STA 470S Sp 24

Zero Inflated & Altered Models for Counts Data

Group B: Cynthia Yan, Jenny Yan, Athena Ru

Count Data

Poisson(λ)

- Non-Negative & Discrete
- Model the number of events occurring in fixed time intervals or space
- $E[X] = Var(X) = \lambda$

Negative Binomial(r, p)

- Number of failures until r successes
- Model count data with overdispersion, where the variance exceeds the mean

Poisson VS. NB

The Poisson model assumes events are independent and have a constant rate of occurrence.

- good for modeling calcium leakage rates – sparks per time

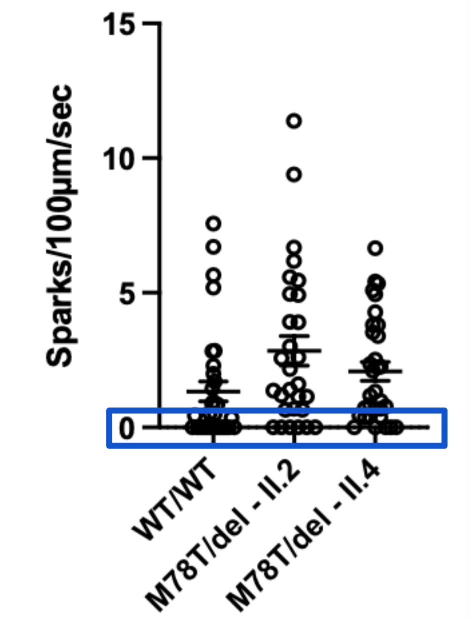
BUT it assumes that the variance is approximately equal to the mean.

- may not be the case with our calcium leakage data

The Negative Binomial model is used to model count data when the variance exceeds the mean.

Model Weakness

- Does not fit well when there is more 0s than what be expected under a Poisson or NB model
    - potentially our client's data
- Failing to account for excess zeros can lead to biased parameter estimates and incorrect inferences



Sources of Zeros

False Zeros/False Negatives

1. Design Error: poor experimental design or sampling practices
2. Observer Error: observer miscounts/misclassifies subjects

Positive Zeros/True Zeros

1. Structural Error: subject is not present because environment is not suitable

Zero-Inflated Models (ZIMs)

Zero-inflation (ZI): we have far more zeros than what would be expected for a Poisson or NB distribution

Zero-Inflated Models (ZIMs) are a mixture model:

- A count model (Poisson or NB)
- A (binomial) logistic model for probability of <u>false zeros</u>

ZIMs assume that the data are generated from two different processes (count and binomial).

Zero Truncated Models

Zero truncated: the response variable cannot have a value of 0

Zero truncated models are a one-stage model:

- The probability of observing zero is excluded from consideration
- Focus is solely on modeling the positive counts

Examples:

1. Dolphin group size
2. Age of an animal in years or months

Zero Altered Models

Zero Altered: like ZI, we have an excess of zeros compared to Poisson or NB expectations, but we propose a different zero-generation mechanism.

Zero altered models are a two-part model:

1. A binomial model for the probability that a zero value is observed
2. A truncated Poisson or truncated NB model for non-zero observations

They are also called hurdle models, conditional models, or compatible models. The model <u>does not distinguish</u> between the different types of zeros.

ZIP VS. ZINB

Zero-Inflated Poisson (ZIP)

$$f(y_i = 0) \quad = \pi_i + (1 - \pi_i) \times e^{-\mu_i}$$
$$f(y_i | y_i > 0) = (1 - \pi_i) \times \frac{\mu^{y_i} \times e^{-\mu_i}}{y_i!}$$

$$E(Y_i) = \mu_i \times (1 - \pi_i)$$
$$\mathrm{var}(Y_i) = (1 - \pi_i) \times (\mu_i + \pi_i \times \mu_i^2)$$

μ_i is mean of response Y_i

π_i is probability that Y_i is a false zero

Zero-Inflated NB (ZINB)

$$f(y_i = 0) \quad = \pi_i + (1 - \pi_i) \times \left(\frac{k}{\mu_i + k}\right)^k$$
$$f(y_i | y_i > 0) = (1 - \pi_i) \times f_{NB}(y)$$

$$E(Y_i) = \mu_i \times (1 - \pi_i)$$

$$\mathrm{var}(Y_i) = (1 - \pi_i) \times (\mu_i + \frac{\mu_i^2}{k}) + \mu_i^2 \times (\pi_i^2 + \pi_i)$$

- allows for overdispersion from the non-zero counts

Zero Truncated Poisson Vs. NB

Zero Truncated Poisson

$$f(y_i;\mu_i|y_i > 0) = \frac{\mu^{y_i} \times e^{-\mu_i}}{(1 - e^{-\mu_i}) \times y_i!}$$

- Use for data with mainly true zeros and rare or negligible excess zeros

Zero Truncated NB

$$f(y_i;\mu_i|y_i > 0) = \frac{\Gamma(y_i+k)}{\Gamma(k)\times\Gamma(y_i+1)} \times \left(\frac{k}{\mu_i+k}\right)^k \times \left(1 - \frac{k}{\mu_i+k}\right)^{y_i} \Big/ \left(1 - (\frac{k}{\mu_i+k})^k\right)$$

- Allows for overdispersion from the non-zero counts

Zero Altered Poisson and NB

$$f_{\text{ZAP}}(y;\beta,\gamma) = \begin{cases} f_{\text{binomial}}(y=0;\gamma) & y=0 \\ (1 - f_{\text{binomial}}(y=0;\gamma)) \times \dfrac{f_{\text{Poisson}}(y;\beta)}{1 - f_{\text{Poisson}}(y=0;\beta)} & y>0 \end{cases} \tag{11.24}$$

$$E_{\text{ZAP}}(Y_i;\pi_i,\mu_i) = \frac{1-\pi_i}{1-e^{-\mu_i}} \times \mu_i$$

$$\text{Var}_{\text{ZAP}}(Y_i;\pi_i,\mu_i) = \frac{1-\pi_i}{1-e^{-\mu_i}} \times (\mu_i + \mu_i^2) - \left(\frac{1-\pi_i}{1-e^{-\mu_i}} \times \mu_i\right)^2$$

$$E_{\text{ZANB}}(Y_i;\pi_i,\mu_i,k) = \frac{1-\pi_i}{1-P_0} \times \mu_i \quad \text{where}\, P_0 = \left(\frac{k}{\mu_i + k}\right)^k$$

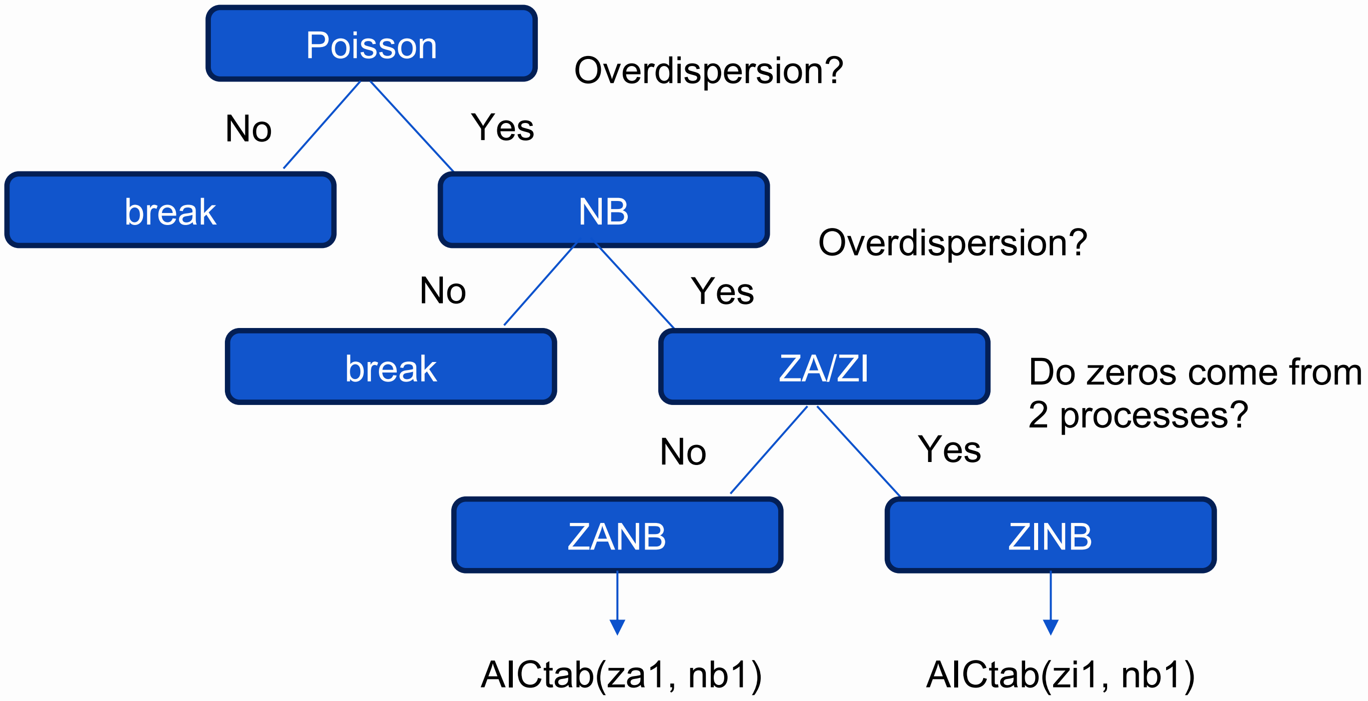
$$\text{Var}_{\text{ZANB}}(Y_i;\pi_i,\mu_i,k) = \frac{1-\pi_i}{1-P_0} \times \left(\mu_i^2 + \mu_i + \frac{\mu_i^2}{k}\right) - \left(\frac{1-\pi_i}{1-P_0} \times \mu_i\right)^2$$

Hurdle model for the probability of presence vs. absence.

Overview of Models

Model	Full Name	Type of Model	Overdispersion
ZIP	Zero-Inflated Poisson	Mixture	Zeros
ZINB	Zero-Inflated Negative Binomial	Mixture	Zeros and counts
ZAP	Zero-Altered Poisson	Two-part	Zeros
ZANB	Zero-Altered Negative Binomial	Two-part	Zeros and counts

Model Selection Flow Chart (updated)



Model Interpretation: ZAP/ZANB

1. Count Component:

- If we are fitting a glm, a positive coefficient signals an expected increase in the count with an increase in the predictor
- Exponentiating the coefficient yields the rate ratio, indicating the multiplicative change in the count for a one-unit shift in the predictor

2. Zero Component:

– A positive coefficient in this component indicates higher odds of observing excess zeros for the associated predictor variable.

Model Interpretation: ZIP/ZINB

1. Count Component:

- Similar to the count component in the ZA model
- The count component estimates the effects of predictor variables on the count outcome when count is positive

2. Inflation Component:

- Similar to the zero component in ZA Model.
- The inflation component of the ZI model models the probability of excess zeros.
- Interpret coefficients for predictor variables in the inflation component.
- A positive coefficient indicates an increase in the odds of excess zeros.

How should our client compare these models based on fit to the data and other factors?

- Residual Deviance – measure of the goodness-of-fit of a statistical model, particularly for models based on maximum likelihood estimation
    - pchisq(model.out$deviance, model.out$df.residual, lower.tail=F)
    - Used with Lack–of–fit p–value
- Pearson's goodness–of–fit statistic – overall residual variation
    - residuals(model.out, type="pearson")
- AIC – Akaike information criterion
    - AICtab(fit_zipoisson,fit_zinbinom,fit_zinbinom1,fit_zinbinom1_bs)

Implementing in R - ZIM

We can use the *glmmTMB* library. This accounts for random effects.

```{r}
library(glmmTMB)
#ZIP
glmmTMB(spark_freq~ genotype + (1|individual) + (1|differentiation) + e, ziformula=~1, data=data, family=poisson)

#ZINB
glmmTMB(spark_freq ~ genotype+ (1|individual) + (1|differentiation) + e, ziformula=~1, data=data, family=nbinom1)
```

Implementing in R - ZAM

For altered models, we can use the *zapoisson* and zanegbinomial functions in combination with the *vglm* function from the VGAM library.

```
```{r}
library(VGAM)

#ZAP
vglm(y ~ x2, family = zapoisson, data = data, trace = TRUE)

#ZANB
vglm(cbind(y1+y2) ~ x2, family = zanegbinomial, data = data, trace = TRUE)
```
```

Implementing in R - Zero Truncated

Import the *VGAM* library with the *vglm* function to implement zero-truncated models.

```
```{r}
library(VGAM)

#Zero-Truncated Poisson
vglm(y ~ x1 + x2, family = pospoisson, data = data)

#Zero-Truncated NB
vglm(y ~ x1 + x2, family = posnegbinomial, data = data)
```
```

We can use the glmmTMB library as well. This accounts for random effects.

```
```{r}
library(glmmTMB)
    update(fit_zinbinom1_bs, ziformula=~., data=data, family=truncated_nbinom1)
```
```

Non-technical Explanation

Client Request

- Client Mission
  - Analyze spark frequencies to determine if the TAX1BP3 gene is significant in causing calcium leakage (and arrhythmias)
- Client Questions
  - Is there a need for a nested analysis?
  - How to do a nested analysis when the data isn't normal (or in this case, there is a potential abundance of 0s)

Poisson VS Negative Binomial (NB)

- Both for Count Data
- Poisson Model assumes variance equals to mean
- NB allows variance to be greater than mean

Excess zeros: having many more zeros than expected under a count model

- Zero Inflated Models (ZIMs)
- Zero Altered Models (ZAMs)

ZI & ZA Poisson VS NB

Domain knowledge

- Identify where the zeros are coming from
- Determine how to distinguish between true zeros and false zeros

Deviance and Residuals:

- Calculate the deviance residuals from your model.

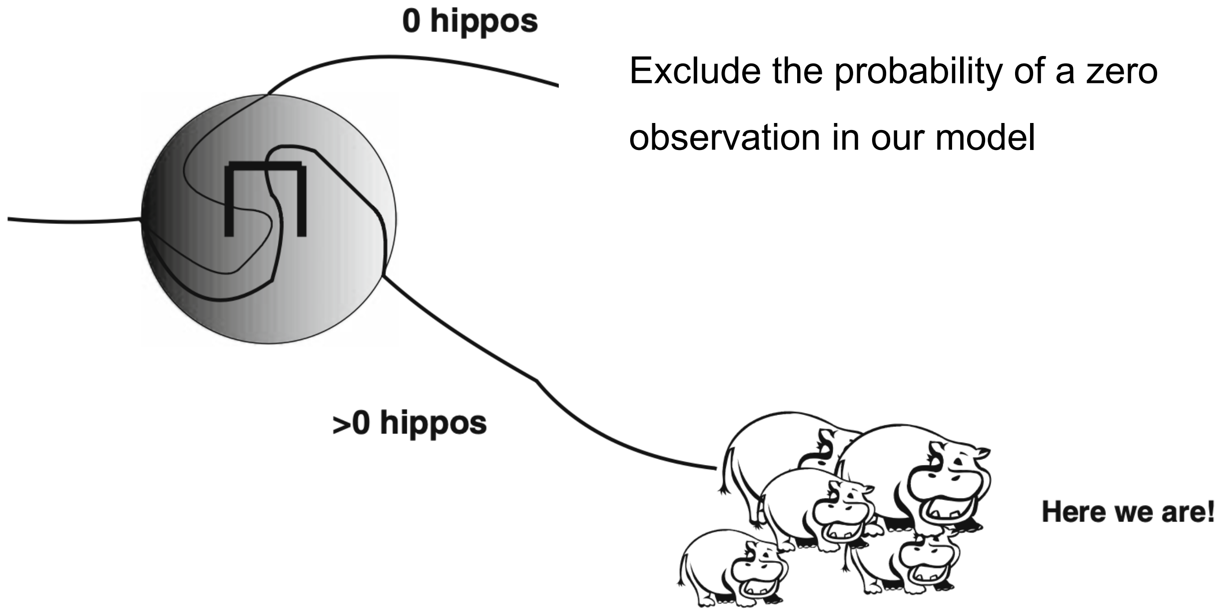
– Overdispersion may manifest as a pattern in the residuals, such as larger-than-expected residuals for certain observations.

Diagnostic Tests:

- Conduct formal statistical tests for overdispersion.

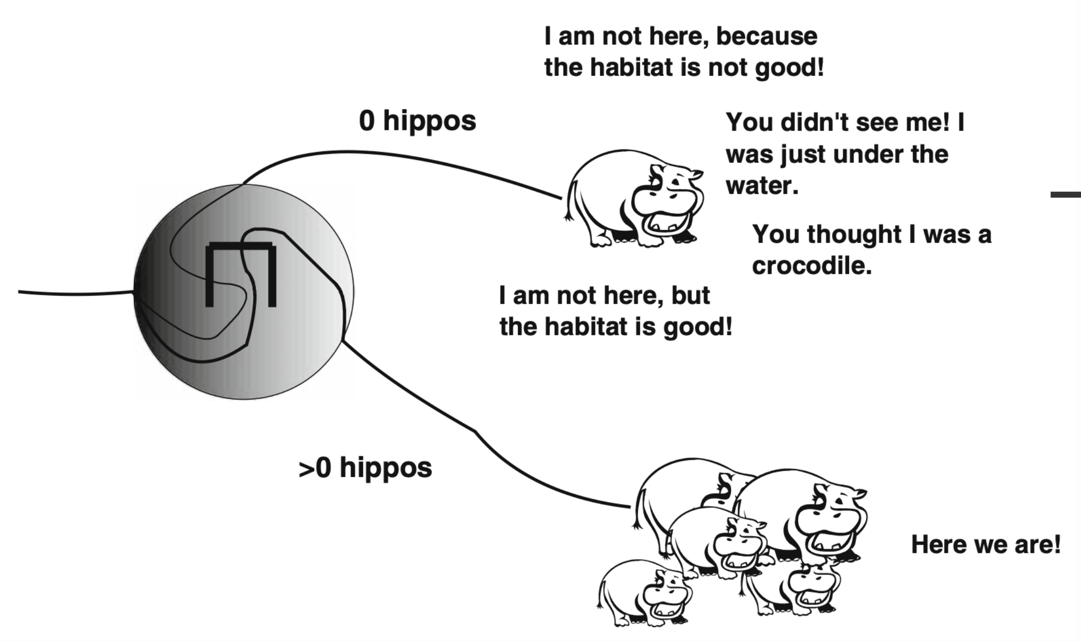
\- Pearson chi-square test: Large Pearson residuals may indicate overdispersion.

Zero Truncated Model



Exclude the probability of a zero observation in our model

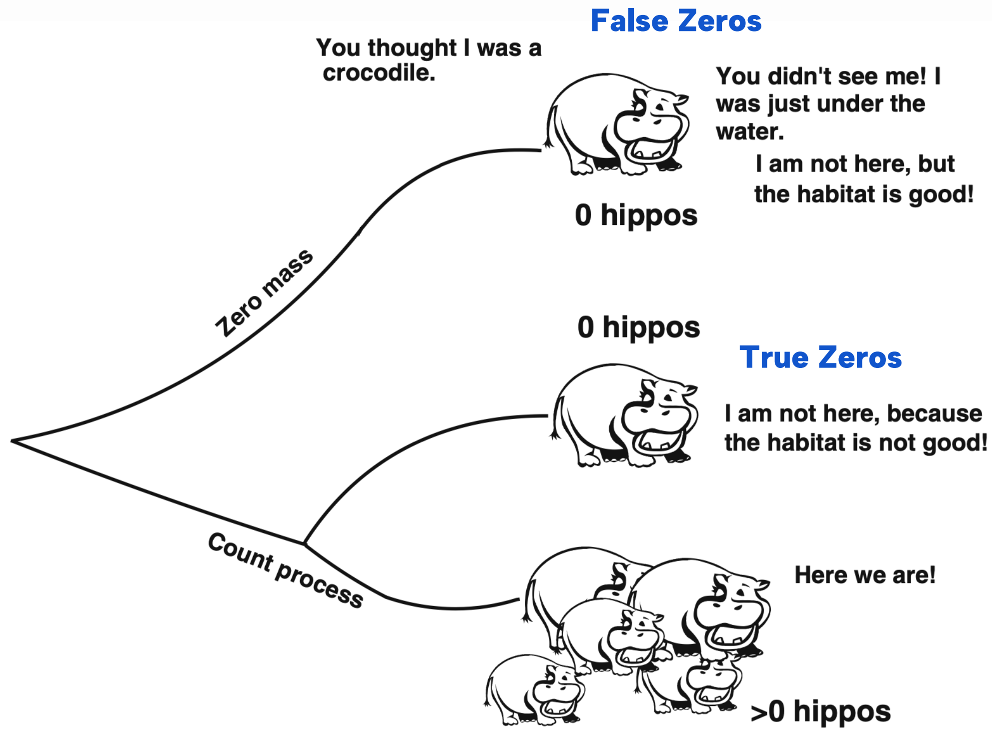
Zero Altered Model



(Binomial) Logistic Model

Zero Truncated Model (Poisson, NB)

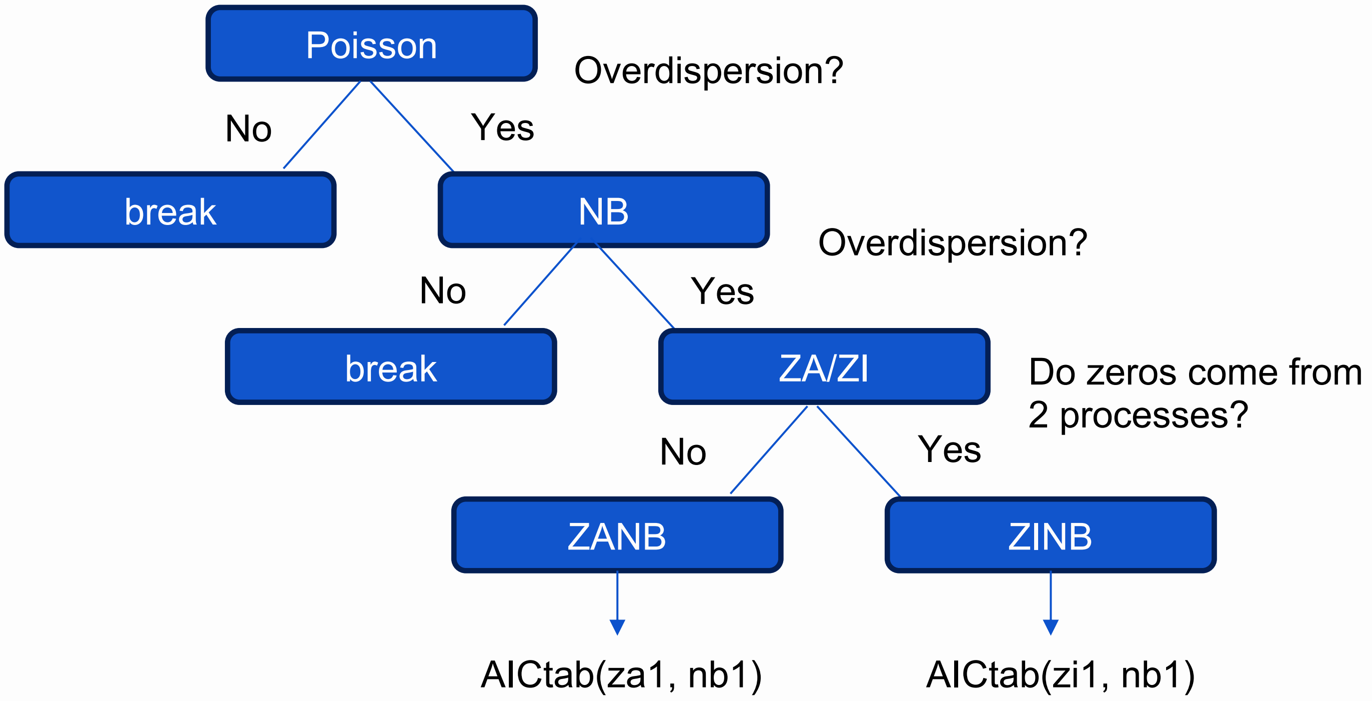
Zero Inflated Model



(Binomial) Logistic Model

Count Model (Poisson, NB)

Model Selection Flow Chart (updated)



Recommendations

- Determine how many zeros are expected under a Poisson or NB model (data may not be actually zero-inflated so a Poisson or NB model will suffice)
- Determine if there is a distinction between false zeros and true zeros
- Follow the Flow Chart to choose a model
- Degree of fitness can be measured by the previously mentioned residual deviance and Pearson's goodness–of–fit statistic
- Compare AIC of ZINB or ZANB model with NB model (hopefully AIC is smaller)
    - Must calculate AIC using function from the same package (`AICtab()`)

References

- https://rdrr.io/cran/VGAM/man/zapoisson.html
- https://rdrr.io/cran/VGAM/man/zanegbinomial.html
- https://search.r-project.org/CRAN/refmans/VGAM/html/zapoisson.html
- https://search.r-project.org/CRAN/refmans/VGAM/html/zanegbinomial.html
- Zuur Et Al. Zero Altered Models Ch11: Zero-Truncated and Zero-Inflated Models for Count Data
- https://cran.r-project.org/web/packages/glmmTMB/vignettes/glmmTMB.pdf
- https://easystats.github.io/performance/reference/check_overdispersion.html