2019

Distributions of Zero-inflated Models with Application to HIV Exposed Infants

Faith Victory Nekesa
Strathmore Institute of Mathematical Sciences (SIMS)
Strathmore University

Distributions of Zero-inflated Models with Application to HIV Exposed Infants

Faith Victory Nekesa

A thesis submitted in partial fulfilment of the requirements for the award of the Degree of Master of Science in Statistical Science at Strathmore University

Statistical Institute of Mathematics (SIMS)
Strathmore University
Nairobi, Kenya.

May 31, 2019

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Name of candidate: Faith Victory Nekesa

Signature: .....

Date:.....

Approval

The thesis of **Faith Victory Nekesa, Reg No. 103017** was reviewed and approved by the following:

Dr. Collins Odhiambo

Lecturer, Strathmore Institute of Mathematical Sciences (SIMS)

Strathmore University

Dr. Linda Chaba

Lecturer, Strathmore Institute of Mathematical Sciences (SIMS)

Strathmore University

Ferdinand Othieno,
Dean, Strathmore Institute of Mathematical Sciences,
Strathmore University.

Professor Ruth Kiraka,
Dean, School of Graduate Studies,
Strathmore University.

Abstract

Background: The instances of data with excess zeros are commonly found in many disciplines, including the public health. Several models have been proposed when analyzing this kind of data. The World Health Organization (WHO) indicates that majority of the 1.8 million children who are at the present with HIV in sub-Saharan Africa got the HIV virus from their mothers probably during delivery, pregnancy or through breastfeeding, but the study shows there is a drop in the the rate of infections due to interventions that have been put in place. Here we attempt to fit zero-inflated models to data in this setting.

Objective: The objective is to systematically compare distributions of the various zero-inflated models with an application to HIV Exposed Infants (HEI).

Methods: We revisit zero-inflated models, conducted the simulations and applied the models to HEI data. The models performance were evaluated by Akaike Information Criteria(AIC).

Results of the study :The simulation results indicated ZAP had the lowest AIC value value of 467.95 at 80% of zeros. The real data showed ZAP as the best fit for the simulation data since it had the lowest AIC value.

Conclussion :From the simulations results of the AIC value and the the real data results, it is clear that ZAP is the best fitting model.

Contents

List Of Abbreviations

AIDS	Acquired Immuno-Deficiency Syndrome
AIC	Akaike Information Criterion
ART	Anti retroviral Therapy
ARV	Anti retroviral drug
BIC	Bayesian Information Criterion
CD4	Cluster of Differentiation 4
DHIS	District Health Information System
EID	Exposed Infant Diagnosis
HEI	HIV Exposed Infants
HIV	Human Immuno-Deficiency Virus
LTFU	Loss To Follow-up
MTCT	Mother-to-child transmission
NASCOP	National AIDS and STI Control Programme
NVP	Nevirapine
PCR	Polymerase Chain Reaction
PMTCT	Prevention of mother-to-child transmission
UNAIDS	United Nations Program on HIV/AIDS
UNICEF	United Nations Childrens Fund

WHO	World Health Organization
ZAP	Zero-altered Poisson model
ZANB	Zero-altered negative binomial model
ZIM	Zero-inflated models
ZINB	Zero-inflated negative binomial model
ZIP	Zero-inflated Poisson model

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background of the Study

Latest assessments done by the Joint United Nations Program on HIV/AIDS (UNAIDS) show that roughly 330,000 children in the world who are below 15 years of age became infected with the virus in 2012. The sub-Saharan Africa region has become the utmost relentlessly affected and accounts for above 90% of pediatric infections (UNAIDS, 2012). Majority of the infections happened in the course of delivery,pregnancy or breastfeeding hence making the PMTCT a priority in the public health sector (Le Coeur S, et al 2003).

Kenya has implemented the World Health Organization(WHO) guidelines based on four divided methodology in order to avoid the mother-to-child spread of HIV. The methods constitute: key prevention of infection by women at the childbearing age; preventing unexpected pregnancies among the females having the HIV virus; averting the spread of HIV from females who live with HIV to their infants and giving the right medication; attention and sustenance to mothers, families and their children living with HIV (Mahy, et al. 2013; NASCOP, 2013; UNAIDS, 2013; UNICEF, 2013).

With the advent of PMTCT interventions across all facilities in Kenya, this has significantly reduced sero-conversion. However, with public health gaps particularly weak systems, there are still pockets of sero-conversion among HIV Exposed Infants (HEI). This gives rise to structured zeros (where PMTCT is effective) and unstructured zeros (where PMTCT is ineffective).

Poisson and negative binomial models have been commonly used for count data (Lambert D. 1992). One assumption of the Poisson regression is that the response variable of mean and variance are equal. Actually in often situations,

the variance is always larger than the mean, which is called over-dispersion. The negative binomial model can be used in the case of over-dispersion since it permits the variance and the mean to be dissimilar.

Both zero inflated model and zero altered model constitute a logistic regression for the zeros in the data and a count regression (either Poisson or negative binomial) for the counts. The difference is how they deal with different types of zeros. While the count process of the zero altered model is a zero-truncated (i.e. the distribution of the response variable cannot have a value of zero), the count process of Zero-inflated can produce zeros (Zuur, et al., 2009).

For count data, depending on an outcomes mean-variance relationship and proportion of zeros, the choices for modeling its distribution range from standard Poisson and negative binomial to zero-inflated Poisson (ZIP), zero-altered Poisson (ZAP), zero-inflated negative binomial (ZINB) model and zero-altered negative binomial (ZANB) models. However, some researchers argue that they have seen cases where ZIP models were inadequate and ZINB also couldn't be reasonably fitted to the data (Famoye and Singh, 2006).The main objective of this study is to determine the best models when dealing with both structured and non-structured zeros in a zero-inflation setting.

1.2 Summary of Zero-inflated and Zero-altered count data models

The ZIP, ZINB, ZAP and ZANB models are mostly used to model zero-inflated and zero-altered count data respectively.

1.2.1 Zero-inflated Poisson (ZIP) Model

The model was suggested by Lambert (1992) with reference to defects in a manufacturing process. In the model, the outcomes $\mathbf{Y}=(\mathbf{Y}_1, (\mathbf{Y}_2..., (\mathbf{Y}_n)^{\mathbf{t}}$ are independent. A postulation of this model is that the only possible observation is 0 given the probability is p , and with probability (**1-p**), a Poisson (λ) random variable is examined in Y.

The mean and variance of ZIP distribution are;

$$E(Yi) = (1-p)^{\lambda_i} \tag{1.1}$$

$$V(Yi) = (1-p_i)(\lambda_1+\lambda_2)((1-p_i)^{\lambda_i})^2 \tag{1.2}$$

The Poisson mean vector has the canonical link $\log(\lambda)$ for a Poisson regression model.

The canonical logit link is regarded for the parameter vector $\mathbf{P} = (p_1, ..., p_n)$ in regression. That is,

$$logit = \log(\frac{p}{1-p}) = \mathbf{G}_\gamma \tag{1.3}$$

$$p = \frac{e^{\mathbf{G}_\gamma}}{1 + \mathbf{G}_\gamma} \tag{1.4}$$

$$(1-p) = \frac{1}{1 + \mathbf{G}_\gamma} \tag{1.5}$$

The ZIP model has two components, one component is to model the probability of being the structural zeros ρ using the logistic regression and the other component is to model the Poisson mean μ.Thus, the presence of structural zeros gives rise not only to a more complex distribution, but also creates an additional link function for modeling the effect of explanatory variables for the occurrence of such zeros. In other words, the ZIP model enables us to better understand the effect of covariates by distinguishing the effects of each specific covariate on structural zeros and on the non-structural zeros

1.2.2 Zero-inflated Negative Binomial (ZINB) Model

The ZINB distribution is a mixture distribution, similar to ZIP distribution, where the probability p for excess zeros and with probability $(1-\rho)$ the rest of the counts followed negative binomial distribution. Negative binomial distribution is given by:

$$P(Y = y) = \frac{\gamma(y+\tau)}{y!\gamma(\tau)}(\frac{\tau}{\lambda+\tau})^\tau(\frac{\lambda}{\lambda+\tau})^y, y = 0, 1, \cdots; \lambda, \tau > 0 \tag{1.6}$$

where $\lambda = E(Y)$, τ is a shape parameter which quantifies the amount of over dispersion, and Y is the response variable of interest. The variance of Y is $\lambda + \lambda^2/\tau$.

The mean and variance of the ZINB distribution are;

$$E(Y) = (1-p)\lambda \tag{1.7}$$

$$V(Y) = (1-p)\lambda(1 - p\lambda + \frac{\lambda}{\tau}) \tag{1.8}$$

This distribution gets closer to the ZIP distribution and the negative binomial distribution as $\tau \to \infty$, and $p \to 0$, respectively.

The ZINB regression model tells about ρ and λ to covariate matrix $\mathbf{X}$ and $\mathbf{Z}$ with regression parameters β and γ as;

$$log(\lambda_i) = x\beta \tag{1.9}$$

$$logit(\lambda_i) = z\gamma, i = 1, 2, \cdots, n \tag{1.10}$$

The ZINB log-likelihood given the observed data is:

$$l(\beta, \tau, \gamma; \mathbf{y}, \mathbf{z}, \mathbf{x}) = \sum_{i=1}^{n} \log(1 + e^{z_i\gamma}) - \sum_{i=1:y_i=0}^{n} \log(e^{z_i\gamma} + (\frac{e^{x_i\beta} + \tau}{\tau})^{-\tau}) + \tag{1.11}$$

$$\sum_{i=1:y_i>0}^{n} (\tau \log(\frac{e^{x_i\beta} + \tau}{\tau}) + y_i \log(1 + e^{-x_i\beta}\tau)) + \sum_{i=1:y_i>0}^{n} (\log\Gamma(\tau) + \log\Gamma(1 + y_i) - \log\Gamma(\tau + y_1)) \tag{1.12}$$

We may use the ZINB model when there is still dispersion in the at-risk subgroup (unstructured data), which is identical to ZIP, except that the NB replaces the Poisson to account for over dispersion for modeling the count response from the at-risk sub-population.

1.2.3 Zero-altered Poisson (ZAP) Regression

Also called the Poisson hurdle (PH) (Hilbe, 2011). ZAP model comprises of a hurdle component which models zero versus non-zero counts, and a truncated Poisson count component that is used for the non-zero counts:

$$p(Y_i = 0) = \rho_i \tag{1.13}$$

$$p(Y_i = k) = (1 - \rho_i), k = 0, 1, \cdots \tag{1.14}$$

ρ_i models all zeros. For ZAP model, the best choice to model probability of zeros is to use a logistic regression model:

$$logit(\rho_i) = x_i B \tag{1.15}$$

The ZAP model does not categorize the zeros in the data as structured zeros or unstructured zeros. It overlooks on that concept which may bring about false interpretations of results and the study findings.

1.2.4 Zero-altered Negative Binomial (ZANB) Regression

It is also known as the negative binomial logit hurdle (NBLH) (Hilbe, 2011). Similarly, ZANB can be used in case of over-dispersion instead of applying the Poisson distribution. The best choice to model probability of excess zeros is to use a logistic regression model (Hilbe, 2011):

$$logit(\rho_i) = x_i B \tag{1.16}$$

The ZANB model which is an extension of ZAP model also assumes the existence of the structured zeros and unstructured zeros. It overlooks on that concept hence may bring about false interpretations of results and the study findings.

1.3 Statement of the Problem

In a typical clinic set-up, there is likely to be many zeros in the data collection of the occurrence of new infections due to improved PMTCT policies. Whereas it is expected that PMTCT policies are implemented uniformly across all public health facilities, implementation naturally differ at different facilities due to differential health systems and infrastructure. This leads to structured zero positive HEI (where implementation is optimum) and non-structured zero positive HEI (where implementation is not optimum). Hence the zero-inflated models will have to be used in the analysis due to the abundance of structured and non-structured zeros in the data. Due to effective PMTCT interventions at different facilities, sero-conversion among HEI has reduced considerably therefore data collected are zero-inflated (contain many zeros) and are difficult to predict. Failure to account for structured and non-structured zero-inflation may result in inference that is not true and also misleading(Lambert D. 1992).

1.4 Objectives

1.4.1 General Objective

To evaluate distributions of the various zero-inflated models with an application to HEI.

1.4.2 Specific Objectives

- To conduct simulations tests to compare performances of various zero-inflated models.
- To compare the performance of the various zero-inflated models when applied to HEI data.
- To determine covariates that are significantly associated with the outcome of interest in the HEI data.

1.5 Significance of the Study

It will enable Ministry of Health (MoH) and health authorities to precisely predict sero-conversion among HEI given the skewed zeros; a situation that is rampant where PMTCT interventions are effective.

Chapter 2

Literature Review

2.1 Introduction

HIV and AIDS is a major difficult in many countries globally and it continues to have disturbing effects in the Sub Saharan Africa which has the majority of HIV infected people at over 90% of all the HIV cases in the world. In spite strategies and education have been put in place to try and eradicate new HIV infections among the infants, no big change has been experienced (UNAIDS, 2013). The unveiling of the global plan in 2011 which aimed to reduce new HIV infection among children by 2015 and quicken efforts towards HIV children and their mothers across Millennium Development Goals (MGDs) which include; improve maternal health, reduce child mortality and combat HIV, tuberculosis and Malaria. This will comprise the four split approach (UNAIDS, 2013; UNICEF, 2013).

This section will provide a theoretical review of existing standard of ZIM. That is; ZIP, ZINB, ZAP, ZANB and the literature about the occurrences of new HIV infections infant HIV.

2.2 Zero-inflated models

Statisticians have come up with new approaches to model zero-inflation in count data. Lambert (1992) suggested an approach to use ZIP model. In his model, two kinds of zeros are said to exist in the data: structured zeros (true zeros) from a non-vulnerable group (an example being people who are healthy and without a disease) and non-structured zeros (false zeros) for those from a susceptible

group (example being those that have a disease in a study that is health-based who may wrongly show a score of zero).

The Poisson hurdle model or ZAP model was first initiated by Mullay (1986) and later it was adjusted by King (1989). It models all zeros as one part and a zero-truncated part for all non-zero studies. The major distinction with ZIP is that hurdle models do not distinguish between the structured and non-structured and all zero observations are assumed to come from a non-vulnerable group.

The ZINB and ZANB models are an extension of the ZIP and ZAP models respectively. Both of the models deal with zero-inflation and over-dispersion at the same time. These types of models have become a bit popular lately and they have been used in several project research which include the following; to analyze number of cigarettes smoked per day (Schunck and Rogge, 2012), dental health status (Wong and Lam, 2012), depressive symptoms (Beydoun, 2012), and alcohol consumption (Atkins, 2012), etc. The main benefit of using these models more specifically when handling zero-inflation is that they do decrease biases that result from the extreme non-normality and also they can model the effect on subjects vulnerability and the magnitude at the same time.

Several studies have been done to give comparison of different performance of models for zero inflated data. The examples of some simulation studies in the literature that compared different model performance for data with many zeros include the studies done by Min Y and Agresti A., 2005, study also by Desjardins CD., 2013 and one done by Miller JM., 2007. Nevertheless, the comparisons made in these studies are limited. For example, Min and Agresti concentrated on comparing parameter estimations of Poisson hurdle with ZIP model; Desjardins evaluated the model performance of ZINB with ZANB models and Miller compared the goodness of fit for Poisson, ZAP and ZIP models.

This thesis we compare the performance of different zero-inflated and zero-altered models using randomly simulated data and the model selection based on AIC.

2.3 Occurrence of New HIV Infections

According to the global plan progress report by UNAIDS (2013), it indicated that there were 210,000 freshly infected infants in Sub Saharan Africa in 2012 which represented a fall of 37% from 2009. In Kenya 13,000 new HIV infections amongst the children were reported in 2012, which showed a reduction of 44% from 2009. This indicated that above half (58%) of the HIV positive mothers received their ARV prophylaxis and also 80% of infants were not on medication despite of them being breast fed. The HIV transmission rate of MTCT report declined to 15% from 26% in 2009. The women in the procreative age (15-49

years) with HIV infection also reduced from 56,000 to 46,000, this indicated that fewer children were exposed to HIV infection. Recently there has been a reduction in the number of HEI. The map in figure 2.1 clearly indicates that.

2.4 Factors Contributing to New Infections in Infants

2.4.1 Feeding options for the Infant

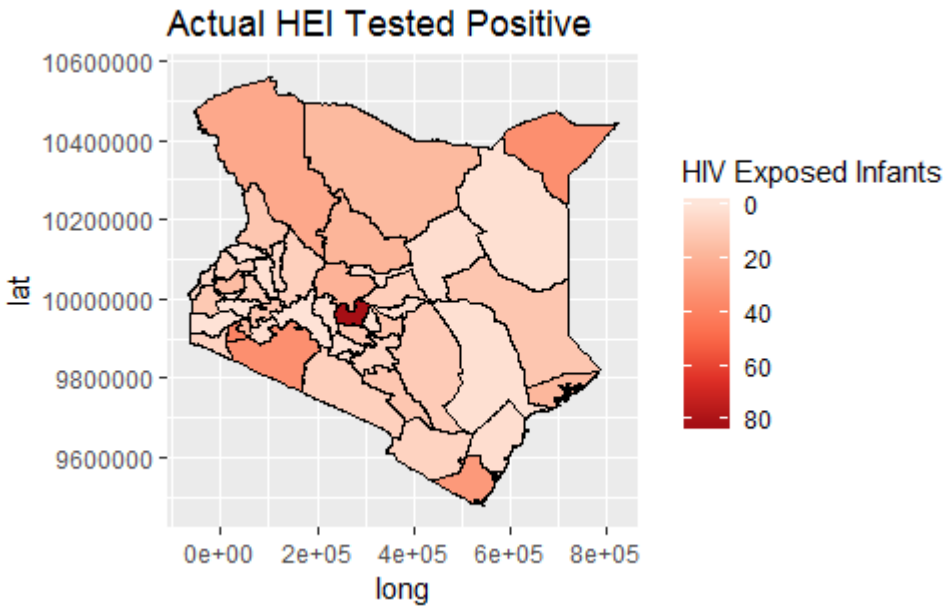
Replacement feeding is suggested to decrease the risk of HIV transmission though sanitation is recommended to reduce ill health and death. The options of feeding are limited to breast feeding with Anti retroviral drugS (ARVs) and exclusive replacement and guided on the selections and the challenges to help them make conversant selection. Those women who choose to breast feed should be supported and encouraged to be exclusive until 6 months and afterward upkeep by complementary until 12 months. The threat to mix feed before 6 months, increases the risk of HIV infections. Exclusive replacement feeding is done for six months and complementary feeds from 6 months. In the maternal conditions where the mother is dead, the infants are recommended exclusive additional feeding (NASCOP, 2012).

2.4.2 Socio Demographic Factors

The socio demographic data gives the distinctiveness of an individual, particularly in terms of maternal age, level of literacy and understanding of taking ART prophylaxis. Cook et al, (2013) study indicated that all health centres offered PMTCT services as routine prenatal care and early infant analysis was at 25% . A study by Gourlay et al., (2013) indicated that maternal age on young women aged 20-25 years were less likely to receive ARV prophylaxis and receive NVP for their infants.

2.4.3 Organizational Barriers

Findings by Cherutich et al, (2008), on barriers in timely realization of early infant diagnosis in among 58 health facilities giving the pediatric HIV services in Kenya showed that utmost health care providers were not familiar with HIV peadriatric guidelines. In some facilities its the children who were brought by their parents for testing were referred to other clinics for diagnostic testing at a cost. Shortage of personnel in performing the test in the laboratory is



HEI.png

Figure 2.1: *Map of kenya showing the number of HEI that were tested positive. (Source: NASCOP EID database)*

showed too (Cherutich et al., 2008). The study clearly show that there is also an urgency to train health care providers, to be overseen and provided with supplies in recognizing the exposed and children infected with HIV.

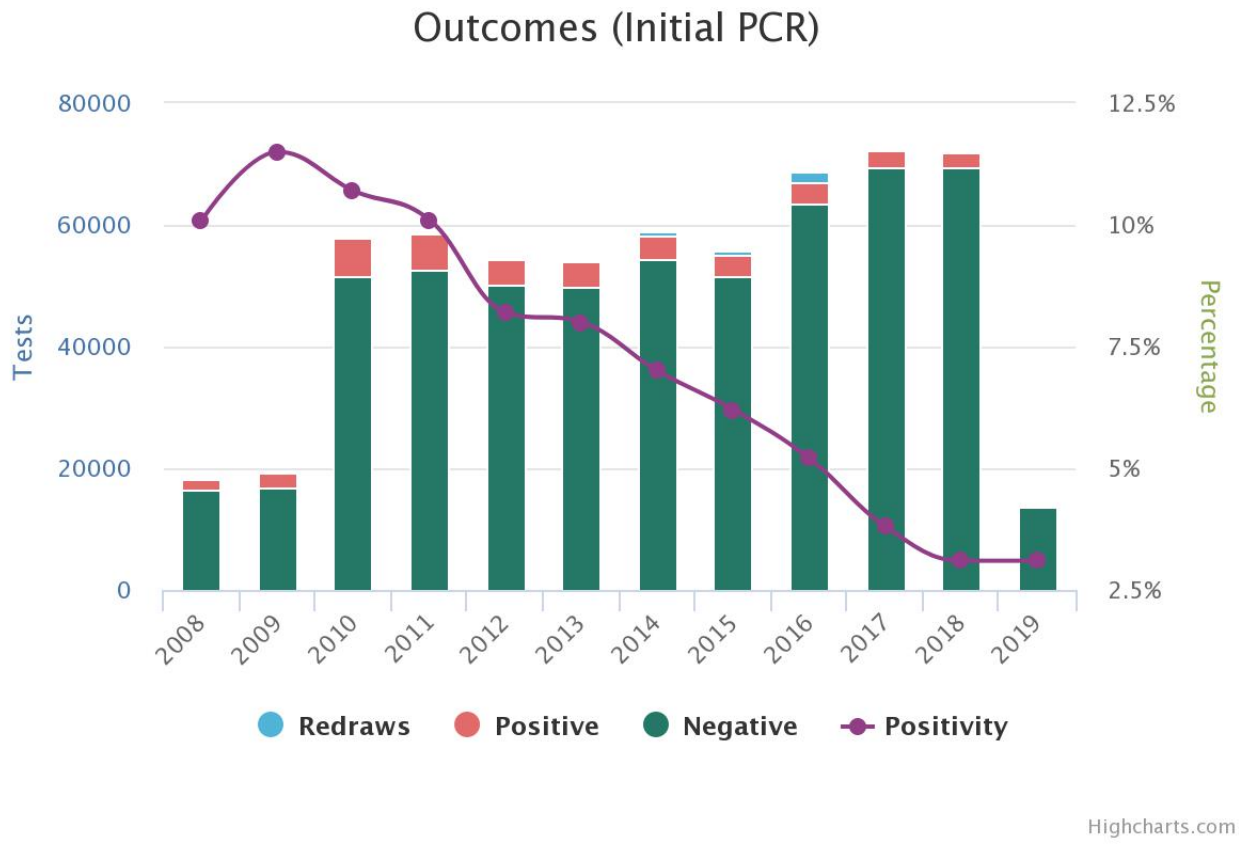


Figure 2.2: *Graph showing the outcomes of the initial pcr type in Kenya from the year 2008 to 2019. The line graph indicates a drop in the positivity in the years. (Source: NASCOP EID database)*

Chapter 3

Research Methodology

3.1 Study setting and design

The PMTCT HEI program in Kenya is overseen by NASCOP through Ministry of Health. Since 2007, there was testing of HIV-exposed symptomatic infants; in 2008 to 2009, as more resources became available for testing, the guidelines modified to test all HIV-exposed infants. The testing of the infant HIV algorithm since 2012 in Kenya was as follows: a maternal or infant HIV antibody test was conducted at first visit for all children of unknown HIV status aged <18 months to establish HIV exposure status. If found positive, an EID test was suggested. (http://eid.nascop.org/).

3.2 Study population

A total of 413 samples were collected from infants visiting health facilities across three counties in Kenya (Mombasa, Kisumu and Nairobi) between January 2016 and July 2017 and tested in seven national laboratories.

3.3 Ethical approval

Data collected from this study is secondary and openly available and hence did not require scientific ethical approval. No identifiable patient information was included in the database.

Table 3.1: **Simulation design for the models**

Factor A: ω	**Factor B:** **k**	**Factor C:** **Models tested on each condition**
0.00	1	Poisson regression model (Poisson)
0.20	10	Negative binomial regression model (NB)
0.40	50	Zero-inflated Poisson model (ZIP)
0.60	100	Zero -inflated negative binomial model (ZINB)
0.80		Zero -altered Poisson model (ZAP)
		Zero -altered negative binomial model (ZANB

3.4 Statistical analysis

Different models were compared using simulations. The simulation set up is described in section 3.5. In order to get the best model which fits the data,and is also a model with lower prediction error, stepwise regression for model selection will be used in selection of variables under each of the models. The models will also be fitted HEI data and comparison of the performance of AIC will done using the AIC.

Demographic data will be summarized with descriptive statistics. The major outcome was number of infant HIV who turned positive. We will examine the health facility attended by the mother, the number of EID Positive, EID Testing Point, PCR Type, testing Point, HEI prophylaxis and the maternal Prophylaxis. Different zero-inflated models will be applied to HEI data. The best model based on AIC approach will be used to determine covariates that are associated with the outcome of interest (EID positive). All analysis in this study was conducted using R Studio version 3.5.3.

3.5 Simulations

Simulated data were created with unpredictable percentages of zeros and a fixed sample size of 500. A condition which had no zero-inflation ($\omega = 0.00$) was tested and used as a standard comparison point. The effect of over-dispersion was observed in the non-zero part. The dispersion parameter k was used with the following values: 1, 10, 50, and 100 which were pre-stipulated. These values represent a range of dispersion which is practical to aid in the assessment of the value of different models under study with varying distributions. The larger the value of k, the less dispersed the variable is and it approaches a Poisson distribution when $k > 10$. Negative binomial distribution was used to generate the response variable with different proportion of zeros added.

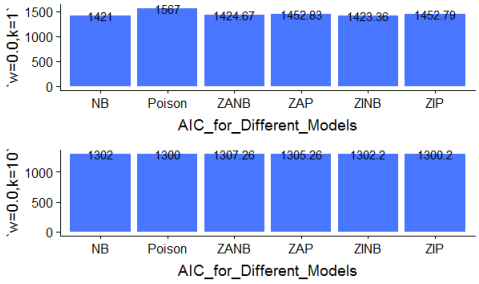


Figure 3.1: AIC for w=0.0,k=1,10. At k=1, the model with the lowest AIC value is NB,hence regarded as best fit and the worst fit model was Poisson.At k=10,the best model was Poisson and the worst model was ZANB

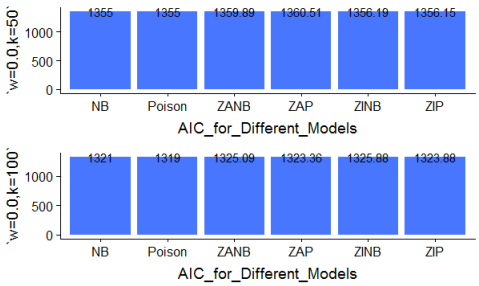


Figure 3.2: AIC for w=0.0,k=50,100. At k=50, the model best model fit is NB and the worst fit model is ZANB. At k=100,the best model is Poisson and the worst model is ZINB

3.5.1 Model selection criteria

To get the best model, the goodness of fit test is done using the AIC (Akaike information criterion). The model with minimum AIC was considered as the best model to fit the data (Lambert, 1992). AIC is given by:

$$AIC = 2logL(\theta) + 2c, \tag{3.1}$$

where $L(\theta)$ is the maximized likelihood function for the estimated model and $L(\theta)$ offers summary information on how much discrepancy exists between the model and the data, where c is the number of free parameters in the model. AIC is used to check the goodness of fit of the model and the complexity of the model.

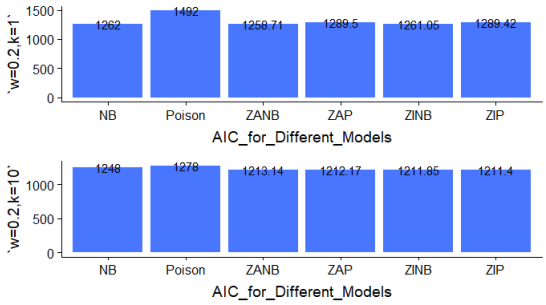


Figure 3.3: AIC for w=0.2,k=1,10. At k=1, the model best model fit is ZANB and the worst fit model is poisson. At k=10,the best model is ZIP and the worst model is Poisson

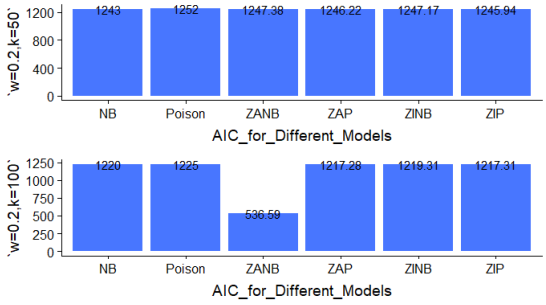


Figure 3.4: AIC for w=0.2,k=50,100. At k=50, the model best model fit is NB and the worst fit model is poisson. At k=100,the best model is ZANB and the worst model is Poisson

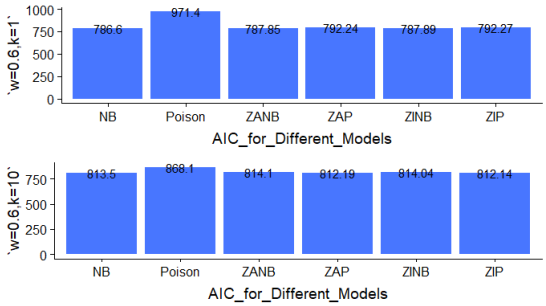


Figure 3.5: AIC for w=0.6,k=1,10. At k=1, the model best model fit is NB and the worst fit model is poisson. At k=10,the best model is ZIP and the worst model is Poisson

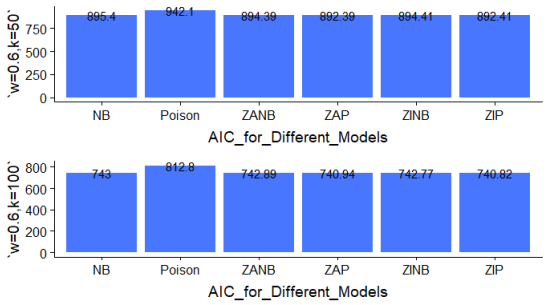


Figure 3.6: AIC for w=0.6,k=50,100. At k=50, the model best model fit is ZAP and the worst fit model is poisson. At k=100,the best model is ZIP and the worst model is Poisson

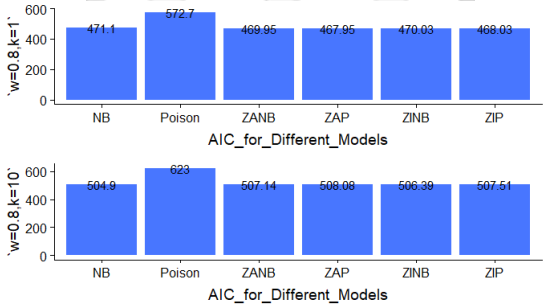


Figure 3.7: AIC for w=0.8,k=1,10. At k=1, the model best model fit is ZAP and the worst fit model is poisson. At k=10,the best model is NB and the worst model is Poisson

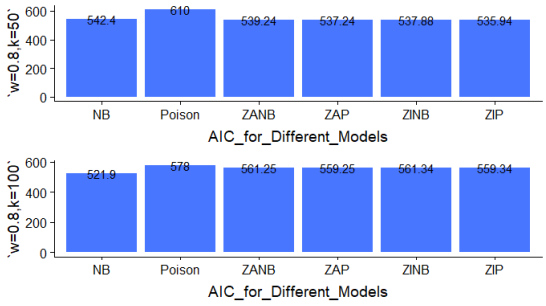


Figure 3.8: AIC for w=0.8,k=50,100. At k=50, the model best model fit is ZIP and the worst fit model is poisson. At k=100,the best model is NP and the worst model is Poisson

Chapter 4

Presentation of Research Findings and Discussions

4.1 Simulation results

The model with the lowest AIC value showed a more preferred model across all simulations for the five levels of zero-inflation combined with four levels of over-dispersion on the six models are presented in Table . Under the condition of no zero inflation,(ω=0.00) a Poisson model was preferred when the dispersion parameter k=10 since it had the lowest AIC value with a low dispersion. When k=1, 50 and 100 under the same condition of no zero inflation, the negative binomial is the most preferred model since it had a lower AIC compared to the other models. When data exhibited 20% of zero inflation, ZIP model was most preferred at k=10. When data exhibited 40% of zero inflation, the best model preferred was a negative binomial with a low dispersion of k=1. When data exhibited 60% of zero inflation, the model with the lowest AIC was ZIP with k=100. With 80% of zeros, the model best preferred was ZAP when k=1, poisson had the highest AIC value hence the least preferred among the models. Generally, ZAP had the lowest AIC value value of 467.95 at 80% of zeros. This showed clearly it was the best fit for the simulation data.

4.2 Results from empirical data analysis

4.2.1 Descriptive statistics for variables

Descriptive statistics which include means, frequencies, and percentages for the variables of EID Positive, County, EID Testing Point, HEI prophylaxis and Maternal Prophylaxis are shown in Table 4.1. 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%). 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 (0.2%)

4.2.2 HEI data outcomes

The HIV exposed infants data is fitted with the zero-inflated models which are; ZIP, ZAP, ZINB and ZANB. The performance of the inflated models will be compared using the AIC values. The results are presented below;

Four models described in Chapter 1 were used to fit the data which had a mixture of structured and non-structured zeros. The AIC values for the different models are presented in Table 4.2. The ZAP model had the lowest AIC value(490.81) indicating the best fit to the data and also works well when we have a mixture of both structured and non-structured zeros. ZINB model had the highest AIC value (492.11) indicating a poor fit for the model. Estimates of the regression coefficients and standard errors are presented separately for all the 4 models in the tables 4.3, 4.4, 4.5 and 4.6. For the Poisson model (referred to as model 1 in the analysis),using the step wise model selection criteria dropped the variables that were not significant (county) and the variables that remained included EID Testing Point, PCR Type, Testing Point, HEI prophylaxis and Maternal Prophylaxis. In the EID Testing Point, the significant testing point is between 2-9 months which shows that the risk is 2.39 times higher for HEI between 2-9 months compared to testing between 0-2 months. For the PCR Type, the chain reaction which was significant was that of 2nd/3rd PCR hence it indicates that a HEI is 2.9 times more likely to detect the HIV virus compared to the initial PCR. In the HEI prophylaxis with comparison to using AZT for 6 weeks+ NVP for over 12 weeks; the risk of using Nevirapine during breastfeeding on HEI is 5 times higher, the risk of using nevirapine for 6 weeks (mother not breastfeeding) is on the HEI is 6.5 times higher, the risk of using other drugs

Table 4.1: **Descriptive statistics for HEI Data**

Variables	**Freq(%)**
No.of EID Positive	N=413
County	
Kisumu	34 (8.2%)
Mombasa	196 (47.5%)
Nairobi	183 (44.3%)
EID Testing Point	
12-24 months	69 (16.7%)
2-9 months	69 (16.7%)
9-12 months	70 (16.9%)
Above 24 months	68 (16.5%)
less 2 months	137 (33.2%)
HEI prophylaxis	
AZT for 6 weeks + NVP for over 12 weeks	123(29.8%)
AZT for 6 weeks + NVP for 12 weeks	16 (3.9%)
NVP during BF	25 (6.1%)
NVP for 12 Wks	16 (3.9%)
NVP for 6 weeks (Mother on HAART or not BF)	30 (7.3%)
NVP+AZT	129 (31.2%)
Others	18 (4.4%)
Sd NVP+AZT+3TC	39 (9.4%)
Sd NVP Only	17 (4.1%)
Maternal Prophylaxis	
AZT (From 14wks or later) +	
sdNVP + 3TC + AZT+3TC for 7 days	17 (4.1%)
AZT+3TC+ATV/r	63 (15.3%)
AZT+3TC+EFV	33 (8.0%)
AZT+3TC+LPV/r	35 (8.5%)
AZT+3TC+NVP	33 (8.0%)
Interrupted HAART (HAART until end of BF)	3 (0.7%)
SdNVP Only	5 (1.2%)
TDF+3TC+ATV/r	33 (8.08%)
TDF+3TC+DTG	1 (0.2%)
TDF+3TC+EFV	124 (30.0%)
TDF+3TC+LPV/r	33 (8.0%)
TDF+3TC+NVP	33 (8.0%)

is 3.2 times high, the risk of using a combination of Sd NVP+AZT+3TC is 4.6 times high and lastly the risk of using Sd NVP only is 3.4 times high. Under the Maternal Prophylaxis, in comparison to the use of AZT (From 14wks or later)+Sd NVP+3TC+AZT+3TC for 7 days the risk of using a combination of AZT+3T+EFV(Efavirenz, which is a capsule and taken by mouth with plenty of water) by the mother to the infant is 5.6 times less, then the risk of using

Table 4.2: **Model fit comparison for HEI Data**
The best model fit for the HEI data is Hurdle Poisson and the worst model fit is ZINB.

Model	**AIC value for HEI Data**
Hurdle Poisson	490.81
Zero-Inflated Poisson	491.18
Hurdle Binomial	491.73
Zero-Inflated Negative Binomial	492.11

combination of AZT+3TC+LPV/r(Lopinavir/Ritonavir, which come in tablet forms) is 3.6 times less, the risk of using a combination of TDF+3TC+ATV/r is 3.5 times less, the risk of using a combination of TDF+3TC+LPV/r is 3.9 times less and lastly the risk of using a combination of TDF+3TC+NVP is 3.3 times lesser. The AIC value after fitting the Poisson model is 474.69,which is the second best fitting model for the EID data.

Negative binomial model(referred to as model 2 in analysis),had the AIC value of 429.19 which had the lowest AIC value hence it was considered as the best model. Using the step wise model selection,the following variables which were considered significant and had an effect on the final AIC value were retained;EID Testing Point, PCR Type, Testing Point, HEI prophylaxis and Maternal Prophylaxis. The PCR type that was significant was that of 2nd/3rd PCR type, which indicates that a HEI is 2 times more likely to detect the HIV virus in comparison to the initial PCR. In the HEI prophylaxis with comparison to using AZT for 6 weeks+ NVP for over 12 weeks; the risk of using NVP during breastfeeding on HEI is 4.2 times higher, then the risk of using nevirapine for 6 weeks (mother not breastfeeding) is on the HEI is 5.4 times higher, the risk of using other drugs is 2.3 times high, the risk of using a combination of Sd NVP+AZT+3TC is 3.3 times high and lastly the risk of using Sd NVP only is 2.4 times high according to the results above. Under the Maternal Prophylaxis, in comparison to the use of AZT (From 14wks or later)+Sd NVP+3TC+AZT+3TC for 7 days, the risk of using a combination of AZT+3T+ATV/r by the mother to the infant is 2.4 times higher, then the risk of using combination of AZT+3TC+LPV/r(Lopinavir/Ritonavir, which come in tablet forms) is 3.1 times less, the risk of using a combination of TDF+3TC+ATV/r is 3.4 times less, the risk of using a combination of TDF+3TC+EFV is 5.9 times lesser, then the risk of using a combination of TDF+3TC+LPV/r is 3.4 times less and lastly the risk of using a combination of TDF+3TC+NVP is 3 times lesser.

In the ZIP model, fitting the data using all the variables and using step wise model selection dropped most of the models and retained EID Testing Point and PCR Type which were the significant variables. Under the EID Testing Point,the risk of testing the infant between 2-9 months is 2.9 times higher to testing between 0-2 months. For the PCR Type in comparison to the initial

Table 4.3: **Zero-inflated Poisson model results**

Count model coefficients (poisson with log link):	Estimate	Std. Error	p-value
EID Testing Point			
2-9 months	0.6228	0.2085	0.00282 ∗ ∗
9-12 months	0.1734	0.2831	0.54027
12-24 months	-0.0601	0.3569	0.86628
Above 24 months	0.2580	0.2530	0.30798
PCR Type			
Confirmatory PCR	-2.8241	1.1495	0.01402∗
2nd/3rd PCR	0.3652	0.5440	0.50204
Zero-inflation model coefficients (binomial with logit link):			
	Estimate	Std. Error	p-value
EID Testing Point			
2-9 months	-0.7319	0.4359	0.09311 .
9-12 months	0.1948	0.5275	0.71198
12-24 months	0.2892	0.5635	0.60772
Above 24 months	-0.2998	0.4730	0.52621
PCR Type			
Confirmatory PCR	-10.5797	82.6014	0.89808
2nd/3rd PCR	-2.0356	0.7455	0.00632 ∗ ∗

Table 4.4: **Zero-inflated Negative Binomial outcomes**

	Estimate	Std. Error	p-value
EID Testing Point			
2-9 months	0.63452	0.23899	0.00793 ∗ ∗
9-12 months	0.17165	0.31957	0.59118
12-24 months	-0.06536	0.39827	0.86965
Above 24 months	0.26338	0.28738	0.35941
PCR Type			
Confirmatory PCR	-2.79575	1.18103	0.01792∗
2nd/3rd PCR	0.37511	0.60213	0.53330
Log(theta)	2.59197	1.19069 7	0.02949 ∗
Zero-inflation model coefficients (binomial with logit link):			
	Estimate	Std. Error	p-value
EID Testing Point			
2-9 months	-0.7150	0.4382	0.10273
9-12 months	0.1998	0.5301	0.70630
12-24 months	0.2866	0.5680	0.61384
Above 24 months	-0.2927	0.4757	0.53840
PCR Type			
Confirmatory PCR	-19.2306	6342.1830	0.99758
2nd/3rd PCR	-2.0276	0.7507	0.00691 ∗ ∗

Table 4.5: **Negative binomial hurdle outcome**

Count model coefficients (truncated negbin with log link):			
	Estimate	Std. Error	p-value
EID Testing Point			
2-9 months	0.65677	0.24183	0.00661∗ ∗
9-12 months	0.20880	0.32211	0.51685
12-24 months	-0.05442	0.39480	0.89036
Above 24 months	0.28004	0.28758	0.33017
PCR Type			
Confirmatory PCR	-9.80622	133.47361	0.94143
2nd/3rd PCR	0.40607	0.60548	0.50244
Log(theta)	2.58135	1.18422	0.02927 ∗
Zero hurdle model coefficients (binomial with logit link):			
	Estimate	Std. Error	p-value
EID Testing Point			
2-9 months	0.7273	0.4259	0.08768 .
9-12 months	-0.2469	0.5167	0.63274
12-24 months	-0.3400	0.5508	0.53702
Above 24 months	0.2962	0.4643	0.52350
PCR Type			
Confirmatory PCR	2.1431	1.3091	0.10161
2nd/3rd PCR	2.0897	0.7344	0.00444 ∗ ∗

PCR, it indicates that the HEI is 2 times less likely to detect the HIV virus during the Confirmatory PCR. Analyzing the model with the 2 variables gave an AIC value of 491.18.

The ZINB model using the stepwise regression,and the direction as backward dropped most of the variables that were not significant and was left with 2 variables which were EID Testing Point and PCR Type. Under the EID Testing Point,the risk of testing the infant between 2-9 months is 2.6 times higher to testing between 0-2 months. For the PCR Type in comparison to the initial PCR, it indicates that the HEI is 2.7 times less likely to detect the HIV virus during the Confirmatory PCR. The AIC value for the ZINB model is 492.11, hence regarded as the worst model fit for the data.

In the hurdle binomial(ZANB),using the step wise regression also dropped down the insignificant variables and was left with EID Testing Point and PCR Type. In the EID Testing Point,the risk of testing the infant between 2-9 months is 2.65 times higher to testing between 0-2 months. For the PCR Type in comparison to the initial PCR, it indicates that the HEI is 2.7 times less likely to detect the HIV virus during the Confirmatory PCR. The AIC value using the 2 variables

Table 4.6: **Hurdle (ZAP) model outcomes**

Count model coefficients (poisson with log link):	Estimate	Std. Error	p-value
EID Testing Point			
2-9 months	0.63982	0.21034	0.00235 ∗ ∗
9-12 months	0.20265	0.28420	0.47582
12-24 months	-0.05207	0.35466	0.88328
Above 24 months	0.27111	0.25325	0.28438
PCR Type			
Confirmatory PCR	-9.26276	103.50896	0.92869
2nd/3rd PCR	0.39130	0.54995	0.47677
Zero hurdle model coefficients (binomial with logit link)			:
	Estimate	Std. Error	p-value
EID Testing Point			
2-9 months	0.7273	0.4259	0.08768
9-12 months	-0.2469	0.5167	0.63274
12-24 months	-0.3400	0.5508	0.53702
Above 24 months	0.2962	0.4643	0.52350
PCR Type			
Confirmatory PCR	2.1431	1.3091	0.10161
2nd/3rd PCR	2.0897	0.7344	0.00444 ∗ ∗

was 491.73,which is the 2nd worst model fit for the data hence not preferred.

To determine the significant covariates, the ZAP model is used since its the best model based on the AIC value. The hurdle poisson (ZAP model), using the step wise model selection dropped the insignificant variables and was left with EID Testing Point and PCR Type. The baseline odds of having a positive count verses zero is 2.12 (exp(0.7556)). This odds is increased by 3.7(exp(2.0897)) times if 2nd/3rd PCR test is done as compared to the initial test. EID Testing point does not have significant effect. Given the response is positive, the average count is 1.97 (exp(0.67656)). This is increased by 1.9 (exp(0.63982)) times if testing is done at 2-9 months compared to less than 2 months. PCR Type does not have significant effect.

4.3 Discussions

Count data with excess zeros are commonly seen in medical research and public health particularly, number of HEI. Yip (1988) and Lambert (1992) proposed zero-inflated Poisson distribution and Heilbronn (1989) used zero-altered Poisson and negative binomial distributions to model this type of data. Li, Lu,

Park, Kim, Brinkley and Peterson (1999) derived multivariate version of the zero-inflated Poisson distribution and applied it to detect equipment problems in electronics manufacturing processes.

Zero-inflated distributions assume that with probability 1 - ρ the only possible observation is 0, and with probability ρ, a random variable describing defect counts is observed. Although different authors have widely used zero-inflated distributions, there is no practical study that systematically compare zero-inflated outcomes in HIV exposed infant setting. Because the zero-inflated model involves both Bernoulli parameter ρ and the state parameter k, we extensively conducted simulations by varying percentage of zeros and these parameters. The results of simulation shows ZAP generally had the lowest AIC value,when the percentage of zero was high. This is consistent with the results from the application data, this is because the percentage of zeros in the HEI dependent variable is 88%. The simulation procedure selected limited important model terms to maximize the ZI likelihood functions.

In all these ZI model, EID testing point and PCR type had significant effects on the response variable. Based on the HEI model, the rate of HIV sero-conversion was high for EID tested between 2-9 months compared to those tested earlier. The results of the studies done recently on patients showed sero-status is not different between boys and girls. This was however, not verifiable in our data because we didn't collect gender covariate.

Although clustered count data with extra zeros often occur, few methods have been developed for correlated data with extra zeros. There are some studies done on the extension of zero inflated models in order to accommodate random effects. In all of these models, there were two separate random effects in the models; therefore, the interpretation of the results was more difficult and sometimes confusing.

Zero-inflated models and zero-altered models give almost similar results from as shown from both simulated data and the HEI data. The decision when choosing between these two according to the study, heavily relied on the AIC value found after the analysis of the work.

Chapter 5

Conclusion and Recommendations

5.1 Conclusion

Data with many zeros are often encountered in many public health applications. Failure to account for the zero-inflation while analyzing such data may result in inferences which are not true. After the simulation study and analysis of EID Data, the negative binomial emerges as the gold-standard for as fitting the data with both structured and non-structured zeros.

5.2 Limitations of the study

There were also some limitations of the study which were observed. One, findings from the simulation study were only based on a limited number of conditions used. Simulation results offer a general idea as to which model is more appropriate, however, more conditions will need to be examined to get a more accurate relationship between the model selection and different levels of zero-inflation. Two, there were inconsistencies observed between the simulation results of the AIC value and the real data results which poses a big challenge to the research findings. Three, explanatory variables for the zero versus non-zero model and the count model were set to be the same. The most attractive advantage of using zero-inflated models is that they allow researchers to have different predictors for two parts of the models, which usually can be justified theoretically.

5.3 Further areas of research

One area for further research is the issue of imbalanced covariates with missing data. From the results of the analysis the variables with missing data affected the final outcome of the AIC values.

Bibliography

[1] Akaike, H. (1973),Information theory and an extension of the maximum likelihood principle: *Second International Symposium on Information Theory.*

[2] Asefa, A and Mitike, G., (2014),Prevention of Mother-to-Child Transmission (PMTCT) of HIV services in Adama town: *BMC Pregnancy and Childbirth.*

[3] Audureau, E., Kahn, J. G., and Besson, M., (2013),Scaling up prevention of mother-to-child HIV transmission programs in Sub Saharan African countries: *a multilevel assessment of site program and country level determinants of performance, BMC Public Health*

[4] Cherutich, P., Inwani,I., Nduati, R, and Mbori-Ngacha, D., (2008),Optimizing pediatric HIV care in Kenya: *challenges in early infant diagnosis*, Bulletin of world health organization.

[5] Cook, R., Ciampa, P., and Sidat, M., Blevin, M., Burlison, J., Davidson, M. A., Arroz, J. A., Vergara, A. E., Vermund, S. H., Moon, T. D., (2011),Predictors of successful early infant diagnosis of HIV in rural district hospital in Zambezia Mozambique: *Journal Acquired Immune Deficiency Syndome.*

[6] Desjardins C. D. (2013), Evaluating the Performance of Two Competing Models of School Suspension under Simulation The Zero-Inflated Negative Binomial and the Negative Binomial Hurdle. *PhD Thesis, University of Minnesota, USA.*

[7] Essomo, M., Meye, J.F., Belembaogo, E., Engoghan, E., Ondo, A. (2008),Prevention of mother-to-child transmission of HIV in Gabon: *The problem of children lost to follow-up.*

[8] Fewtrell, M.S., Kennedy, K., Singhal, A., Martin, R.M., Ness, A., Hadders-Algra, M. (2008),How much loss to follow-up is acceptable in long-term randomized trials and prospective studies?: *Arch Dis Child.*

[9] Gourlay, A, A., Birdthistle, I., Mburu, G., Iorpenda, K., and Wringe, A., (2013), Barriers and facilitating factors to the uptake of antiretroviral drugs for prevention of mother-to-child transmission of HIV in sub-Saharan Africa: *Journal of the International AIDS Society.*

[10] Heilbron, D. C. (1994), Zero-Altered and Other Regression Models for Count Data with Added Zeros, *Biometrical Journal, Journal of Mathematical Methods in Biosciences*, 36, 531-547.

[11] Hilbe, J. M. (2011),Negative Binomial Regression: *Cambridge University Press*,USA.

[12] Hu M, Pavlicova M, and Nunes E. (2011), Zero-inflated and hurdle models of count data with extra zeros: *Examples from an HIV-risk reduction intervention trial.* Am J Drug Alcohol Abuse 37: 367375. doi: 10.3109/00952990.2011.597280 PMID 21854279

[13] Jones, S.A., Sherman, G.G., Varga, C.A. (2005), Exploring socio-economic conditions and poor follow-up rates of HIV-exposed infants in Johannesburg: *AIDS Care*, South Africa.

[14] Kalembo, F.W., Zgambo, M. (2012),A Major Challenge to Successful Implementation of Prevention of Mother-to-Child Transmission of HIV-1 Programs in Sub-Saharan Africa: *Loss to Follow-up.*

[15] Lambert, D. (1992), Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing, *Technometrics*, 34, 1-14. doi: 10.2307/1269547.

[16] Le Coeur, S., Kanshana, S., Jourdain, G. (2003), HIV-1 transmission from mother to child and its prevention: *Mdecine Tropicale: Revue du Corps de Sant Colonial.*

[17] Le May, A., and Holmes, S. (2012), Introduction to nursing research: *Hodder Arnold*,London.

[18] Li, C. S., Lu, J. C., Park, J., Kim, K. M., Brinkley P. A., and Peterson, J. (1999), A Multivariate Zero-inflated Poisson Distribution and Its Inferences, *Technometrics*, 41(1), 29-38.

[19] Loeys T, Moerkerke B, Smet OD, et al. (2012), The analysis of zero-inflated count data: *Beyond zero inflated Poisson regression.* Br J Math Stat Psych 65: 163180. doi: 10.1111/j.2044-8317.2011.02031.x

[20] Mahy, M., Stover, J, Kiragu, K, Hayashi, C., Akwara, P., Luo, C., CStanecki, K., Ekpini, K., and Shaffer, N., (2010),What will it take to achieve virtual elimination of mother to child transmission of HIV?:*An assessment*

of current progress and future needs,Journal of sexually transmission infection.

[21] Miller J.M. (2007), Comparing Poisson, hurdle and ZIP model fit under varying degrees of skew and zero-inflation. *PhD Thesis, University of Florida, USA*.

[22] Ministry of Health , Nairobi, Kenya (2012),Kenya AIDs indicator Survey : *National AIDs and STI Control programme*.

[23] Min Y and Agresti A. (2005), Random effect models for repeated measures of zero-inflated count data. Stat Model 5: 119. doi: 10.1191/1471082X05st084oa.

[24] Mullahy J. (1986). Specification and testing of some modified count data models. J Econometrics 33: 341365. doi: 10.1016/0304-4076(86)90002-3.

[25] Ngwede, S., Gombel, N., Midzi, S., Tshimanga, M., Shambira, G., and Chadambuka, A., (2013), Factors associated with HIV infection among children born to mothers on the prevention of mother to child transmission programme at Chitungwiza Hospital: *BMC Public Health*,Zimbabwe.

[26] Rivero-Mndez, M., Dawson-Rose, C. S., and Sols-Bez, S. S., (2010),A qualitative study of providers perception of Adherence of women Living with HIV AIDS in Puerto Rico 15(2) : 232251.

[27] Sibanda, E.L., Weller, I.V.D., Hakim, J.G., Cowan, F.M. (2013), The magnitude of loss to follow-up of HIV-exposed infants along the prevention of mother-to-child HIV transmission continuum of care: *A systematic review and meta-analysis*.

[28] Sileshi G, Hailu G and Nyadzi GI. (2009), Traditional occupancyabundance models are inadequate for zero-inflated ecological count data. Ecol Model 220: 17641775. doi: 10.1016/j.ecolmodel.2009.03.024

[29] Tindyebwa, D., Kayita, J., Musoke, P., Eley, B., Nduati, R., Coovadia, H., Bobart, R., MboriNgacha, D., and Kieffer, M. P., (editors) (2011), Handbook on Peadriatric AIDS in Africa: *African Network for the care of children affected by HIV/AIDS*,(ANECCA), Kampala, Uganda.

[30] UNAIDS(2012), *joint United Nations Programme on HIV/AIDs*, Global report UNAIDS report on global AIDs epidemic: Geneva,2012.

[31] UNICEF Newyork (2015),United Nation Children's Fund : *Towards an AIDs-free generation-Children and AIDs*,Stock taking report.

[32] WHO (2012), World Health Organization :Programmatic Update *Use of Antiretroviral drugs for treating pregnant women and preventing HIV in-*

fection in infants.

[33] Xia, Y., Morrison-Beedy, D., Ma, J., Feng, C., Cross, W. and Tu, X.M. (2012), Modeling count outcomes from HIV risk reduction interventions: *a comparison of competing statistical models for count responses*. AIDS Research and Treatment, Article ID 593569, 11 pages. doi: 10.1155/2012/593569.

[34] Yip, P. (1988), Inference about the Mean of a Poisson distribution in the Presence of a Nuisance Parameter, *Australian Journal of Statistics*, 30, 299-306.

Appendices

Simulation R codes

library(pscl)

library(ZIM)

library(MASS)

library(psych)

library(VGAM)

Negative Binomial

size 500

count1< − rzinb(n = 500, k = 1, lambda = 1, omega = 0.2)

barplot(table(count1), col='lightblue')

x1< − rnbinom(100, mu = 2, size = 10)

x2< − rnbinom(100, mu = 5, size = 10)

data=data.frame(count1,x1,x2)

model1< − glm(count1 ∼ x1 + x2, family = poisson, data=data) Poisson regression

model2< − glm.nb(count1 ∼ x1 + x2, data=data) negative binomial regression

model3< − zeroinfl(count1 ∼ x1 + x2,dist = "poisson", data=data) zero-inflated Poisson regression

model3 < − AIC(model3)

model4< − zeroinfl(count1 ∼ x1 + x2, dist="negbin", data =data) zero-inflated negative binomial regression

model4 < − AIC(model4)

model5< − hurdle(count1 ∼ x1 + x2, data=data) Poisson hurdle

model5< − AIC(model5)

model6< − hurdle(count1 ∼ x1 + x2, dist="negbin", data=data) negative binomial hurdle

model6< −AIC(model6)

w=0.4,k=10

count2< −rzinb(n = 500, k = 10, lambda = 1, omega = 0.4) barplot(table(count2), col='lightblue')

x1< − rnbinom(100, mu = 2, size = 10)

x2< − rnbinom(100, mu = 5, size = 10)

data=data.frame(count2,x1,x2)

model1< − glm(count2 ∼ x1 + x2, family = poisson, data=data) Poisson regression model1

model2< − glm.nb(count2 ∼ x1 + x2, data=data) negative binomial regression model2

model3< − zeroinfl(count2 ∼ x1 + x2,dist = "poisson", data=data) zero-inflated Poisson regression

model3 < − AIC(model3)

model4< − zeroinfl(count2 ∼ x1 + x2, dist="negbin", data =data) zero-inflated negative binomial regression

model4 < − AIC(model4)

model5< − hurdle(count2 ∼ x1 + x2, data=data) Poisson hurdle

model5< − AIC(model5)

model6< − hurdle(count2 ∼ x1 + x2, dist="negbin", data=data) negative binomial hurdle

model6 < − AIC(model6)

w=0.6,k=50

count3< − rzinb(n = 500, k = 50, lambda = 1, omega = 0.6)

barplot(table(count3), col='lightblue')

x1< − rnbinom(100, mu = 2, size = 10)

x2< − rnbinom(100, mu = 5, size = 10)

data=data.frame(count3,x1,x2)

model1< − glm(count3 ∼ x1 + x2, family = poisson, data=data) Poisson regression model1

model2< − glm.nb(count3 ∼ x1 + x2, data=data) negative binomial regression model2

model3< − zeroinfl(count3 ∼ x1 + x2,dist = "poisson", data=data) zero-inflated Poisson regression

model3 < − AIC(model3)

model4< − zeroinfl(count3 ∼ x1 + x2, dist="negbin", data =data) zero-inflated negative binomial regression

model4 < − AIC(model4)

model5< − hurdle(count3 ∼ x1 + x2, data=data) Poisson hurdle

model5< − AIC(model5)

model6< − hurdle(count3∼x1 + x2, dist="negbin", data=data) negative binomial hurdle

model6 < − AIC(model6)

w=0.8,k=100

count4< − rzinb(n = 500, k = 100, lambda = 1, omega = 0.8)

barplot(table(count4), col='lightblue')

x1< − rnbinom(100, mu = 2, size = 10)

x2< − rnbinom(100, mu = 5, size = 10)

data=data.frame(count4,x1,x2)

model1< − glm(count4 ∼x1 + x2, family = poisson, data=data) Poisson regression model1

model2< − glm.nb(count4∼x1 + x2, data=data) negative binomial regression model2

model3< − zeroinfl(count4∼x1 + x2,dist = "poisson", data=data) zero-inflated Poisson regression

model3 < − AIC(model3)

model4< −zeroinfl(count4∼x1 + x2, dist="negbin", data =data) zero-inflated negative binomial regression

model4< −AIC(model4)

model5< −hurdle(count4 ∼x1 + x2, data=data) Poisson hurdle

model5< −AIC(model5)

model6< −hurdle(count4∼x1 + x2, dist="negbin", data=data) negative binomial hurdle

model6< −AIC(model6)

w=0.2,k=10

count5< −rzinb(n = 500, k = 10, lambda = 1, omega = 0.2)

barplot(table(count5), col='lightblue')

x1< −rnbinom(100, mu = 2, size = 10)

x2< − rnbinom(100, mu = 5, size = 10)

data=data.frame(count5,x1,x2)

model1< −glm(count5∼x1 + x2, family = poisson, data=data) Poisson regression model1

model2< −glm.nb(count5∼x1 + x2, data=data) negative binomial regression model2

model3< −zeroinfl(count5 ∼x1 + x2,dist = "poisson", data=data) zero-inflated Poisson regression

model3< −AIC(model3)

model4< −zeroinfl(count5∼x1 + x2, dist="negbin", data =data) zero-inflated

negative binomial regression

model4< − AIC(model4)

model5< − hurdle(count5∼x1 + x2, data=data) Poisson hurdle

model5< − AIC(model5)

model6< −hurdle(count5 ∼x1 + x2, dist="negbin", data=data) negative binomial hurdle

model6< −AIC(model6)

w=0.4,k=50

count6< − rzinb(n = 500, k = 50, lambda = 1, omega = 0.4)

barplot(table(count6), col='lightblue')

x1< − rnbinom(100, mu = 2, size = 10)

x2< −rnbinom(100, mu = 5, size = 10)

data=data.frame(count6,x1,x2)

model1< −glm(count6 ∼x1 + x2, family = poisson, data=data) Poisson regression model1

model2< −glm.nb(count6 ∼x1 + x2, data=data) negative binomial regression model2

model3< −zeroinfl(count6∼x1 + x2,dist = "poisson", data=data) zero-inflated Poisson regression

model3< −AIC(model3)

model4< −zeroinfl(count6∼x1 + x2, dist="negbin", data =data) zero-inflated negative binomial regression

model4 < −AIC(model4)

model5< −hurdle(count6∼x1 + x2, data=data) Poisson hurdle

model5< −AIC(model5)

model6< −hurdle(count6∼x1 + x2, dist="negbin", data=data) negative binomial hurdle

model6< −AIC(model6)

w=0.8,k=1

count7< −rzinb(n = 500, k = 1, lambda = 1, omega = 0.8)

barplot(table(count7), col='lightblue')

x1< −rnbinom(100, mu = 2, size = 10)

x2< −rnbinom(100, mu = 5, size = 10)

data=data.frame(count7,x1,x2)

model1< −glm(count7 ∼x1 + x2, family = poisson, data=data) Poisson regression model1

model2< −glm.nb(count7∼x1 + x2, data=data) negative binomial regression model2

model3< − zeroinfl(count7∼x1 + x2,dist = "poisson", data=data) zero-inflated Poisson regression

model3 < −AIC(model3)

model4< −zeroinfl(count7∼ x1 + x2, dist="negbin", data =data) zero-inflated negative binomial regression

model4 < − AIC(model4)

model5< −hurdle(count7∼x1 + x2, data=data) Poisson hurdle

model5< − AIC(model5)

model6< −hurdle(count7∼x1 + x2, dist="negbin", data=data) negative binomial hurdle

model6< −AIC(model6)

w=0.2,k=50

count8< − rzinb(n = 500, k = 50, lambda = 1, omega = 0.2)

barplot(table(count8), col='lightblue')

x1< − rnbinom(100, mu = 2, size = 10)

x2< − rnbinom(100, mu = 5, size = 10)

data=data.frame(count8,x1,x2)

model1< − glm(count8 ∼x1 + x2, family = poisson, data=data) Poisson regression model1

model2< −glm.nb(count8∼x1 + x2, data=data) negative binomial regression model2 model3< −zeroinfl(count8 ∼x1 + x2,dist = "poisson", data=data) zero-inflated Poisson regression

model3 < − AIC(model3)

model4< − zeroinfl(count8∼ x1 + x2, dist="negbin", data =data) zero-inflated negative binomial regression

model4 < − AIC(model4)

model5< − hurdle(count8∼x1 + x2, data=data) Poisson hurdle

model5< − AIC(model5)

model6< − hurdle(count8 ∼ x1 + x2, dist="negbin", data=data) negative binomial hurdle

model6 < − AIC(model6)

w=0.4,k=100

count9< − rzinb(n = 500, k = 100, lambda = 1, omega = 0.4)

barplot(table(count9), col='lightblue')

x1< − rnbinom(100, mu = 2, size = 10)

x2< − rnbinom(100, mu = 5, size = 10)

data=data.frame(count9,x1,x2)

model1< − glm(count9 ∼x1 + x2, family = poisson, data=data) Poisson regression model1

model2< − glm.nb(count9 ∼ x1 + x2, data=data) negative binomial regression model2

model3< − zeroinfl(count9∼ x1 + x2,dist = "poisson", data=data) zero-inflated Poisson regression

model3 < − AIC(model3)

model4< − zeroinfl(count9 ∼x1 + x2, dist="negbin", data =data) zero-inflated negative binomial regression

model4 < − AIC(model4)

model5< − hurdle(count9 ∼ x1 + x2, data=data) Poisson hurdle

model5< − AIC(model5)

model6< − hurdle(count9 ∼ x1 + x2, dist="negbin", data=data) negative binomial hurdle

model6 < − AIC(model6)

w=0.6,k=1

count10< − rzinb(n = 500, k = 1, lambda = 1, omega = 0.6)

barplot(table(count10), col='lightblue')

x1< −rnbinom(100, mu = 2, size = 10)

x2< −rnbinom(100, mu = 5, size = 10)

data=data.frame(count10,x1,x2)

model1< −glm(count10 ∼ x1 + x2, family = poisson, data=data) Poisson regression model1

model2< − glm.nb(count10 ∼ x1 + x2, data=data) negative binomial regression model2

model3< − zeroinfl(count10 ∼ x1 + x2,dist = "poisson", data=data) zero-inflated Poisson regression

model3 < − AIC(model3)

model4< − zeroinfl(count10 ∼ x1 + x2, dist="negbin", data =data) zero-inflated negative binomial regression

model4 < − AIC(model4)

model5< − hurdle(count10 ∼ x1 + x2, data=data) Poisson hurdle

model5< − AIC(model5)

model6< − hurdle(count10 ∼ x1 + x2, dist="negbin", data=data) negative binomial hurdle

model6 < − AIC(model6)

w=0.8,k=10

count11< − rzinb(n = 500, k = 10, lambda = 1, omega = 0.8)

barplot(table(count11), col='lightblue')

x1< − rnbinom(100, mu = 2, size = 10)

x2< − rnbinom(100, mu = 5, size = 10)

data=data.frame(count11,x1,x2)

model1< − glm(count11 ∼ x1 + x2, family = poisson, data=data) Poisson regression model1

model2< − glm.nb(count11 ∼ x1 + x2, data=data) negative binomial regression model2

model3< − zeroinfl(count11 ∼ x1 + x2,dist = "poisson", data=data) zero-inflated Poisson regression

model3 < −AIC(model3)

model4< −zeroinfl(count11 ∼ x1 + x2, dist="negbin", data =data) zero-inflated negative binomial regression

model4 < − AIC(model4)

model5< − hurdle(count11 ∼x1 + x2, data=data) Poisson hurdle

model5< − AIC(model5)

model6< − hurdle(count11 ∼ x1 + x2, dist="negbin", data=data) negative binomial hurdle

model6 < − AIC(model6)

w=0.4,k=1

count12< − rzinb(n = 500, k = 1, lambda = 1, omega = 0.4)

barplot(table(count12), col='lightblue')

x1< − rnbinom(100, mu = 2, size = 10)

x2< − rnbinom(100, mu = 5, size = 10)

data=data.frame(count12,x1,x2)

model1< − glm(count12 ∼x1 + x2, family = poisson, data=data) Poisson regression model1

model2< − glm.nb(count12 ∼ x1 + x2, data=data) negative binomial regression model2

model3< − zeroinfl(count12 ∼ x1 + x2,dist = "poisson", data=data) zero-inflated Poisson regression

model3 < − AIC(model3)

model4< − zeroinfl(count12∼ x1 + x2, dist="negbin", data =data) zero-inflated negative binomial regression

model4 < − AIC(model4)

model5< − hurdle(count12 ∼x1 + x2, data=data) Poisson hurdle

model5< − AIC(model5)

model6< − hurdle(count12 ∼ x1 + x2, dist="negbin", data=data) negative binomial hurdle

model6 < − AIC(model6)

w=0.6,k=10

count13< − rzinb(n = 500, k = 10, lambda = 1, omega = 0.6)

barplot(table(count13), col='lightblue')

x1< − rnbinom(100, mu = 2, size = 10)

x2< − rnbinom(100, mu = 5, size = 10)

data=data.frame(count13,x1,x2)

model1< − glm(count13 ∼x1 + x2, family = poisson, data=data) Poisson regression model1

model2< − glm.nb(count13 ∼ x1 + x2, data=data) negative binomial regression model2

model3< − zeroinfl(count13 ∼ x1 + x2,dist = "poisson", data=data) zero-inflated Poisson regression

model3 < − AIC(model3)

model4< − zeroinfl(count13 ∼ x1 + x2, dist="negbin", data =data) zero-inflated negative binomial regression

model4 < − AIC(model4)

model5< − hurdle(count13 ∼ x1 + x2, data=data) Poisson hurdle

model5< − AIC(model5)

model6< − hurdle(count13 ∼ x1 + x2, dist="negbin", data=data) negative binomial hurdle

model6 < − AIC(model6)

w=0.8,k=50

count14< − rzinb(n = 500, k = 50, lambda = 1, omega = 0.8)

barplot(table(count14), col='lightblue')

x1< − rnbinom(100, mu = 2, size = 10)

x2< − rnbinom(100, mu = 5, size = 10)

data=data.frame(count14,x1,x2)

model1< − glm(count14 ∼ x1 + x2, family = poisson, data=data) Poisson regression model1

model2< − glm.nb(count14 ∼ x1 + x2, data=data) negative binomial regression model2

model3< − zeroinfl(count14 ∼ x1 + x2,dist = "poisson", data=data) zero-inflated Poisson regression

model3 < −AIC(model3)

model4< − zeroinfl(count14 ∼ x1 + x2, dist="negbin", data =data) zero-inflated negative binomial regression

model4 < −AIC(model4)

model5< − hurdle(count14 ∼ x1 + x2, data=data) Poisson hurdle

model5< − AIC(model5)

model6< − hurdle(count14 ∼ x1 + x2, dist="negbin", data=data) negative binomial hurdle

model6 < − AIC(model6)

w=0.2,k=100

count15< − rzinb(n = 500, k = 100, lambda = 1, omega = 0.2)

barplot(table(count15), col='lightblue')

x1< − rnbinom(100, mu = 2, size = 10)

x2< − rnbinom(100, mu = 5, size = 10)

data=data.frame(count14,x1,x2)

model1< − glm(count15 ∼ x1 + x2, family = poisson, data=data) Poisson regression model1

model2< − glm.nb(count15 ∼ x1 + x2, data=data) negative binomial regression model2

model3< − zeroinfl(count15 ∼ x1 + x2,dist = "poisson", data=data) zero-inflated Poisson regression

model3 < − AIC(model3)

model4< − zeroinfl(count15 ∼ x1 + x2, dist="negbin", data =data) zero-inflated negative binomial regression

model4 < − AIC(model4)

model5< − hurdle(count15∼ x1 + x2, data=data) Poisson hurdle

model5< − AIC(model5)

model6< − hurdle(count14 ∼ x1 + x2, dist="negbin", data=data) negative binomial hurdle

model6 < − AIC(model6)

w=0.6,k=100

count16< − rzinb(n = 500, k = 100, lambda = 1, omega = 0.6)

barplot(table(count16), col='lightblue')

x1< − rnbinom(100, mu = 2, size = 10)

x2< − rnbinom(100, mu = 5, size = 10)

data=data.frame(count16,x1,x2)

model1< − glm(count16 ∼ x1 + x2, family = poisson, data=data) Poisson regression model1

model2< − glm.nb(count16 ∼ x1 + x2, data=data) negative binomial regression

model2

model3< − zeroinfl(count16 ∼ x1 + x2,dist = "poisson", data=data) zero-inflated Poisson regression

model3 < − AIC(model3)

model4< − zeroinfl(count16 ∼ x1 + x2, dist="negbin", data =data) zero-inflated negative binomial regression

model4 < − AIC(model4)

model5< − hurdle(count16 ∼ x1 + x2, data=data) Poisson hurdle

model5< −AIC(model5)

model6< − hurdle(count16 ∼ x1 + x2, dist="negbin", data=data) negative binomial hurdle

model6 < − AIC(model6)

w=0.0,k=1

count17< − rzinb(n = 500, k = 1, lambda = 1, omega = 0.0)

barplot(table(count17), col='lightblue')

x1< − rnbinom(100, mu = 2, size = 10)

x2< − rnbinom(100, mu = 5, size = 10)

data=data.frame(count17,x1,x2)

model1< − glm(count17 ∼ x1 + x2, family = poisson, data=data) Poisson regression model1

model2< −glm.nb(count17 ∼ x1 + x2, data=data) negative binomial regression model2

model3< − zeroinfl(count17 ∼ x1 + x2,dist = "poisson", data=data) zero-inflated Poisson regression

model3< − AIC(model3)

model4< − zeroinfl(count17 ∼ x1 + x2, dist="negbin", data =data) zero-inflated negative binomial regression

model4 < − AIC(model4)

model5< − hurdle(count17 ∼x1 + x2, data=data) Poisson hurdle

model5< − AIC(model5)

model6< − hurdle(count17 ∼ x1 + x2, dist="negbin", data=data) negative binomial hurdle

model6 < − AIC(model6)

w=0.0,k=10

count18< − rzinb(n = 500, k = 10, lambda = 1, omega = 0.0)

barplot(table(count18), col='lightblue')

x1< − rnbinom(100, mu = 2, size = 10)

x2< − rnbinom(100, mu = 5, size = 10)

data=data.frame(count18,x1,x2)

model1< − glm(count18 ∼ x1 + x2, family = poisson, data=data) Poisson regression model1

model2< − glm.nb(count18 ∼ x1 + x2, data=data) negative binomial regression model2

model3< − zeroinfl(count18 ∼ x1 + x2,dist = "poisson", data=data) zero-inflated Poisson regression

model3 < − AIC(model3)

model4< − zeroinfl(count18 ∼ x1 + x2, dist="negbin", data =data) zero-inflated negative binomial regression

model4 < − AIC(model4)

model5< − hurdle(count18 ∼ x1 + x2, data=data) Poisson hurdle

model5< − AIC(model5)

model6< − hurdle(count18 ∼ x1 + x2, dist="negbin", data=data) negative binomial hurdle

model6 < −AIC(model6)

w=0.0,k=50

count19< − rzinb(n = 500, k = 50, lambda = 1, omega = 0.0)

barplot(table(count19), col='lightblue')

x1< − rnbinom(100, mu = 2, size = 10)

x2< − rnbinom(100, mu = 5, size = 10)

data=data.frame(count6,x1,x2)

model1< − glm(count19 ∼ x1 + x2, family = poisson, data=data) Poisson regression model1

model2< − glm.nb(count19 ∼x1 + x2, data=data) negative binomial regression model2

model3< − zeroinfl(count19 ∼ x1 + x2,dist = "poisson", data=data) zero-inflated Poisson regression

model3 < − AIC(model3)

model4< − zeroinfl(count19 ∼ x1 + x2, dist="negbin", data =data) zero-inflated negative binomial regression

model4 < −AIC(model4)

model5< − hurdle(count19 ∼ x1 + x2, data=data) Poisson hurdle

model5< − AIC(model5)

model6< − hurdle(count19∼ x1 + x2, dist="negbin", data=data) negative binomial hurdle

model6 < − AIC(model6)

w=0.0,k=100

count20< − rzinb(n = 500, k = 100, lambda = 1, omega = 0.0) barplot(table(count20), col='lightblue') x1< − rnbinom(100, mu = 2, size = 10) x2< − rnbinom(100, mu = 5, size = 10) data=data.frame(count6,x1,x2) model1< − glm(count20 ∼ x1 + x2, family = poisson, data=data) Poisson regression model1

model2< − glm.nb(count20 ∼ x1 + x2, data=data) negative binomial regression model2

model3< − zeroinfl(count20 ∼ x1 + x2,dist = "poisson", data=data) zero-inflated Poisson regression

model3 < − AIC(model3)

model4< − zeroinfl(count20 ∼ x1 + x2, dist="negbin", data =data) zero-

inflated negative binomial regression

model4 < − AIC(model4)

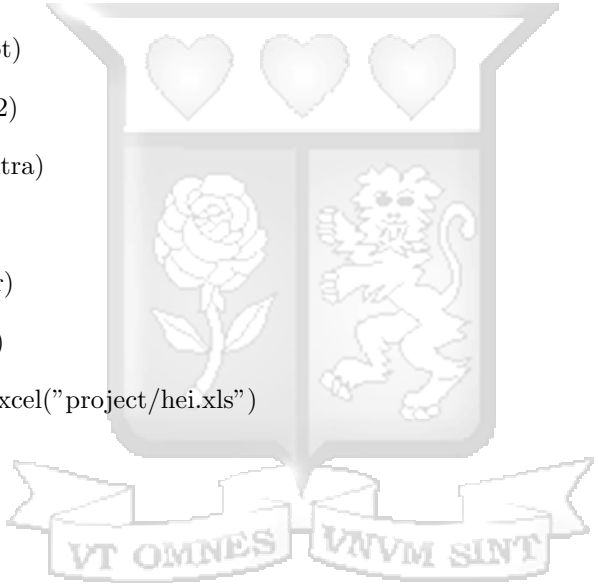
model5< − hurdle(count20 ∼ x1 + x2, data=data) Poisson hurdle

model5< − AIC(model5)

model6< −hurdle(count20∼ x1 + x2, dist="negbin", data=data) negative binomial hurdle

model6 < − AIC(model6)

bar graph plots



library(cowplot)

library(ggplot2)

library(gridExtra)

library(grid)

library(ggpubr)

library(readxl)

hei< − read excel("project/hei.xls")

attach(hei)

View(hei)

dev.off()

w=0.0

p1=ggplot(hei[2,], aes(AIC for Different Models,‘w=0.0,k=1‘)) + geom col(fill = "royalblue1") + geom text(aes(label = ‘w=0.0,k=1‘))

p2=ggplot(Data2, aes(AIC for Different Models,‘w=0.0,k=10‘)) + geom col(fill = "royalblue1") + geom text(aes(label = ‘w=0.0,k=10‘))

p3=ggplot(Data2, aes(AIC for Different Models,‘w=0.0,k=50‘)) + geom col(fill = "royalblue1") + geom text(aes(label = ‘w=0.0,k=50‘))

p4=ggplot(Data2, aes(AIC for Different Models,‘w=0.0,k=100‘)) + geom col(fill = "royalblue1") + geom text(aes(label = ‘w=0.0,k=100‘))

ggarrange(p1, p2, ncol = 1, nrow = 2)

ggarrange(p3, p4, ncol = 1, nrow = 2)

w=0.2

p11=ggplot(Data3w2, aes(AIC for Different Models,‘w=0.2,k=1‘)) + geom col(fill = ”royalblue1”) + geom text(aes(label = ‘w=0.2,k=1‘))

p12=ggplot(Data3w2, aes(AIC for Different Models,‘w=0.2,k=10‘)) + geom col(fill = ”royalblue1”) + geom text(aes(label = ‘w=0.2,k=10‘))

p13=ggplot(Data3w2, aes(AIC for Different Models,‘w=0.2,k=50‘)) + geom col(fill = ”royalblue1”) + geom text(aes(label = ‘w=0.2,k=50‘))

p14=ggplot(Data3w2, aes(AIC for Different Models,‘w=0.2,k=100‘)) + geom col(fill = ”royalblue1”) + geom text(aes(label = ‘w=0.2,k=100‘))

ggarrange(p11, p12, ncol = 1, nrow = 2)

ggarrange(p13, p14, ncol = 1, nrow = 2)

w=0.6

Data3 w6 < − read csv(”project/Data3 w.6.csv”)

attach(Data3 w6)

View(Data3 w6)

p21=ggplot(Data3 w6, aes(AIC for Different Models,‘w=0.6,k=1‘)) + geom col(fill = ”royalblue1”) + geom text(aes(label = ‘w=0.6,k=1‘))

p22=ggplot(Data3 w6, aes(AIC for Different Models,‘w=0.6,k=10‘)) + geom col(fill = ”royalblue1”) + geom text(aes(label = ‘w=0.6,k=10‘))

p23=ggplot(Data3 w6, aes(AIC for Different Models,‘w=0.6,k=50‘)) + geom col(fill = ”royalblue1”) + geom text(aes(label = ‘w=0.6,k=50‘))

p24=ggplot(Data3 w6, aes(AIC for Different Models,‘w=0.6,k=100‘)) + geom col(fill = ”royalblue1”) + geom text(aes(label = ‘w=0.6,k=100‘))

ggarrange(p21, p22, ncol = 1, nrow = 2)

ggarrange(p23, p24, ncol = 1, nrow = 2)

w=0.8

Data3 w8 < − read csv(”project/Data3 w8.csv”)

attach(Data3 w8)

View(Data3 w8)

p31=ggplot(Data3 w8, aes(AIC for Different Models,‘w=0.8,k=1‘)) + geom col(fill = ”royalblue1”) + geom text(aes(label = ‘w=0.8,k=1‘))

p32=ggplot(Data3 w8, aes(AIC for Different Models,‘w=0.8,k=10‘)) + geom col(fill = ”royalblue1”) + geom text(aes(label = ‘w=0.8,k=10‘))

p33=ggplot(Data3 w8, aes(AIC for Different Models,‘w=0.8,k=50‘)) + geom col(fill = ”royalblue1”) + geom text(aes(label = ‘w=0.8,k=50‘))

p34=ggplot(Data3 w8, aes(AIC for Different Models,‘w=0.8,k=100‘)) + geom col(fill = ”royalblue1”) + geom text(aes(label = ‘w=0.8,k=100‘))

ggarrange(p31, p32, ncol = 1, nrow = 2)

ggarrange(p33, p34, ncol = 1, nrow = 2)

R codes for the EID data

Model 1:Poisson regression

model1=glm(EIDPositive ∼,family=poisson, data = EIDData2)

pois.model1< −step(model1, direction = ”both”)

Start: AIC=474.69

EIDPositive ∼ EIDTestingPoint + PCRType + TestingPoint + HEIprophylaxis + MaternalProphylaxis

summary(model1)

Call: glm(formula = EIDPositive ∼., family = poisson, data =EIDData2)

Model 2:Negative binomial

$fm - nb < -$ MASS::glm.nb(EIDPositive ∼., data = EIDData2)

summary($fm - nb$)

Call:

MASS::glm.nb(formula = EIDPositive ∼., data = EIDData2, init.theta = 0.6083719267, link = log)

nb.model1< −step($fm - nb$, direction = "both")

EIDPositive ∼ EIDTestingPoint + PCRType + TestingPoint + HEIprophylaxis + MaternalProphylaxis

summary(model2)

glm.nb(formula = EIDPositive ∼ EIDTestingPoint + PCRType + TestingPoint + HEIprophylaxis + MaternalProphylaxis, data = EIDData2, init.theta = 0.6083719267, link = log)

Model 3:Zero-inflated Poisson

model3=zeroinfl(EIDPositive ∼ EIDTestingPoint+PCRType, dist = "poisson", data = EIDData2)

zero.model3< −step(model3, direction = "backward")

Start: AIC=491.18

EIDPositive ∼ EIDTestingPoint + PCRType

summary(model3)

Call: zeroinfl(formula = EIDPositive ∼ EIDTestingPoint + PCRType, data = EIDData2, dist = "poisson")

Model 4:Zero-inflated negative binomial

model4=zeroinfl(EIDPositive ∼ EIDTestingPoint+PCRType,dist = "negbin", data = EIDData2)

zero.model4< −step(model4, direction = "backward") Start: AIC=492.11 EIDPositive < − EIDTestingPoint + PCRType

summary(model4)

zeroinfl(formula = EIDPositive ∼ EIDTestingPoint + PCRType, data = EIDData2, dist = "negbin")

Model 5:Negative binomial hurdle

model5=hurdle(EIDPositive ∼EIDTestingPoint+PCRType, data=EIDData2, dist = "negbin")

hurdle.model5∼step(model5, direction = "backward")

Start: AIC=491.73 EIDPositive ∼ EIDTestingPoint + PCRType

summary(model5)

hurdle(formula = EIDPositive EIDTestingPoint + PCRType, data = EIDData2, dist = "negbin")

Model 6:Poisson hurdle

model6=hurdle(EIDPositive ~EIDTestingPoint+PCRType, data = EIDData2, dist = "poisson")

hurdle.model6~step(model6, direction = "backward")

Start: AIC=490.81

EIDPositive ~ EIDTestingPoint + PCRType

summary(model6)

hurdle(formula = EIDPositive ~ EIDTestingPoint + PCRType, data = EIDData2, dist = "poisson")

R script for the map

library(maptools)

library(raster)

library(plyr)

library(ggplot2)

library(rgdal)

library(ggmap)

library(scales)

Kenya< −getData("GADM", country="KE", level=0)

Kenya1< −getData("GADM", country="KE", level=1)

plot(Kenya1)

$Kenya1_UTM <-$ spTransform(Kenya1, CRS("+init=EPSG:32737"))

$Kenya1_UTMdataNAME_1$

View($Kenya1_UTMdataNAME_1$)

HEI TestedPositive< −c(8,8,16,12,3,3,2,39,1,10,30,3,20,18,7,12,34,17,9,7,1,12,7,2,1,9,34,16,3,82,19,11,7,5,2,3,1,2

theme opts< −list(theme(panel.grid.minor = element blank(),

panel.grid.major = element blank(),

panel.background = element blank(),

plot.background = element blank(),

axis.line = element blank(),

axis.text.x = element blank(),

axis.text.y = element blank(),

axis.ticks = element blank(),

axis.title.x = element blank(),

axis.title.y = element blank(),

plot.title = element blank()))

NAME 1< −Kenya1 UTM data NAME 1

HEI TestedPositive df< −data.frame(NAME 1, HEI TestedPositive)

HEI TestedPositive df

Kenya1 UTM data*id* < −rownames (Kenya1 UTM data)

Kenya1 UTM data < − join(Kenya1 UTM data, HEI TestedPositive df, by=”NAME 1”)

Kenya1 df < − fortify(Kenya1 UTM)

Kenya1 df < − join(Kenya1 df,Kenya1 UTM data, by=”id”)

theme opts< −list(theme(panel.grid.minor = element blank(),

panel.grid.major = element blank(),

panel.background = element blank(),

plot.background = element blank(),

axis.line = element blank(),

axis.text.x = element blank(),

axis.text.y = element blank(),

axis.ticks = element blank(),

axis.title.x = element blank(),

axis.title.y = element blank(),

plot.title = element blank()))

ggplot() +

geom polygon(data = Kenya1 df, aes(x = long, y = lat, group = group, fill =

HEI TestedPositive), color = "black", size = 0.25) + theme(aspect.ratio=1)+ scale fill distiller(name="HIV Exposed Infants", palette = "Reds", trans = "reverse", breaks = pretty breaks(n = 5))+ labs(title="Actual HEI Tested Positive")