Bayesian Lognormal-Logit Hurdle Modeling

10 total points

Christopher Cahill

6 October 2023

NOTE: date above is the due date

Background

Fishery trawl data are often plagued by a large number of zeros–in essence, one often catches zero or no fish and then occassionally encounters many fish. As a general rule, when > 20-30% of the data are zeros, traditional modeling techniques often fail and more specialized tools are needed to model the data.

One way to model excessive zeros in data is via the hurdle model, which is sometimes referred to as a "delta-GLM" or a "zero-altered model." In a hurdle model, it is assumed that two independent processes generate the data, and hurdle models have subtly different assumptions from a zero-inflated model (Hilbe et al. 2017; see also this link).

The binary component of the model is often modeled using a Bernoulli distribution, while the other component is either a count or continuous distribution. Because hurdle models are two-part models, each component of the model can be estimated separately. Hurdle models take advantage of the axiom of conditional probability of two events X and Y:

$$\Pr(X, Y) = \Pr(Y \mid X) \Pr(X)$$

$$\begin{aligned} X &= \Pr\left(d_i > 0\right) \\ Y &= \Pr\left(d_i = D\right) \end{aligned}$$

where d_i is in this case the density of data point i. We can re-write this hurdle model as:

$$\Pr\left(d_i = D\right) = \Pr\left(d_i = D \mid d_i > 0\right) \Pr\left(d_i > 0\right)$$

where we use separate models for $\Pr\left(c_i > 0\right)$ and $\Pr\left(c_i = C \mid c_i > 0\right)$.

The question

- What is the mean density of Pacific Cod off the west coast of Canada, and does depth affect fish density?

Download the `pcod_data.rds` data and fit a Lognormal-Logit Hurdle model (equations below) to the trawl data. Check model diagnostics, conduct graphical evaluations of MCMC chain performance and posterior predictive checks, and summarize posterior distributions for key parameters. Determine the mean density of Pacific Cod (after accounting for excessive zeros) and the effect of depth on cod density. Submit all R and Stan code required to conduct these analyses, along with a brief write up of your findings. Record all warnings and error messages that you receive. As this is a highly constrained model, you may get exceptions thrown during the warmup phase, which is okay as long as they do not occur after the warmup is complete.

The model

$$\Pr\left(D = d_i\right) = \begin{cases} \text{Bernoulli}(p) & \text{if } d_i = 0 \\ \text{Lognormal}\left(\lambda_i, \sigma\right) & \text{if } d_i > 0 \end{cases}$$

where

$$\text{logit}(\theta) = p$$

and

$$\log\left(\lambda_i\right) = \mathbf{X}_{ij}\beta_j$$

Priors

Assume diffuse normal priors for all parameters: $\sigma \sim N(0, 5)$, $\beta_j \sim N(0, 10)$, and $\text{logit}(\theta) \sim N(0, 10)$

References:

Hilbe et al. 2017.

Hints

This is a tricky model, so don't get too discouraged if at first everything is explosions and sadness. You should read the data in as a vector (call this `y_obs`), and to model the probability of zeros you should have a chunk of code like:

```
for each data point i
  // calculate the probability of y_obs[i] when no fish detected:
  (y_obs[i] == 0) ~ bernoulli(..stuff here..);
  if(y_obs[i] > 0){
    // calculate the likelihood for y_obs[i] > 0
    ...stuff here...
  }
end i
```

- Also, check out inv_logit() in Stan. It should make your life easier.

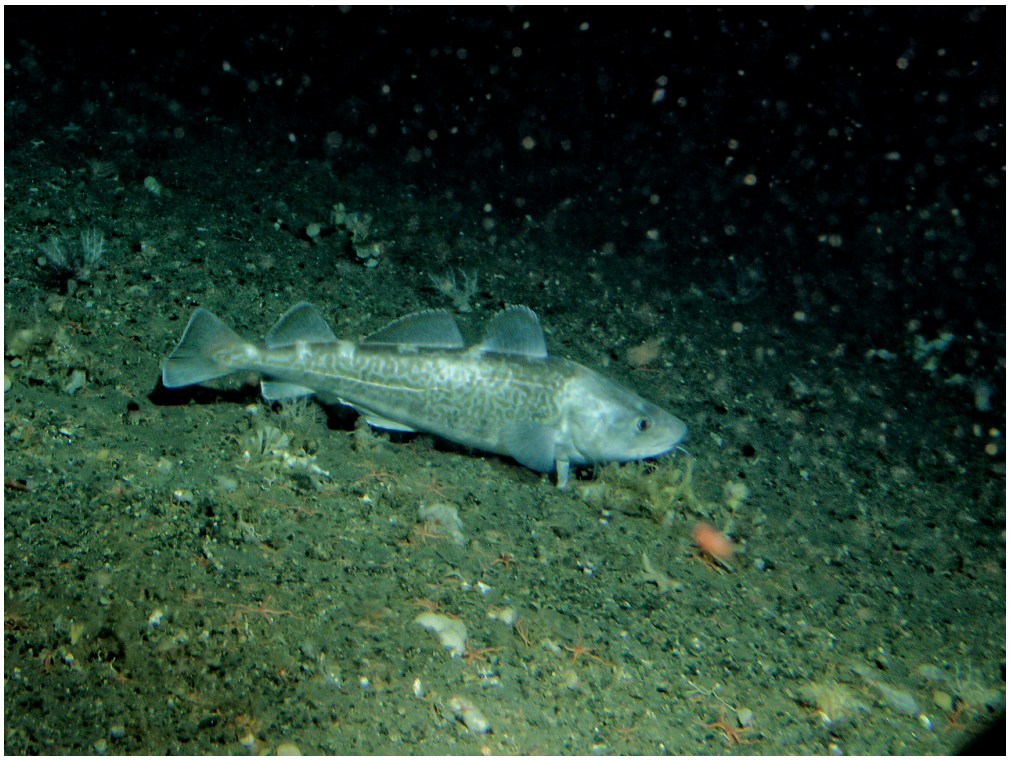

Figure 1: photo credit: NOAA fisheries