Introduction to Generalised Linear Models

Dr Niamh Mimnagh

niamh@prstats.org

https://github.com/niamhmimnagh

PR STATS

Why do We Build Models?

- Statistical models are tools for understanding patterns in data.
- They help us answer questions such as:
  - Does habitat quality affect bird breeding success?
  - Does temperature influence flowering time?
  - Does disease risk change with herd size?
- Models can serve different purposes: they may be explanatory, aiming to understand causal relationships, or predictive, aiming to forecast new observations.
- Regardless of the purpose, the model provides a structured way to link observed data to underlying processes.

Simple Linear Models

- A standard linear regression model relates a response variable Y to one or more predictors X through a straight-line relationship:

$$Y_i = \beta_0 + \beta_1 x_i + \cdots + \beta_p x_p + \epsilon_i$$

- where ϵ_i is a normally distributed error term with mean zero and constant variance.
- Each β coefficient represents the expected mean change in y for a 1-unit increase in its associated predictor
- This model works well when the response is continuous, roughly normally distributed, and has constant variability.
- However, many ecological and biological datasets do not meet these assumptions. This is especially true when dealing with binary, count, or skewed data.

Simple Linear Models

We have seen that a simple linear model may be written:

$$Y_i = \beta_0 + \beta_1 x_i + \cdots + \beta_p x_p + \epsilon_i$$

$$\text{with } \epsilon_i \sim N(0, \sigma^2)$$

We can show from the equation above that the expected value of Y_i is:

$$E[Y_i] = \beta_0 + \beta_1 x_i + \cdots + \beta_p x_p$$

And the variance is:

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

We can write:

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

$$\mu_i = \beta_o + \beta_i x_i$$

When Linear Regression Fails

- Consider trying to model species presence (yes/no) using linear regression.
- Predicted values can fall outside the 0 - 1 range, which is nonsensical for probabilities.
- Similarly, when modelling count data (e.g. the number of nests per site), the normality assumption fails because counts are non-negative integers and often have variance that depends on the mean.
- Using a linear model on such data can lead to biased estimates, poor predictions, and misleading inference.

Generalised Linear Models

- Generalised Linear Models (GLMs) extend ordinary linear regression to handle a broader range of response types.
- They replace the normal distribution with a more appropriate family, such as the binomial for binary outcomes or the Poisson for counts.
- They also introduce a link function that connects the predictors to the mean of the response in a way that respects its natural constraints (e.g., probabilities between 0 and 1).
- This flexible framework makes GLMs essential for analysing ecological data, epidemiological outcomes, and many other applied problems.

GLM Components

- Every GLM has three key components:
  1. Random component - specifies the probability distribution of the response (e.g. binomial, Poisson, normal).
  2. Systematic component - the linear predictor, a weighted combination of covariates ($\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$).
  3. Link function - a mathematical function that relates the expected response μ to the linear predictor η.
- Together, these components allow us to model non-normal data while retaining the interpretability and structure of linear models.

The Linear Model as a GLM

Random Component

$$Y_i \sim N(\mu_i, \sigma^2)$$
$$\mu_i = \beta_o + \beta_i x_i$$

Systematic Component

What about the link function?

The linear model uses an identity link function.

The identity link just means that the expected response μ_i is equal to the linear predictor. We don't require a transformation between the mean of the response and the linear predictor. This is because, for normally distributed data, the response can take on any real value, so we don't need a link function to constrain it.

Example

- We want to examine the relationship between mean annual plant biomass in grassland plots and average annual rainfall.
- - Biomass measured in g/m² across 50 plots.
- - Rainfall measured in mm/year from nearby weather stations.
- - We expect higher rainfall to lead to higher biomass, but with natural variation due to other factors (soil type, species mix, grazing pressure).
- - Response variable (biomass) is continuous and assumed normally distributed.

Example

- Model structure in GLM terms:
  - Random component: $Y_i \sim N(\mu_i, \sigma^2)$
  - Systematic component: $\eta_i = \beta_o + \beta_1 \text{Rainfall}_i$
  - Link function: $\mu_i = \eta_i$ (identity link)
- Why use GLM instead of just LM?
  - The Gaussian family with identity link is a GLM — showing that the normal model fits naturally into the GLM framework.
  - This lets us use the same GLM machinery for more complex families later (Poisson, binomial, etc.).
- Goal:
  - Fit and interpret the model using both lm() and glm() in R, and confirm they give identical results.

Coding Demo

Modelling Different Data Types

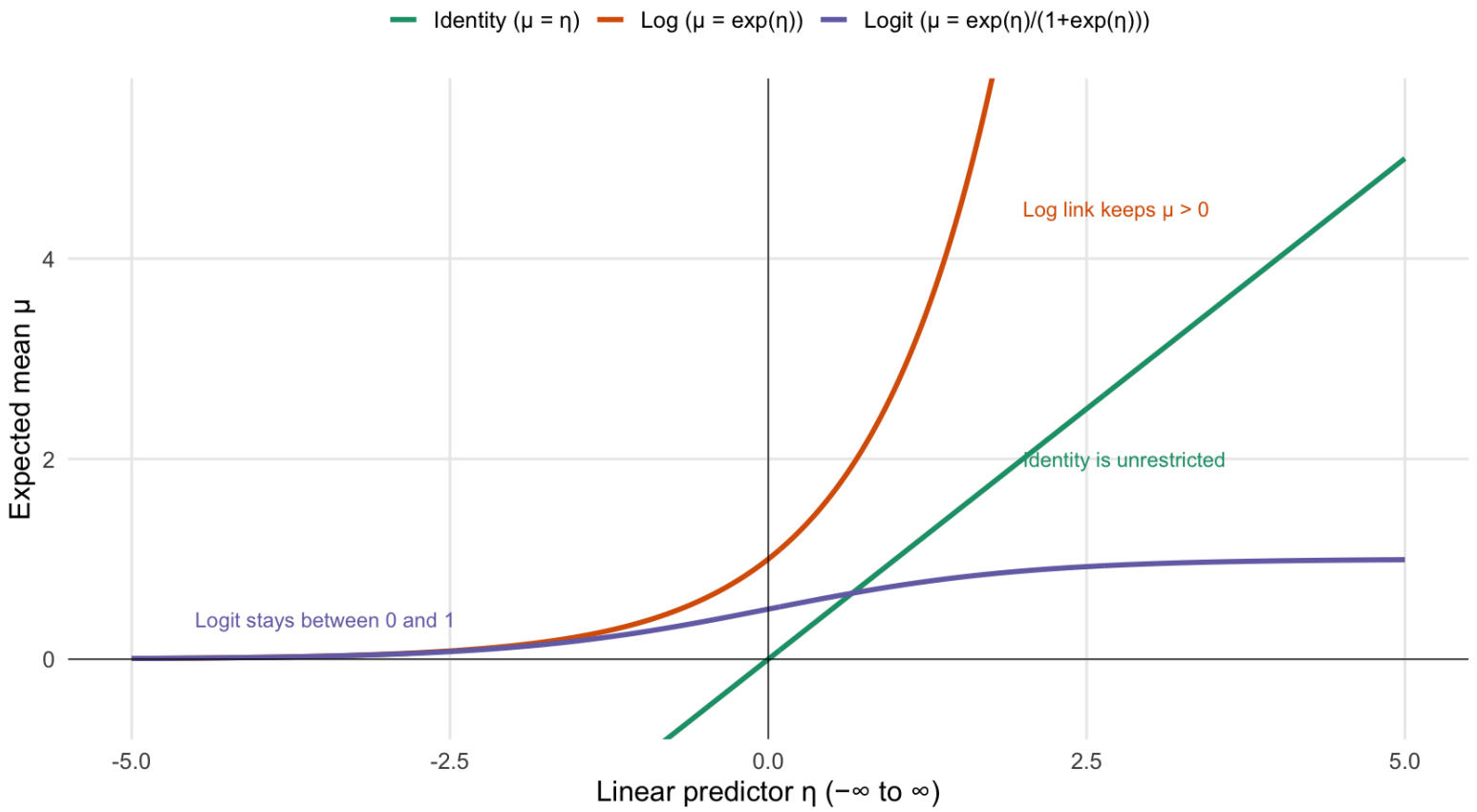
- Different response types arise naturally in ecological studies. GLMs unify these cases under one common framework. The choice of distribution and link function reflect the type of response you have.

Response	Example	Distribution	Link
Continuous Data	Body mass	Normal	Identity
Binary Data	Presence/absence	Binomial	Logit/probit
Counts	Number of individuals	Poisson	Log

Link Functions

- The link function defines how μ relates to the linear predictor, η which can take any real value, from $-\infty$ to $+\infty$. But the expected response μ is often constrained by the nature of the data. For example, it might need to stay positive or be bounded between 0 and 1.
- The link function bridges this gap by transforming μ to a scale where a linear model makes sense:
  - The identity link leaves the relationship unchanged, $\mu = \eta$, which works when μ itself can take any real value.
  - The log link ensures fitted values are always positive after back-transformation.
  - The logit link maps probabilities $(0 - 1)$ onto the real line, so they can be modelled with a linear predictor.
- In essence, the link function allows us to use a simple linear structure for η while ensuring that the modelled mean μ respects the natural scale and constraints of the response variable.

Link Functions



Importance of GLMs

- Binary and/or count outcomes are common in ecological data
  - Predicting species occurrence from environmental covariates.
  - Modelling animal counts in surveys or camera traps.
  - Analysing disease prevalence in wildlife populations.
- If we use the wrong model (e.g. a linear regression for binary data), the standard errors and p-values will often be incorrect.
- This can lead to overconfident or misleading conclusions.
- GLMs, by using the correct distributional assumptions, provide valid inference for a wider range of data types.

When the Outcome is Binary

Many real-world problems involve binary (yes/no, success/failure) outcomes:

- Did a patient survive? (yes/no)
- Was the animal infected? (yes/no)
- Did the student pass the course? (yes/no)

$$Y_i = \begin{cases} 1, if\ success \\ 0, if\ failure \end{cases}$$

What is 'success'?

It's whatever you want it to be! It's not necessarily the 'best' outcome.

When the Outcome is Binary

When data is binary, there are only two possible outcomes, and so when we talk about the probability of each outcome, we have:

$$P(success) = P(Y_i = 1) = \pi_i$$
$$P(failure) = P(Y_i = 0) = 1 - \pi_i$$

Y_i has a Bernoulli distribution:

$$Y_i \sim Bernoulli(\pi_i)$$

Why Not Use Linear Regression?

Linear regression assumes:

- The response variable is continuous and unbounded.
- The relationship between predictors and the response is linear.

Problems with using it on binary data:

- Predictions can fall outside [0,1]
- Error terms are heteroscedastic (non-constant variance)
- Residuals are not normally distributed

This leads to poor model performance and invalid inference.

The Bernoulli GLM

We want to model the success probabilities π_i as a function of predictors.

Can we simply write $\pi_i = \beta_0 + \beta_1 x_{1i} + \ \ldots + \beta_p x_{pi}$?

No!

Since the β coefficients are unbounded (they can take any real value from $-\infty$ to ∞), this would result in unbounded π_i values

But π_i are probabilities, and so have to be bounded in the $(0,1)$ interval

So we need a link function that maps the $(0,1)$ interval to the real line.

The Bernoulli GLM

$$Y_i \sim Bernoulli(\pi_i)$$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{1i} + \ \ldots + \beta_p x_{pi}$$

The function $\log\left(\frac{\pi_i}{1-\pi_i}\right)$ can also be written as $logit(\pi_i)$

'logit' stems from the words **log**istic un**it**, since its based on the cumulative distribution function of the logistic distribution

It is simply the natural logarithm of the odds

This is how we ensure predicted probabilities stay between 0 and 1.

The Logistic Function

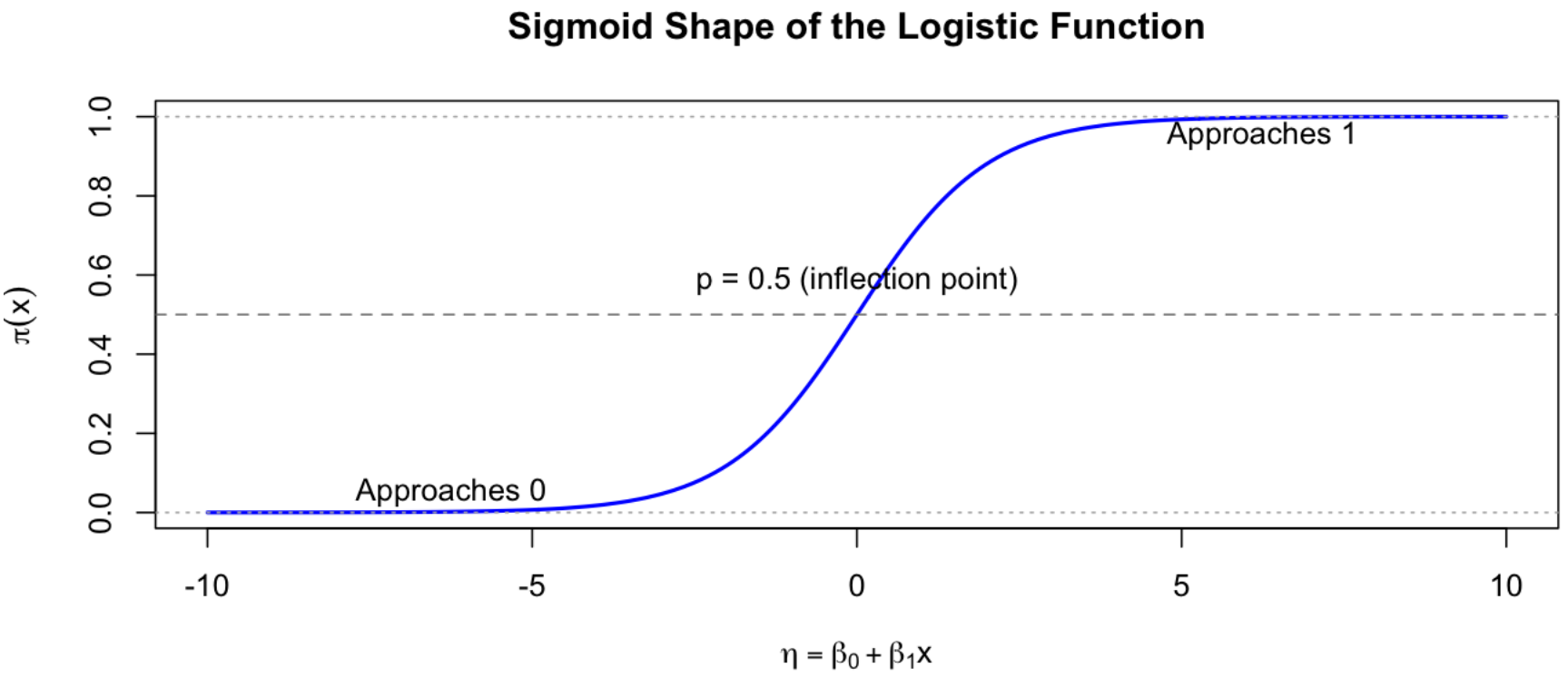
Rewriting the logit model:

$$\pi = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

This is a sigmoid-shaped curve. The logistic curve:

- Approaches 0 and 1 asymptotically.
- Has an inflection point at p = 0.5.
- Is nonlinear in probability space but linear in log-odds space.

The Logistic Function



Odds, Log-Odds and Probability

- Probability (π) is the chance of an event occurring (range: 0 to 1) e.g., π = 0.8 means an 80% chance of success
- Odds is the ratio of probability of success to probability of failure

$$Odds = \frac{\pi}{1-\pi}$$

e.g. if $\pi = 0.8$, then the $odds = \frac{0.8}{0.2} = 4$

- The log of the odds, called the logit, stretches the $0-1$ probability range into the full real number line:

$$logit(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$$

- This transformation ensures that we can model log-odds as a straight line in the predictors.

Log Odds: Why Use Them?

- Logistic regression models the log of the odds:

$$logit(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots \beta_p x_{pi}$$

- The coefficients β are on the log-odds scale
- Each β_j represents the change in log-odds for a 1-unit change in x_j.
- β_0 :log-odds of the outcome when $x = 0$
- β_1 :change in log-odds for a one-unit increase in x
- By exponentiating a coefficient, we obtain an odds ratio:
  - An odds ratio > 1 means the predictor increases the odds of the event.
  - An odds ratio < 1 means the predictor decreases the odds of the event.
- For example, if the coefficient for vegetation cover is 0.5, then each unit increase multiplies the odds of species presence by exp(0.5) ≈ 1.65.

Example: Interpreting Odds

- If the predicted probability of species presence at a site is 0.8, the odds of presence are $\frac{0.8}{0.2} = 4$.
- This means the species is four times more likely to be present than absent.
- If instead the odds were 8, the probability is now 0.89 - showing that increases in odds correspond to smaller absolute changes in probability near the extremes.

Probability π from η:

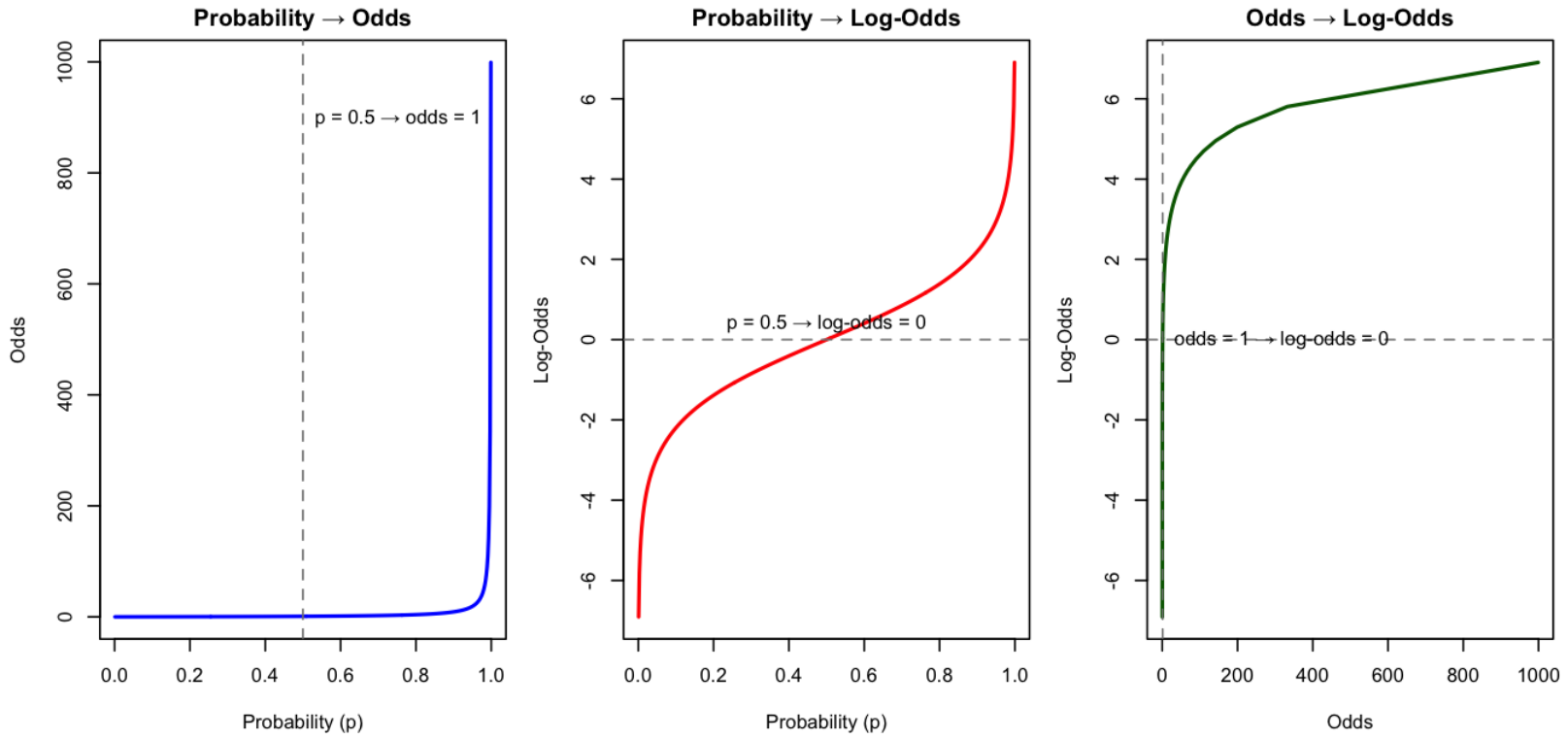
$$\pi = \frac{e^{\eta}}{1 + e^{\eta}}$$

If $\eta = 1.2$, $\pi = \frac{3.32}{1+3.32} \approx 0.77$

The Logistic Curve

- The logistic curve starts near 0 for very negative predictor values, transitions smoothly through 0.5 at the midpoint, and approaches 1 for very large predictor values.
- This S-shape reflects the biological reality that probabilities cannot exceed the natural bounds of 0 and 1.
- It also models diminishing returns - extreme changes in predictors have less effect when probabilities are already near 0 or 1.

The Logistic Curve



Example: Disease Presence

Suppose we model whether an animal has a certain disease (1 = yes, 0 = no) based on age:

$$Y_i \sim Bernoulli(\pi_i)$$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 Age_i$$

β_0: log-odds of infection at age 0

β_1: Change in log-odds of success for each additional year.

Exponentiate the coefficient: e^{β_1} to give the multiplicative effect of an increase of 1 year of age on disease.

Coding Demo

Extending Logistic Regression

- Logistic regression can easily include multiple predictors, interactions, and categorical variables.
- For example:

```
glm(presence ~ veg_cover * predator_abundance, family =
binomial, data = wetlands)
```

- This models the interaction between vegetation cover and predator abundance, allowing us to see if the vegetation effect changes depending on predator pressure.

Count Data

- Count data are everywhere in ecological research.
- Examples include:
  - Number of nests per colony
  - Number of individuals observed in a transect
  - Number of infections in a herd
  - Number of flowers per plant
- Count data are always non-negative integers, and often the variability increases with the mean.
- This makes them poorly suited for standard linear regression.

The Poisson Distribution

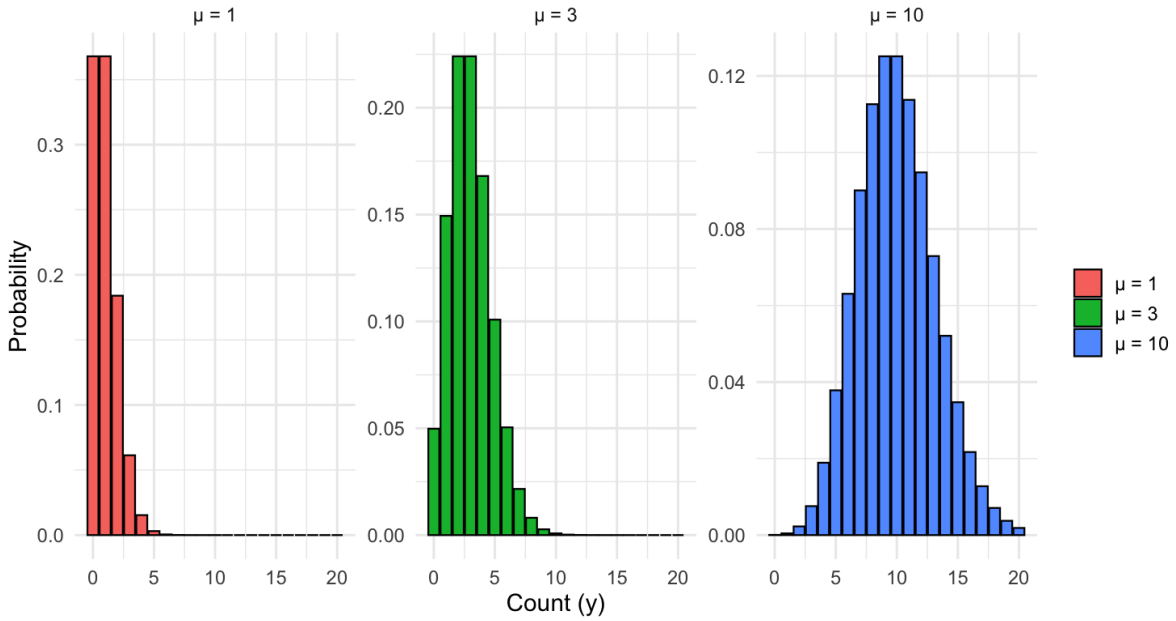
- The Poisson distribution is commonly used to model counts.
- It has a single parameter, the mean μ, which is also the variance:

$$P(Y = y) = \frac{e^{-\mu}\mu^{y}}{y!}$$

- This property - mean equals variance - is called equidispersion.
- While real data sometimes deviate from this, the Poisson is a natural first choice for counts.

Visualising Poisson Distributions

- The Poisson distribution is controlled by a single parameter λ.
- As μ increases, both the mean and variance increase.



Poisson Regression as a GLM

- In a Poisson GLM, the expected count μ_i is linked to predictors via the log link:

$$\log(\mu_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots$$

- Equivalently:

$$\mu_i = e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots}$$

- This guarantees that the predicted counts are always positive.
- Linear predictors η can take any real value, including negatives.
- But counts must be non-negative.
- The log link transforms η so $\mu = e^{\eta} > 0$, ensuring valid predictions.

The Poisson Model

- For each observation i, the response follows a Poisson distribution:

$$Y_i \mid x_i \sim Poisson(\mu_i),$$

where μ_i is the expected count for that observation.

- The log link relates μ_i to the predictors:

$$\log(\mu_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots$$

- So the full model is:

$$Y_i \sim Poisson(e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots})$$

Interpreting Poisson Regression Coefficients

- A coefficient β_j in a Poisson regression represents the log change in the expected count for a one-unit change in the predictor.
- Exponentiating the coefficient gives an incident rate ratio:
  - If $e^{\beta} > 1$, the predictor increases the expected count.
  - If $e^{\beta} < 1$, the predictor decreases the expected count.
- For example, if $\beta = 0.3$, then each unit increase multiplies the expected count by $e^{0.3} \approx 1.35$ (a 35% increase).

Example: Nest Counts

- How does vegetation density at a site affect the number of bird nests found there?

- **Response variable:** Number of bird nests at each site.
- These are count data: non-negative integers (0, 1, 2, …).
- The variance often increases with the mean for count data.
- **Predictor variable:** Vegetation density (measured as, say, % cover, or biomass, or vegetation index).
- Continuous, positive values.
- Ecological reasoning: denser vegetation may provide more nesting sites and better cover from predators.

Coding Demo

Comparing GLM Families

- At this point, we've seen three important GLM families:
  - Binomial for binary/proportion outcomes
  - Poisson for count outcomes
  - Normal for continuous outcomes
- All share the same structure:

1. A random component (distribution)
2. A systematic linear predictor
3. A link function connecting them

Choosing the Right GLM

- To choose the right GLM:

1. Identify the response type (binary, count, continuous).
2. Look at the distributional properties (variance structure, skewness).
3. Select the appropriate family and link function.

- This decision is guided by both the data's natural constraints and your scientific understanding of the process.

Beyond Basic GLMs

- GLMs are flexible but can be extended further to handle:
  - Overdispersion with quasi-models or negative binomial regression
  - Zero-inflation for count data with excess zeros
  - Mixed-effects models for hierarchical or grouped data
  - Generalised Additive Models (GAMs) for non-linear effects
- These extensions build on the same GLM principles.

A Practical GLM Workflow

- When working with GLMs, the process generally follows these steps:

1. Understand your data - type of response, predictors, data structure.
2. Choose the appropriate GLM family and link function.
3. Fit the model using glm() or an equivalent tool.
4. Check the model fit and diagnostics.
5. Interpret coefficients and predictions.
6. Communicate results in a clear, meaningful way.

- This workflow helps ensure robust and interpretable results.

Example

- Let's consider a small example dataset with species presence/absence (1/0), nest counts, vegetation cover (%), distance to water (m), and predator activity index.
- Question A (occurrence): Which site features influence species presence vs absence ?
- Question B (abundance): At sites where it occurs, how do habitat features affect nest counts?
- Presence and abundance answer different questions
- Species may occur at a site but with low nesting numbers
- Two-stage workflow: 1) Predict occurrence, 2) Predict abundance

Example

Logistic regression (presence)

- Binary response: presence/absence

$$\text{logit}(\pi) = \beta_0 + \beta_1 \text{veg} + \beta_2 \text{water} + \beta_3 \text{predators}$$

- Coefficients: change in log-odds of presence
- Report: odds ratios, partial effects, ROC/AUC

Poisson regression (abundance)

- Count response: number of nests

$$\log(\mu) = \gamma_0 + \gamma_1 \text{ veg} + \gamma_2 \text{water} + \gamma_3 \text{predators}$$

- Coefficients: multiplicative change in expected nests
- Report: incidence rate ratios (IRRs)

GLM Workflow

1. Explore the data:

- What is the response variable? Is it binary, count, or continuous? How is it distributed? Check histograms or summary stats.
- Exploratory data analysis guides the choice of GLM family and link function.

2. Choose the model family:

- Think carefully about what makes sense for your response:
  - Presence/absence: binomial family with logit link
  - Counts: Poisson family with log link
  - Heights/weights: Normal family with identity link
- Matching the family to the data ensures predictions respect natural constraints and inference remains valid.

3. Fit the model

GLM Workflow

4. Interpret the output
   - Coefficients - on the link scale (log-odds, log-counts).
   - Standard errors & p-values - testing each coefficient.
   - Deviance - a measure of fit, similar to residual sum of squares.
- Remember to exponentiate coefficients for easier interpretation as odds ratios or rate ratios.

5. Make predictions
   - For logistic regression → predicted probabilities (0–1).
   - For Poisson regression → expected counts (positive integers).
- These predictions can then be visualised as curves or effect plots to aid interpretation.

GLM Workflow

6. Check diagnostics

- Even with GLMs, diagnostics matter. Key checks include:
  - Residual plots for systematic patterns.
  - Overdispersion tests for Poisson models.
  - Influence measures like Cook's distance for outliers.
- Diagnostics ensure the model is appropriate and robust.
- Typical problems include:
  - Overdispersion in count models: consider quasi-Poisson or negative binomial.
  - Non-linearity in predictors: consider transformations or splines.
  - Perfect separation in logistic regression: may require penalised methods.
- Recognising these issues is the first step to improving the model.

GLM Workflow

7. Communicate results

- Visualisation is crucial for communicating GLM results. For example:
  - Plot predicted probability curves with confidence intervals for logistic regression.
  - Plot expected counts against a key predictor for Poisson regression.
- These plots let stakeholders see how predictors influence outcomes.
- Always present results in terms that your audience can understand:
  - For logistic regression, explain effects as changes in probability or odds ratios.
  - For Poisson regression, explain effects as multiplicative changes in expected counts.
  - Avoid technical jargon unless your audience is statistically trained.

Example Summary Statement

Instead of saying:

“The coefficient for vegetation cover is 0.45 ($p < 0.01$).”

Say:

“Each additional unit of vegetation cover increases the odds of species presence by 57%, making sites with more vegetation significantly more likely to support the species ($p < 0.01$).”

- Framing results in plain language makes your analysis impactful.

Case Study Results

- For our wetland example:
  - Logistic regression showed vegetation cover strongly increased species presence.
  - Poisson regression showed nest counts increased with vegetation but decreased with distance to water.
- Visualising both models gave clear ecological insights that could inform conservation planning.

How the Full Course Expands This

- In the full course, we'll go beyond the basics by:
  - Examining overdispersed and zero-inflated data
  - Examining how to model proportions
  - Diving deeper into diagnostics, model selection, and validation.
  - Exploring mixed models for hierarchical and temporal data.
  - Practising with complex ecological datasets.
  - Discussing Bayesian approaches for flexibility.

Why Mastering GLMs Matters

- GLMs are the foundation of modern statistical modelling in ecology, epidemiology, and beyond.
- They allow you to correctly analyse non-normal data, make robust predictions, and draw meaningful inferences.
- Once you're comfortable with GLMs, it's much easier to understand more advanced models.

Thank You

- Thank you for joining this GLM mini-course!
- You've seen the core ideas, key GLM families, and a practical workflow.
- In the full course, we'll dive much deeper, with real ecological datasets, more hands-on coding, and more advanced topics like overdispersion, zero-inflation and proportional data.