Introduction to Generalised Linear Models for Ecologists

Dr Niamh Mimnagh

niamh@prstats.org

https://github.com/niamhmimnagh/glmm01

PR STATS

Example: Binary (Yes/No) Data

- Lets say we are studying 100 sites, and collecting data on the detection/non-detection of pine martens at each site.
- The covariates recorded during this study are habitat (grassland, forest, wetland), elevation (m) and effort (minutes spent at each site)
- Treat Y_i as the response:

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

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

Example: Binary (Yes/No) Data

- **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.

Example: Binary (Yes/No) Data

- e^{β_0} is the odds of detectability at the baseline predictor values
- $\pi_0 = \frac{e^{\beta_0}}{1+e^{\beta_0}}$ is the probability of detectability at those baseline 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)

Example: Binary (Yes/No) Data

- If predictors are centred, then $elevation_c = elevation - \overline{elevation}$, and $effort_c = effort - \overline{effort}$
- 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.

Coding Demo

Example: Binary (Yes/No) Data

- 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.

Example: Binary (Yes/No) Data

- 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 (viewing an individual), but we are not accounting for probability of occupancy. If probability of occupancy is less than 1, the thing you want (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

Example: Occupancy Model in brms

- 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)

$$Y_{ij} \sim Bernoulli(Z_i \pi_i)$$
$$logit(\pi_{ij}) = \beta_0 + \beta_1 x_{1j} + \cdots$$
$$Z_i \sim Bernoulli(\phi_i)$$
$$logit(\phi_i) = \alpha_0 + \alpha_1 w_1 + \ \ldots$$

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

Example: Occupancy Model in brms

- However, 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 $Z_i\pi_{ij} = 0$.

Example: Occupancy Model in brms

- 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)

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.

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.

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:

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

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

Definitions

- A raster is a grid of cells (pixels) used to represent spatial data. Each cell has a value, representing information like elevation, temperature etc.
- A Coordinate Reference System (CRS) defines how the two-dimensional raster grid relates to locations on earth, and ensures that layers align correctly when combined. Without a common CRS, rasters wont ‘line up’ correctly in space.
- 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

Workflow

1. Assemble predictors (rasters) with shared CRS, extent, and resolution
2. Collect response data
3. Extract values of predictors at points (to create a modelling table)
4. Specify the GLM
5. Fit and evaluate the model
6. Predict to rasters (probability of occurrence)
7. Communicate results, uncertainty and limitations

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

1. The output is a probability of occurrence under the model assumptions
2. High values indicate predicted suitability, not guaranteed presence
3. Consider thresholding only for certain decisions (e.g., management) and report sensitivity to threshold choice
4. Provide legends, scales, CRS and metadata

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

SDMs in R

1. terra: create and process rasters, extraction and predictions
2. sf: vector geometries
3. dplyr: data wrangling
4. ggplot2: plotting
5. pROC: AUC/ROC
6. blockCV: spatial CV

Example

- The data we are looking at today has three rasters:

1. Elev: a smooth elevation gradient
2. Bio1: temperature
3. Bio2: precipitation

- Our data contains a species whose probability of occurrence follows a quadratic response to these variables

Coding Demo