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

Binary Logistic Regression Review

- Previously, we learned about binary logistic regression, where each trial is either a success (1) or failure (0).
- The model estimates the log-odds of success using a linear combination of predictors:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$$

- Each observation corresponds to one trial. This is appropriate for datasets where each row represents a unique individual outcome.

Binomial Models for Proportional Data

- What if instead, we have proportional data? For example, we may know:
  - The number of seeds that germinated out of 20
  - The number of surviving individuals out of a release
  - The number of infected animals per herd

- These outcomes are grouped binomial variables. They are not single yes/no, 0/1 events like in binary regression, but aggregated over multiple trials. We model this using a binomial distribution.

Proportional Data

- $\frac{8}{20}$ seeds germinated in one flowerpot, and $\frac{15}{20}$ in another.
- This outcome is not a binary event, but a discrete proportion from multiple trials.
- We could look at it seed-by-seed and model germinated/did not germinate as a binary outcome.
- But that would ignore the grouped structure and result in inefficient estimation.
- Binomial logistic regression allows us to model the number of successes out of a fixed number of trials.

The Binomial GLM for Proportional Data

- In a binomial GLM, each observation consists of a number of successes and a number of failures. The model assumes:

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

- Where Y_i is the number of successes out of n_i trials, and π_i is the probability of success. The logit link function transforms this probability so it can be modelled using a linear predictor:

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

Ways to Specify the Response in R

- In R, you can specify a binomial logistic regression model using either a matrix of counts or a proportion with weights. For example:

Option 1: Using counts

```
glm(cbind(successes, failures) ~ x1 + x2, family = binomial, data = df)
```

Option 2: Using proportions with weights

```
glm(prop ~ x1 + x2, weights = total_trials, family = binomial, data = df)
```

- Both approaches are equivalent and produce the same estimates, provided the response variable and weights are defined correctly.

Weights in Binomial Models

- When using the proportion-and-weights formulation, the weights argument specifies the number of trials used to calculate each proportion.
- This ensures that the model recognises that a proportion based on 100 trials is more reliable than one based on just 5 trials.
- Internally, the model uses this information to assign greater influence to high-information observations during estimation.

Link Functions

- While the logit link is the most common, other link functions are available:
- The probit link uses the inverse cumulative distribution function of the normal distribution.
- The complementary log-log (cloglog) link is asymmetric and often used when the outcome is very rare.

- The choice of link function affects model interpretation and the shape of the predicted response curve but not the underlying data structure.

Example: Germination Study

- Suppose we conduct an experiment where 20 seeds are planted in each of several pots, with different levels of light. For each pot, we record the number of seeds that germinated.

Pot	Light	Germinated
A	Full	2
B	Partial	10
C	Shade	7

- We want to model the effect of light condition on probability of germination.

Coding Demo

Overdispersion in Binomial Models

- The binomial model assumes that the variance is a function of the mean: $Var(Y) = n\pi\ (1\ - \pi)$.
- However, in practice, we often observe more variability than expected, a phenomenon known as overdispersion.
- This suggests that additional unmodelled variation exists. To account for this, you can:
  - Fit a quasibinomial model
  - Use a beta-binomial model
  - Add random effects using mixed models.

Multinomial Logistic Regression

- In many applications, the outcome variable can take on more than two categories, and these categories are not ordered.
- For example, an animal may choose between three habitats: forest, wetland, or grassland. These choices are distinct but have no natural order.
- Binomial logistic regression cannot be used when the outcome has more than two unordered categories. In such cases, we use multinomial logistic regression.
- Multinomial logistic regression models the probability of each possible outcome as a function of predictor variables.
- Each outcome is treated as a distinct category, and we model the log-odds of each category relative to a chosen baseline category.

Data Format for Multinomial Models

- The data should have one row per observation, and one column that contains the outcome variable (the factor with at least 3 levels), along with columns for any additional predictor variables.

Site	Habitat	Temperature	Slope
1	Grassland	18	2
2	Wetland	21	1.5
3	Heath	20	3

Model Structure

$$Y_i \sim Categorical(\pi_i)$$

- For outcome Y ∈ {A, B, C}, there are three probabilities:
  - π_{iA}: the probability that Y belongs to category A ($P(Y = A)$)
  - π_{iB}: the probability that Y belongs to category B
  - π_{iC}: the probability that Y belongs to category C

- "A" = reference level (we let $\eta_{iA} = 0$)
- Model estimates separate logits for each other category vs. baseline

Model Structure

$$Y_i \sim Categorical(\pi_i)$$

Log-risk 'C vs. A' $$\log\left(\frac{\pi_{ic}}{\pi_{iA}}\right) = \eta_{iC} = \beta_{C0} + \beta_{C1}\, x_1 + \ldots$$

Log-risk 'B vs. A' $$\log\left(\frac{\pi_{iB}}{\pi_{iA}}\right) = \eta_{iB} = \beta_{B0} + \beta_{B1}\, x_1 + \ldots$$

- "A" = reference level (we let $\eta_{iA} = 0$)
- Model estimates separate logits for each other category vs. baseline

From Logits to Probabilities

- Given baseline A and linear predictors, $\log\left(\frac{\pi_{iB}}{\pi_{iA}}\right) = \eta_{iB}$, $\log\left(\frac{\pi_{iC}}{\pi_{iA}}\right) = \eta_{iC}$
- We can calculate probabilities by inverting with softmax:

$$\pi_{iA} = \frac{1}{1 + e^{\eta_{iB}} + e^{\eta_{iC}}}$$

$$\pi_{iB} = \frac{e^{\eta_{iB}}}{1 + e^{\eta_{iB}} + e^{\eta_{iC}}}$$

$$\pi_{iC} = \frac{e^{\eta_{iC}}}{1 + e^{\eta_{iB}} + e^{\eta_{iC}}}$$

- Special case K=2 reduces to the usual logistic.

$$\pi_{iA} = \frac{1}{1+e^{\eta_{iB}}}, \quad \pi_{iB} = \frac{e^{\eta_{iB}}}{1+e^{\eta_{iB}}} = \pi_{iB} = \frac{1}{1+e^{-\eta_{iB}}}$$

Why Use a Baseline?

In a binomial GLM we use: $\log\left(\frac{\pi}{1-\pi}\right) = \eta$.

Why not do the same thing now?

$$\log\left(\frac{\pi_B}{1-\pi_B}\right) = \eta_B \rightarrow \pi_B = \frac{1}{1+e^{-\eta_B}}$$

This is equivalent to fitting different binomial models for each level of your factor, and it doesn't constrain the resulting probabilities to summing to 1, which probabilities must.

Why Use a Baseline?

For example: Lets choose values for $\eta_B = 2.1, \eta_C = 0.1, \ \eta_A = 0$

- $\pi_{iA} = \frac{1}{1+e^{\eta_{iB}}+e^{\eta_{iC}}} = \frac{1}{1+e^{2.1}+e^{0.1}} = 0.097358$
- $\pi_{iB} = \frac{e^{\eta_{iB}}}{1+e^{\eta_{iB}}+e^{\eta_{iC}}} = \frac{e^{2.1}}{1+e^{2.1}+e^{0.1}} = 0.795044$
- $\pi_{iC} = \frac{e^{\eta_{iC}}}{1+e^{\eta_{iB}}+e^{\eta_{iC}}} = \frac{e^{0.1}}{1+e^{2.1}+e^{0.1}} = 0.107597$

$$0.097358 + 0.795044 + 0.107597 = 1$$

Using a baseline category ensures all categories are comparable (their probabilities sum to 1).

Why Use a Baseline?

- What if we don't use a baseline, and compare the probability of each category to the probability of 'not that category' as we did for the binomial case?

- $\pi_A = \frac{1}{1+e^{-\eta_A}} = \frac{1}{1+e^{0}} = 0.5$
- $\pi_B = \frac{1}{1+e^{-\eta_B}} = \frac{1}{1+e^{2.1}} = 0.89$
- $\pi_C = \frac{1}{1+e^{-\eta_C}} = \frac{1}{1+e^{0.1}} = 0.52$
- If we don't use a baseline level, and instead compare 'one-vs-all' the probabilities are not constrained to sum to one, because we are treating each probability as coming from a different binomial model.

Model Assumptions

- Independence of observations (there is no unmodelled clustering, temporal or spatial correlation present).
- Correct link and linear predictor: log relative risks are linear/additive in x.
- No (near) perfect multicollinearity among predictors; stable estimation.
- Independence of irrelevant alternatives (IIA) holds.

Independence of Irrelevant Alternatives

- For any two categories B and A, the ratio $\frac{\pi_B}{\pi_A}$ does not depend on the presence or attributes of other categories.
- Equivalently: adding or removing a third category (or changing its attributes) does not change $\frac{\pi_B}{\pi_A}$.
- In ecology: if two habitats are very similar (share similar characteristics, and might offer similar benefits as a habitat), Independence of Irrelevant Alternatives may be questionable.

Red Bus/Blue Bus Paradox

- Suppose we have two travel options: Car and Red bus. We can think in terms of 'utility' (an unobserved score each individual assigns to each alternative) – think of it in terms of attractiveness/suitability/preference.
- $P(Car) = 2/3$,
- $P(Red) = 1/3$.
- So, the ratio of car to red bus = $2:1$
- Now introduce a Blue bus that is identical to Red bus. This means that we must assume $P(Blue) = P(Red)$.
- Under the Independence of Irrelevant Alternatives (IIA) property, the odds of Car vs Red bus must stay $2:1$, regardless of Blue's existence.

Red Bus/Blue Bus Paradox

- $P(Car) = 1/2$,
- $P(Red) = 1/4$,
- $P(Blue) = 1/4$
- The Car share drops from $2/3$ to $1/2$ even though nothing about Car or "bus-ness" changed - only the bus colour!
- Intuitively, we expected Red and Blue to split the original bus users, not to siphon off extra drivers from Car.
- This is the paradox: IIA enforces proportional substitution across *all* options, so duplicating one option (creating a very similar alternative) inflates its total share.

When IIA is Plausible

- Plausible if categories are clearly distinct and alternatives are not close substitutes.
- Distinct habitat types: Nest-site choice among rock crevice vs. reed bed vs. tree cavity in a landscape where these options are physically and functionally very different.
- Different prey guilds: Predator choosing among insects vs. small mammals vs. fish where handling/search strategies are unrelated.
- Disparate substrates: Spawning on gravel bars vs. submerged logs vs. macrophytes with strong mechanical/flow differences.
- Trap type selection: Invertebrates choosing among pitfall vs. flight-intercept vs. light traps deployed identically; traps are not close substitutes.

When IIA is Questionable

- Near substitutes: Foraging patch choice among three grass meadows of similar height and distance; removing one meadow should shift choices mostly to the most similar meadow, not proportionally to all habitats.
- Shared unobserved factors: Reef selection where two reefs share predator pressure or human disturbance not in the data; these hidden attributes couple their utilities.
- Spatially clustered options: Stopover site choice along a flyway where adjacent wetlands are close substitutes; adding/removing one nearby wetland changes odds to its neighbours disproportionately.
- Duplicate alternatives: Flower choice among two cultivars of the same plant vs. a different species; duplication breaks IIA if pollinators treat cultivars as near-identical.

Hausman-McFadden Test: "drop one alternative"

- If the assumption of independence under irrelevant alternatives, holds, then coefficients estimated on the full set (all categories) and on a restricted set (after dropping one category), should be statistically the same.
- Fit the full multinomial logit on all K alternatives, and assemble all your coefficients together into a vector: $\widehat{\boldsymbol{\beta}}_{Full}, Var(\widehat{\boldsymbol{\beta}}_{Full})$
- Pick an alternative category whose affect you want to examine. Remove that alternative and all observations where it was chosen.
- Refit on the remaining K−1 alternatives, and assemble all the coefficients together again: $\widehat{\boldsymbol{\beta}}_{Restricted}, Var(\widehat{\boldsymbol{\beta}}_{Restricted})$

Hausman-McFadden Test: "drop one alternative"

- Form a Hausman statistic on the overlapping parameters:

$$H = \left(\widehat{\boldsymbol{\beta}}_{Restricted} - \widehat{\boldsymbol{\beta}}_{Full}\right)^{T} \left[Var\left(\widehat{\boldsymbol{\beta}}_{Restricted}\right) - Var\left(\widehat{\boldsymbol{\beta}}_{Full}\right)\right]^{-1} \left(\widehat{\boldsymbol{\beta}}_{Restricted} - \widehat{\boldsymbol{\beta}}_{Full}\right)$$

Here we are calculating the distance between the coefficient estimates in the full vs. restricted model. If IIA is true, the $\left(\widehat{\boldsymbol{\beta}}_{Restricted} - \widehat{\boldsymbol{\beta}}_{Full}\right)$ should be zero on average.

This formula is the Mahalanobis distance. It tells you how far the two estimates are from each other, and if that distance is more than you'd expect just due to noise.

Small H: the two estimates differ no more than expected noise → IIA holds
Large H: they differ more than expected → evidence against IIA.

Hausman-McFadden Test: "drop one alternative"

- Form a Hausman statistic on the overlapping parameters:

$$H = \left(\widehat{\boldsymbol{\beta}}_{Restricted} - \widehat{\boldsymbol{\beta}}_{Full}\right)^{T} \left[Var\left(\widehat{\boldsymbol{\beta}}_{Restricted}\right) - Var\left(\widehat{\boldsymbol{\beta}}_{Full}\right)\right]^{-1} \left(\widehat{\boldsymbol{\beta}}_{Restricted} - \widehat{\boldsymbol{\beta}}_{Full}\right)$$

Here we are calculating the distance between the coefficient estimates in the full vs. restricted model. If IIA is true, the $\left(\widehat{\boldsymbol{\beta}}_{Restricted} - \widehat{\boldsymbol{\beta}}_{Full}\right)$ should be zero on average.

This formula is the Mahalanobis distance. It tells you how far the two estimates are from each other, and if that distance is more than you'd expect just due to noise.

Small H: the two estimates differ no more than expected noise → IIA holds
Large H: they differ more than expected → evidence against IIA.

What to Do

- Model the similarity: Include alternative-specific attributes (e.g., distance, canopy cover, predator index) so options are less “exchangeable.”
- Use richer choice models: Nested logit (e.g., shrub vs. open then species), mixed logit (random coefficients), or multinomial probit to allow flexible substitution patterns.
- Collapse or redesign categories: Combine near-substitutes or redesign the sampling to ensure meaningful differences.
- Sanity checks: Compare MNL with a nested/mixed version, inspect cross-elasticities / share shifts after removing an option, or use a Hausman–McFadden style test (with caution in small samples).

Fitting the Model in R

- You can use the multinom() function from the nnet package in R to fit multinomial logistic regression models:

```
library(nnet)
model <- multinom(outcome ~ x1 + x2, data = mydata)
```

- Each row of the dataset represents a single observation with its outcome and predictor values. Each set of coefficients compares a class vs. baseline

Interpreting Coefficients

- Risk: (π_A, π_B, π_C) are the probabilities of choosing each category
- Relative risk (vs. A): $\frac{\pi_B}{\pi_A}, \frac{\pi_C}{\pi_A}$
- Log relative risk is what the multinomial logit models linearly.
- β_{k0}: log relative risk of k vs. A when predictors are 0 (or at their mean value).
- $e^{\beta_{k0}}$= relative risk of k vs. A at x=0.
- β_{k1} : change in log relative risk for a one-unit increase in x.
- $e^{\beta_{k1}}$ = Relative Risk Ratio (RRR) is the multiplicative factor applied to the relative risk per 1-unit increase in x.
- RRR > 1: higher x shifts probability toward k (relative to A), while RRR < 1 shifts it toward A.

Predicting Probabilities

- You can use the predict() function with type = 'probs' to get the predicted probability of each outcome category for each observation:

```
predict(model, type = 'probs')
```

- This returns a matrix of probabilities for the training data that can be used for plots and summaries. These probabilities sum to one and can be visualised to understand how predictors influence outcome likelihoods.

```
predict(model, newdata=df, type = 'probs')
```

This will return a matrix of probabilities for a new dataset.

Visualising Effects

- Visualising predicted probabilities across levels of a categorical predictor or over a range of a continuous predictor helps in understanding model implications.
- Packages such as `ggeffects`, `effects`, and `emmeans` provide useful tools for this.
- For example, we can plot how the probability of choosing each habitat changes with soil pH.

Example

- We have 600 animals who each make a habitat choice. The response Y_i is the habitat chosen by individual i; a factor with $K = 4$ categories: Mudflat, Seagrass, Rocky, Sand.
- Predictors (x_i): individual- and site-level covariates,
- body_mass (g; heavier birds tolerate deeper water → more Seagrass),
- age_class (juvenile/adult; juveniles avoid Rocky due to handling risk),
- prey_density (g/m²; higher increases odds of Mudflat/Seagrass),

$$Y_i \sim Categorical(\boldsymbol{\pi}_i)$$
$$\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, \pi_{i3}, \pi_{i4})$$
$$\log\left(\frac{\pi_{ij}}{\pi_{ik}}\right) = \beta_0 + \beta_1 body\ mass_i + \beta_2 age\ class_i + \beta_3 prey\ density_i$$

Coding Demo

Ordinal Logistic Regression

- An ordinal outcome is categorical with a natural order (e.g. poor < fair < good < excellent)
- However, we cannot assume equal distances between categories e.g., Is the distance between ‘poor’ and ‘fair’ equal to the distance between ‘good’ and ‘excellent’?
- Ordinal GLMs are for outcomes with a natural order, treating categories as ranks rather than numbers.
- They estimate how predictors shift probability toward higher or lower categories while not assuming equal spacing between categories.
- Example contexts: disease severity (mild < moderate < severe), habitat quality (poor < fair < good < excellent).

Cumulative Odds

- For K ordered levels ($C_1 < \ldots < C_K$), we can form $K - 1$ binary questions:

- $Y \leq C_1$ vs $Y > C_1$;
- $Y \leq C_2$ vs $Y > C_2$;
- …
- $Y \leq C_{K-1}$ vs $Y > C_{K-1}$.

Poor, Fair, Good, Excellent:

Is Y in the Poor category, or in a higher category?
Is Y in the Poor/Fair categories, or in a higher category?
Is Y in the Poor/Fair/Good categories or the Excellent category?

- Splitting the data this way respects the order. Modelling events cumulatively like this avoids treating categories as unrelated, and guarantees ordered probabilities.

Modelling the Outcome

- We have K ordered categories $C_1 < C_2 < \ldots < C_K$.
- For each observation i, define
- $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{ik})$ with $\pi_{ij} = P(Y_i = C_j)$, $\sum_j \pi_{ij} = 1$, $\pi_{ij} \geq 0$.

$$Y_i \sim Categorical(\boldsymbol{\pi}_i)$$

- Because the categories are ordered, we will not model each π_{ij} as unrelated. Instead, we will work with cumulative probabilities that encode the order.

Modelling the Outcome

- Define $\Gamma_{ij} = P(Y_i \leq C_j) = \sum_{k=1}^{j} \pi_{ik}$ for j = 1,…,K−1.
- Interpretation: π_{ij} is the probability of being in category j. Γ_{ij} is the probability of being in category j or any lower category.
- For a fixed i, the sequence is non-decreasing:

$$\Gamma_{i1} \leq \Gamma_{i2} \leq \cdots \leq \Gamma_{iK-1} \leq \Gamma_{iK}$$

- Because $\Gamma_{ij+1} = \Gamma_{ij} + \pi_{ij+1}$

Modelling the Outcome

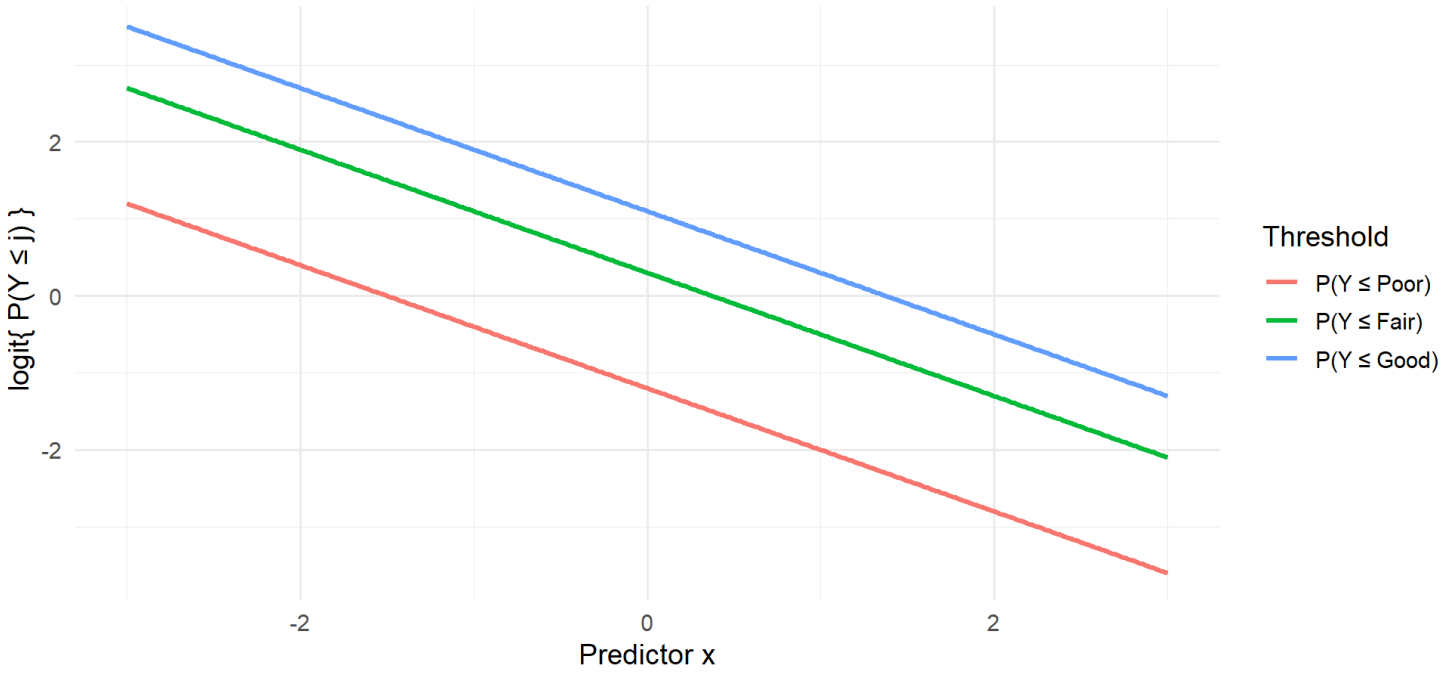
$$Y_i \sim Categorical(\boldsymbol{\pi}_i)$$

$$log\left(\frac{\Gamma_{ij}}{1-\Gamma_{ij}}\right) = \beta_{0\mathrm{j}} - x_{\mathrm{i}}{}^{\mathrm{T}}\beta_1, \qquad j = 1, \ldots, K-1.$$

- $\beta_{0\mathrm{j}}$: threshold/cutpoint for level j (an intercept per threshold).
- β_1 : slope vector shared across thresholds.
- With a single predictor x: one slope β_1 across all thresholds (parallel slopes).
- Interpretation: an increase of 1 in the predictor x multiplies the odds of a higher category by e^{β_1} at every threshold. i.e., Add 1 to x and the odds of landing in a higher category go up by the same factor for all the 'poor vs higher', 'fair vs higher', and 'good vs higher' splits.

The Proportional Odds Assumption

- The proportional odds assumption states that the relationship between the predictors and the outcome is constant across all cumulative logits.
- This means that the effect of the predictor is the same regardless of which category cutoff we consider.
- Violations of this assumption suggest that the predictor affects different levels of the outcome differently.

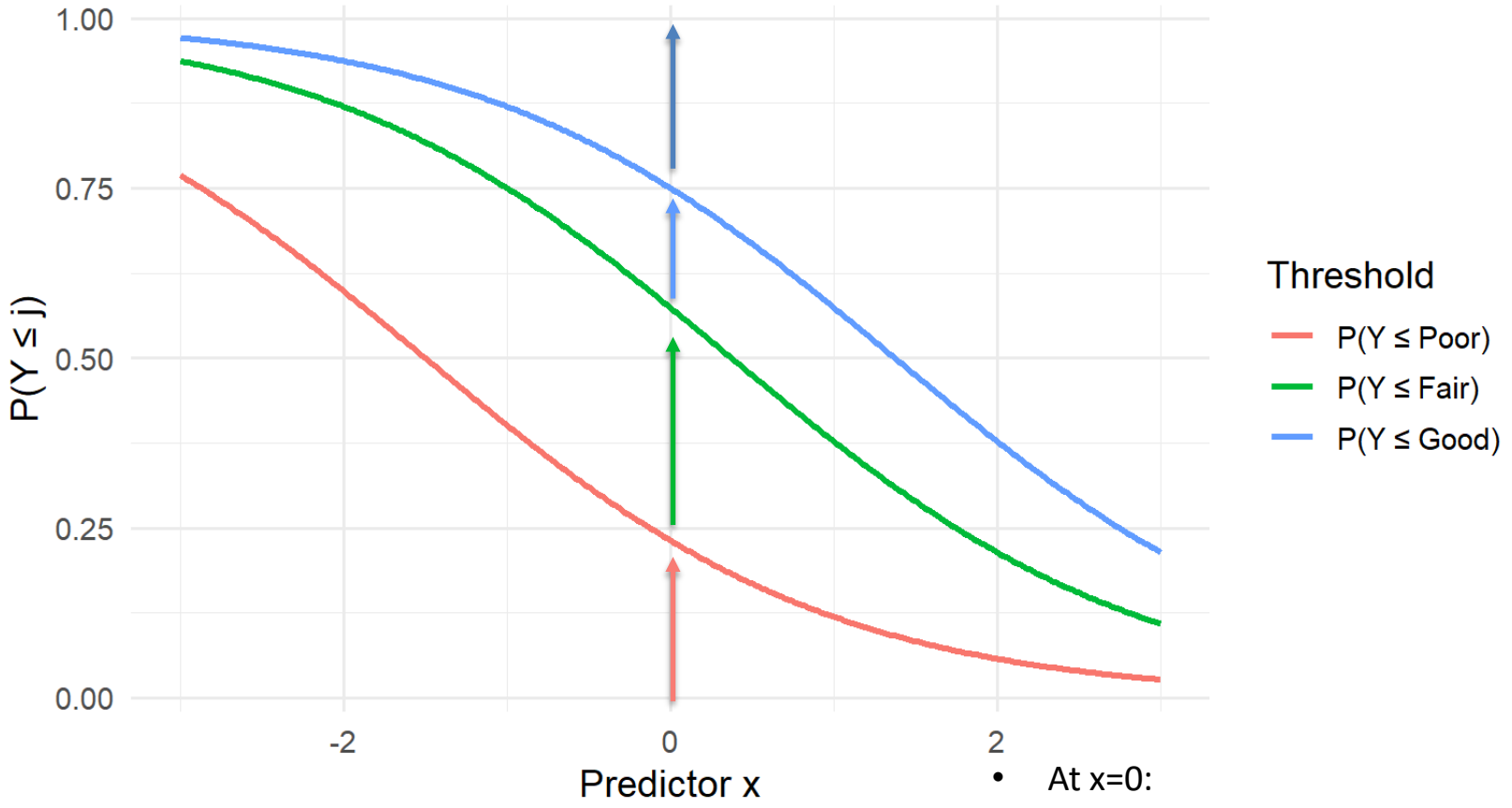


- Treat every line as a yes/no question — “is the outcome at or below Poor? at or below Fair? at or below Good?”
- Intercepts: the vertical offsets of the lines where the model draws the boundaries between categories.
- Slope: lines are parallel because all boundaries respond to x the same way. If the lines slant downward, increasing x pushes probability to higher categories; if they slant upward, increasing x pushes probability to lower categories.

Probabilities

- We are working with cumulative probabilities when modelling the predictors, but the category probabilities can be recovered from the cumulative ones:

- $\pi_{i1} = \Gamma_{i1} = P(Y_i \leq C_1) = P(Y_i = C_1)$,
- $\pi_{ij} = \Gamma_{ij} - \Gamma_{i,j-1} = P\left(Y_i \leq C_j\right) - P\left(Y_i \leq C_{j-1}\right) = P(Y_i = C_j)$
- $\pi_{ik} = 1 - \Gamma_{i,k-1} = 1 - P(Y_i \leq C_{k-1}) = P(Y_i = C_k)$



- At x=0:
  - P(Y='Poor')=0.25,
  - **P(Y='Fair')=0.55-0.25=0.35**,
  - P(Y='Good')=0.75-0.55=0.2,
  - P(Y='Excellent')=1-0.75=0.25

Cumulative Logit Model

- At each boundary j (between categories j and $j+1$) we ask a yes/no question for every observation: Is the observed Y_i at or below C_j?
- The model predicts $\Gamma_{ij} = P(Y_i \leq C_j \mid x_i)$: the probability the answer is "yes."
- Each boundary has its own intercept β_{0j}. Increasing it makes "Y ≤ C_j" more likely for everyone; decreasing it makes it less likely.
- During model fitting, the intercept is nudged upwards if class membership is underpredicted and nudged downwards if class membership is overpredicted.
- The algorithm iterates these nudges until predictions align with the observed labels.

Interpreting Coefficients

- Cutpoints β_{0j} (intercepts/thresholds)
  - Locate category boundaries on the latent scale; larger $\beta_{0j} \Rightarrow$ higher cumulative probability.
  - Not usually interpreted individually beyond ordering and where most mass lies.
- Slopes β_1 (common across j)
  - For a one-unit increase in x:
  - Odds of being at or below C_j multiply by $e^{-\beta_1}$ for every j
  - Equivalently, odds of being in a higher category (above C_j) multiply by e^{β_1}
  - Thus $e^{\beta_1} > 1$ shifts probability toward higher categories; $e^{\beta_1} < 1$ shifts toward lower categories.

Model Fitting in R

- To fit these models in R, your outcome needs to be an ordered factor (R needs to know the order of your categories in order to preserve that order).

```
factor(score, levels = c("low", "medium", "high"), ordered = TRUE)
```

- Use the ordinal package to fit these models (it is more robust than the polr package in terms of diagnostics):

```
library(ordinal)
model2 <- clm(abundance ~ treatment + soil, data = veg)
```

Example

- We are predicting habitat suitability ratings for a ground-nesting bird across 300 wetland sites.
- Response Suitability ∈ {Poor < Fair < Good < Excellent}.
- Predictors are VegCover (% cover, scaled), DistRoad (log-km), Management (Protected vs Unprotected).

$$Y_i \sim Categorical(\pi_i),$$

$$\log\left(\frac{P(Y_i \leq C_j)}{P(Y_i > C_j)}\right) = \beta_{0j} + \beta_1 VegCover_i + \beta_2 DistRoad_i + \beta_3 Management_i$$

Coding Demo