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

Zero Counts

- Many real-world count data show more zeros than typical count models expect.
- Examples include:
  - Ecology: survey sites without a rare species
  - Public health: individuals with zero doctor visits
  - Insurance: policyholders with no claims
- A standard Poisson model predicts the proportion of zeros as $P(Y=0)=e^{(-\lambda)}$, but often the observed proportion of zeros is much larger.

Zero-Inflation

- Zero-inflation occurs when the number of zero-counts in a dataset are larger than can be accounted for by typical models. Observed counts often show far more zeros than a Poisson distribution would predict. This mismatch suggests an extra process is generating zeros, beyond random chance.

- Zero-inflation occurs when two processes generate zeros:

1. Structural zeros: the event truly cannot occur (e.g., a pond with no fish)
2. Sampling zeros: the event could occur but did not (e.g., no fish caught despite fish being present).

- Zero-inflated models explicitly model both sources.

What if We Ignore Zero-Inflation?

- If you fit a standard Poisson model to zero-inflated data:
  - The model underestimates the frequency of zeros.
  - The variance appears too large (overdispersion).
  - Standard errors for covariates are biased.
  - Predictions are misleading, especially for low counts.
- Therefore, zero-inflated models are essential when an additional zero-generating mechanism exists.

Standard Count Models Recap

- Poisson
  - $Y_i \sim Poisson(\lambda_i)$
  - $E[Y_i] = \lambda_i$
  - $Var(Y_i) = \lambda_i$
  - Good when variance ≈ mean and zeros are not excessive.
- Negative Binomial
  - $Y_i \sim Negative\ Binomial\ (\lambda_i, \theta)$
  - $E[Y_i] = \lambda_i$
  - $Var(Y_i) = \lambda_i + \frac{\lambda_i{}^2}{\theta}$
  - θ controls extra-Poisson variation (smaller θ means more dispersion).
  - As $\theta \rightarrow \infty$ the NB approaches the Poisson.

Zero-Inflated Models: Two-Part Thinking

- If our data contains more zeros than expected under other count models, we say that there is an extra zero-generating process at play, that is not being accounted for. Some systems produce zeros for two different reasons.

1. Always-zero (structural) group: units that cannot generate counts at all (e.g., no host plants at a site, trap not deployed, unsuitable habitat).
2. Sampling zero group: units that could generate counts but happened to be zero this time by chance.

- Zero-inflated models assume:
  – A Bernoulli trial decides if the observation is in the always-zero group.
  – Otherwise, the count is drawn from a Poisson or Negative Binomial
- So total zeros = Structural zeros + Random zeros from count distribution.

Zero-Inflated Poisson (ZIP)

- A zero-inflated Poisson model is a mixture model.
- A Bernoulli distribution decides whether the observation is a structural zero (always zero, not susceptible to counts at all)
- If its not a structural zero, then the observation follows a standard Poisson distribution, which can itself produce zeros (sampling zeros) or positive counts.

$$zero_i \sim Bernoulli(\pi_i)$$
$$logit(\pi_i) = \gamma_0 + \gamma_1 z_{1i} + \cdots + \gamma_q z_{qi}$$
$$count_i \sim Poisson(\lambda_i)$$
$$\log(\lambda_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{p1}$$
$$Y_i = \begin{cases} 0, & if\ zero_i = 1\ (with\ probability\ \pi_i) \\ count_i, & if\ zero_i = 0\ (with\ probability\ 1 - \pi_i) \end{cases}$$

In short,

$$Y_i \sim ZIP(\pi_i, \lambda_i)$$

Zero-Inflated Negative Binomial (ZINB)

- If the data exhibit both excess zeros and overdispersion, a ZINB is more appropriate than ZIP.

$$zero_i \sim Bernoulli(\pi_i)$$
$$count_i \sim negative\ binomial(r, \mu_i)$$
$$Y_i = \begin{cases} 0, & if\ zero_i = 1\ (with\ probability\ \pi_i) \\ count_i, & if\ zero_i = 0\ (with\ probability\ 1 - \pi_i) \end{cases}$$

In short,

$$Y_i \sim ZINB(r, \pi_i, \mu_i)$$

ZIP vs. ZINB

ZIP

- Use when Poisson fit shows too many zeros but mild dispersion and DHARMa dispersion test OK.

ZINB

- Use when NB beats Poisson on AIC; DHARMa dispersion test fails for ZIP but passes for ZINB.

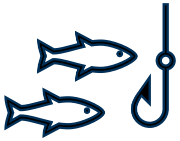
How to compare models

1. AIC/BIC
2. Vuong test
3. Always check DHARMa residuals (uniformity/QQ, dispersion, zero-inflation tests) to confirm the winner actually fits.

Interpreting Coefficients

- The coefficients from the count model (using a log link) are interpreted the same way we interpret coefficients for the Poisson or negative binomial model.
- For example, if $\beta_1 = 0.3$, then a one unit increase in the predictor x increases the expected count (or rate if you're using an offset term) by a factor of $e^{0.3} \approx 1.35$, conditional on being in the count process
- The coefficients from the zero model (using a logit link) are interpreted the same way we interpret coefficients for the binomial model.
- For example, if $\gamma_1 = 0.5$, then a one unit increase in the predictor increases the odds of being a structural zero/excess zero by a factor of $e^{0.5} \approx 1.65$

Example:
Fish in Lakes



- Let's say we're going fishing in multiple lakes over multiple days.
- Y_{ij}: number of fish caught in lake i on day j.
- Structural zeros: some lakes truly have no fish. We will never be able to catch any fish in those lakes.
- Sampling zeros: Maybe we aren't good at fishing, or there are issues with weather etc. so even fishy lakes still have days with 0 catch.
- We will include effort E_{ij} (e.g., hours netted) as an offset.
- We can use a ZIP or a ZINB model for this.

Coding Demo

Goodness of Fit: Vuong Test

- The Vuong test is used to compare two models M_1 and M_2 fitted to the *same data* that are non-nested (e.g., Poisson vs ZIP, NB vs ZINB). The test asks which model is closer to the data-generating process.
- For each observation i, compute the pointwise log-likelihood ratio

$$m_i = log\, f_1\left(y_i \middle| \hat{\theta}_1\right) - log\, f_2(y_i|\hat{\theta}_2)$$

$$V = \frac{\bar{m}\sqrt{n}}{s_m}$$

H_0: models are equally close in distance to the data generating process

H_a: One model is closer

$V > 1.96 \;\rightarrow$ favour M_1

$V < -1.96 \rightarrow$ favour M_2

Residual Diagnostics

- Standard residuals are tricky for zero-inflated counts.
- Counts are discrete and heteroscedastic → Pearson/deviance residuals look banded, skewed, and depend on the mean.
- Zero-inflated mixtures combine two processes (structural zeros + counts), so a single residual scale can hide misfit.
- Result: visual checks can be ambiguous; p-values based on Normality assumptions are unreliable.

DHARMa Residuals

- Simulate many replicate responses from the fitted model for each observation
- Compute the rank of the observed value within its simulated distribution
- Residuals are Uniform(0,1) under a correct model

- **Uniformity / QQ plot:** flat line indicates a good global fit; systematic deviation indicates misfit.
- **Residuals vs fitted plot**: patterns indicate the wrong mean/variance structure or missing terms.
- **Dispersion test:** detects over/under-dispersion in the count part.
- **Zero-inflation test**: remaining extra zeros beyond the model (even for ZIP/ZINB).

When to Use DHARMa

- Use DHARMa tests with likelihood-based, generative models: Poisson/Negative binomial GLMs, ZIP/ZINB, hurdle models, GLMMs
- These cannot be used for quasi models (quasi-Poisson/quasi-Binomial): as these models have no full likelihood, and DHARMa simulates from the likelihood, DHARMa cannot simulate correctly for quasi-models.
- Tip: If using pscl::zeroinfl(), refit in glmmTMB for DHARMa diagnostics.

Communicating Results

- State both processes
- Zero process: logit link; report odds ratios with CIs.
  "Altitude increasing by 100m multiplies odds of a lake being fishless by 1.35."
- Count process: log link; report rate ratios with CIs, and the offset unit.
  "+1 °C in temperature multiples catch rate by 1.20 per hour of effort."
- Report population-relevant effects
  "+1 °C increases expected catch by 0.42 fish per lake-day on average."
- Separate the probability of zero into parts:
  "Altitude mainly raises the structural-zero probability (π), not the sampling-zero part."

Common Misinterpretations

“$\boldsymbol{\pi}$ represents the proportion of zeros.”

- π is the probability of being in the always-zero state (given covariates).
- The observed zero rate also includes sampling zeros: $\text{Prop}(Y = 0) \neq \pi$ in general and varies with covariates.

“A covariate’s effect is the same in both parts.”

- A predictor may increase μ (more counts) while also increasing π (more structural zeros), or it may increase μ while decreasing π.

Hurdle Models

- Sometimes the process generating zeros is entirely separate from the process generating positive counts.
- Examples:
  - Doctor visits: Zero means ‘didn’t visit at all.’ Once you visit at least once, you can’t be zero anymore.
  - Technology adoption: First hurdle is the decision to adopt; only adopters have counts.
  - Species surveys (presence-abundance): If a species is observed at a site, their count cannot be zero.

- A hurdle/Zero-Altered Poisson (ZAP) model reflects this two-step process.

Hurdle Models

- A Bernoulli distribution decides whether the observation is a zero.
- If its not a zero, then the observation follows a truncated Poisson distribution, which cannot produce zeros.

$$zero_i \sim Bernoulli(\pi_i)$$
$$logit(\pi_i) = \gamma_0 + \gamma_1 z_{1i} + \cdots + \gamma_q z_{qi}$$
$$count_i \sim truncated\ Poisson(\lambda_i)$$
$$\log(\lambda_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{p1}$$
$$Y_i = \begin{cases} 0, & if\ zero_i = 1\ (with\ probability\ \pi_i) \\ count_i, & if\ zero_i = 0\ (with\ probability\ 1 - \pi_i) \end{cases}$$

In short,

$$Y_i \sim Hurdle(\pi_i, \lambda_i)$$

Hurdle Models

- A Bernoulli distribution decides whether the observation is a zero.
- If its not a zero, then the observation follows a truncated Poisson distribution, which cannot produce zeros.

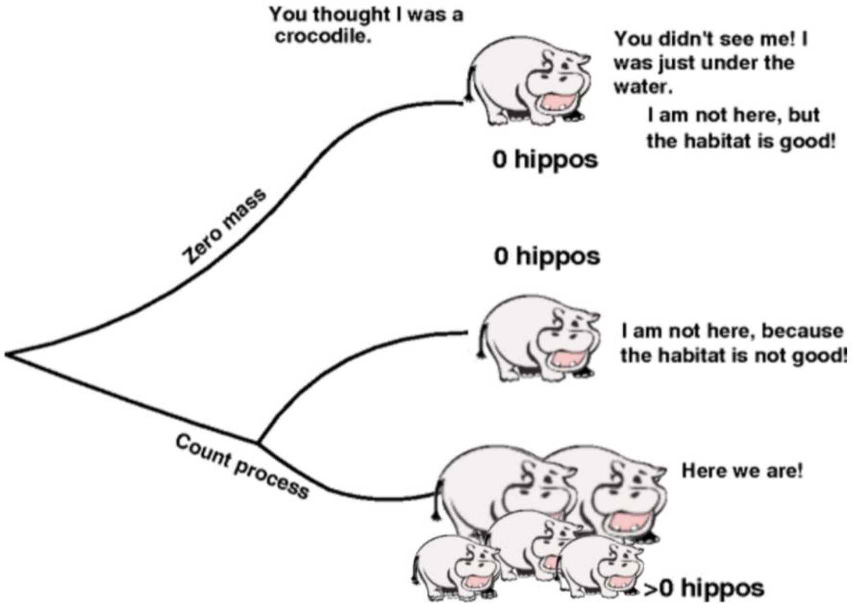
$$zero_i \sim Bernoulli(\pi_i)$$
$$logit(\pi_i) = \gamma_0 + \gamma_1 z_{1i} + \cdots + \gamma_q z_{qi}$$
$$count_i \sim truncated\ Poisson(\lambda_i)$$
$$\log(\lambda_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{p1}$$
$$Y_i = \begin{cases} 0, & if\ zero_i = 1\ (with\ probability\ \pi_i) \\ count_i, & if\ zero_i = 0\ (with\ probability\ 1 - \pi_i) \end{cases}$$

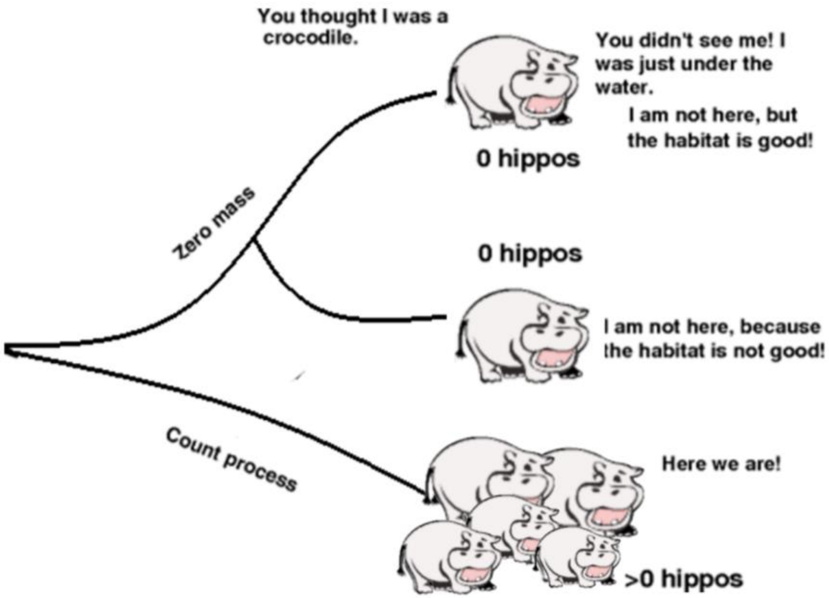
In short,

$$Y_i \sim Hurdle(\pi_i, \lambda_i)$$

ZIP vs ZAP Models

- A ZIP model allows zero counts to come from the count process, whereas a ZAP (Hurdle) model forces all zero counts to come from the zero process.





Example:
Fish in Lakes

- Let's say we're going fishing in multiple lakes over multiple days.
- Y_{ij}: number of fish caught in lake i on day j.
- Structural zeros: some lakes truly have no fish. We will never be able to catch any fish in those lakes.
- Sampling zeros: Maybe we aren't good at fishing, or there are issues with weather etc. so even fishy lakes still have days with 0 catch.
- We will include effort E_{ij} (e.g., hours netted) as an offset.
- We can use a ZIP or a ZINB model for this.

Coding Demo