a GLM

Probability of spotting a sparrow $p_i$ as a function of vegetation height $H_i$

$$
\begin{aligned}
\text{data distribution:} \quad & y_i \sim \text{Bernoulli}(p_i) \\
\text{link function:} \quad & \text{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = \mu_i \\
\text{linear predictor:} \quad & \mu_i = \alpha + \beta H_i
\end{aligned}
$$

# Example of a GLM

Survival of salmon from parr to smolt $s_i$ as a function of water temperature $T_i$

$$\text{data distribution: } y_i \sim \text{Binomial}(N_i, s_i)$$

$$\text{link function: } \text{logit}(s_i) = \log\left(\frac{s_i}{1 - s_i}\right) = \mu_i$$

$$\text{linear predictor: } \mu_i = \alpha + \beta T_i$$

# Summary

There are three important components to GLMs

1. Distribution of the data
2. Link function $g$
3. Linear predictor $\eta$