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

Data Wrangling and Preparation for GLMs

- Good models start with clean, structured data
- GLMs are sensitive to how you handle missing values, factor levels, and scaling

Rubbish In, Rubbish Out.

- GLMs assume:

1. Predictors are correctly specified
2. No hidden inconsistencies in data
3. Factors are correctly encoded

- Data preparation affects interpretability, convergence, and statistical power

Long vs. Wide Data Formats

- Long format means each row is a single observation - for example, one site–year–species combination—with the variables in separate columns (e.g., count, effort, temperature, species, year).
- This layout is what most GLMs and GLMMs expect, because each row represents one outcome with its own covariates and offset. It also works smoothly with tools like ggplot2 and dplyr.
- Wide format means each row is a single unit (such as a site), while repeated measurements appear as separate columns (e.g., count_2019, count_2020, … or one column per species).
- This layout is convenient for matrix-based methods, quick cross-tabs, and algorithms that require a rectangular matrix, but it’s awkward for regression models because the repeated measures are spread across columns rather than rows.

Long Format

- Every measurement carries its own covariates and an offset, so it maps naturally to regression rows.
- Grouping, interactions, and repeated measures are straightforward to specify.
- GLMs/GLMMs, plotting, and summaries all expect or work best with this layout.

Site	Year	Species	Effort	Temp	Count
A	2019	Sparrow	1.5	14.8	15
A	2019	Warbler	1.3	14.8	6
A	2020	Sparrow	1.7	15.3	12
B	2019	Sparrow	2.0	13.9	9

Wide Format

- It’s convenient for **matrix-based analyses** (e.g., distance matrices, ordination/PCA, clustering, correlation heatmaps) where you need a rectangular matrix of variables.
- GLMs expect **one row per observation** with covariates and offsets; in wide format, repeated measures are spread across columns, so interactions, offsets, and random effects are awkward.

Site	Sparrow_2019	Sparrow_2020	Warbler_2019	Warbler_2020
A	15	12	6	3
B	9	0	0	0

Missing Data

- GLMs do not accept NAs; rows with missing values in the formula are dropped.
- Decide how you’ll handle missingness before fitting the model.
- Aim to preserve validity (unbiased estimates) and efficiency (power).

Missingness Mechanisms

MCAR — Missing Completely at Random

- Missingness is unrelated to any data (observed or unobserved).
- Complete-case analysis is **unbiased**; you only lose power.

MAR — Missing At Random

- Missingness depends **only on observed variables** (after conditioning).
- Complete-case can be biased; prefer **multiple imputation** or ML methods.

MNAR — Missing Not At Random

- Missingness depends on the **unobserved value itself**.
- Standard GLMs and routine imputation are generally biased; need explicit models of missingness and **sensitivity analyses**.

Practical Plan for Analysis

- Diagnose patterns: who/when is missing; visualise by covariates/time.
- Argue mechanism: MCAR/MAR/MNAR based on context and checks.
- Choose strategy:
  - MCAR → Complete cases acceptable.
  - MAR → Imputation (mean, median, last value carried forward)
  - MNAR → Jointly model the outcome and probability of missingness.
- Include predictors of missingness in imputation/models to make MAR plausible.
- Report clearly what was missing, assumptions, and sensitivity results.

Categorical Predictors

- Categorical predictors may be included as factors in linear models.
- This is done using dummy coding.
- Linear models require numeric inputs
- A categorical variable with k levels cannot be directly included
- Solution: represent categories using dummy (indicator variables)
- Each dummy variable = 1 if the observation is in that category, and 0 otherwise

Dummy Coding

- Suppose habitat type has three categories: forest, grassland and wetland. What do we do with these?
- Answer: Create two dummy variables (Not three!)
- Grassland: 1 if grassland, 0 otherwise
- Wetland: 1 if wetland, 0 otherwise
- Reference level: forest (when both dummy variables are 0)

Reference Levels

- The reference category is the one with all dummies = 0
- By default (in R), it's the first level alphabetically
- You can change the reference level to:
  - Improve interpretation (e.g., compare against the most common habitat)
  - Highlight a specific category of interest

Habitat	Grassland	Wetland
Grassland	1	0
Forest	0	0
Wetland	0	1
Forest	0	0
Wetland	0	1
Grassland	1	0

Categorical Predictors

$$Movement_i = \beta_0 + \beta_1 Grassland_i + \beta_2 Wetland_i + \epsilon$$

- β_0: Mean movement in the reference habitat (forest)
- β_1: difference in movement between grassland and forest
- β_2: difference in movement between wetland and forest

Numeric Covariates

- First, confirm that variables intended to be numeric are actually stored as numeric rather than as factors or character strings.
- Make sure measurement units are consistent across your dataset, because mixed units (for example, °C and K, or mm and cm) will distort estimates.
- Inspect the distributions of each predictor and look for skew, heavy tails, and outliers, since these can unduly influence model fit and diagnostics.
- Identify near-zero variance predictors, because variables that hardly vary add noise without contributing signal.

Centring Numeric Predictors

- Centring means subtracting the mean from a variable so that zero represents an average value in your data.
- After centring, the model intercept becomes the expected outcome at the mean of the centred predictors, which is usually more interpretable than an outcome at zero.
- Centring can reduce collinearity, especially in interaction terms and polynomial expansions, which often leads to more stable coefficient estimates.
- Centring does not change slopes or overall model fit in linear models; it only changes the reference point for interpretation.
- Avoid centring binary indicators by default, as shifting their 0/1 meaning can make interpretation less intuitive; only centre them if you have a specific reason.

Scaling Numeric Predictors

- Scaling (standardising) means dividing a centred variable by its standard deviation so that it is measured in standard-deviation units.
- Scaling places predictors on comparable scales, which can improve numerical stability and optimiser convergence in many fitting procedures.
- Always compute scaling parameters (means and standard deviations) on the training data only and apply the same parameters to validation and test sets to avoid data leakage.
- When outliers are a concern, consider robust alternatives such as scaling by the median and MAD, or use min–max scaling if an algorithm expects inputs within a fixed range.

Model Extensions: Interactions and Non-Linear Terms

- Simple additive models assume each predictor affects the outcome independently and by a constant amount, which is often unrealistic.
- Interaction terms let the effect of one predictor depend on the level of another; for example, the influence of temperature on growth may be stronger when rainfall is high than when it is low.
- When you include interactions, remember that each main-effect coefficient describes the effect when the other interacting variables are zero (or at their reference levels), so centre continuous predictors and choose meaningful references to make interpretation clearer.
- For factor-by-numeric interactions, the interaction tests whether slopes differ between groups; for factor-by-factor interactions, it tests whether differences between categories change across levels of another factor.

Model Extensions: Interactions and Non-Linear Terms

- Nonlinear terms capture curvature, thresholds, and diminishing returns that linear terms miss; quadratic polynomials are a simple option.
- Because interactions and nonlinearities add complexity and can overfit, include them for substantive reasons, ensure you have data support across the combined predictor space, and evaluate models with cross-validation rather than relying only on single p-values.
- Interactions operate on the link scale (e.g., log-odds or log-rate), so translate effects back to probabilities or rates to communicate them accurately.
- Centering continuous predictors can reduce multicollinearity between interaction terms and their components, improving numerical stability.

Interaction Model Structure

$$Movement_i = \beta_0 + \beta_1 HabitatArea_i + \beta_2 Grassland_i + \beta_3 Wetland_i + \beta_4(HabitatArea_i \times Grassland_i) + \beta_5(HabitatArea_i \times Wetland_i) + \epsilon_i$$

β_0 : Baseline mean richness (forest, nitrogen = 0)

β_1 : Habitat Area slope for reference habitat

β_2 : Baseline difference Grassland vs Forest at Nitrogen=0

β_3 : Baseline difference Wetland vs Forest at Nitrogen=0

β_4 : Change in Habitat Area slope for Grassland vs Forest

β_5 : Change in Habitat Area slope for Wetland vs Forest

Interaction Model Structure

$$Movement_i = \beta_0 + \beta_1 HabitatArea_i + \beta_2 Grassland_i + \beta_3 Wetland_i + \\ \beta_4 (HabitatArea_i \times Grassland_i) + \\ \beta_5 (HabitatArea_i \times Wetland_i) + \varepsilon_i$$

Forest (reference): $Richness = \beta_0 + \beta_1 HabitatArea$

Grassland: $Richness = (\beta_0 + \beta_2) + (\beta_1 + \beta_4) HabitatArea$

Wetland: $Richness = (\beta_0 + \beta_3) + (\beta_1 + \beta_5) HabitatArea$

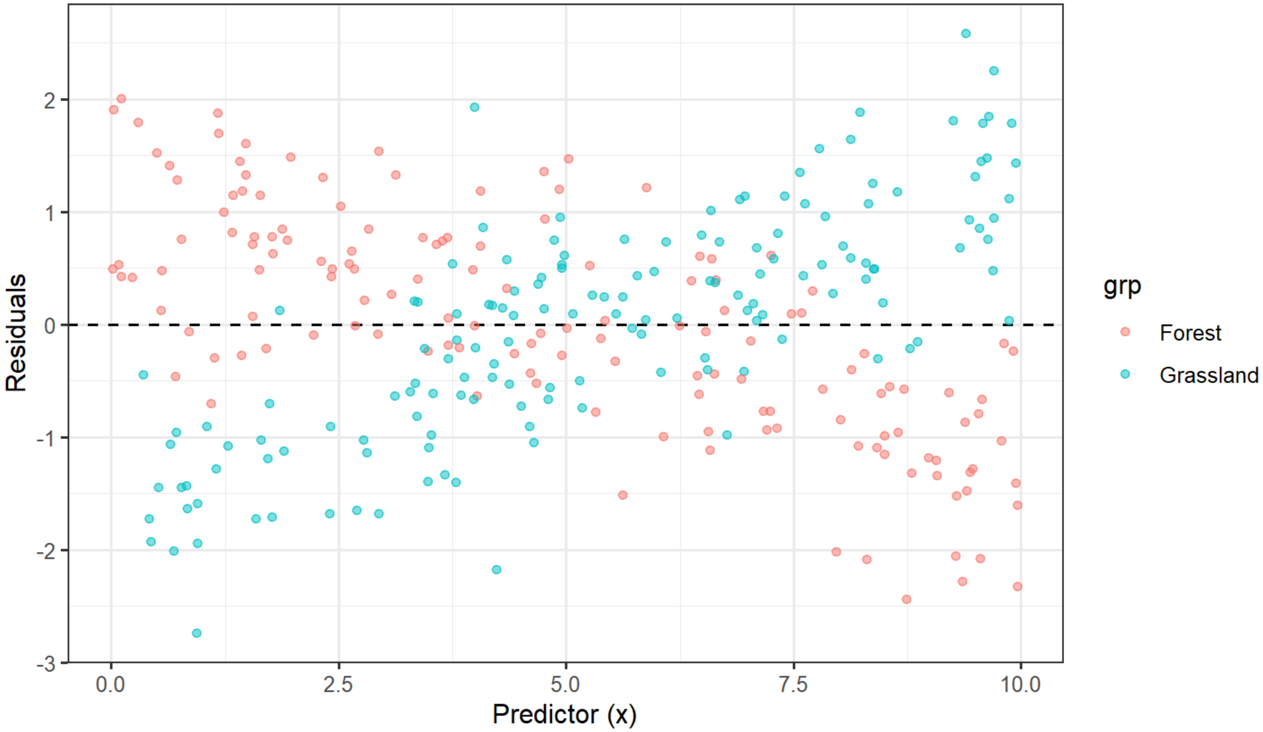
If $\beta_4 = 0$ → The effect of habitat area in the grassland is the same as the effect in the forest.

If $\beta_4 \neq 0$ → Habitat area effect on species richness is different in grassland versus forest

Why Include Interactions?

- Interaction terms are realistic: different habitats may respond differently to nutrients.
- Interaction terms can model fit when true effects are not additive.
- Caution: Interactions increase model complexity and must be justified by theory or data.

Why Include Interactions?



Specifying Interactions in R

- Check axis scales so that small effects are not exaggerated.
- Show sample sizes if groups are imbalanced so viewers can judge reliability.
- Avoid hiding extrapolation beyond the observed data range.
- In R formulas, use * to include both main effects and their interaction (e.g., richness ~ area * habitat), and use : to add only the interaction when main effects are specified elsewhere.
- To include all two-way interactions among several predictors, use the expansion (x1 + x2 + x3)^2, and be cautious with higher-order expansions because complexity grows rapidly.

Examples

Factor x Factor Interactions

- Soil_type (A/B) × Fertiliser (Yes/No)
- Soil A, No Fertiliser = 10; Soil A, Yes Fertiliser = 20
- Soil B, No Fertilizer = 15; Soil B, Yes Fertiliser = 16
- Fertiliser helps in Soil A but not in Soil B → interaction!

Factor x Numeric Interactions

- Temperature effect differs by species

- shows how the slope of temperature changes across species

Interpreting Interaction Terms

- Without an interaction, the model assumes a single slope for the predictor across all groups or values of the other variable.
- With an interaction, the slope of one predictor changes with the level or value of the other predictor, so each group (or value) can have its own slope.
- In a model with $A \times B$, the main effect of A is the effect of A when B is at its reference level or at zero/its centered mean.
- The interaction coefficient quantifies how the slope of A changes with B.
- Keep main effects when you include an interaction, even if their p-values are not significant, because dropping them violates the hierarchy and changes the meaning of the interaction term.
- You may include an interaction even when main effects are not significant, since a predictor can matter only at certain levels of the other predictor.

Nonlinear Terms

- Many predictor–response relationships are curved rather than straight, so a single linear slope can miss important structure and bias estimates.
- Classic examples include hump-shaped species richness along elevation gradients, temperature–disease risk with thresholds and diminishing returns of yield with increasing rainfall.
- Nonlinear patterns can be modelled with polynomial terms (e.g., quadratic or cubic) to capture curvature and peaks/valleys in a simple, interpretable way.
- Centering predictors before adding polynomial terms improves numerical stability and makes the baseline (zero) value interpretable; avoid extrapolating curves far beyond the observed range.

Polynomial Terms

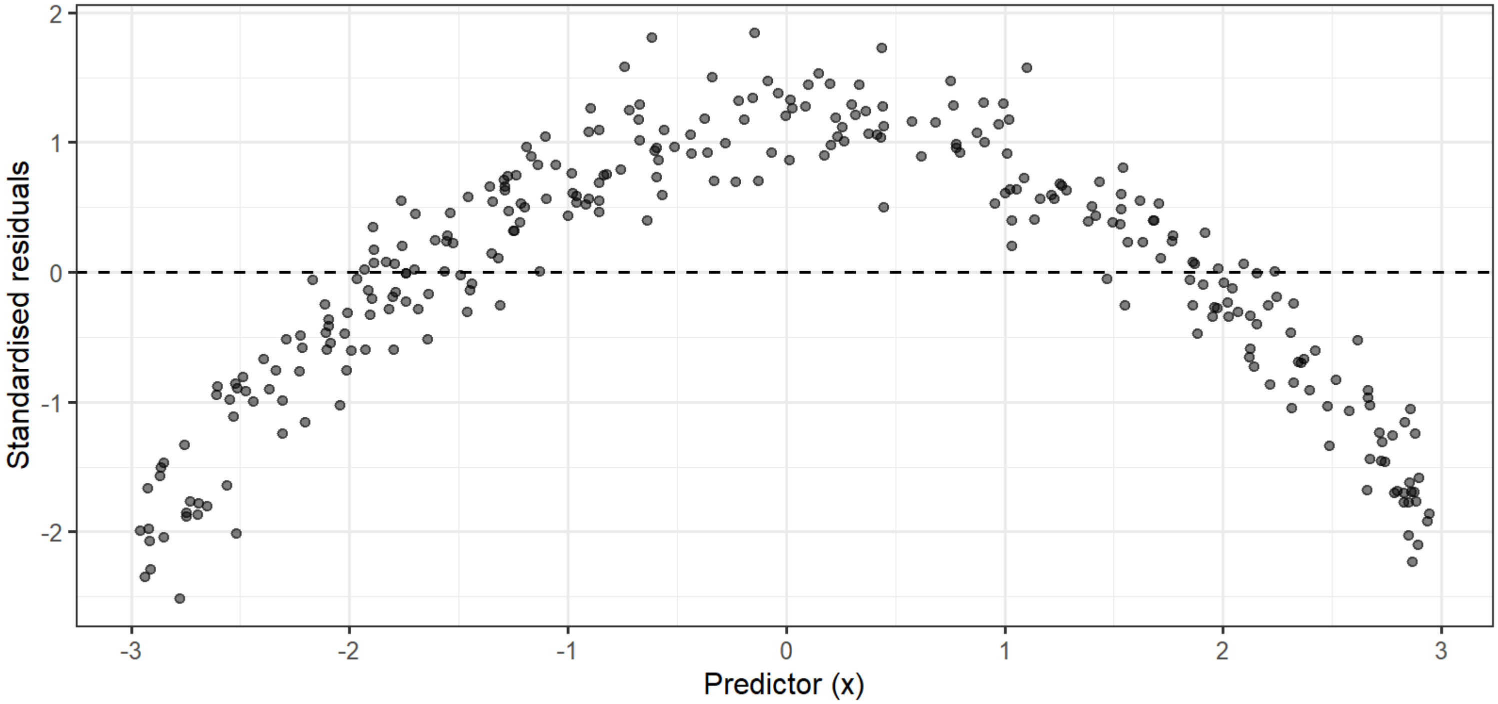
- Polynomial terms extend a linear predictor by adding powers of a variable, allowing curved relationships while keeping a simple parametric form.
- A quadratic term (x^2) captures a single bend (U- or inverted-U), while a cubic term (x^3) can capture an inflection; higher degrees add flexibility but also increase the risk of overfitting.
- Respect model hierarchy: if you include x2, you should also include xx; interpret the linear and quadratic coefficients jointly rather than in isolation.

```
glm(y ~ x + I(x^2), family = poisson)
```

When to Use Nonlinear Terms

- Use nonlinear terms when residual plots show systematic curvature or patterns that a straight line cannot explain.
- Let domain knowledge guide you: if biology suggests thresholds, peaks, or diminishing returns, a curved form is appropriate.
- Start with a simple polynomial (often a quadratic) when you expect a single bend, and centre the predictor to stabilise estimates.

When to Use Nonlinear Terms



Combining with Interactions

- Nonlinear interactions occur when the shape of a continuous effect changes across levels of another variable, not just its slope.
- For example, the quadratic temperature–abundance relationship may peak at different temperatures and have different curvature at different sites.
- Interpret the model with predicted curves for each site on the response scale rather than by reading individual coefficients, and retain the main effects to respect model hierarchy even if their p-values are small.

Caution: Overfitting

- Adding interaction or high-degree polynomial terms can inflate standard errors and make coefficient estimates unstable.
- Models that are too complex often learn noise in the training data and therefore predict poorly on new data.
- Prefer the simplest model that explains the pattern, and add complexity only with a clear biological or substantive rationale.
- Evaluate any added terms using information criteria (e.g., AIC).
- Ensure you have adequate sample size and coverage for the combinations of predictors; sparse cells make interaction estimates unreliable.

Diagnostic Plots

- Diagnostic plots help you assess whether model assumptions hold and where the fit is going wrong before you interpret coefficients.
- A residuals-versus-fitted plot should look like a random cloud around zero; visible curves, funnels, or group patterns suggest nonlinearity, heteroscedasticity, or missing interactions.
- A QQ-plot compares residual quantiles to a theoretical distribution; systematic departures from the line indicate violations of distributional assumptions (e.g., non-normal errors in linear models).
- For GLMs, choose an appropriate residual type (deviance, Pearson, or randomized quantile residuals).
- Use these plots to flag outliers and high-influence points and to guide targeted fixes (transforms, alternative families, added terms), then re-check diagnostics after refitting.

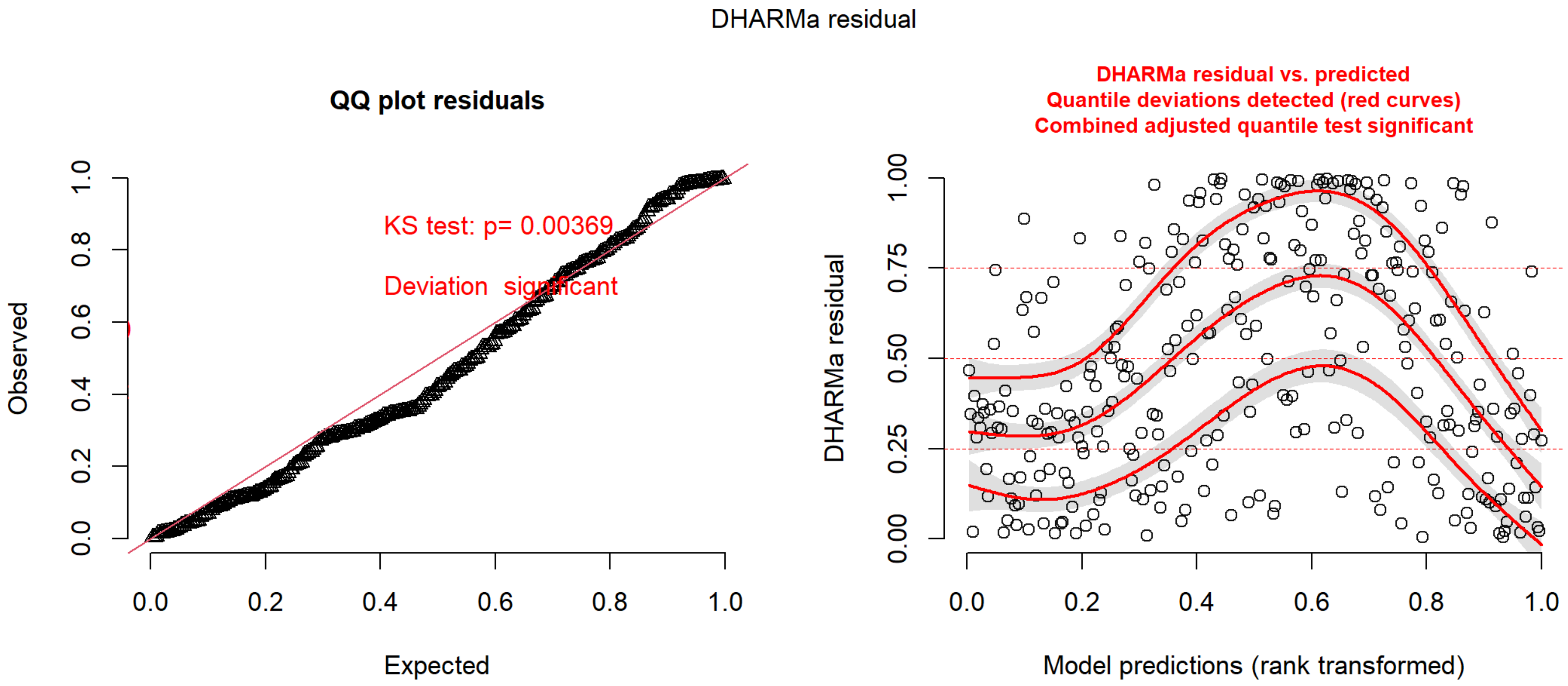
DHARMa Residuals

- Diagnostics for Hierarchical (Multilevel) Regression Models (DHARMa) provides simulation-based residuals that are valid for GLMs.
- Itsimulates datasets from the fitted model, ranks the observed response within those simulations, and converts that rank to a Uniform(0,1) residual. If the model is well specified, these residuals should be uniform and pattern-free.
- Use DHARMa for non-normal or discrete outcomes (counts, binomial, zero-inflated, over/under-dispersed) where ordinary residuals can be misleading.

DHARMa Residuals

- The default DHARMa plot shows a uniform QQ plot and residuals versus fitted values; additional functions plot residuals against predictors, leverage grouping factors, and check for temporal or spatial autocorrelation.
- Points close to the 45° line in the QQ-uniform plot and a flat residuals-vs-fitted cloud indicate a good fit.
- Systematic bends, funnels, or group patterns suggest misspecification, heteroscedasticity, or missing terms.

DHARMa Residuals



Outcome type	Residuals / Tools	Key plots and tests
Continuous (Normal)	Standardised residuals	Residuals vs fitted; Normal QQ; Cook's distance/leverage
Binary (Bernoulli)	Deviance/Pearson; DHARMa	Calibration plot; influence
Binomial (trials)	Deviance/Pearson; DHARMa	Residuals vs fitted; overdispersion check
Multinomial	DHARMa	Confusion matrix; IIA check
Ordinal	DHARMa	Residuals vs predictors; proportional-odds (Brant);
Count (Poisson)	Deviance/Pearson; DHARMa	residuals vs fitted; dispersion test
Overdispersed count (NB / quasi-Poisson)	DHARMa	residuals vs predictors; dispersion tests
Zero-inflated count	DHARMa	Zero-inflation test, dispersion test; Vuong test

Plotting Confidence Intervals

- Always plot uncertainty rather than only the fitted line, so the audience can see how precise - or imprecise - the estimate is across the predictor range.
- Confidence intervals show a range of plausible values for the mean response; without them, small wiggles in the line are easily overinterpreted as real effects.
- Use shaded ribbons for continuous curves and error bars for point estimates, clearly labelling the level (e.g., 95%) and plotting GLM results on the response scale for communication.

Marginal Effect vs. Raw Data

- The marginal effect curve shows the model's average predicted response as one predictor varies, averaging over the other variables (PDP).
- The raw data display the observed outcomes and their variability, including noise, confounding, and any imbalance in where observations occur.
- Plotting both together clarifies what the model believes versus what was observed, and highlights regions where the model is smoothing, extrapolating, or missing structure.
- Always add uncertainty bands to the marginal effect and show the predictor's distribution (rug or bins) so viewers don't overinterpret sparse areas.
- Remember that marginal effects describe model behaviour, not causality; large gaps between the curve and the raw data suggest misspecification or influential covariate correlations.

Avoiding Misleading Visuals

- Use clearly labelled and consistent axis scales so small differences are not exaggerated; for bar charts the y-axis should start at zero, while for lines/scatter a non-zero baseline can be acceptable if it is clearly indicated.
- Report group sample sizes (e.g., “n = 42”) in legends, captions, or facet titles so viewers can judge the reliability of each estimate.
- Mark extrapolation visibly—fade or shade regions beyond the observed data range—and avoid drawing confident ribbons where there are no data.
- Always display uncertainty (confidence or prediction intervals) rather than only a fitted line, and state the interval level in the caption.
- Keep scales, limits, and colour mappings consistent across facets so slopes and magnitudes are comparable.