

Introduction to Generalised Linear Models for Ecologists

Dr Niamh Mimmagh

niamh@prstats.org

[https://github.com/niamhmimmagh/GLME01---
Introduction-to-Generalised-Linear-Models-for-
Ecologists](https://github.com/niamhmimmagh/GLME01---Introduction-to-Generalised-Linear-Models-for-Ecologists)

Count Data

- Count data are non-negative integers that often have a skewed distribution and a variance that increases with the mean.
- Examples:
 - The number of birds observed in a plot
 - The number of disease cases reported per day
 - The number of accidents per year.
- Using standard linear regression for such data can lead to nonsensical predictions (e.g. negative counts), and incorrect inferences because linear regression assumes constant variance and normally distributed errors.
- Poisson regression solves these issues by explicitly modelling the distribution of counts.

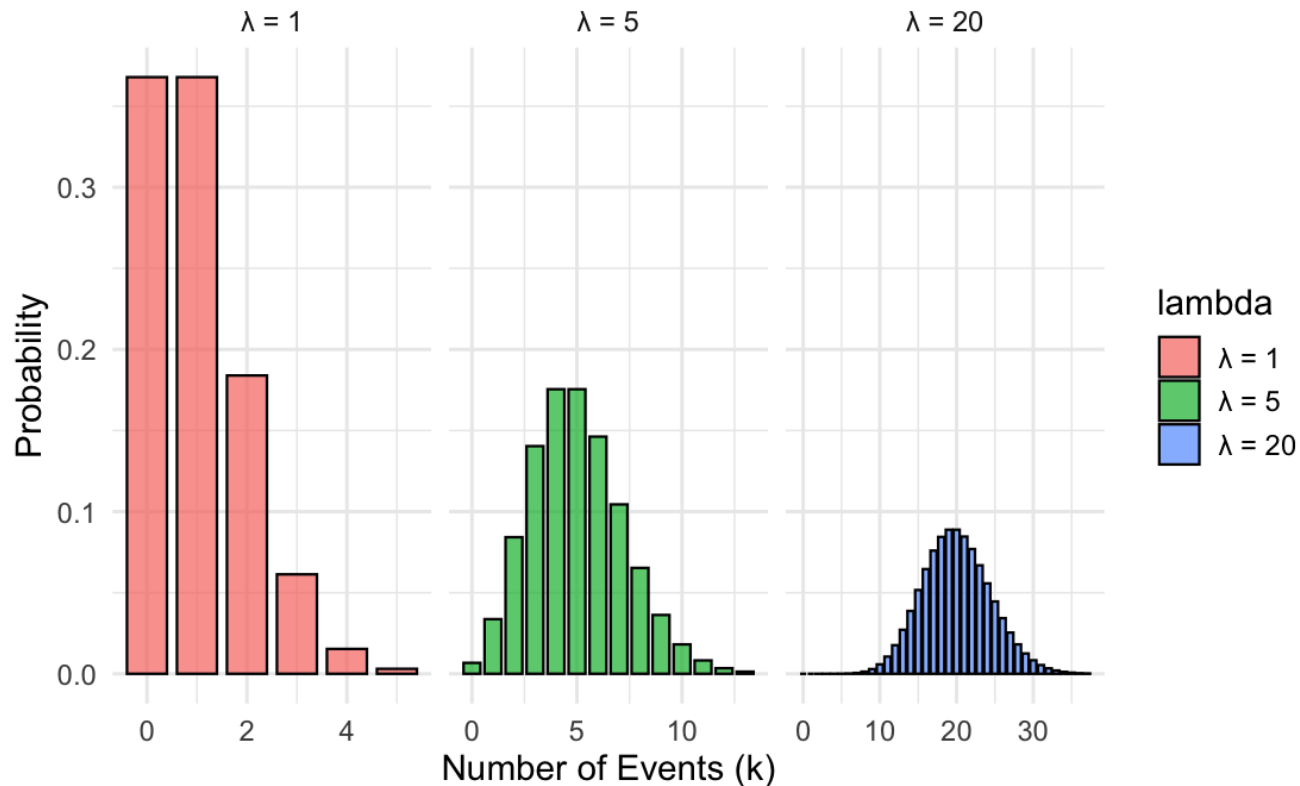
Why Not Use Linear Regression?

- Count data violates the assumptions of linear regression.
- Linear regression assumes that residuals are normally distributed and have constant variance.
- When applied to count data, linear models can predict negative values, which are meaningless for counts.
- Count data typically exhibit heteroscedasticity, where the variance grows with the mean. We need a model that respects the nature of count data.

The Poisson Distribution

- The Poisson distribution gives the probability of observing a count (0, 1, 2, ...) of events in a fixed time or space window.
- It assumes events occur independently and at a constant average rate λ (equivalently, counts in non-overlapping intervals are independent).
- A key property of this distribution is equidispersion: the mean and variance are both λ .
- Use it for rare-event counts (e.g., calls per minute, nests per plot).
- If the variance \gg mean or there are many zeros, consider a negative binomial or zero-inflated model.

The Poisson Distribution



The Poisson Distribution

- When λ is small the Poisson distribution is strongly right-skewed, with most of the probability at zero.
- As λ increases to around five, the distribution becomes less skewed, and counts spread out and cluster around the mean λ .
- For large λ values (about 20 or more), the distribution looks nearly symmetric and is well approximated by a Normal distribution with mean and variance equal to λ .
- In general, skewness decreases as λ grows, and the most likely count is roughly λ rounded down.

Poisson Regression as a GLM

- We model each count Y_i as Poisson with mean λ_i .

$$Y_i \sim \text{Poisson}(\lambda_i)$$

- The model links the mean λ_i to a linear predictor through the natural log function:

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots \beta_p x_{pi}$$

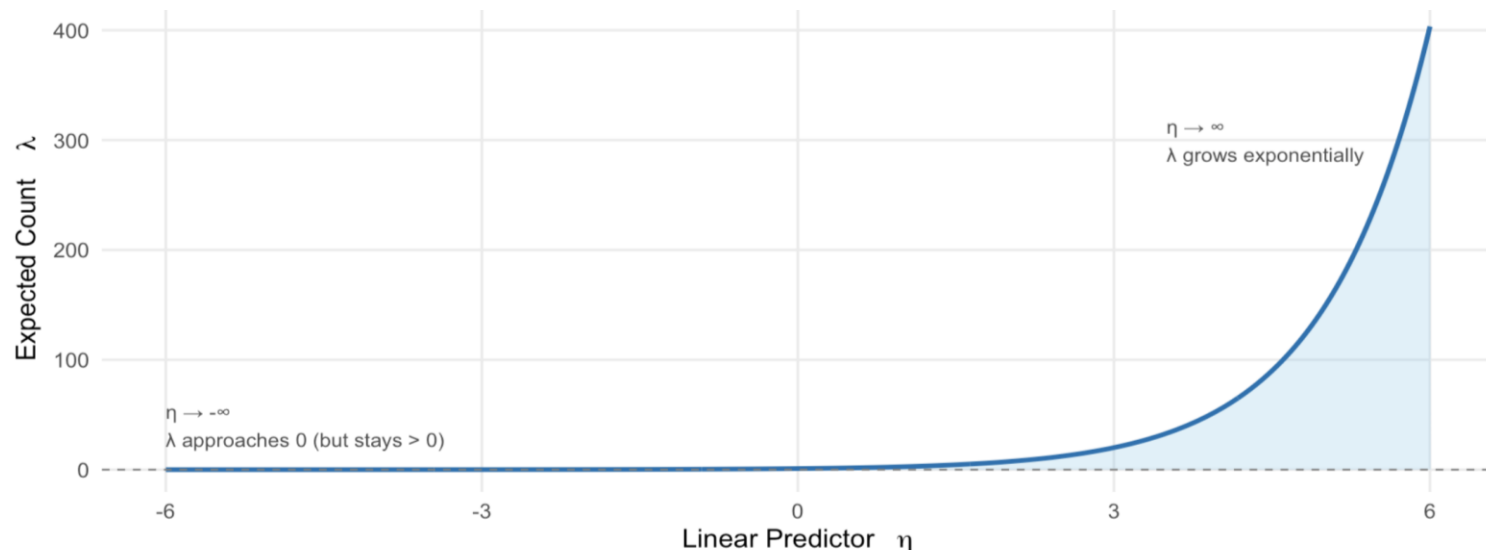
- This log link keeps λ_i positive and treats effects additively on the log scale (multiplicative on the count scale).

The Log Link Function

- The log link function transforms the linear predictor into a positive expected count. It does this by exponentiating the linear predictor.

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{1i} \rightarrow \lambda_i = e^{\beta_0 + \beta_1 x_{1i}}$$

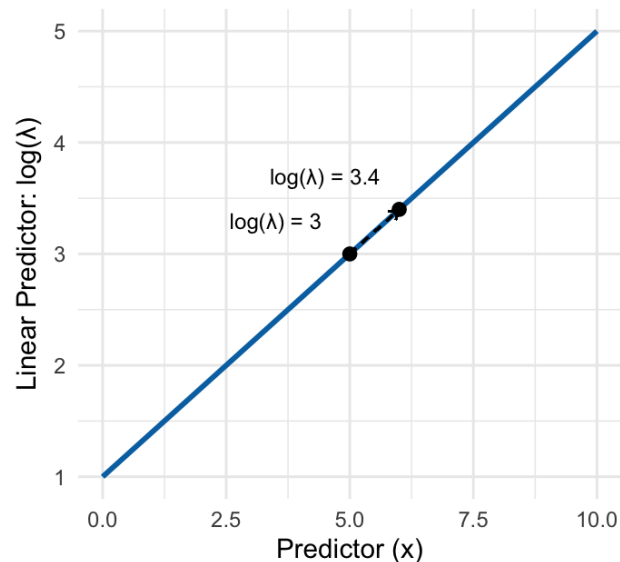
- For any real number, the predicted mean count is always positive.



The Log Link Function

Additive Effect on Log Scale

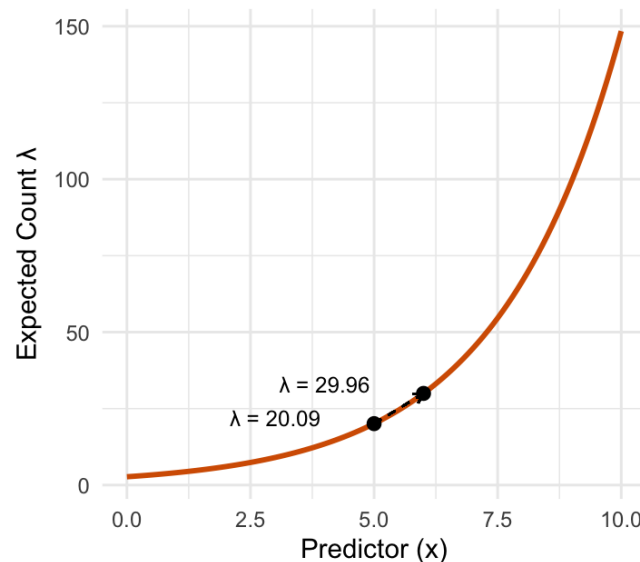
$$\log(\lambda) = \beta_0 + \beta_1 x$$



Adding 0.4 to 3 gives 3.4

Multiplicative Effect on Original Scale

$$\lambda = \exp(\beta_0 + \beta_1 x)$$



Multiplying 20.09 by $e^{0.4}$ gives 29.96

The Log Link Function

- In a Poisson regression with a log link:

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{1i}$$

- On the log scale, a one-unit increase in x_1 simply adds β_1
- On the original scale, taking the exponential gives:

$$\lambda_i = e^{\beta_0 + \beta_1 x_i} = e^{\beta_0} e^{\beta_1 x_i}$$

- So, each 1-unit increase in x multiplies λ by e^{β_1} .
- This impacts the way that we interpret the coefficients of a Poisson model.

Interpreting Coefficients

- β_1 is the slope on the log scale. It is the log rate ratio (sometimes called the log incidence rate ratio). A 1-unit increase in x adds β_1 to $\log(\lambda)$
- e^{β_1} is called the rate ratio. It tells you how many times larger (or smaller) the expected count is for a one-unit increase in x .
- If $\beta_1 = 0.5$, then the rate ratio is $e^{0.5} \approx 1.65$, meaning there is a 65% increase in the expected count, per unit increase in x
- If $\beta_1 = -0.3$, then $e^{-0.3} \approx 0.74$, meaning that for a one unit increase in x , we multiply the expected count by 0.74. (there is a 26% decrease in the expected count, per unit increase in x)
- β_0 is the intercept on the log scale. It is the expected count when all predictors are 0 (or at average predictor values, if they have been centered).

Example

- Consider an example where the count of bird species depends on habitat area.
- Y_i is the number of bird species observed at site i
- $Area_i$ is the habitat area for site i in hectares.
- The response variable Y_i follows a Poisson distribution

$$Y_i \sim \text{Poisson}(\lambda_i)$$
$$\log(\lambda_i) = \beta_0 + \beta_1 Area_i.$$

- Each additional hectare of habitat changes the expected count of bird species by a multiplicative factor of e^{β_1}

Coding Demo

Poisson Regression Assumptions

1. Each observed count is assumed to arise from an independent process.

- Violations:
 - Temporal or spatial autocorrelation
 - Repeated measures on the same unit

2. The expected mean equals the variance (equidispersion):

$$E[Y_i] = \lambda_i,$$
$$Var(Y_i) = \lambda_i$$

- The Poisson distribution assumes variance grows in lockstep with the mean.

Adjusting for Exposure

- In many real-world datasets, not all observations are equally exposed.
- This could be due to:
 - Unequal observation windows (e.g. different follow-up times)
 - Different population at risk (cases per person-year)
 - Variable measurement effort (e.g. number of traps checked)
- For example, bird surveys may be of different lengths, disease counts may be over different population sizes, or accident counts may vary by the number of kilometres travelled.
- If we ignore this, we may draw misleading conclusions about the underlying rates.

Counts vs. Rates

- Sometimes we model rates instead of raw counts.
- Counts are the number of events observed in a fixed space, time or population (e.g., 10 bird species recorded at a certain site). These are modelled as raw counts Y_i
- Rates are counts that are standardised by an exposure variable, such as time or area (e.g., bird species per hectare, or cases per year).

$$Rate_i = \frac{Count_i}{Exposure_i}$$

- We model rates because different observations may have different levels of exposure. If you just model counts, a longer observation time or larger area will naturally have more events, and this can confound the effect of interest.

Offsets

- An offset is a known quantity that we include in the model with a fixed coefficient of one.
- In Poisson regression, the offset is typically the log of the exposure variable. Including it in the linear predictor adjusts the expected counts proportionally to the exposure.
- Using an offset turns the model into a rate model: we explain counts per unit exposure (e.g., per trap-night, per km surveyed, per hour observed).
- A site watched twice as long is expected to have twice the count, all else equal.
- Practical notes: offsets must be positive and entered untransformed except for the log (don't centre/scale them).

How Poisson Regression Handles Rates

- Standard Poisson regression for counts:

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$$

- For rates, we include the log of exposure E_i as an offset:

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \log(E_i)$$

- Rearranging:

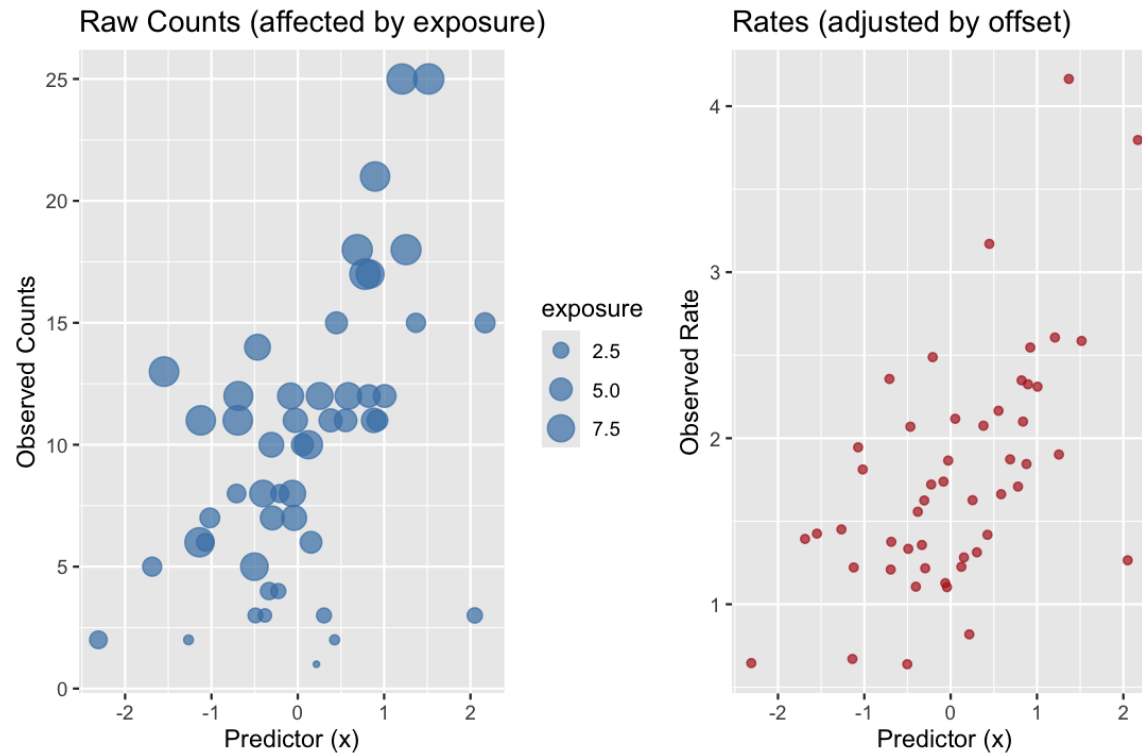
$$\log\left(\frac{\lambda_i}{E_i}\right) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$$

- So we're modelling the rate directly:

$$Rate_i = \frac{\lambda_i}{E_i} = e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}$$

- The offset forces the expected count to scale linearly with exposure

Visualising Offsets



Interpreting Coefficients with an Offset

- Without an offset: coefficients describe a multiplicative effect on the raw counts.
- With an offset: coefficients describe a multiplicative effect on the rate.
- Intercept β_0 : baseline log rate per unit exposure at the reference covariate levels.
- e^{β_0} = expected rate (e.g., animals per hour).
- β_1 (continuous x_j): a one-unit increase in x_j multiplies the rate by e^{β_1}
- β_1 (factor x_j): e^{β_1} is the rate ratio vs the reference level (per unit exposure).

Example 1

- Suppose you're counting the number of bird species in each site, but site areas differ wildly (1-50 hectares).
- Larger sites will naturally have more species simply due to the greater area. Rather than modelling species counts, you model rate per hectare by including an offset:

$$\log(\lambda_i) = \beta_0 + \beta_1 \text{HabitatQuality}_i + \log(\text{Area}_i)$$

- Which is equivalent to:

$$\log\left(\frac{\lambda_i}{\text{Area}_i}\right) = \beta_0 + \beta_1 \text{HabitatQuality}_i$$

- Now, β_1 affects the density of species per hectare, not just raw counts

Example 2

- Imagine you're studying disease across different animal populations. Y_i is the number of infected individuals in population i .
- Clearly, larger populations can be expected to have more infected individuals, simply because there are more individuals present to become infected.
- In this case, it makes sense to examine the infection rate per individual, rather than just counts of infected individuals

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{1i} + \log(Pop_i)$$

$$\log\left(\frac{\lambda_i}{Pop_i}\right) = \beta_0 + \beta_1 x_{1i}$$

- So the model is actually for the infection rate per individual

Coding Demo

When a Single Offset Isn't Enough

- Sometimes more than one exposure factor affects how counts scale
- Example scenarios:
 - Accident counts depend on both time on the road AND distance travelled
 - Wildlife survey counts depend on survey duration AND area covered
 - Hospital infections depend on patient-days AND bed capacity
- We may need to include multiple offset terms to properly standardise rates

Example: Road Accidents

- Modelling the number of road accidents in different regions
- Accident counts scale with:
 - Population size (more people leads to more accidents)
 - Vehicle kilometres travelled (more driving leads to more accidents)
- Ignoring one exposure risks biasing covariate effects
- Solution: include both $\log(\text{Population})$ and $\log(\text{distance travelled})$ as offsets

Multiple Offsets in the Linear Predictor

- Offsets add linearly inside the log link function:

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \log(E_{1i}) + \log(E_{2i})$$

By log rules: $\log(E_{1i}) + \log(E_{2i}) = \log(E_{1i} \times E_{2i})$

So multiple offsets combine multiplicatively: $\lambda_i \propto E_{1i} \times E_{2i}$

- You're still modelling a rate, but per unit of BOTH exposures

Example: Wildlife Survey

- Imagine we are counting animals in plots of varying size and survey duration.
- Expected count should scale with both area surveyed (we expect to see more animals if we're examining a larger habitat) and survey effort (the longer we survey, the more likely we are to see animals)

$$\log(\lambda_i) = \beta_0 + \beta_1 \text{HabitatQuality}_i + \log(\text{Area}_i) + \log(\text{Effort}_i)$$

$$\log\left(\frac{\lambda_i}{\text{Area}_i \times \text{Effort}_i}\right) = \beta_0 + \beta_1 \text{HabitatQuality}_i$$

- β_1 affects density per hectare per hour, not raw counts

Interpreting Multiple Offset Models

- β_0 is the baseline log density (animals per hectare-hour) at the reference level of habitat quality.
- e^{β_0} is the expected density when habitat quality is at its reference (e.g., 0 for continuous variables or baseline category).
- β_1 interprets the effect on density per hectare per hour.
- If habitat quality is categorical, e^{β_1} is the density (rate) ratio for that category vs the reference category (e.g., density in “High” habitat compared to “Low”), holding other predictors and offsets constant.
- If habitat quality is continuous, A one-unit increase in habitat quality multiplies density by e^{β_1} .

What an Offset Really Does

- An offset is a known adjustment with a fixed coefficient of 1.
- It says: 'If the exposure doubles, the expected count also doubles.'
- There is no uncertainty: the effect is assumed perfectly proportional.

$$\begin{aligned}\log(\lambda_i) &= \beta_0 + \beta_1 x_i + \log(E_i) \rightarrow \\ \lambda_i &= E_i \times \exp(\beta_0 + \beta_1 x_i)\end{aligned}$$

- E_i is NOT estimated. It just scales the expected mean.

What a Predictor Really Does

- A predictor is treated as an unknown effect estimated from data.
- Instead of forcing the coefficient to be 1, you estimate it:
$$\log(\lambda_i) = \beta_0 + \beta_1 x_i + \beta_2 \log(E_i)$$
- β_2 can be 1 (proportional) but may be $\beta_2 < 1$ or $\beta_2 > 1$
- This lets you test if exposure has a different-than-proportional effect.

When Should it Be an Offset?

Use an offset if:

- It's a known property of the process that counts scale exactly with exposure
 - More time → proportionally more events
 - Larger population → proportionally more infections
- You're not interested in testing its effect, just standardising for it

Example: Counting accidents per 10,000 km → km is just exposure

When Should it Be a Predictor?

Use a predictor if:

- You want to test whether the relationship is truly proportional
- You suspect nonlinear or non-proportional scaling
- It's a variable of scientific interest, not just nuisance scaling

Example: Does larger hospital size change infection risk per patient-day?

Examples

Disease counts:

- Offset population if just comparing infection rates per individual
- Predictor population if testing herd immunity effects

Bird surveys:

- Offset survey hours if assuming linear effort → counts
- Predictor if longer surveys yield diminishing returns

Road accidents:

- Offset vehicle km for rates per km
- Predictor if more traffic volume alters risk per km

Model Fitting in R

```
glm(cases ~ covariate, family = poisson, offset =  
log(time) + log(population))
```

Equivalently:

```
glm(cases ~ covariate, family = poisson, offset =  
log(time * population))
```

Coding Demo

Workflow for Practical Poisson Modelling

The practical workflow is:

1. Start with raw counts.
2. Identify appropriate exposure.
3. Include $\log(\text{exposure})$ as an offset.
4. Fit the model.
5. Visualise and check fit.

This ensures fair comparisons and valid inference.

Common Pitfalls

- Forgetting some exposure components - e.g., including $\log(\text{area})$ but forgetting $\log(\text{time})$, so you model counts per hectare, not per hectare-hour.
- Treating exposure as a covariate instead of an offset – the model estimates a slope for exposure (not fixed at 1), breaking the rate interpretation and biasing effects.
- Double-counting exposure – don't use an offset and divide the response by exposure, or include the same component twice (e.g., time in days and hours).
- Misinterpreting coefficients as effects on counts - with offsets you're modelling rate. Predicted counts still scale with exposure.

Overdispersion

- If the variance is greater than the mean, this is called overdispersion, and violates the Poisson assumption of equidispersion. This can lead to underestimated standard errors, and inflated Type I errors.
- Possible causes of overdispersion are:
 - Unobserved heterogeneity (missing covariates)
 - Clustering or repeated measures
 - Zero inflation
- We can check for overdispersion by comparing residual deviance to the residual degrees of freedom.

$$\varphi = \frac{\text{Residual Deviance}}{\text{Residual } df}$$

- $\varphi \approx 1 \rightarrow$ No overdispersion
- $\varphi > 1 \rightarrow$ Overdispersion