Models for Data with Zero Inflation

FW8051 Statistics for Ecologists

Department of Fisheries, Wildlife and Conservation Biology

Learning Objectives

- Be able to fit models to response data with lots of zeros (hurdle and zero-inflated models)

Learning Objectives

- Be able to fit models to response data with lots of zeros (hurdle and zero-inflated models)
- Be able to describe these models and their assumptions using equations and text and match parameters in these equations to estimates in computer output.

Zero-Inflation

Zero-inflation deals with response data, Y_i, not predictors, X_i.

Zero-Inflation

Zero-inflation deals with response data, Y_i, not predictors, X_i.

Zero inflation has received the most attention for count data:

- Covered in Zuur et al. Ch 11
- Kery Ch. 14

Zero-Inflation

Zero-inflation deals with response data, Y_i, not predictors, X_i.

Zero inflation has received the most attention for count data:

- Covered in Zuur et al. Ch 11
- Kery Ch. 14

Also relevant to:

- Binary data (occupancy models, Kery Ch 20)

Zero-Inflation

Zero-inflation deals with response data, Y_i, not predictors, X_i.

Zero inflation has received the most attention for count data:

- Covered in Zuur et al. Ch 11
- Kery Ch. 14

Also relevant to:

- Binary data (occupancy models, Kery Ch 20)
- Continuous data (e.g., Friederichs et al. 2011. Oikos 120:756-765)

Abundance Data and Zero Inflation

{From of Matt Russell, UMN}

Top 4 reasons why you might get a 0 when counting critters?

Abundance Data and Zero Inflation

{From of Matt Russell, UMN}

Top 4 reasons why you might get a 0 when counting critters?

- Sites are not suitable for the species

Abundance Data and Zero Inflation

{From of Matt Russell, UMN}

Top 4 reasons why you might get a 0 when counting critters?

- Sites are not suitable for the species
- Density effects: a site is suitable, but unoccupied

Abundance Data and Zero Inflation

{From of Matt Russell, UMN}

Top 4 reasons why you might get a 0 when counting critters?

- Sites are not suitable for the species
- Density effects: a site is suitable, but unoccupied
- Design errors: sampling for too short of a time period, or during the wrong times

Abundance Data and Zero Inflation

{From of Matt Russell, UMN}

Top 4 reasons why you might get a 0 when counting critters?

- Sites are not suitable for the species
- Density effects: a site is suitable, but unoccupied
- Design errors: sampling for too short of a time period, or during the wrong times
- Observer error: some species are difficult to identify/detect

Sampling and modeling macroinvertebrates

- Mayflies sampled using stratified random sampling along the Upper Mississippi River

- Characterized by a low-flow environment

- Samples collected with a 23 cm x 23cm sampler

- 43% of sample locations yielded zero mayflies

Gray 2005. *Ecol Model* 185:1-12

Univ. of Michigan

Center for Coastal Resources Management

Some examples:
ingrowth of trees in a forest inventory

US Forest Service

- We don't measure all trees when sampling
- Typically establish a minimum diameter to sample (say 5.0 inches DBH)

PLOTID	ForestType	Year1	Year2	Number of ingrowth trees ha^{-1}
1	Aspen	2010	2015	0
2	Red pine	2010	2015	0
3	Aspen	2010	2015	20
4	Red Pine	2010	2015	0
5	Red Pine	2010	2015	15

Zeros and common statistical distributions

Count data:

- Poisson and Negative Binomial distributions allow for zeros, i.e., $P(Y = 0) \neq 0$.

Zeros and common statistical distributions

Count data:

- Poisson and Negative Binomial distributions allow for zeros, i.e., $P(Y = 0) \neq 0$.
- Need to ask, are there more zeros than expected for a Poisson($\hat{\lambda}$) or NegBin($\hat{\lambda}, \hat{\theta}$) distribution?

Zeros and common statistical distributions

Count data:

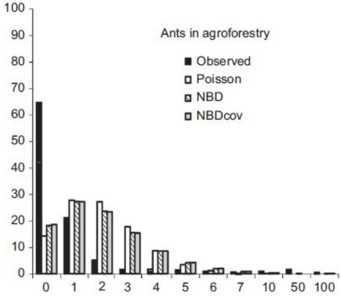
- Poisson and Negative Binomial distributions allow for zeros, i.e., $P(Y = 0) \neq 0$.
- Need to ask, are there more zeros than expected for a Poisson($\hat{\lambda}$) or NegBin($\hat{\lambda}, \hat{\theta}$) distribution?

For continuous data:

- We do not expect a “piling” up of zeros
- We can apply “mixture models” (similar to the models you will here see for count data)
- For an example, see: Friederichs et al. 2011. Oikos 120:756-765. (on Moodle)

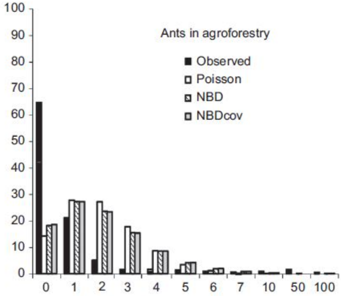
How can we determine if we have excess zeros?

How can we determine if we have excess zeros?



Sileshi 2008 (on Moodle)

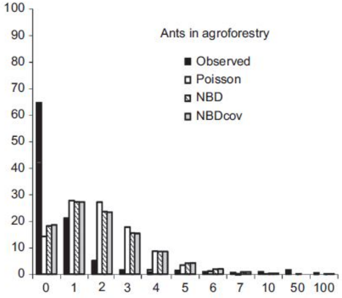
How can we determine if we have excess zeros?



Sileshi 2008 (on Moodle)

- Compare predicted and observed number of 0's (could use for a Goodness-of-fit test)

How can we determine if we have excess zeros?



Sileshi 2008 (on Moodle)

- Compare predicted and observed number of 0′s (could use for a Goodness-of-fit test)
- Can also test for overdispersion (variation > mean?)

Modeling Zero-Inflated Data

What do we do if we have zero-inflation?

- Hurdle models: model presence-absence (0 non-zero) and counts given presence

Modeling Zero-Inflated Data

What do we do if we have zero-inflation?

- Hurdle models: model presence-absence (0 non-zero) and counts given presence
- Mixture models: allow for multiple ways to get a 0

Modeling Zero-Inflated Data

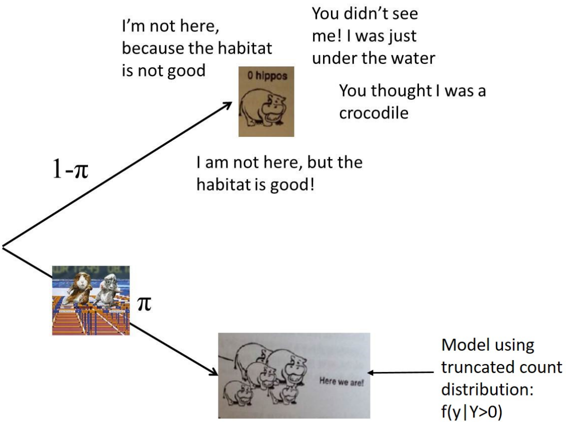
What do we do if we have zero-inflation?

- Hurdle models: model presence-absence (0 non-zero) and counts given presence
- Mixture models: allow for multiple ways to get a 0

For the in-class exercise, we will focus on the latter approach.

Hurdle Models

Group all 0′s into a single category:



Hurdle: positive counts arise if you exceed some threshold (with probability π)

Hurdle Models

1. Presence-absence subcomponent:

$$Z_i = \left\{ \begin{array}{ll} 0 \text{ when } y = 0 & \text{occurs with probability } (1-\pi) \\ 1 \text{ when } y > 0 & \text{occurs with probability } \pi \end{array} \right\}$$

Can model Z_i using using logistic regression to allow presence-absence to depend on covariates

Hurdle Models

1. Presence-absence subcomponent:

$$Z_i = \left\{ \begin{array}{ll} 0 \text{ when } y = 0 & \text{occurs with probability } (1-\pi) \\ 1 \text{ when } y > 0 & \text{occurs with probability } \pi \end{array} \right\}$$

Can model Z_i using using logistic regression to allow presence-absence to depend on covariates

2. Count model subcomponent:

Model the non-zero data (using truncated distribution models)

- Poisson or negative binomial, modified to exclude the possibility of a 0

Hurdle Models

1. Presence-absence subcomponent:

$$Z_i = \left\{ \begin{array}{ll} 0 \text{ when } y = 0 & \text{occurs with probability } (1-\pi) \\ 1 \text{ when } y > 0 & \text{occurs with probability } \pi \end{array} \right\}$$

Can model Z_i using using logistic regression to allow presence-absence to depend on covariates

2. Count model subcomponent:

Model the non-zero data (using truncated distribution models)

- Poisson or negative binomial, modified to exclude the possibility of a 0

Can do this in two steps or use a single modeling framework (see Hurdle models Ch 11.5 in Zuur et al).

The non-zeros

Truncated distributions for non-zero count data:

$P(Y = y|Y > 0) = \frac{P(Y=y)}{P(Y>0)} = \frac{f(y)}{(1-f(0))}$

remember, $P(A|B)$=P(A and B)/P(B)

The non-zeros

Truncated distributions for non-zero count data:

$P(Y = y|Y > 0) = \frac{P(Y=y)}{P(Y>0)} = \frac{f(y)}{(1-f(0))}$

remember, $P(A|B)$=P(A and B)/P(B)

A truncated Poisson would look like…

The non-zeros

Truncated distributions for non-zero count data:

$P(Y = y|Y > 0) = \frac{P(Y=y)}{P(Y>0)} = \frac{f(y)}{(1-f(0))}$

remember, $P(A|B)$=P(A and B)/P(B)

A truncated Poisson would look like...

$$P(Y = y|y > 0) = \frac{\frac{e^{-\lambda}\lambda^{y}}{y!}}{1-e^{-\lambda}}$$

The non-zeros

Truncated distributions for non-zero count data:

$P(Y = y|Y > 0) = \frac{P(Y=y)}{P(Y>0)} = \frac{f(y)}{(1-f(0))}$

remember, $P(A|B)$=P(A and B)/P(B)

A truncated Poisson would look like…

$$P(Y = y|y > 0) = \frac{\frac{e^{-\lambda}\lambda^y}{y!}}{1-e^{-\lambda}}$$

We can incorporate covariates, using: $\log(\lambda) = \beta_0 + \beta_1 x + \ldots$

The non-zeros

Truncated distributions for non-zero count data:

$P(Y = y|Y > 0) = \frac{P(Y=y)}{P(Y>0)} = \frac{f(y)}{(1-f(0))}$

remember, $P(A|B)$=P(A and B)/P(B)

A truncated Poisson would look like...

$$P(Y = y|y > 0) = \frac{\frac{e^{-\lambda}\lambda^y}{y!}}{1-e^{-\lambda}}$$

We can incorporate covariates, using: $\log(\lambda) = \beta_0 + \beta_1 x + \ldots$

Note, however:

- We are modeling $E[Y|X, Y > 0] = \lambda_i$ and not $E[Y|X]$

The non-zeros

Truncated distributions for non-zero count data:

$P(Y = y|Y > 0) = \frac{P(Y=y)}{P(Y>0)} = \frac{f(y)}{(1-f(0))}$

remember, $P(A|B)$=P(A and B)/P(B)

A truncated Poisson would look like...

$$P(Y = y|y > 0) = \frac{\frac{e^{-\lambda}\lambda^y}{y!}}{1-e^{-\lambda}}$$

We can incorporate covariates, using: $\log(\lambda) = \beta_0 + \beta_1 x + \ldots$

Note, however:

- We are modeling $E[Y|X, Y > 0] = \lambda_i$ and not $E[Y|X]$
- Need to be careful when plotting fitted model or constructing Bayesian p-values

The non-zeros

Truncated distributions for non-zero count data:

$P(Y = y|Y > 0) = \frac{P(Y=y)}{P(Y>0)} = \frac{f(y)}{(1-f(0))}$

remember, $P(A|B)$=P(A and B)/P(B)

A truncated Poisson would look like...

$$P(Y = y|y > 0) = \frac{\frac{e^{-\lambda}\lambda^y}{y!}}{1-e^{-\lambda}}$$

We can incorporate covariates, using: $\log(\lambda) = \beta_0 + \beta_1 x + \ldots$

Note, however:

- We are modeling $E[Y|X, Y > 0] = \lambda_i$ and not $E[Y|X]$
- Need to be careful when plotting fitted model or constructing Bayesian p-values
- See Zuur et al. p. 288 for expressions for $E[Y|X]$ and $Var[Y|X]$

The non-zeros

For continuous data:

- Log-normal, gamma distributions live on $(0, \infty)$ (so no need to truncate these)

The non-zeros

For continuous data:

- Log-normal, gamma distributions live on $(0, \infty)$ (so no need to truncate these)
- Or, can use truncated distributions (e.g., Normal) $= \frac{f(y)}{1-F(0)}$ where $F(y) = P(Y \leq y)$

The non-zeros

For continuous data:

- Log-normal, gamma distributions live on $(0, \infty)$ (so no need to truncate these)
- Or, can use truncated distributions (e.g., Normal) $= \frac{f(y)}{1-F(0)}$ where $F(y) = P(Y \leq y)$

Which function in R is used to determine $F(Y)$?

The non-zeros

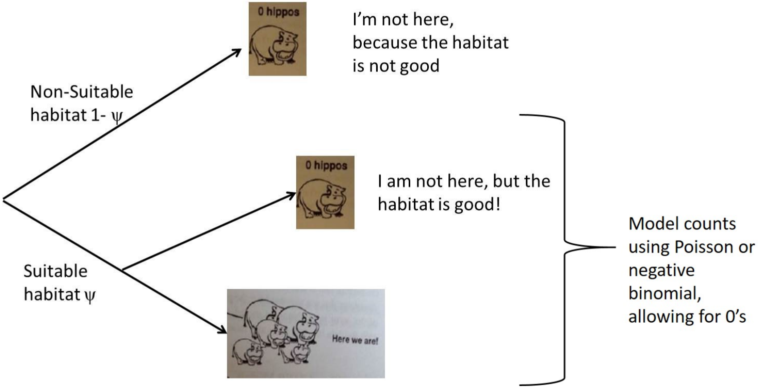
For continuous data:

- Log-normal, gamma distributions live on $(0, \infty)$ (so no need to truncate these)
- Or, can use truncated distributions (e.g., Normal) $= \frac{f(y)}{1-F(0)}$ where $F(y) = P(Y \leq y)$

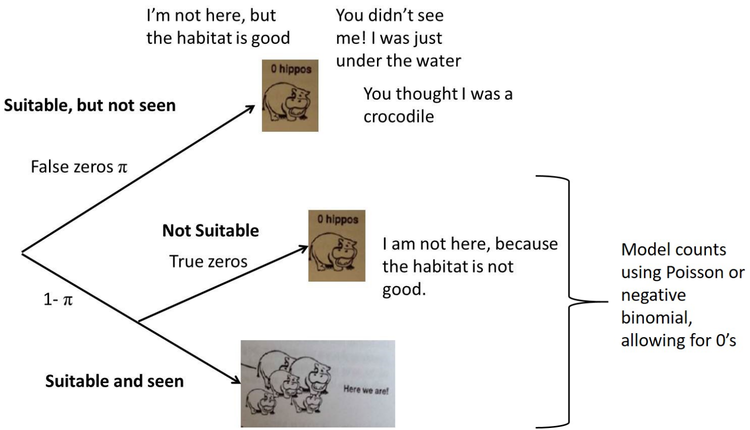
Which function in R is used to determine $F(Y)$? `pnorm`!

Mixture Model: Suitable and Non-Suitable Habitat (Kery)

Two ways to get a 0:



Mixture Models: true and false zeros (Zuur et al)



Reality

Zero-inflation:

- Kery suggests we think of the extra zeros as arising from non-suitable habitat
- Zuur et al. suggests we view the extra zeros as suitable habitat where species are not detected

Reality

Zero-inflation:

- Kery suggests we think of the extra zeros as arising from non-suitable habitat
- Zuur et al. suggests we view the extra zeros as suitable habitat where species are not detected

Assigning meaning to the zero-inflation process can in some cases be useful, but it also requires a leap of faith!

Reality

Zero-inflation:

- Kery suggests we think of the extra zeros as arising from non-suitable habitat
- Zuur et al. suggests we view the extra zeros as suitable habitat where species are not detected

Assigning meaning to the zero-inflation process can in some cases be useful, but it also requires a leap of faith!

See comments on this blog:
https://statisticalhorizons.com/zero-inflated-models

ZIP model: Zero-inflated Poisson

Probability Mass Function: $f(y) = \frac{e^{-\lambda}\lambda^{y}}{y!}$

Let: π be the probability of a zero-inflated response

ZIP model: Zero-inflated Poisson

Probability Mass Function: $f(y) = \frac{e^{-\lambda}\lambda^y}{y!}$

Let: π be the probability of a zero-inflated response

ZIP model (Zuur):

$$P(Y = y) = f(y) = \begin{cases} \pi + (1-\pi)e^{-\lambda} & \text{if } y = 0 \\ (1-\pi)\frac{e^{-\lambda}\lambda^y}{y!} & \text{if } y = 1, 2, 3, \ldots \end{cases}$$

ZIP model: Zero-inflated Poisson

Probability Mass Function: $f(y) = \frac{e^{-\lambda}\lambda^{y}}{y!}$

Let: π be the probability of a zero-inflated response

ZIP model (Zuur):

$$P(Y = y) = f(y) = \begin{cases} \pi + (1-\pi)e^{-\lambda} & \text{if } y = 0 \\ (1-\pi)\frac{e^{-\lambda}\lambda^{y}}{y!} & \text{if } y = 1, 2, 3, \ldots \end{cases}$$

Get a 0 two ways:

- Zero-inflated process leads to a 0, occurs with probability π
- Non-zero inflated 0, occurs with probability $(1-\pi)f(0)$

ZIP model: Zero-inflated Poisson

Probability Mass Function: $f(y) = \frac{e^{-\lambda}\lambda^y}{y!}$

Let: π be the probability of a zero-inflated response

ZIP model (Zuur):

$$P(Y = y) = f(y) = \begin{cases} \pi + (1-\pi)e^{-\lambda} & \text{if } y = 0 \\ (1-\pi)\frac{e^{-\lambda}\lambda^y}{y!} & \text{if } y = 1, 2, 3, \ldots \end{cases}$$

Get a 0 two ways:

- Zero-inflated process leads to a 0, occurs with probability π
- Non-zero inflated 0, occurs with probability $(1-\pi)f(0)$

Non-zero responses: $(1-\pi)f(y)$

ZIP model: Zero-inflated Poisson

Zuur and `zeroinfl` function in `pscl` R package:

- Parameterizes in terms of π = the probability of a zero-inflated response

Kery:

- Parameterizes in terms of $\psi = 1 - \pi$ = the probability of a NON zero-inflated response

ZIP model: Zero-inflated Poisson

Zuur and `zeroinfl` function in `pscl` R package:

- Parameterizes in terms of π = the probability of a zero-inflated response

Kery:

- Parameterizes in terms of $\psi = 1 - \pi$ = the probability of a NON zero-inflated response

ZIP model (Zuur and `zeroinfl`):

$$P(Y = y) = f(y) = \begin{cases} \pi + (1 - \pi)e^{-\lambda} & \text{if } y = 0 \\ (1 - \pi)\frac{e^{-\lambda}\lambda^y}{y!} & \text{if } y = 1, 2, 3, \ldots \end{cases}$$

ZIP model: Zero-inflated Poisson

Zuur and `zeroinfl` function in `pscl` R package:

- Parameterizes in terms of π = the probability of a zero-inflated response

Kery:

- Parameterizes in terms of $\psi = 1 - \pi$ = the probability of a NON zero-inflated response

ZIP model (Zuur and `zeroinfl`):

$$P(Y = y) = f(y) = \begin{cases} \pi + (1-\pi)e^{-\lambda} & \text{if } y = 0 \\ (1-\pi)\frac{e^{-\lambda}\lambda^y}{y!} & \text{if } y = 1, 2, 3, \ldots \end{cases}$$

ZIP model (Kery):

$$P(Y = y) = f(y) = \begin{cases} 1 - \psi + \psi e^{-\lambda} & \text{if } y = 0 \\ \psi\frac{e^{-\lambda}\lambda^y}{y!} & \text{if } y = 1, 2, 3, \ldots \end{cases}$$

ZINB model: Zero-inflated Negative Binomial

Probability Mass Function: $f(y) = \binom{y+\theta-1}{y} \left(\frac{\theta}{\mu+\theta}\right)^{\theta} \left(\frac{\mu}{\mu+\theta}\right)^{y}$

ZINB model (Zuur et al):

$$f(y) = \begin{cases} \pi + (1-\pi)\left(\frac{\theta}{\mu+\theta}\right)^{\theta} & \text{if } y = 0 \\ (1-\pi)\binom{y+\theta-1}{y} \left(\frac{\theta}{\mu+\theta}\right)^{\theta} \left(\frac{\mu}{\mu+\theta}\right)^{y} & \text{if } y = 1, 2, 3, \ldots \end{cases}$$

ZINB model: Zero-inflated Negative Binomial

Probability Mass Function: $f(y) = \binom{y+\theta-1}{y} \left(\frac{\theta}{\mu+\theta}\right)^{\theta} \left(\frac{\mu}{\mu+\theta}\right)^{y}$

ZINB model (Zuur et al):

$$f(y) = \begin{cases} \pi + (1-\pi)\left(\frac{\theta}{\mu+\theta}\right)^{\theta} & \text{if } y = 0 \\ (1-\pi)\binom{y+\theta-1}{y}\left(\frac{\theta}{\mu+\theta}\right)^{\theta}\left(\frac{\mu}{\mu+\theta}\right)^{y} & \text{if } y = 1, 2, 3, \ldots \end{cases}$$

ZINB model (Kery):

$$f(y) = \begin{cases} 1 - \pi + \pi\left(\frac{\theta}{\mu+\theta}\right)^{\theta} & \text{if } y = 0 \\ \pi\binom{y+\theta-1}{y}\left(\frac{\theta}{\mu+\theta}\right)^{\theta}\left(\frac{\mu}{\mu+\theta}\right)^{y} & \text{if } y = 1, 2, 3, \ldots \end{cases}$$

Fitting Models in R

We can use the `zeroinfl` function in the `pscl` package in R to fit:

- Both types of models (Hurdle model, mixture)
- With both the Poisson and Negative Binomial distributions (see in class exercise)

Fitting Models in R

We can use the `zeroinfl` function in the `pscl` package in R to fit:

- Both types of models (Hurdle model, mixture)
- With both the Poisson and Negative Binomial distributions (see in class exercise)

Can also code models in JAGS (see Kery Ch 14) and fit using other packages (e.g. `glmmTMB`)

zeroinfl versus Kery

Remember:

- `zeroinf`: models probability of a zero-inflated response (i.e., "false" zero) = π_i
- Kery: models the probability of a NON zero-inflated response (i.e., probability of a "true" zero or a count > 0) = ψ_i

As a result, the sign of the coefficients will differ between the two approaches.

Model Comparisons

Can compare Poisson, Negative Binomial, Zero-inflation models

- Using AIC
- Graphs of observed vs expected proportion of zeros in a dataset
- Graphs of the sample mean–variance relationship.

Model Comparisons

Can compare Poisson, Negative Binomial, Zero-inflation models

- Using AIC
- Graphs of observed vs expected proportion of zeros in a dataset
- Graphs of the sample mean–variance relationship.

My experience, and that of others, is that a Negative Binomial model (without zero-inflation) often "wins" (but not always)

- See Warton (2005) on Canvas, as well as Gray (2005), Sileshi (2008)

Model Comparisons

Can compare Poisson, Negative Binomial, Zero-inflation models

- Using AIC
- Graphs of observed vs expected proportion of zeros in a dataset
- Graphs of the sample mean–variance relationship.

My experience, and that of others, is that a Negative Binomial model (without zero-inflation) often "wins" (but not always)

- See Warton (2005) on Canvas, as well as Gray (2005), Sileshi (2008)

Also, zero-inflated negative binomial models can sometimes be difficult to fit (past homework problem)