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 Data

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.

Binary Data

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

$$P(success) = P\left(Y_{ij} = 1\right) = \pi_{ij}$$
$$P(failure) = P\left(Y_{ij} = 0\right) = 1 - \pi_{ij}$$

Y_i has a Bernoulli distribution:

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

Why Not Use a Linear Model?

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

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

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.

Probability vs. Odds

- 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 compare the probability of success to the probability of failure.

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

- For example, if $\pi = 0.8$ then $odds = \frac{0.8}{0.2} = 4$

  (4 to 1 odds of success)

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

- If odds = 4, then $\pi = \frac{4}{1+4} = \frac{4}{5} = 0.8$

Interpreting Fixed-Effect Coefficients

$$logit(\pi) = \beta_0 + \beta_1 x = -0.5 + 1.2x$$

- $\beta_0 = -0.5$: when $x = 0$, the log-odds of success are -0.5.
- Odds = $e^{-0.5} \approx 0.607$: when $x = 0$, the odds of success are 0.61 to 1 (success is less likely than failure).
- $\pi = \frac{0.607}{1+0.607} \approx 0.38$: when $x = 0$, the probability of success is about 38%.

- $\beta_1 = 1.2$: for each 1-unit increase in x, the log-odds increase by 1.2.
- Odds ratio = $e^{1.2} \approx 3.32$: each 1-unit increase in x multiplies the odds of success by about 3.3.
- At $x = 1$: $logit(\pi) = -0.5 + 1.2(1) = 0.7 \rightarrow Odds = e^{0.7} \approx 2.01$
- $\pi = \frac{2.01}{1+2.01} \approx 0.67$: increasing x by one unit raises probability from 38% to 67%.

Assumptions: Independence of Observations

- Each Y_i is an independent Bernoulli trial given the predictors (no residual correlation across observations).
- If this assumption is violated, it will lead to underestimated SEs, too-small p-values, overconfident CIs, misleading inference.

Common violations:

- Clusters/groups: students within classes, animals within herds, patients within hospitals
- Repeated measures: multiple rows per subject over time
- Spatial/temporal autocorrelation: nearby in space/time more similar than distant

Assumptions:
No Perfect Multicollinearity

- Predictors should not be perfectly correlated with each other.
- High collinearity leads to unstable coefficients and inflated standard errors.
- Check Variance Inflation Factor (VIF), and drop redundant predictors if needed. The VIF measures how much the variance of a regression coefficient is increased due to correlation with other predictors. A higher VIF means the predictor is more redundant with others. A rule of thumb is:
  - VIF > 5 → moderate collinearity
  - VIF > 10 → high collinearity
- Collinearity doesn’t bias the model, but makes it hard to interpret.

Assumptions: Sample Size

- Logistic regression needs enough events for stable estimates.
- The rule of thumb is ≥10 events per variable.
- For example, if we have 5 predictors, we need at least 50 events.

- Why?
- Small sample sizes can lead to large standard errors (wide confidence intervals), unstable odds ratios (sensitive to small data changes), and convergence issues (failure to estimate coefficients or infinite estimates).

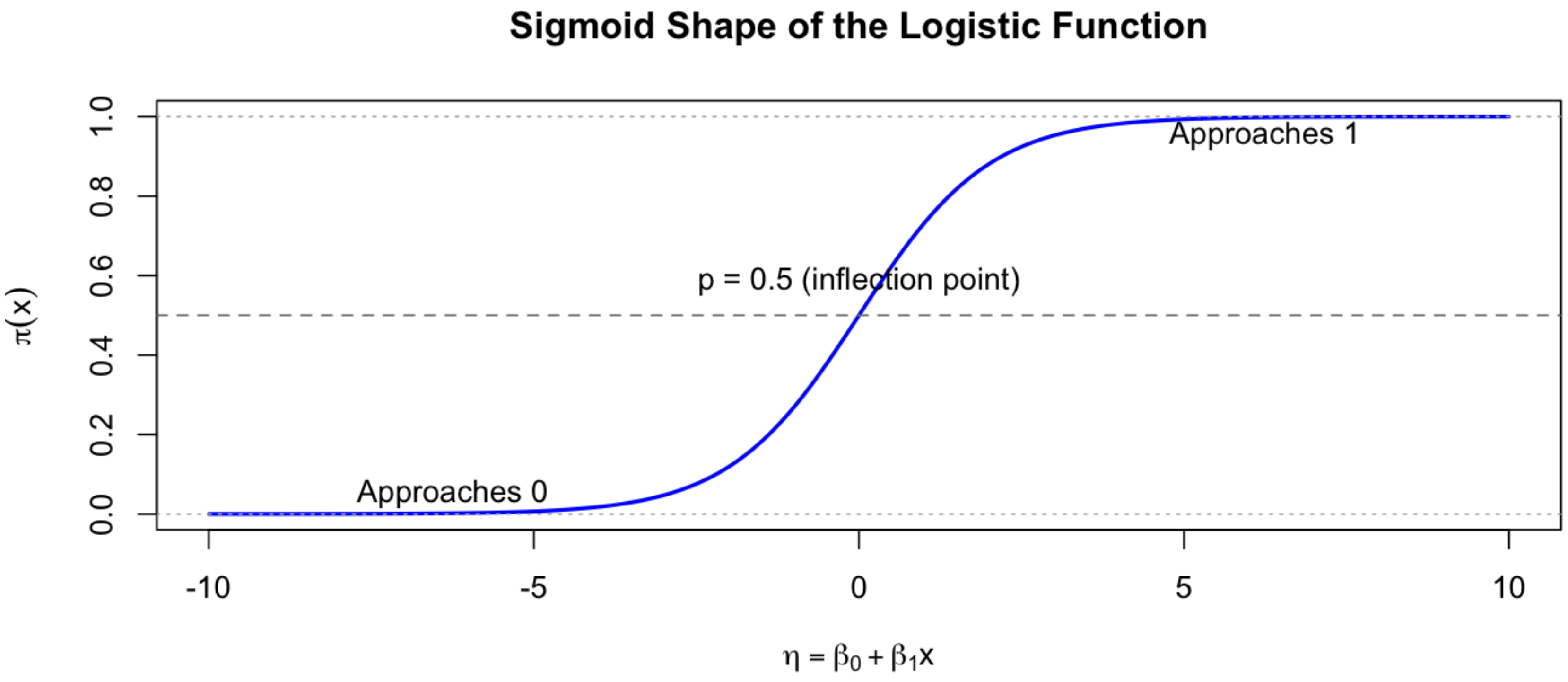
The Logistic Function

- 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



Other Link Functions

- The logistic function is most commonly used for binary data, but it isn’t the only choice. Other choices include:
  - Probit (normal CDF): assumes the latent propensity follows a normal distribution. It tends to give very similar results to logit but is sometimes preferred in fields like toxicology, psychometrics, or genetics where a normal latent variable is a natural assumption.
  - Complementary log-log (cloglog): asymmetric S-shape - changes faster near 0 than near 1. It’s useful when the probability of an event increases rapidly and then levels off (e.g. survival models, time-to-event data, rare events).
- All are S-shaped, but with subtle differences

Example:
Disease Presence

- We want to estimate how an animal’s age relates to the probability of disease presence and produce usable, age-specific risk estimates.

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

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

- Predictor: Age in years (we’ll centre at 2 years so the intercept is interpretable)
- Baseline odds of disease presence at age 2 years: e^{β_0}
- Age effect: each additional year multiplies odds of disease by e^{β_1} e.g., a 5-year difference $\Rightarrow e^{5\beta_1}$ times the odds

Coding Demo

Nested Models

- Nested models are a pair of models where one is a special case of the other - i.e., the smaller model is completely contained within the larger model.
- The simpler model has fewer predictors, while the more complex model includes all the terms in the simple model plus additional predictors.

- If Model A is: $logit(\pi) = \beta_0 + \beta_1 x$
- And Model B is: $logit(\pi) = \beta_0 + \beta_1 x + \beta_2 z$

- Then Model A is nested inside Model B.

Nested Models

- Comparing nested models lets us test whether the extra variables in the larger model significantly improve the model's fit.
- It helps us answer the question: "Do these new predictors help explain the outcome, beyond what we already had?"

- We can answer this question using the Likelihood Ratio Test, which compares how well each model explains the data, using the difference in log-likelihoods.

$$G^2 = -2(\log L_A - logL_B)$$

Coding Demo

Analysing Model Fit

- Fitting a model is only the beginning.
- We need to verify that:
  - The model adequately describes the data.
  - The assumptions are not violated.
  - No single observation unduly influences results.

Model Comparison: Pseudo-R^2 Measures

- Logistic models use pseudo-R^2 measures to approximate explanatory power:

McFadden's R^2:

$$R^2 = 1 - \frac{logL_{Model}}{logL_{null}}$$

- McFadden's R^2 compares the likelihood of the fitted model against the null (intercept-only) model. Values closer to 1 indicate better improvement over the null, but they are usually much lower than R^2 in linear regression. 0.2-0.4 already signals a well-fitting logistic model.

Model Comparison: Deviance

- Deviance measures how well the model fits the data.
- It's based on the log-likelihood of the model.

$$Deviance = -2(logLikelihood - logLikelihood\ of\ saturated\ model)$$

- A lower deviance indicates a better fit.
- Deviance compares the likelihood of your model to a saturated model (a model that fits the data perfectly).
- Deviance differences between models can be used for hypothesis testing: comparing the deviance of a full model and a reduced model. A large drop in deviance means the added predictors improve model fit significantly

Model Comparison: Deviance

In the model output, you often see:

- **Null deviance**: Fit of the intercept-only model (no predictors).
- **Residual deviance**: Fit of the model with predictors.

$$Improvement = Null\ deviance\ - Residual\ deviance$$

If residual deviance is much lower than null deviance, your predictors are explaining the variation in the response, and adding predictors improves the model fit compared to the intercept-only model.

Model Comparison:
AIC

AIC (Akaike Information Criterion) helps compare non-nested models or models with different numbers of predictors.

$$AIC = -2(\log Likelihood) + 2k$$

Where, the log likelihood measures how well the model fits the data, and k = number of parameters (complexity penalty).

Lower AIC is better.
It balances:

- **Fit** (how well the model explains the data)
- **Parsimony** (fewer predictors is better if fit is similar)

Residuals

- Residuals measure the difference between observed and predicted outcomes. In logistic regression, residuals are calculated from the difference between the observed outcome (0/1) and the fitted probability.
- Logistic regression uses several types of residuals:
  - Deviance residuals
  - Pearson residuals
  - Standardised residuals
- Each highlights different types of model issues.

Deviance Residuals

- Deviance residuals measure how much each observation contributes to the model's deviance - a larger residual means the observation is poorly explained by the model.
- Because outcomes are 0/1, deviance residuals naturally fall into two curved bands:
  - Observations with $y = 0$ have negative residuals.
  - Observations with $y = 1$ have positive residuals.
    This banding is expected and not, by itself, a sign of misfit.
  - Large residuals (in either band) highlight individual observations the model predicts poorly.
- What matters is whether there are patterns within or across the bands:
  - A systematic curve or clustering may suggest missing predictors, a wrong link function, or a nonlinear effect not captured by the model.
  - A few isolated large points may suggest outliers.
- In a well-fitting model, the two bands should look roughly balanced around zero with no extra structure. Strong asymmetry, curvature, or groups of unusually large residuals indicate possible model misspecification.

Pearson Residuals

- Pearson residuals are analogous to residuals in linear regression.

$$r_i = \frac{y_i \; - \pi_i}{\sqrt{\pi_i(1 - \pi_i)}}$$

- They measure the difference between observed and expected values. They can be used to assess overall model fit.
- Large Pearson residuals highlight observations that deviate strongly from the model's predictions.
- In logistic regression, the response is binary, so residuals cannot look like a cloud around 0. Instead, the Pearson residuals will always fall into two curved bands (one for observed 0's, one for observed 1's). A "good" model is one where:
  - The bands are symmetrical around 0 (no systematic bias toward positive or negative).
  - The spread of residuals matches the variance implied by the model (no evidence of overdispersion).
  - There are no obvious trends with fitted values or predictors.

Standardised Residuals

- Standardised residuals are residuals divided by their estimated standard deviation.
- This rescales residuals to be approximately comparable across observations. Raw residuals aren’t comparable because their variance differs across observations (due to heteroscedasticity and leverage).
- Useful for detecting outliers and influential data points.
- Values greater than |2| suggest potential outliers.
- Values greater than |3| are often considered highly unusual.

When to Use Each Residual

- **Deviance residuals:** best for checking individual fit and influence.
  - They are derived from the model's deviance and reflect how much each observation contributes.
  - Large absolute values suggest poor fit or potential leverage points.
- **Pearson residuals:** helpful for overall goodness-of-fit statistics.
  - They are based on the difference between observed and expected counts, scaled by variance.
  - Useful for testing model adequacy in aggregate.
- **Standardised residuals:** great for detecting outliers.
  - They rescale residuals to account for differing variances across observations.
  - Values beyond |2| or |3| may signal unusual observations worth investigating.
- No single residual tells the full story – use multiple diagnostics together.

Leverage

- Leverage measures how far an observation's predictor values are from the average predictor values.
- High-leverage points have unusual predictor values (not necessarily unusual response values).
- These points can exert strong influence on the model's fitted values.
- Examining leverage helps identify observations that disproportionately affect regression estimates.
- As a rule of thumb, leverage values greater than 2p/n (where *p* is the number of predictors including the intercept, and *n* the sample size) may be considered high.

Influence

- Influence considers both leverage (unusual predictor values) and residual size (poor fit).
  An influential point is one that, if removed, would substantially change the model.
- Influence is assessed using measures like Cook's Distance.
- We look at influential points because they can disproportionately affect parameter estimates and conclusions.
- They are not automatically "bad" - they may reflect real structure or unusual but valid cases.
- Best practice is to investigate them: check for data errors, assess context, and compare models with and without them.

Cook’s Distance

- Measures the influence of a single observation on all fitted values.
- Combines residual size (fit) and leverage (position in predictor space).
- An observation with a large residual and a high leverage will have a high Cook’s Distance.
- Interprets how much model coefficients would change if that observation were removed.
- Rule of thumb: values > 0.5 or 1 may indicate influential points.
- Should not be used mechanically - investigate influential cases for possible data issues or meaningful outliers.

Coding Demo