Extended Version of Zero-inflated Negative Binomial Distribution with Application to HIV Exposed Infant Count Data

Dr. Collins Odhiambo and Stella kibika

September 12, 2019

Table of contents

Introduction

- Count models such as Poisson and negative binomial are preferred. In cases where they are not able to handle zeros, they are extended to zero-inflated models.
- The zero inflated models introduce a second link function, the logit link, which allows the zeros to be distinguished as either structured or unstructured.
- The hurdle models do not distinguish the zeros and are applicable when all zeros arising from the data are structured.
- More flexible models i.e. ZINB-Crack (ZINB-CR), ZINB-Sushila (ZINB-S) and ZINB-Generalized exponential (ZINB-GE) that have been developed. These models have been shown to perform better than the ZINB with fewer parameters.

Background

- We seek to extend the ZINB model to allow for greater flexibility by including more parameters.
- Structured zeros are inevitable while unstructured zeros occur by chance.
- In this study the focus will be on mothers who undergo prenatal and postnatal care in facilities that are equipped to prevent transmission(structured) and mothers who do not receive prenatal and postnatal care or visit facilities that are not equipped to prevent transmission(unstructured).
- Creating more flexible models is important for the bias-variance tradeoff and creating models that have better predictive ability.

Problem Statement

Why ZINB: Because of overdispersion

The Problem

The ZINB applies weights to the structured and random zeros. It gives a weight π to the structured zeros and $(1-\pi)$ to the random zeros and other count values greater than zero.

$$f(Y=y)=\begin{cases}\pi+(1-\pi)(1+\frac{\mu}{r})^{-r} & y=0\\ (1-\pi)\frac{\Gamma(y+r)}{y!\Gamma(r)}(1+\frac{\mu}{r})^{-r}(1+\frac{r}{\mu})^{-y} & y>0\end{cases}$$

However,it is not clear what the proportion of zeros is allowable, after which the data should be considered as zero-inflated. This paper seeks to extend the ZINB to be allow more flexibility to the model.

Objectives

- To create the ZINB-Shanker mixed distribution, find its properties and estimate the parameters.
- To conduct simulations to compare the standard ZINB and ZINB-SH.
- To apply real data i.e. HIV exposed infants data to validate the extended ZINB distribution.

Methodology

1. The new distribution is made of the negative binomial and Shanker distributions and then adjusted for zero inflation.
2. The probability distribution function is given by (NB):

$$f(x) = \binom{m+x-1}{x} p^r (1-p)^x \quad x = 0, 1, 2, \ldots \quad m > 0, \quad 0 < p <$$

3. The Shanker distribution is a one parameter distribution which is a mixture of $exponential(\theta)$ and $gamma(2, \theta)$ distributions. This mixture gives the final probability of the shanker as:

$$f(\lambda; \theta) = \frac{\theta^2}{\theta^2 + 1}(\theta + \lambda)e^{(-\theta\lambda)}; \quad \lambda > 0, \quad \theta > 0.$$

NB-SH Distribution

1. This is a compound distribution; a mixture of Negative Binomial and Shanker.
2. The probability distribution function of the NB-SH as :

$$h(x;m,\theta)=\binom{m+x-1}{x}\sum_{k=0}^{m}\binom{m}{k}(-1)^k(1+\frac{\theta(x+k)}{\theta^2+1})(1+\frac{x+k}{\theta})^{-}$$

3. This distribution is a special case of the generalized negative binomial-Shanker(GNB-SH) with the parameter $\beta=1$.

Extended ZINB Version

We create a ZINB-Shanker distribution by distinguishing the structured and random zeros. The model is a mixture of Bernoulli and Negative binomial-Shanker. The model for zero is as below:

$$g(x) = \begin{cases} \pi + (1-\pi)((1+\frac{\theta x}{\theta^2+1})(1+\frac{x}{\theta})^{-2} & x \\ (1-\pi)\binom{m+x-1}{x}\sum_{k=0}^{m}\binom{m}{k}(-1)^k(1+\frac{\theta(x+k)}{\theta^2+1})(1+\frac{x+k}{\theta})^{-2} & x \end{cases}$$

Extended ZINB Version

To get the properties of the ZINB-SH, some general rules on finding mean and variance of zero inflated models are used.
For a zero inflated model with random variable(y),
$E(y) = (1-\pi)E(Y|Z=0)$.
$Var(y) = (1-\pi)var(y|z=0) + \pi(1-\pi)E(y|z=0)^2$.
The mean of ZINB-SH is therefore,
$E(x) = (1-\pi)\frac{m\theta^2}{\theta^2+1}(\omega-\delta)$
and variance can be found by,
$var(x|z=0) = \frac{m(m+1)\theta^2}{\theta^2+1}(\rho-2\omega+\delta) - [\frac{m\theta^2}{\theta^2+1}(\omega-\delta)]^2$
which results in,
$var(z) = (1-\pi)\frac{m(m+1)\theta^2}{\theta^2+1}(\rho-2\omega+\delta) - [\frac{m\theta^2}{\theta^2+1}(\omega-\delta)]^2 + \pi(1-\pi)\frac{m(m+1)\theta^2}{\theta^2+1}(\rho-2\omega+\delta)$.

Simulations

We employ iterative method and use R program to run the algorithm.
The goal is to generate random numbers following the ZINB-SH(m,θ).
The steps are:

- Generate U from the Uniform(0,1) distribution.
- Let λ come from the shanker distribution(θ).
- Generate Y from the NB(m,p) distribution.
- Generate U^* fron the Uniform(0,1) distribution.
- if $U^* > \phi$ set X=Y, otherwise X=0

Data Description

- 8.2% of the facilities sampled were from Kisumu county, 47.5% from Mombasa county and 44.3% from Nairobi county.
- Testing of HIV for exposed infants were mainly done when they were less than 2 months(33.2%) since early detection of HIV infection to the child could assist in early treatment and special care be given to the child.
- The HEI Prophylaxis mostly prescribed at the facilities for the infants was NVP+AZT(31.2%) and the least prescribed was NVP for 12 weeks(3.9%). (0.2%)
- For the case of maternal prophylaxis, the most prescribed ARV dose for the mothers was AZT+3TC+ATV/r(15.3%) and the least prescribed as TDF+3TC+DTG

Results

This is still an ongoing M.Sc project some results from real data are yet to be established.
The HIV exposed infants data will be fitted to the ZINB-SH model and performance compared to other inflated models using the AIC values.