Introduction to Generalised Linear Models for Ecologists

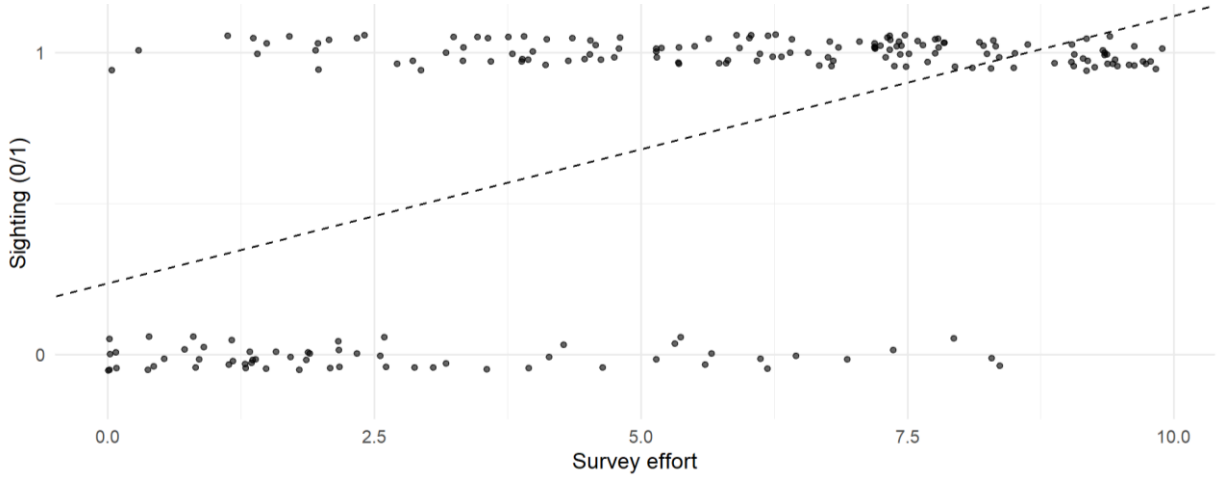
Dr Niamh Mimnagh

niamh@prstats.org

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

PR STATS

Presence/Absence Data



- If we try to fit a straight line to this data, we are asking the line to directly explain the observed successes and failures.
- This is saying: “the response itself (0 or 1) changes linearly with effort.” But this does not make sense: the outcome is discrete, so the line will not pass through the points in a meaningful way.
- What we really care about is not the individual 0s and 1s, but the chance of seeing an animal given the level of effort.

Presence/Absence Data

- Rather than trying to predict the raw binary outcomes directly, we aim to model the probability of a sighting.
- This shifts the focus from “can we exactly predict each 0 or 1” to “can we explain the underlying chance of success.”
- Probabilities are continuous between 0 and 1, which allows us to draw a smooth curve that describes how the chance of success changes with effort.
- A straight line still does not work on the probability scale (because it can go below 0 or above 1), but this gives us the right target: we want to model the probability, not the binary outcome itself.

Presence/Absence Data

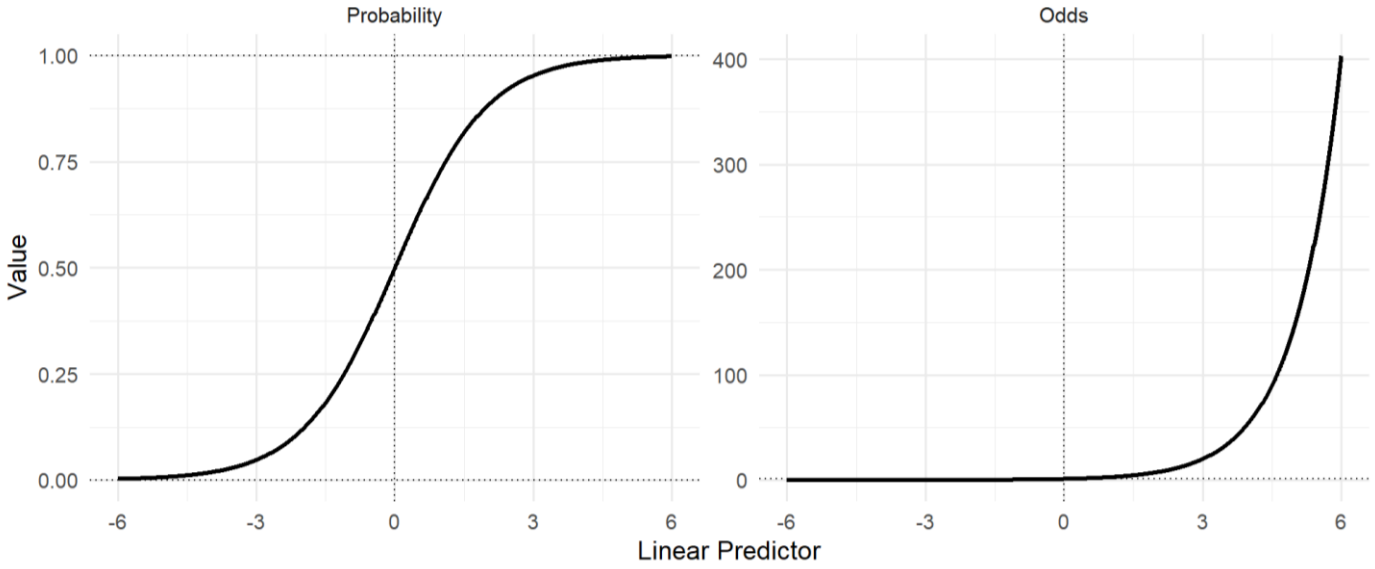
$$Y_i \sim Bernoulli(\pi_i)$$

- π_i is the probability, and it describe the chance of success as a number between 0 and 1.
- Now we want to use our predictor variables to model the probability of success, but the linear predictor $\beta_0 + \beta_1 x$ can take any real value.
- If we directly equate probability with a linear predictor ($\pi_i = \beta_0 + \beta_1 x$), the model could easily predict probabilities less than 0 or greater than 1, which makes no sense.
- We need to transform the probability scale (0 to 1) onto the whole real line, so that the linear predictor can work properly.

Presence/Absence Data

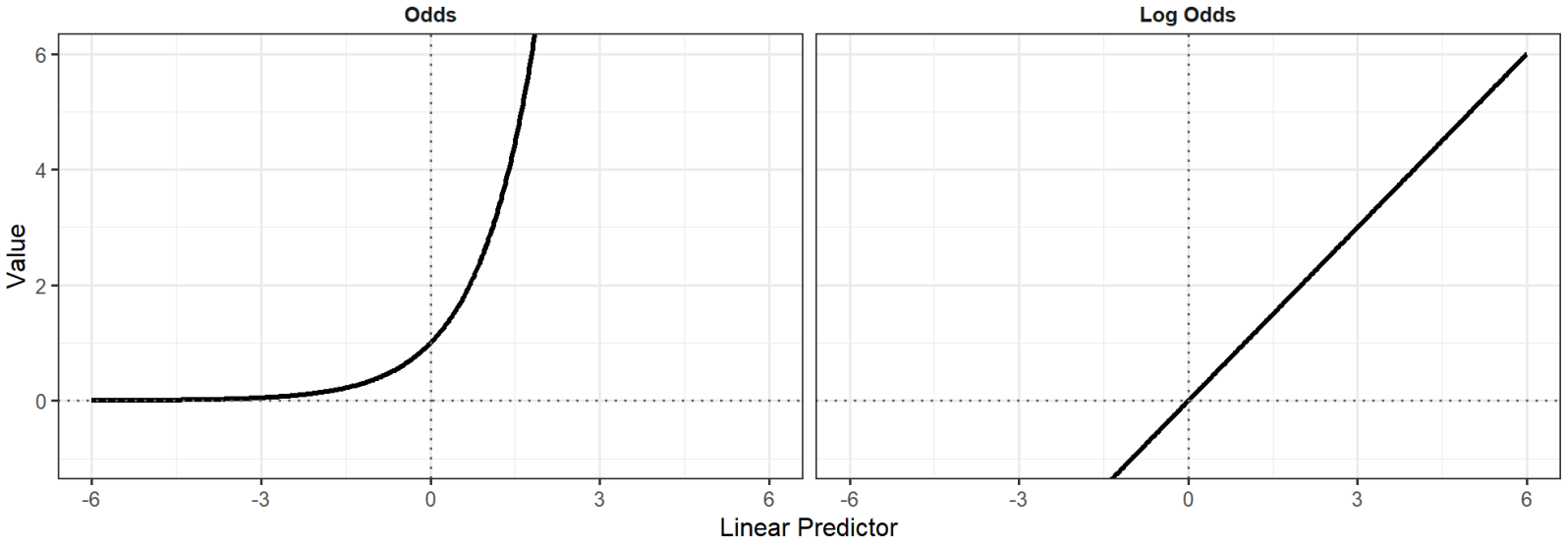
How do we do this?

If instead of modelling probability, we model the odds ($\frac{\pi}{1-\pi}$), this gets us partway there:



Presence/Absence Data

However, the odds still don't allow the linear predictor to be negative so we take the log(odds)



Modelling Species Occupancy

- Lets say we are studying 100 sites, and collecting data on whether we see any pine martens at each site.
- We visit each site once a month, and record a ‘1’ if pine martens are seen, and a ‘0’ otherwise.

Modelling Species Occupancy

- The covariates recorded during this study are habitat (grassland, forest, wetland), elevation (m) and effort (minutes spent at each site)
- **Habitat**: ecological suitability (cover, prey, den sites): affects both occupancy and detectability.
- **Elevation**: harsher climate/poorer resources at high elevation: lower occupancy and slightly harder detection.
- **Effort**: longer searches / more time for encounters: higher detection given presence. Effort affects detectability, not presence.

Modelling Species Occupancy

- Treat Y_i as the response:

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

$$logit(\pi_i) = \beta_0 + \beta_1\, elevation_i + \beta_2\, habitat_i + \beta_3 effort_i\ .$$

- e^{β_0} is the odds of detectability at the baseline predictor values
- If predictors are not centred, the baseline predictor values are elevation = 0m, effort = 0 minutes, habitat = grassland (reference)
- So β_0 is the log-odds for a grassland site at sea level with 0 survey minutes (often meaningless/extrapolative if 0 isn't in your data – e.g. 0 minutes of effort)

Modelling Species Occupancy

- If predictors are centred, then

$$Y_i \sim Bernoulli(\pi_i)$$
$$logit(\pi_i) = \beta_0 + \beta_1\, elevation[c]_i + \beta_2\, habitat_i + \beta_3 effort[c]_i\ .$$

- Now β_0 is the log-odds at the average continuous covariates and the reference habitat, and e^{β_0} is the baseline odds for a typical site.
- If you only centre elevation and not effort, then β_0 is the log-odds at mean elevation but effort of 0.

Choosing the Reference

- **Most common / status-quo category:** The intercept can describe the situation you see most in your data. If most sites are grassland, make grassland the reference. Then ‘forest’ and ‘wetland’ coefficients are contrasts to the landscape people survey most.

- **Scientifically neutral baseline:** Use a category that feels like a sensible ‘starting point’ ecologically, so positive/negative shifts feel intuitive. If pine martens prefer cover, you might choose grassland as the reference so ‘forest’ is a positive shift.

Choosing the Reference

- **Management relevance:** Put the category decision-makers care about as the baseline so effects are frames as improvements or declines from policy reality.

- **Data stability:** If one habitat has very few observations, don't use it as the reference. The intercept becomes poorly pinned down and all other contrasts inherit that uncertainty.

Modelling Species Occupancy

- For every one-unit increase in elevation, the odds of detectability are multiplied by the odds ratio e^{β_1}.
  - If $\beta_1 < 0$: then higher elevation lowers the odds of detectability.
  - If $\beta_1 > 0$: then higher elevation raises the odds of detectability.
- Moving from grassland to forest, the odds of detectability are multiplied by e^{β_2} ,and moving from grassland to wetland, the odds of detectability are multiplied by e^{β_3}.
  - If $\beta_2 > 0$ or $\beta_3 > 0$ then forest/wetland sites have higher odds than grassland sites.
- For each additional survey minute, the odds of detectability are multiplied by e^{β_4}

  If $\beta_4 > 0$: each extra survey minute increases detection odds.

Wait, What Are We Actually Modelling Here?

It’s Not Occupancy

Modelling Species Occupancy

- Its important to note here, that we are modelling detection/non-detection of animals, not true presence/absence.
- Why?
- Because we are accounting for probability of success (of viewing an individual), but we are not accounting for probability of occupancy. If probability of occupancy is less than 1, the state you're interested in (true presence) is latent.
- You can say you're modelling presence/absence if:
  - Detection probability is effectively 1 (you have perfect detection – if the animal is there, you will definitely see it)
  - Or if you explicitly model detection probability

Can We Ever Assume Perfect Detection?

- In ecology, the assumption that all individuals present are detected is very strong and usually unrealistic, but there are a few cases where it can hold.

1. Complete censuses in closed environments: counting fish in an aquarium, trees in a bounded plot, or tagged animals in a fenced reserve
2. Large, immobile, or easily identifiable organisms: counting trees in a forest inventory plot, or elephant herds in open savannah
3. Automated detection with perfect coverage: camera traps or acoustic sensors that cover the entire area, with individuals tagged or uniquely identifiable

- However, in most wildlife surveys, animals move, hide or vocalise unpredictable, and environmental conditions affect detection, which means that detection probabilities are usually far below 1.

Modelling Species Occupancy

- If we have J visits to each site, and we assume that π_i, the detection probability per site is constant (i.e. it does not change over visits)

$$Y_{ij} \sim Bernoulli(Z_i \pi_i)$$
$$logit(\pi_i) = \beta_0 + \beta_1 x_{1j} + \cdots$$

$$Z_i \sim Bernoulli(\phi_i)$$
$$logit(\phi_i) = \alpha_0 + \alpha_1 w_1 + \ \ldots$$

- Here, Y_{ij} models detection/non-detection, and it has a probability of detection/non detection π_i.

Modelling Species Occupancy

- There is now an additional term Z_i. This is another Bernoulli variable (so it can only take values of 0 or 1) and it represents the true (unknown) occupancy at site i.
- If $Z_i = 1$, then the site is truly occupied, and Y_{ij} is free to estimate detection using its probability π_{ij}. However, if $Z_i = 0$, then the site is unoccupied, and so the probability of detection term $Z_i\pi_{ij} = 0$.

Modelling Species Occupancy

- Now our model has two stages.
- The first stage asks: 'Are there any animals at site i?' and Z_i tells us 'Yes' or 'No' with probability ϕ_i.
- If there are animals there, the second stage asks: 'Did we see them on each of our j visits?' and Y_{ij} tells us 'Yes' or 'No' with probability π_i.

Modelling Species Occupancy

- We have visited site i a total of J_i times, and each time, recorded detections $Y_{ij} \in \{0,1\}$. The site-level detection count is then $S_i = \sum_{j=1}^{J_i} Y_{ij}$
- This is just: 'how many times did we detect the species at site i, over all our visits?'. We assume visit detections are independent given presence, and that detection is constant across visits at site i.
- If a site is unoccupied, no detections are possible, so we can only have $S_i = 0$. If a site is occupied, then we use $Y_{ij} \sim Bernoulli(\pi_{ij})$ which can produce 0 or 1.
- So, a zero at a site can happen in two ways:
  - The site is truly absent (structural zero)
  - The site is present, but missed on all visits (sampling zero)

What Does This Model Remind You Of?

- We have a two-stage process:
- Occupancy model: is the site occupied?

$$Z_i \sim Bernoulli(\phi_i)$$

- Detection model: given occupancy, did we detect it on visit j?

$$Y_{ij} \sim Bernoulli(\pi_i)$$

- Zeroes can arise in two ways:
- Structural zeros: site truly unoccupied ($Z_i = 0$)
- Sampling zeros: site occupied but not detected on any visit ($Z_i = 1$, all $Y_{ij} = 0$)
- This looks very much like a…Zero-inflated model! Specifically, a zero-inflated binomial (ZIB).

Zero-Inflated Binomial Models for Species Occupancy

- We can write the site-level detection count as:

$$S_i \sim ZIB(J_i, \pi_i, \phi_i)$$

- Where:
  - $S_i = \sum_{j=1}^{J_i} Y_{ij}$ is the number of detections at site i.
  - J_i = number of visits to site i.
  - π_i =detection probability per visit (conditional on presence).
  - ϕ_i = occupancy probability (probability site is truly occupied).
- **Zero inflation arises because:**
- With probability $1 - \phi_i$,site is unoccupied → $S_i = 0$ (structural zero).
- With probability ϕ_i ,site is occupied → $S_i \sim Binomial(J_i, \pi_i)$.

Coding Demo

Predicting Abundance from Count Data

- Count data are common in ecology (e.g., bird surveys, insect traps, fish counts).
- We can just fit a Poisson or negative Binomial GLM directly to the observed counts, but this does not take into account the issue of imperfect detection (the probability of detection is unlikely to be 1, and so when we model the counts, we are modelling detected counts, and not taking into account that the abundance may not align completely.
- What if we want to use our collected counts to estimate the true abundance?
- The N-mixture model (Royle, 2004) is a model that separates abundance from detection by taking into account that detection is not perfect

The GLM Approach We Have Already Seen

- The standard GLM for counts is:

$$Y_i \sim Poisson(\lambda_i)$$
$$\log(\lambda_i) = \beta_0 + \beta_1 x_{1i} + \ \ldots$$

- This model is easy to fit in glm() or brms(), and it works if we can assume every individual is perfectly detected.
- However, it conflates true abundance with detection probability.

N-Mixture Models

- In reality, the detection probability is likely less than 1.
- For example, lets say we go into a site where 10 birds are present, but only detect 3-7 of them per survey.
- A GLM will systematically underestimate abundance if the probability of detection is less than 1.
- N-mixture models handle this by introducing a latent abundance N_i at each site. We now have an abundance process:

$$N_i \sim Poisson(\lambda_i)$$

And an observation process:

$$Y_i \sim Binomial(N_i, \pi_i)$$

Assumptions of the N-Mixture Model

- Surveys occur within a short 'closed' window: N_i is constant across visits (no births, deaths, immigration, emigration, or permanent movement between sites).
- Sites are independent, and individuals are not shared across sites during the closure period.
- Conditional independence of counts. Given N_i and detection π_{ij}, the y_{ij} from different visits are independent.
- Binomial detection within a visit: each individual present can be counted at most once, and detections are independent across individuals.
- Detection probability is homogeneous across individuals within a visit, or any heterogeneity is fully explained by modelled covariates.
- Survey effort is equal or included as a covariate/offset.

Extensions to the N-Mixture Model

- Overdispersed abundance: Negative Binomial N_n
- Excess zeros: ZIP for true absences beyond Poisson expectation.
- Heterogeneous detection: Beta-binomial detection, individual/visit random effects, or finite mixtures for π.
- Double-observer/replicated observers: Observer effects and dependence during a visit.
- Dynamic/open populations
- Spatial structure: Site-level spatial random effects to model autocorrelation in λ or π.
- Multi-species/community models: Hierarchical shrinkage across species for λ and π; shared covariates.

Coding Demo

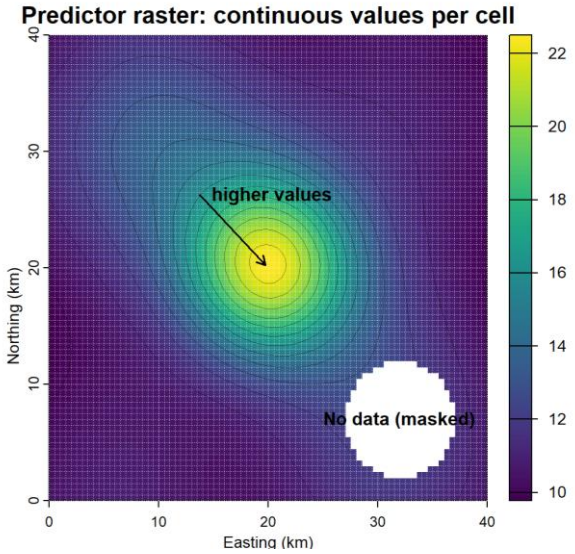
Modelling Species Distributions

- Species Distribution Models (SDMs) can answer questions like:
  - Where is a species likely to occur today? (habitat suitability)
  - What environmental gradients define its niche?
  - How might distributions shift under climate scenarios?
  - Where should be survey or conserve next?

- The key idea behind an SDM is to link occurrence to environment, so that we can predict occurrence in unsampled space and time

Rasters

- A raster is a grid of cells (pixels) used to represent spatial data. Each cell has a value, representing information like elevation, temperature etc.
- The extent defines the rectangular area covered by a raster, specified by the minimum and maximum X and Y values.
- The resolution is the size of each cell (pixel) in real-world units. This determines the level of detail in the raster. Higher resolution means more detail and a larger file size. Resolution must match across rasters when used as predictors in models.



What Counts as an SDM?

- An SDM is a statistical model that estimates the relationship between species occurrence/abundance and environmental predictors, and then projects that relationship across space and time.

Core ingredients

1. Response data: presence absence or counts
2. Predictors: environmental rasters (climate, topography, land cover, soil, distance)
3. Model: GLM, GLMM, GAM
4. Projection grid: a raster stack with the same CRS, extent and resolution

Assumptions

1. Stationarity: the species-environment relationship generalises across the mapping area
2. Representative sampling: occurrences and absences reflect available environments (beware survey bias)
3. Predictor coverage: training covers the environmental range you'll project to (avoid uncontrolled extrapolation)
4. Independence: observations are independent (spatial autocorrelation can inflate performance)
5. Data quality: geolocation accuracy, consistent CRS, plausible covariates

Turning a GLM into an SDM

- A GLM becomes an SDM when you:

1. Use spatially referenced predictors (rasters) aligned for projection
2. Extract predictors at occurrence locations to build the training table
3. Fit a model suitable for the response
4. Project to the raster stack using the same transformations used in the model
5. Handle extrapolation (clamping, warnings)
6. Evaluate with spatial CV to avoid overly optimistic metrics
7. Map and interpret probabilities, thresholds and uncertainties

Interpreting Maps

Occurrence models (binary response):

- The output is a probability of occurrence under the model assumptions.
- High values indicate predicted suitability, not guaranteed presence.

Count/abundance models (e.g., Poisson, N-mixture):

- The output is an expected abundance (mean count) under the model assumptions.
- Values represent relative density across the landscape, not exact headcounts.
- High predicted abundance indicates favourable conditions, but uncertainty should be acknowledged.

Common Mistakes

1. Using mismatched rasters (different CRS or resolution)
2. Leakage: calculating transforms (e.g., scaling) differently for rasters vs. training data
3. Overfitting with high-order polynomials/interactions without cross-validation
4. Ignoring spatial autocorrelation
5. Projecting far beyond the training environmental space without warnings

Coding Demo