2022

Improving performance of hurdle models using Rare-Event Weighted Logistic Regression: application to maternal mortality data.

Okello, Sharon Awuor
Strathmore Business School
Strathmore University

Improving Performance of Hurdle Models using Rare-Event Weighted Logistic Regression: Application to Maternal Mortality Data

Sharon Awuor Okello

Submitted in partial fulfilment of the requirements for the Degree of Master of Science in Statistical Sciences of Strathmore University

Institute of Mathematical Sciences

Strathmore University

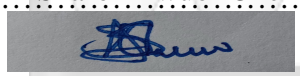
Nairobi, Kenya

June 6, 2022

Declaration

I declare that this work has not been previously submitted and approved for 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: Sharon Awuor Okello

Signature: 

Date: August 23, 2022

Approval

The thesis of Sharon Awuor Okello was reviewed and approved by the following:

Dr. Collins Ojwang' Odhiambo

Supervisor,

Institute of Mathematical Sciences, Strathmore University.

Dr. Evans Otieno Omondi

Supervisor,

Institute of Mathematical Sciences, Strathmore University.

Dr. Godfrey Madigu

Dean,

Institute of Mathematical Sciences, Strathmore University.

Dr. Bernard Shibwabo

Director,

Office of Graduate Studies, Strathmore University.

Abstract

Hurdle models, which are commonly used alongside zero-inflated models to analyze dispersed zero-inflated count data, employ a logit link function to predict whether an observation takes a positive count or a zero count based on a set of covariates. However, the logit model tends to be biased toward the majority zero class in cases involving rare events, and may underestimate the positive counts when their proportion is significantly smaller than that of the zero counts. This research aimed to improve the performance of hurdle models by incorporating rare-event weighted logistic regression model. Poisson and Negative Binomial (NB) Hurdle Rare Event Weighted Logistic Regression (REWLR) model estimates were developed and fit on various simulation conditions and maternal mortality data for performance evaluation using Akaike Information Criterion (AIC) and Area Under Curve (AUC). The Negative Binomial Hurdle REWLR emerged to be the best performing among all the evaluated models due to the ability to handle dispersion and adjust for class imbalance. The research findings will provide reliable estimates of the maternal mortality ratio in Nairobi without the risk of over-fitting zero counts.

Table of contents

List of figures

List of tables

List of abbreviations

OLS	Ordinary Least Squares	WHO	World Health Organization
MDSR	Maternal Death Surveillance and Response	MMR	Maternal Mortality Ratio
SDG	Sustainable Development Goals	KNH	Kenyatta National Hospital
ANC	Antenatal Care	REWLR	Rare-Event Weighted Logistic Regression
GLM	Generalized Linear Model	MNCH	Maternal, Newborn and Child Health

Acknowledgement

I want to express my sincere gratitude to my academic supervisors, Dr Collins Odhiambo and Dr Evans Omondi, for the valuable advice and support throughout the research period and guidance during the thesis write-up. I am also grateful to the Strathmore Institute of Mathematical Sciences and the faculty who have imparted knowledge and offered support throughout this academic program.

Dedication

This thesis is dedicated to God for giving me the gift of life, knowledge and perseverance. To my mum Prisca Atieno Muga for her love and support. To my beloved son Haris Hawi Nyangi for whom I am motivated to be the best version of myself.

Chapter 1

Introduction

1.1 Background to the study

Count data are generated by enumeration processes that produce discrete non-negative numbers. Due to the heteroskedastic and skewed nature of these data, the standard OLS models are not suitable for parameter estimations (Hutchinson and Holtman, 2005). Count models provide a better fit. Poisson and Negative Binomial regression are the most commonly used models for count data estimations. The Poisson model, considered the standard count model, assumes that the sample variance and sample mean are equal, a condition referred to as equidispersion. However, this is seldom the case. In practice, the sample variance is often either greater than (overdispersed) or less than the mean (under-dispersed). Poisson models provide a poor fit for such data.

Overdispersion in count data may arise due to several reasons, including the presence of excess zero counts in the data (Hilbe, 2011). The negative binomial model offers a better fit for overdispersed data but may also suffer overdispersion limitations. Overdispersion in a negative binomial model could occur when the observed model variance is greater than NB's expected variance, at times, due to more zeros than the model can accommodate (Hilbe, 2014). Zero-inflated mixture models are better suitable for modelling count data with more zeros than can be accounted for by the regular count models. These models - Zero-inflated Poisson (ZIP), Zero-inflated Negative Binomial (ZINB), Poisson Hurdle (PH) and Negative Binomial Hurdle (NBH) - propose separate data-generating processes for zero and positive counts.

Zero-inflated models introduced by Lambert (1992) and Greene (1994) propose a mixture distribution, where data is generated from Bernoulli and Poisson or Negative Binomial

processes. The most outstanding feature of the zero-inflated models, as explained by (Rose et al., 2006) is the assumption of the existence of an at-risk group that can never experience an event (structural zeros), and an at-risk group that may still not experience the event (sampling zeros). Zeros can thus be estimated using a mixture of a binary distribution - which estimates the probability of structural zeros, and a count model - which estimates all counts, including zeros.

The general structure of the zero-inflated model is given by:

$$P(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i)p(y_i;\lambda\,|y_i = 0) & y_i = 0 \\ \\ (1-\pi_i)\,p(y_i;\lambda) & y_i > 0; \end{cases} \tag{1.1}$$

Where π_i is the probability of being a structural zero, $p(y_i;\lambda)$ is the probability mass function of the count model, and $p(y_i;\lambda\,|y_i = 0)$ is the probability mass function of a count model for the count zero.

A typical example that would motivate the application of zero-inflated models is the case of modelling the weekly number of cigarettes smoked by a group of people, following, say, a policy implementation that aims to reduce cigarette smoking. Participants who respond that they have smoked 'zero' cigarettes may either be non-smokers who cannot have any value other than zero (structural zeros) or smokers who have reduced their weekly consumption to zero (sampling zeros).

Hurdle models, also called zero-altered models, provide an alternative means of modelling zero-inflated data. The distinctive feature between these models and the zero-inflated models is that hurdle models assume the existence of a single structural source of zeros. The general concept of the hurdle models is that a binomial probability model determines whether a count response variable takes a zero or a positive number. If the response variable returns a positive value, the 'hurdle' is crossed, and a zero-truncated model determines the magnitude of the positive counts (Mullahy, 1986).

In a maternal mortality setting, as in this study, the number of zeros reported is often excessive. From a perspective of the total number of live births, maternal deaths can be viewed as a

rare event. Based on WHO recommendation, data collected on maternal death through the Maternal Death Surveillance and Response (MDSR) systems include zero-reporting where weekly statistics are submitted even if no death has occurred (Smith et al., 2017). These are reported as 'zero' deaths. However, the zero deaths reported aren't distinguishable as from a structural or sampling source. One can't divide the population of women giving birth into a *risk* and a *not-at-risk* group and be certain the *not-at-risk* group will only report zero cases of death. Because of this, the current research focuses on Hurdle models for estimation of maternal deaths.

The logistic regression model is vital in the formulations of zero-inflated mixture models. In Hurdle models, either logistic or probit regression models are used to estimate the probability of obtaining a positive count (Hilbe, 2014). However, logistic regression models show limitations when predicting probabilities in imbalanced classes, e.g., prediction of zero versus positive counts for the binary component of Hurdle models. In the case of maternal deaths reported where the zero class may always be significantly larger than the non-zero class, logistic regression will tend to underestimate the probability of crossing the 'hurdle'.

Maternal death is the death of a woman while pregnant or within 42 days of pregnancy termination, irrespective of the duration and site of the pregnancy, from any cause related to or aggravated by the pregnancy or its management but not from accidental or incidental causes (Organization et al., 2019). According to 2017 WHO estimates, MMR declined by 38% globally between 2000 and 2017 from 342 deaths to 211 deaths per 100,000 live births. Kenya reported an impressive 52% reduction in MMR from 708 to 342 for the same period (Organization et al., 2019). This significant reduction in maternal death cases can be attributed to policies and initiatives implemented by the Kenyan government, such as Free Maternity Program, Beyond Zero, Linda Mama Campaign, among others. Despite all the strides towards reducing the number of maternal deaths, MMR is still high in Kenya. Reducing the number of maternal deaths remains a national priority (Mwangi et al., 2019).

The third SDG of the UN launched in 2015 aims for global MMR reduction to 70 or less, or at most 140, by 2030. With the current pace of progress, Kenya may fall short of this

target despite the programs and initiatives in place. More research is needed to guide new initiatives and support existing policies.

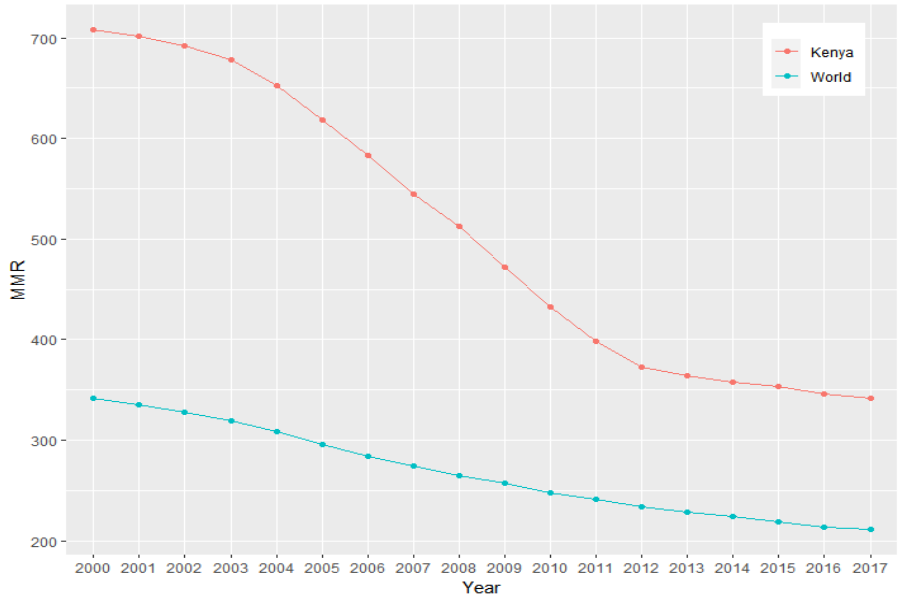


Figure 1.1: MMR Trends between 2000 - 2017: Source Organization et al. (2019)

Research into maternal mortality has involved establishing incidences, analyzing trends, identifying factors that may influence maternal deaths. The research has also involved developing and comparing models to determine the best-suited model for estimating and predicting maternal deaths.

1.2 Statement of the Problem

Rare events in count data results in class imbalance, where the proportion of zero counts is greater than that of positive counts. In such cases, the standard logistic regression may not be optimal, Maalouf and Siddiqi (2014). Hence the need for better performing models to estimate the probability of non-zero counts for Hurdle models. The current research incorporates Rare-Event Weighted Logistic Regression (REWLR) in our Hurdle models' binary component to improve the model performance.

1.3 Objective of the study

1.3.1 Objectives

- To investigate the performance of Hurdle-REWLR models using simulation analysis and with the application to maternal mortality data.
- To assess the performances of Hurdle-REWLR models for various proportions of zero-inflation.

1.3.2 Research Questions

i. Does incorporating REWLR in Hurdle models improve the performance of the models?

ii. Does the degree of class imbalance between zero and non-zero classes influence model performance?

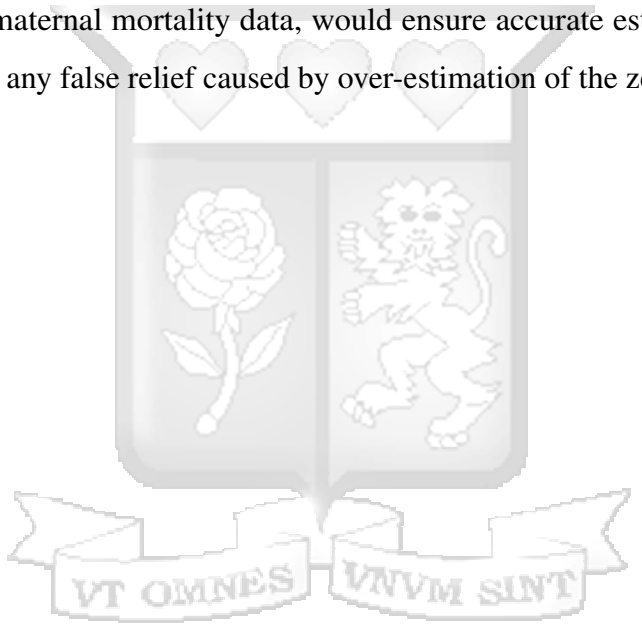
1.4 Justification

The goal of any statistical study is often to predict an outcome of interest or provide inference about the same. Practical inferences about a population require the sample statistics and estimates to be generalizable to a broader population, hence the need for models best suited for the population from which a sample is drawn. Statisticians and researchers have adjusted distributions and modified models to find the distribution that best explains their data and models that offer the best fit. Subsequently, count data models were extended to accommodate the excess zeros that arise in situations where the event of interest is rare. Formulations of these zero-modified count models involved nesting logistic regression to estimate the probabilities of a structural zero or a zero count for zero-inflated and hurdle models, respectively. GLM literature suggests the logit function is symmetric, so the response curve approaches zero and one at the same rate. This feature makes logistic regression inefficient due to the risk of underestimating the probability of a rare event.

Maternal deaths in Kenya are already under-reported due to the inefficient data collection systems and the high number of women who give birth outside healthcare facilities. Employing models that may underestimate the already under-reported maternal death cases would harm maternal health policies, as it would offer a false sign of relief. Hence the need for count estimation models adjusted to deal with rare events.

1.5 Significance of the Study

This study proposes a model extension to improve the performance of Hurdle models. The model, applied to maternal mortality data, would ensure accurate estimation of the death cases and eliminate any false relief caused by over-estimation of the zero-death cases.



Chapter 2

Literature review

2.1 Introduction

Variations of two statistical approaches have been used in modelling count data characterized by excess zeros in the outcome variable. This chapter provides an overview of these statistical approaches, the concepts behind them, and their applications in past research. We also review the logistic regression model and its limitations in estimating probabilities of rare events.

2.2 Models

2.2.1 Hurdle Models

Mullahy (1986) developed the Poisson hurdle model to handle zero-inflated count data in cases where sampling and structural zeros were not distinguishable. His proposed 2-part model analyzed zero counts separately from positive counts. He applied the model to study peoples' daily consumption of beverages based on certain socio-demographic factors. The study results revealed that the hurdle model allowed for more flexibility in model specification than the basic model. The models proposed in this research could also account for both under-dispersion and over-dispersion.

King (1989) separately developed hurdle models in an application to a political science study. His research aimed to develop an approach that models the onset of war separately from its escalation. The model was developed following Mullahy (1986) theory of data generation mechanism, where certain factors determine whether a country goes or does not go into war,

and once a country crosses the hurdle, factors such as alliances will determine the number of wars with which the country will be involved. This model proved to be an improvement of Mullahy's hurdle model.

Rose et al. (2006) applied the Poisson and Negative Binomial hurdle models in estimating the number of adverse events reported for each subject following a vaccination injection. They assumed a single source of zeros (sampling) because their study design made it such that all subjects were at risk of experiencing at least one adverse event. This assumption favoured hurdle over zero-inflated models. The goodness of fit statistics for ZINB and NBH were indistinguishable. A quasi-experimental study by Chaudhari et al. (2012) utilized hurdle models in the estimation of the total dental utilization using data obtained from dental claims. The model allowed them to decompose the hurdle likelihood function to allow for individual estimation of the probability of dental care, type of dental care and level of utilization. The likelihood decomposing feature gave the Negative Binomial Hurdle model the edge over the other models.

Hurdle models have also been widely applied in mortality estimation studies. Fenta and Fenta (2020) determined NBH model over ZIP, ZINB and PH for estimating risk factors of child mortality in Ethiopia. In a different study, NBH emerged to be the best statistical model for estimating predictors of under-five mortality in Ethiopia. The hurdle model was also selected as the best fitting model in the Mamun (2014) study to estimate under-five deaths. Both pieces of research involved comparing the Hurdle models to the zero-inflated models and, in some cases, the standard count models.

Besides the application of hurdle models in the various research fields, researchers have also developed modified versions of the models to provide better fit for their data. One of the hurdle model extensions was by Min and Agresti (2005), to accommodate correlated data. The authors modified the Hurdle model to include a random effect for their research to estimate the number of episodes of side effects recorded at each visit and compared two treatments. Fitting the random effects hurdle models proved less complex than fitting a zero-inflated random-effects model. In addition, the model provided more straightforward

interpretations. A two-part model meant the two parts could be fitted and estimated separately, hence reducing complexity.

2.2.2 Zero-inflated Models

Since their introduction, zero-inflated models have been continually modified and applied in many fields. Aryuyuen et al. (2014) developed the ZINB - Generalized Exponential distribution to provide a better fit for heavy-tailed over-dispersed zero-inflated data. He assessed the new model's performance compared to ZIP and ZINB, applied on simulated data and actual data for hospital stays by senior US residents. The resultant model proved to be a better fit than the ZIP and ZINB distributions. Kibika (2020) developed the ZINB - Shanker distribution by combining Zero-inflated Negative Binomial and Shanker distribution. The goal of developing the new model was to allow greater flexibility by increasing randomness in the ZINB probability distribution function. The model was used to model HIV cases among infants exposed to HIV through breastfeeding, etc. Overall fit tests revealed ZINB to offer the best fit. ZINB-Shanker distribution proved competitive for larger sample sizes.

Diop et al. (2021) proposed a modification to ZIP which involved the use of the quantile function of the Generalized Extreme Value (GEV) distribution as a link function for zero-inflated data with rare events. The approach was proposed to curb the drawbacks of logistic regression when dealing with imbalanced data, where the probability of a rare event is underestimated. Ali (2020) did a comparison study between ZIP, ZIP-GEV, ZIP-clog log and ZIP-probit. The analysis results revealed the Zero-inflated Poisson with a GEV link function to be the best performing model.

Zero-inflated models have also been widely utilized in maternal health studies. Arefaynie et al. (2022) used ZIP regression in a study to determine the number of antenatal care and associated factors in Ethiopia. Fitriani et al. (2019) and Loquiha et al. (2013) used ZINB regression to model maternal mortality in Malang and Mozambique.

2.2.3 Logistic Regression Model

The logistic regression model is the most commonly used statistical model for classifying binary data. It estimates the probability of a binary outcome, independently or dependent on a set of predictors. It has been broadly utilized in many fields including healthcare Yego et al. (2014), epidemiology Tolles and Meurer (2016), education Mason et al. (2018), economics Jabeur (2017), etc. Hurdle models also employ logistic regression to assign the probability that governs whether a count takes on a zero or a positive value. Models based on various link functions, including logit, probit, log-log, clog-log, have been proposed for the binary response estimation, but logistic regression remains the most popular Desjardins (2013). Its convenient interpretation and implementation makes it an ideal method for modelling binary response variable Ali (2020). Logistic regression, however, has drawbacks when applied to the classification of imbalanced binary events.

Research has highlighted this limiting feature of the logistic regression and proposed solutions to account for class imbalance in binary data. Rahim et al. (2019) applied Synthetic Minority Over-sampling Technique (SMOTE) sampling to Logistic regression, intending to improve its classification accuracy in bankruptcy detection. The study results showed that the SMOTE logistic regression outperformed the standard logistic regression with imbalanced data. In his study, Wang (2020) investigated the sampling-based interventions for imbalanced binary classes. The two approaches considered were undersampling the majority class or oversampling the minority class. The results reveal that undersampling the majority class did not always penalize the estimations, and oversampling the minority class did not consistently reduce estimation efficiency.

King and Zeng (2001) proposed a different approach for dealing with imbalanced binary classes, which involved applying weights and prior correction in the estimation of probabilities and regression coefficients. Their study results showed that the models implementing the recommended corrections outperformed the existing standard methods. However, the study's recommended approach turned out to be over-correcting bias in Maximum Likelihood Estimations.

Maalouf and Siddiqi (2014) developed the Rare Event Weighted Logistic Regression (REWLR) for classifying large imbalanced data with a rare event. The proposed algorithm applied weights and regularization terms to achieve better predictive accuracy, counter over-fitting and reduce bias and variance. Weighted Logistic Regression Approach for Rare Events was used in a study by Zare et al. (2013) to determine risk factors for female breast cancer where the choice of REWLR over Logistic Regression was influenced by the rarity of events of interest in their research. In a comparison study, REWLR proved to perform better than other algorithms, including the Truncated-Regularized Iteratively Re-weighted Least Squares algorithm and Truncated-regularized Prior Correction Maalouf et al. (2018). The authors recommended the application of appropriate corrections and adjustments to Logistic Regression when data is imbalanced.

2.3 Maternal Mortality in Kenya

Past studies on maternal mortality have aimed at establishing incidences, analyzing trends, identifying factors that may influence maternal deaths or specific causes of death with the goal of reducing cases of maternal deaths. One of those studies by Nyaboga (2009) described the trends, magnitude, contributing factors and causes of maternal mortality in Kenya's national referral hospital, KNH. His research identified age, parity, place of delivery, contraceptive use, ANC attendance, and socioeconomic status as the influential factors for maternal mortality in the national referral hospital. The specific maternal death causes were outlined in his paper as HIV, abortion complications, eclampsia, sepsis and postpartum haemorrhage.

Recommendations based on the research were in line with implementing BEmONC or CEmONC interventions in all healthcare facilities. Emergency Obstetric and Newborn Care (EmONC) describes a set of interventions that treat leading causes of perinatal and maternal mortality (Tecla et al., 2017). Basic EmONC (BEmONC) services include: administration of antibiotics to counter sepsis, anticonvulsants for hypertension disorders, uterotonic for postpartum haemorrhage, Manual placenta removal, Assisted vaginal delivery, retained products of conception extraction and neonatal resuscitation. Comprehensive EmONC

(CEmONC) services include all BEmONC components in addition to Caesarean section surgical capability and blood transfusions (Odhiambo and Kinoti, 2019).

2.4 Our Research

From the reviewed research, the question that remains to be explored is how a modified logistic regression would affect the performance of hurdle models. To the best of the author's knowledge, no study has attempted to improve the predictive performance of hurdle models' binary component by accommodating class imbalance. Based on the problem we have described so far, the objective of the current research is to improve the performance of hurdle models nested with rare-event weighted logistic regression when applied to maternal mortality data.

2.5 Conclusion

This chapter highlighted past research involving zero-inflated data, which elucidated the decision behind the choice model for the current research and the need for more robust classification models. The researchers concurred that decision about whether to apply hurdle models or zero-inflated models in modelling count data with excessive zeros should be guided by the beliefs about the data-generating mechanism of the zeros Min and Agresti (2005); Miller (2007); Desjardins (2013). Rose et al. (2006) proposed hurdle models be considered if there is a chance of zero deflation in the data.

The previous research work also supported the need for better-performing classification techniques in the hurdle model's binary component for data with rare events.

Chapter 3

Methodology

3.1 Introduction

This chapter details the development of the modified hurdle model which is achieved by incorporating REWLR for binary component estimations. The model is applied to simulated data to assess model performance with various proportions of zero counts and a real dataset to assess factors that influence maternal mortality in Nairobi.

3.2 Research Design

This study aims to improve the performance of hurdle models, and assess the effects of select obstetric and demographic factors on the number of maternal deaths in Nairobi. The maternal mortality data utilized for this study was pulled from JPHES, a portal of District Health Information Software (DHIS2), that streamlines health data reporting. The data contains the number of maternal deaths and other obstetric and demographic factors recorded in MNCH facilities in Nairobi between October 2021 and January 2022.

The study also introduces a modified Hurdle model that is based on Mullahy (1986)'s Hurdle models, and Maalouf and Siddiqi (2014)'s Rare-Events Weighted Logistic Regression model.

The general structure of a hurdle model as proposed by Mullahy (1986) is given by:

$$P(Y_i = y_i) = \begin{cases} (1-p_i) & y_i = 0 \\ \\ (p_i)\,\frac{p(y_i;\lambda_i)}{1-p(y_i;\lambda_i|y_i=0)} & y_i > 0; \end{cases} \tag{3.1}$$

This is the two-part model which uses a logistic regression model to estimate p_i and a zero-truncated count model for the estimation of the zero-truncated count model.

$$logit(p_i) = x_{1i}\beta_1 \text{ and } log(\lambda_i) = x_{2i}\beta_2 \tag{3.2}$$

We obtain the zero-truncated model by excluding the probability that $y_i = 0$ from the count distribution, which is achieved by dividing the probability mass function of the count model by 1 minus the probability of a zero count i.e.,

$$\frac{p(y_i;\lambda_i)}{1-p(y_i;\lambda_i|y_i=0)}. \tag{3.3}$$

The probability p_i of a positive count in hurdle models is typically modeled using a logistic regression model, presented as:

$$p_i = \frac{e^{\mathbf{X}\beta_i}}{1+e^{\mathbf{X}\beta_i}} = \frac{1}{1+e^{-\mathbf{X}\beta_i}} \tag{3.4}$$

where β_i's are the vector of coefficients, and X is a vector of predictors.

We use MLE to find the parameter estimates of the hurdle model; this is obtained by separately maximizing the log-likelihood functions of the binary and the zero-truncated distributions.

The log-likelihood function of the Hurdle model using a logistic regression model for the binary component is given by:

$$
\begin{aligned}
\ell(\beta_1,\beta_2) &= \ln\prod_{i=1}^{n}\left[(p_i)^{y_i}(1-p_i)^{(1-y_i)}\times\frac{p(y_i;\mu_i)}{1-p(y_i;\mu_i|y_i=0)}\right]\\
&= \sum_{i=1}^{n}\left[y_i\ln p_i+(1-y_i)\ln(1-p_i)+\ln\frac{p(y_i;\mu_i)}{1-p(y_i;\mu_i|y_i=0)}\right]\\
&= \sum_{i=1}^{n}\ln 1-p_i+\sum_{i=1}^{n}y_i\ln\frac{p_i}{1-p_i}+\sum_{i=1}^{n}\ln\frac{p(y_i;\mu_i)}{1-p(y_i;\mu_i|y_i=0)}\\
&= \sum_{i=1}^{n}\ln 1-p_i+\sum_{i=1}^{n}y_i(x\beta)+\sum_{i=1}^{n}\ln\frac{p(y_i;\mu_i)}{1-p(y_i;\mu_i|y_i=0)}\\
&= \sum_{i=1}^{n}-\left(\ln 1+e^{x\beta}\right)+\sum_{i=1}^{n}y_i(x\beta)+\sum_{i=1}^{n}\ln\frac{p(y_i;\mu_i)}{1-p(y_i;\mu_i|y_i=0)} \qquad (3.5)
\end{aligned}
$$

The maximum likelihood estimate for the binary component is the mean of the y variable from the n draws, i.e.

$$p_i=\frac{1}{n}\sum_{i=1}^{n}y_i \qquad (3.6)$$

There is no closed form solution to obtain the maximum likelihood estimates for the zero-truncated component. MLE are therefore obtained by using IRLS method of Newton-Raphson algorithm to solve the score equations.

3.3 Hurdle-REWLR Model

The proposed model overcomes logistic regression, and hence hurdle models', weakness in the case of imbalanced data by adopting regularization, weighting, and bias correction on logistic regression's log likelihood function.

The log-likelihood function of the REWLR model introduced by Maalouf and Siddiqi (2014) is given by:

$$
\begin{aligned}
\ell(\beta) &= \mathrm{In}\prod_{i=1}^{n}(p_i)^{w_1y_i}(1-p_i)^{w_0(1-y_i)}-\frac{\lambda}{2}\|\beta\|^2\\
&= -w_0\sum_{i=1}^{n}\left(\ln 1+e^{x\beta}\right)+(w_1-w_0)\sum_{i=1}^{n}y_ix\beta-\frac{\lambda}{2}\|\beta\|^2 \qquad (3.7)
\end{aligned}
$$

where:

i. w_s are the weights applied to counter imbalance in the data, which penalize the misclassification made by setting a higher class weight to the minority class (positive counts) while reducing weight for the majority class (zeros).

$$w_1 = \frac{\tau}{\bar{y}};\ \ w_0 = \frac{(1-\tau)}{(1-\bar{y})} \tag{3.8}$$

   (a) τ is the proportion of (non-zero) events in the population

   (b) $\bar{y}$ is the proportion of (non-zero) events in the sample;

ii. $\frac{\lambda}{2}\|\beta\|^2$ is a regularization term that introduces a penalty for large values of β hence avoids overfitting.

The log-likelihood of the binary logistic component and the zero-truncated Poisson or NB component are estimated separately and then combined for model fit assessments.

Neither the binary nor the zero-truncated components have closed form solutions for maximum likelihood estimation. MLEs are thus obtained by using IRLS method of Newton-Raphson algorithm to solve REWLR and zero truncated Poisson or zero truncated NB score equations.

3.3.1 Poisson Hurdle-REWLR Model

The Probability Mass Function of the Poisson Hurdle-REWLR Model is given by:

$$P(Y_i = y_i) = \begin{cases} (1-p_i) & y_i = 0 \\ (p_i)\,\frac{e^{-\lambda_i}\lambda_i^{y_i}}{(1-e^{-\lambda_i})y_i!} & y_i > 0 \end{cases} \tag{3.9}$$

Model estimates are obtained by maximizing the MLE function of the Poisson Hurdle-REWLR distribution:

$$\begin{aligned}\ell(\beta_1,\beta_2) &= \ln\prod_{i=1}^{n}\left((p_i)^{w_1 y_i}(1-p_i)^{w_0(1-y_i)} - \frac{\lambda}{2}\|\beta\|^2 + \frac{e^{-\lambda_i}\lambda_i^{y_i}}{\left(1-e^{-\lambda_i}\right)y_i!}\right)\\ &= -w_0\sum_{i=1}^{n}\left(\ln 1 + e^{x_{1i}\beta_1}\right) + (w_1-w_0)\sum_{i=1}^{n} y_i x_{1i}\beta_1 - \frac{\lambda}{2}\|\beta\|^2\\ &\quad + \sum_{i=1}^{n}\left(-\lambda + y_i x_{2i}\beta_2 - \ln(1-e^{-e^{x_{2i}\beta_2}}) - \ln(y_i!)\right) \qquad (3.10)\end{aligned}$$

$$\frac{\partial\ell(\beta_1,\beta_2)}{\partial\beta_1} = -w_0\sum_{i=1}^{n}\left(0+e^{x_{1i}\beta_1}\right)x_{1i} + (w_1-w_0)\sum_{i=1}^{n} y_i x_{1i} - 0 = 0 \qquad (3.11)$$

$$\frac{\partial\ell(\beta_1,\beta_2)}{\partial\beta_2} = \sum_{i=1}^{n}\left(-0 + y_i x_{2i} - e^{x_{2i}\beta_2}(x_{2i}) - 0\right) = 0 \qquad (3.12)$$

$$\hat{\beta}_1 = \frac{(w_0-w_1)}{n x_{1i}} n\ln(y_i) \qquad (3.13)$$

$$\hat{\beta}_2 = \frac{1}{x_{2i}}\ln y_i \qquad (3.14)$$

Since both components have no closed form solutions, MLEs are thus obtained by using IRLS method of Newton-Raphson algorithm.

3.3.2 Negative Binomial Hurdle-REWLR Model

The Probability Mass Function of the Negative Binomial Hurdle-REWLR Model is given by:

$$P(Y_i = y_i) = \begin{cases} (1-p_i) & y_i = 0\\ \frac{p_i}{1-\left(\frac{k}{\mu_i+k}\right)^k}\frac{\Gamma(y_i+k)}{y_i!\Gamma(k)}\left(\frac{\mu_i}{\mu_i+k}\right)^{y_i}\left(\frac{k}{\mu_i+k}\right)^k & y_i > 0; \end{cases} \qquad (3.15)$$

where the dispersion parameter k is given by $\frac{1}{\alpha}$.

Model estimates are obtained by maximizing the MLE function of the Negative Binomial Hurdle-REWLR distribution:

$$\begin{aligned}\ell(\beta_1,\beta_2) &= \ln\prod_{i=1}^{n}\left((p_i)^{w_1 y_i}(1-p_i)^{w_0(1-y_i)} - \frac{\lambda}{2}\|\beta\|^2\right) + \\ &\quad \frac{1}{1-\left(\frac{1}{1+\alpha\mu_i}\right)^{\alpha^{-1}}}\frac{\Gamma\left(y_i+\alpha^{-1}\right)}{y_i!\Gamma(\alpha^{-1})}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}\left(\frac{1}{1+\alpha\mu_i}\right)^{\alpha^{-1}} \\ &= -w_0\sum_{i=1}^{n}\left(\ln 1+e^{x\beta}\right) + (w_1-w_0)\sum_{i=1}^{n} y_i x\beta - \frac{\lambda}{2}\|\beta\|^2 \\ &\quad + \sum_{i=1}^{n}\Big(\ln\Gamma\left(y_i+\alpha^{-1}\right) - \ln\Gamma(\alpha^{-1}) - \ln y_i! - \left(y_i+a^{-1}\right)\ln\left(1+\alpha\mu_i\right) \\ &\quad + y_i\ln\alpha\mu_i - \ln\left[1-(1+\alpha\mu_i)^{-a^{-1}}\right]\Big) \qquad (3.16)\end{aligned}$$

$$\frac{\partial\ell(\beta_1,\beta_2)}{\partial\beta_1} = -w_0\sum_{i=1}^{n}\left(0+e^{x_{1i}\beta_1}\right)x_{1i} + (w_1-w_0)\sum_{i=1}^{n} y_i x_{1i} - 0 = 0 \qquad (3.17)$$

$$\frac{\partial\ell(\beta_1,\beta_2)}{\partial\beta_2|\mu} = \sum_{i=1}^{n}\left[\frac{y_i}{\mu(1+\alpha\mu)} - \frac{(1+\alpha\mu)^{\alpha^{-1}-1}}{(1+\alpha\mu)^{\alpha^{-1}}-1}\right] = 0 \qquad (3.18)$$

$$\frac{\partial\ell(\beta_1,\beta_2)}{\partial\beta_2|\alpha} = \sum_{i=1}^{n}\left[\sum_{v=0}^{y_i-1}\left(\frac{v}{v+\alpha v^{-1}}\right) + \frac{y_i}{\mu(1+\alpha\mu)} + \frac{\alpha^{-2}(1+\alpha\mu)^{-1}\log(1+\alpha\mu)}{(1+\alpha\mu)^{\alpha^{-1}}-1}\right] = 0 \qquad (3.19)$$

Since both components have no closed form solutions, MLEs are thus obtained by using IRLS method of Newton-Raphson algorithm.

3.4 Simulations

Data is simulated under PH and NBH distributions. Simulations are performed using a combination of sample size and proportion of zeros observed.

The following experimental conditions are applied for the simulation study:

- Zero inflation - 50%, 60%, 75%, and 90%.

- Sample size - 200, 1000
- The dispersion parameter value is set at 3

The simulation analysis involves generating data from four different distributions: Negative Binomial Hurdle, Poisson Hurdle-REWLR, and Negative Binomial Hurdle-REWLR. For each model, we generate 200 and 1000 random samples with varying zero proportions from the true model, and then all the models are fit to the simulated datasets. In addition, a predictor variable is simulated from the Poisson distribution, with a constant mean of 3 across all simulation conditions. The simulated covariate mimics the type of covariates for the real maternal mortality data, e.g., number of pregnant women attending at least 4 ANC visits, number of assisted Vaginal deliveries, etc. The dispersion parameter k is used with a pre-stipulated value of 3.

We use AIC to compare the true and misspecified models in terms of the percentage of the differences in the AICs for the misspecified models and true model $(\%\Delta AIC)$. Where $\Delta AIC = AIC(\text{Misspecified model}) - AIC(\text{True model})$. We also compute AUC statistics to achieve and compare an aggregate measure of performance across all possible classification thresholds, for the two binary components.

Data is generated in R using the rpois(), rbinom(), rhpois(), rhnbinom() functions from the *stats*, *actuar* and *countreg* packages. Analyses for the simulated and real data is performed in R using hurdle() from the *pscl* package, glm from *stats* package and vglm() from *VGAM* package.

3.5 Maternal Mortality data

The data contains information on obstetric outcomes, including maternal deaths for public and private facilities in Nairobi that offer MNCH services. It covers the duration of October 2021 to January 2022, containing records for 222 MNCH facilities.

Data is available for at least one facility in all the 17 sub-counties in Nairobi: Westlands, Dagoretti North, Dagoretti South, Langata, Kibra, Roysambu, Kasarani, Ruaraka, Embakasi

South, Embakasi North, Embakasi Central, Embakasi East, Embakasi West, Makadara, Kamukunji, Starehe and Mathare. Nairobi is a cosmopolitan county and Kenya's capital city hence the data offers a good representation of the Kenyan population.

Table 3.1: Variable Definition

Variable	Description
MaternalDeaths	Number of Maternal deaths in MNCH Nairobi facilities between October 2021 - January 2022
AssistedDeliveries	Number of women who had assisted vaginal deliveries
BreechDelivery	Number of women who had breech delivery
CS	Number of women who gave birth by caesarian sections
LiveBirths	Number of live births
EarlyTeenPreg	Number of adolescents (10-14 years) pregnant at 1st ANC visit
LateTeenPreg	Number of adolescents (15-19 years) pregnant at 1st ANC visit
NormalDeliveries	Number of women who had normal deliveries
ANC4Visits	Number of women who have attended at least 4 ANC visits
Uterotonics3stg	Number of women giving birth who received uterotonics in the third stage of (or immediately after birth)
Carbatosin	Number of Mothers given uterotonics within 1 minute (Carbatosin)
Oxytocin	Number of Mothers given uterotonics within 1 minute (Oxytocin)
AntHaemorrage	Number of women who had Ante partum Haemorrage
PostHaemorrage	Number of women who had Post Partum Haemorrage
ObstructedLabour	Number of women who had Obstructed Labour
Eclampsia	Number of women who had Eclampsia
RupturedUterus	Number of women who had Ruptured Uterus
Sepsis	Number of women who had sepsis
FGMComplicatons	Number of Mothers with delivery complications associated with FGM
Stillbirth	Number of women who had Macerated stillbirth

The response variable for this research is the number of maternal deaths reported in MNCH Nairobi facilities between October 2021 - January 2022. The predictors consist of obstetric factors, maternal complications, and demographic factors that previous literature suggested influence maternal deaths.

3.6 Model selection

The study uses Akaike information criterion (AIC) to compare the model fit between the modified hurdle models and the standard Hurdle models.
AIC is computed as:

$$AIC = -2log(L) + 2K;$$

where L is the likelihood, and K is the number of parameters in the model.
AIC evaluates how well a model fits the data from which it was generated. The best-fitting model yields the lowest AIC values.

Area under the curve (AUC), of the receiver operating characteristic (ROC) is also computed for both binary components to show the performance of the classification models. AUC measures the ability of the classification algorithm to distinguish between classes. Higher values of AUC imply better model performance.

Chapter 4

Results and Interpretation

4.1 Simulation

In this section, we present the theoretical results of the study models through simulation analyses performed on the Poisson Hurdle-REWLR, NB Hurdle-REWLR and Poisson Hurdle, and NB Hurdle models.

For the simulation analysis, we first evaluated the performance of HNB, HNB-REWLR and HP-REWLR models when the data are simulated from a Hurdle Poisson model. The Hurdle-REWLR models reported the lowest AIC values among all the other models. The percentage differences in the AICs between the misspecified models and the Poisson Hurdle models increased as the proportions of zero in the data increased. This was the trend for both the small (n=200) and large (n=1000) sample sizes.

Analysis on NB Hurdle data resulted in NB Hurdle REWLR outperforming NB Hurdle at 60%, 75% and 90% zero inflation for both small and large sample sizes. NB Hurdle outperformed the other models only when the data had a 50% zero inflation.

Data generated by the study model distributions, Poisson Hurdle-REWLR and NB Hurdle-REWLR, performed best on the true models, based on AIC statistics. In the Poisson Hurdle REWLR generated data, the lowest AIC values were recorded by the same model, for all the simulation conditions. In a small sample size, the model performed best in data with 75% zero inflation while in large sample size, Poisson Hurdle-REWLR performance is best in 60% inflation data. The least percentage change in AIC was achieved by the NB Hurdle REWLR model. Models fit on the NB Hurdle-REWLR simulated data achieved the lowest AIC. The least percentage change in AIC was achieved by the NB Hurdle model.

Table 4.1: AIC (Percentage Change in AIC) for Misspecified and Actual Models

Reference	n	Zeros	PH	PHRE	NBH	NBHRE
PH	200	0.50	418	402 (-3.98)	420 (0.43)	403 (-3.63)
PH	200	0.60	512	479 (-6.88)	514 (0.3)	480 (-6.64)
PH	200	0.75	519	478 (-8.68)	521 (0.37)	480 (-8.16)
PH	200	0.90	582	527 (-10.43)	584 (0.42)	529 (-9.92)
PH	1000	0.50	2252	2136 (-5.44)	2254 (0.09)	2140 (-5.23)
PH	1000	0.60	2473	2302 (-7.43)	2474 (0.04)	2303 (-7.36)
PH	1000	0.75	2730	2513 (-8.65)	2732 (0.07)	2515 (-8.56)
PH	1000	0.90	2855	2592 (-10.17)	2857 (0.07)	2593 (-10.1)
PHRE	200	0.50	578 (9.53)	523	580 (9.84)	525 (0.36)
PHRE	200	0.60	599 (9.13)	544	600 (9.29)	545 (0.26)
PHRE	200	0.75	562 (10.03)	506	564 (10.35)	509 (0.67)
PHRE	200	0.90	587 (9.61)	531	589 (9.92)	534 (0.5)
PHRE	1000	0.50	2807 (9.61)	2537	2809 (9.68)	2542 (0.18)
PHRE	1000	0.60	2952 (9.4)	2675	2954 (9.44)	2676 (0.06)
PHRE	1000	0.75	2976 (9.32)	2699	2978 (9.38)	2701 (0.07)
PHRE	1000	0.90	2921 (9.5)	2644	2923 (9.54)	2645 (0.04)
NBH	200	0.50	374 (7.56)	380 (8.98)	346	352 (1.61)
NBH	200	0.60	499 (5.84)	488 (3.76)	470	460 (-2.27)
NBH	200	0.75	562 (7.87)	540 (4.11)	518	496 (-4.47)
NBH	200	0.90	649 (6.26)	617 (1.39)	608	576 (-5.57)
NBH	1000	0.50	2054 (5.9)	2069 (6.56)	1933	1948 (0.76)
NBH	1000	0.60	2161 (5.13)	2144 (4.37)	2050	2033 (-0.85)
NBH	1000	0.75	2864 (8.81)	2767 (5.61)	2612	2514 (-3.88)
NBH	1000	0.90	3128 (9.65)	2984 (5.29)	2826	2682 (-5.38)
NBHRE	200	0.50	614 (12.66)	588 (8.78)	562 (4.58)	536
NBHRE	200	0.60	685 (11.4)	649 (6.45)	643 (5.58)	607
NBHRE	200	0.75	679 (14.28)	640 (9.12)	621 (6.26)	582
NBHRE	200	0.90	680 (12.82)	645 (8.02)	629 (5.69)	593
NBHRE	1000	0.50	3337 (12.58)	3180 (8.27)	3074 (5.11)	2917
NBHRE	1000	0.60	3044 (11.13)	2896 (6.59)	2853 (5.18)	2705
NBHRE	1000	0.75	3426 (14.11)	3245 (9.3)	3125 (5.81)	2943
NBHRE	1000	0.90	3279 (14.71)	3111 (10.08)	2966 (5.69)	2797

Overall, NB Hurdle REWLR outperformed the Poisson Hurdle, Poisson Hurdle REWLR and NB Hurdle models. Plots of the resulting AIC values are presented in figure 4.1 and figure 4.2 for Poisson Hurdle simulated data, figure 4.3 figure 4.4 for Poisson Hurdle REWLR simulated data, figure 4.5 and figure 4.6 for NB Hurdle simulated data, figure 4.7 and figure 4.7 for NB Hurdle REWLR simulated data.

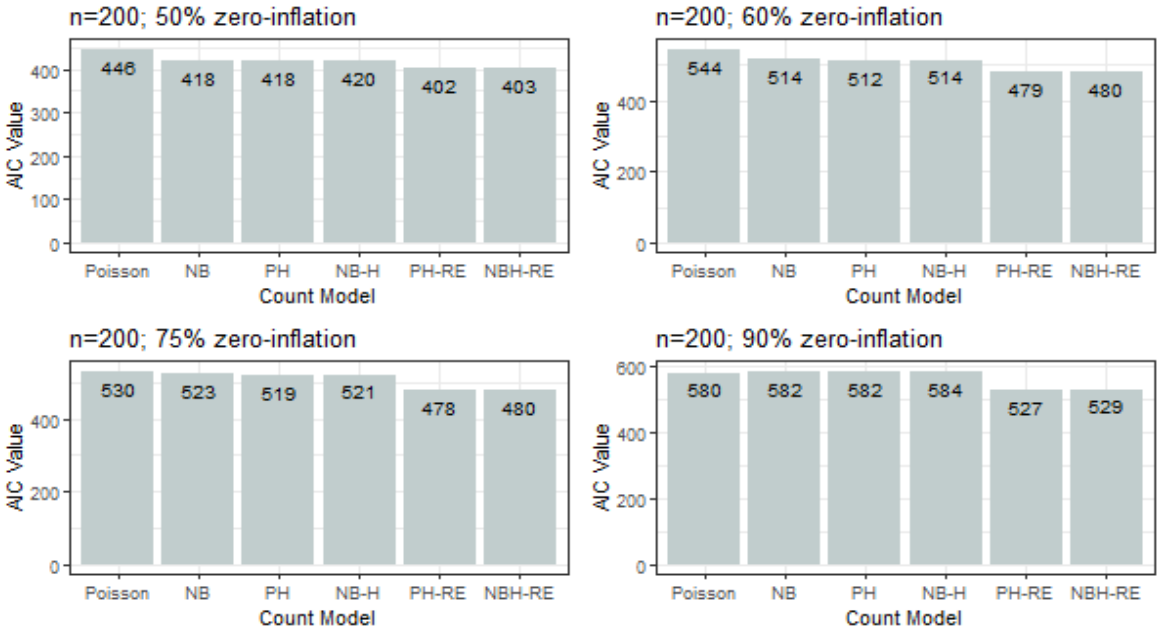


Figure 4.1: AICs from models fit on Poisson Hurdle simulated data, n = 200

Figure 4.2: AICs from models fit on Poisson Hurdle simulated data, n = 1000

Figure 4.3: AICs from models fit on Poisson Hurdle-RE simulated data, n = 200

Figure 4.4: AICs from models fit on Poisson Hurdle-RE simulated data, n = 1000

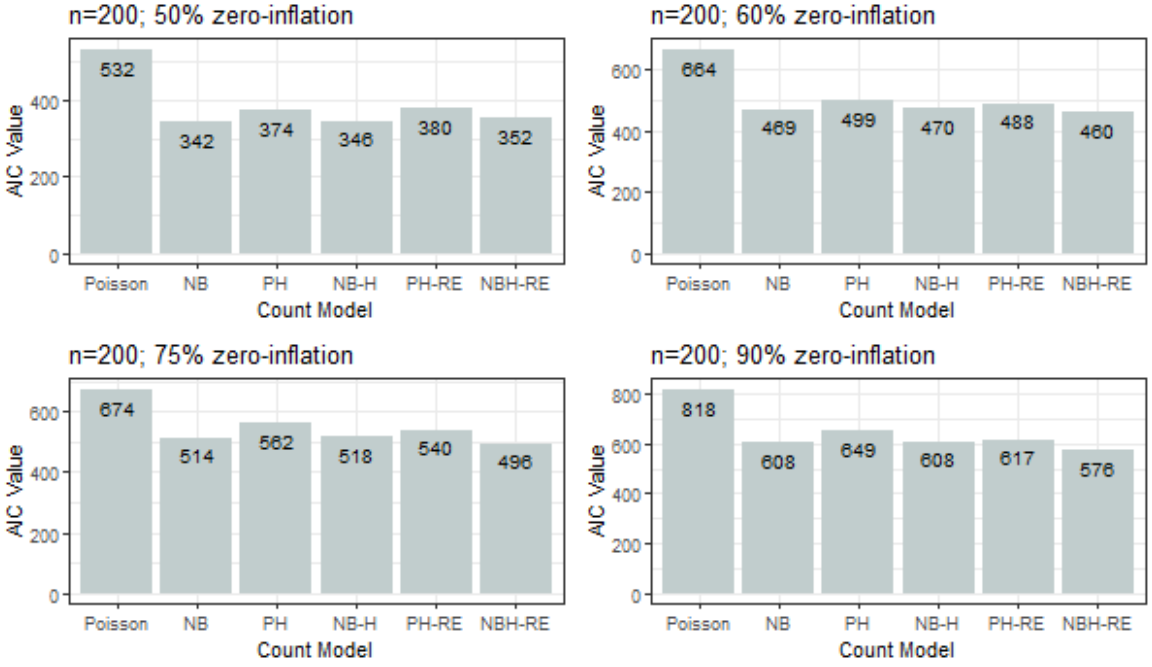


Figure 4.5: AICs from models fit on NB Hurdle simulated data, n = 200

Figure 4.6: AICs from models fit on NB Hurdle simulated data, n = 1000

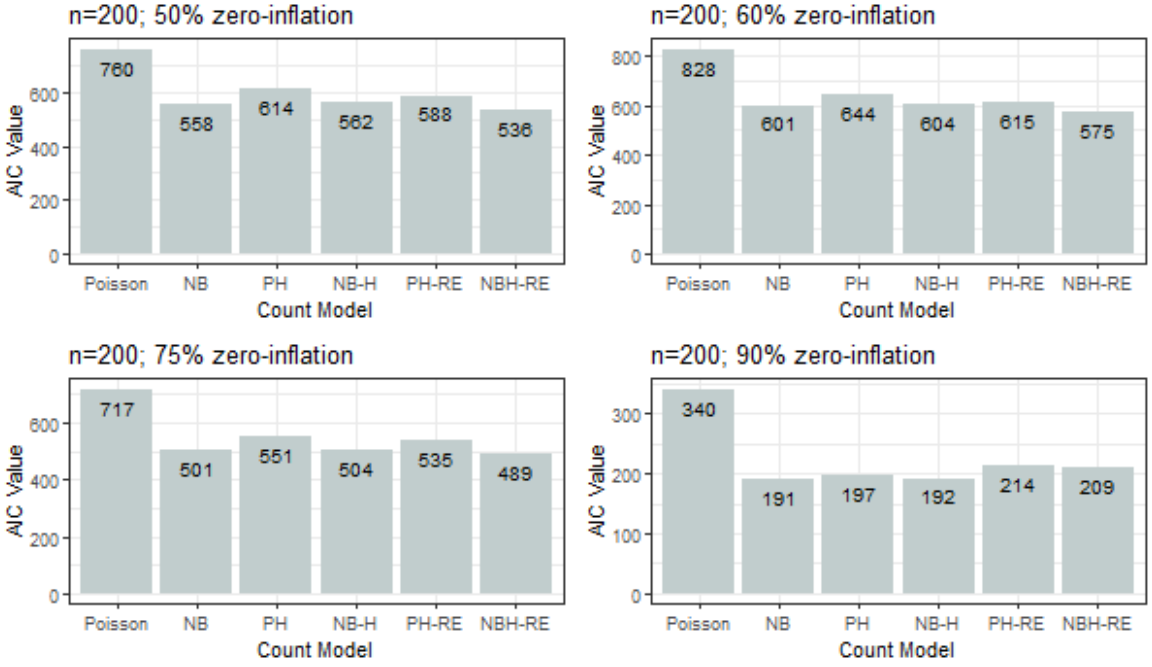


Figure 4.7: AICs from models fit on NB Hurdle-RE simulated data, n = 200

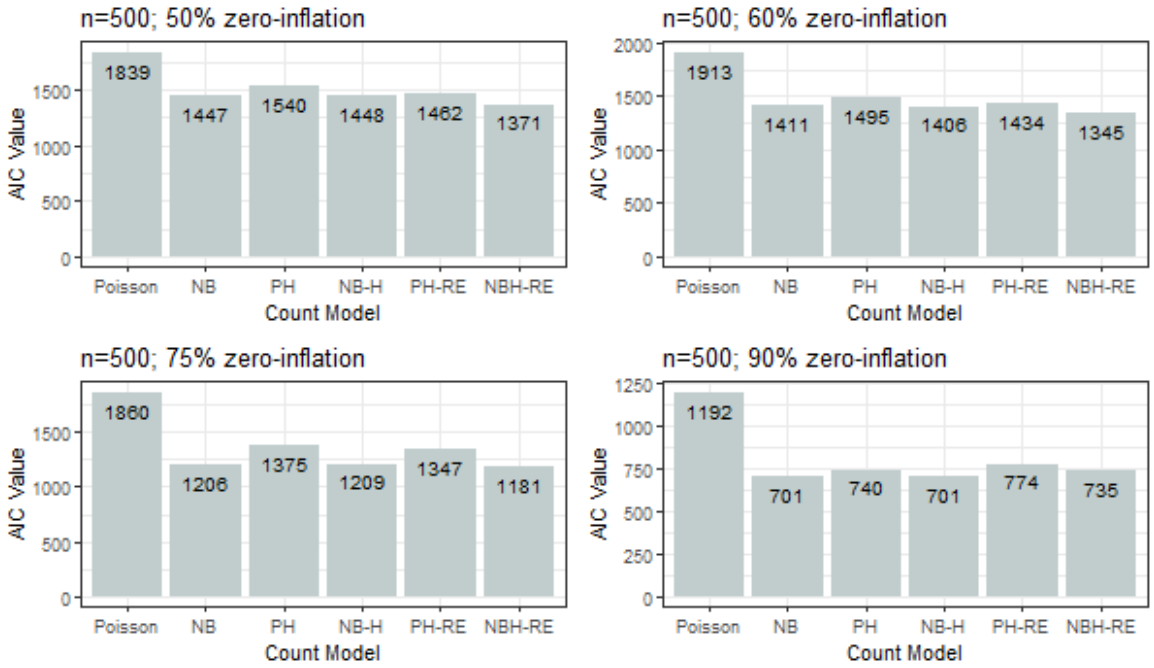


Figure 4.8: AICs from models fit on NB Hurdle-RE simulated data, n = 1000

4.2 Application to Maternal Deaths Data

4.2.1 Descriptive Statistics

The study sample data reported 293 maternal deaths of the 53792 recorded live births. The sample variance of 3.758 exceeds the sample mean of 1.32, indicating overdispersed data. The data also exhibits zero inflation as 61.71% of the dependent variable counts are zero.

Table 4.2 below exhibits the average counts for some of the obstetric factors used as covariates in this study. The facilities which reported maternal deaths had higher average counts of all the conditions. For instance, the number of Mothers given uterotonics was higher in the group that experienced maternal deaths. Stillbirth occurrence was also primarily associated with maternal death. Correlation analysis between the maternal deaths and the predictors revealed that Maternal deaths was highly correlated with BreechDelivery (r = 0.7342), Uterotonics3stg (r = 0.9615), Oxytocin (r = 0.9615), Carbatosin (r = 0.9615), AntHaemorrage (r = 0.9743), Eclampsia (r = 0.9742), ObstructedLabour (r = 0.9741), PostHaemorrage (r = 0.9741),

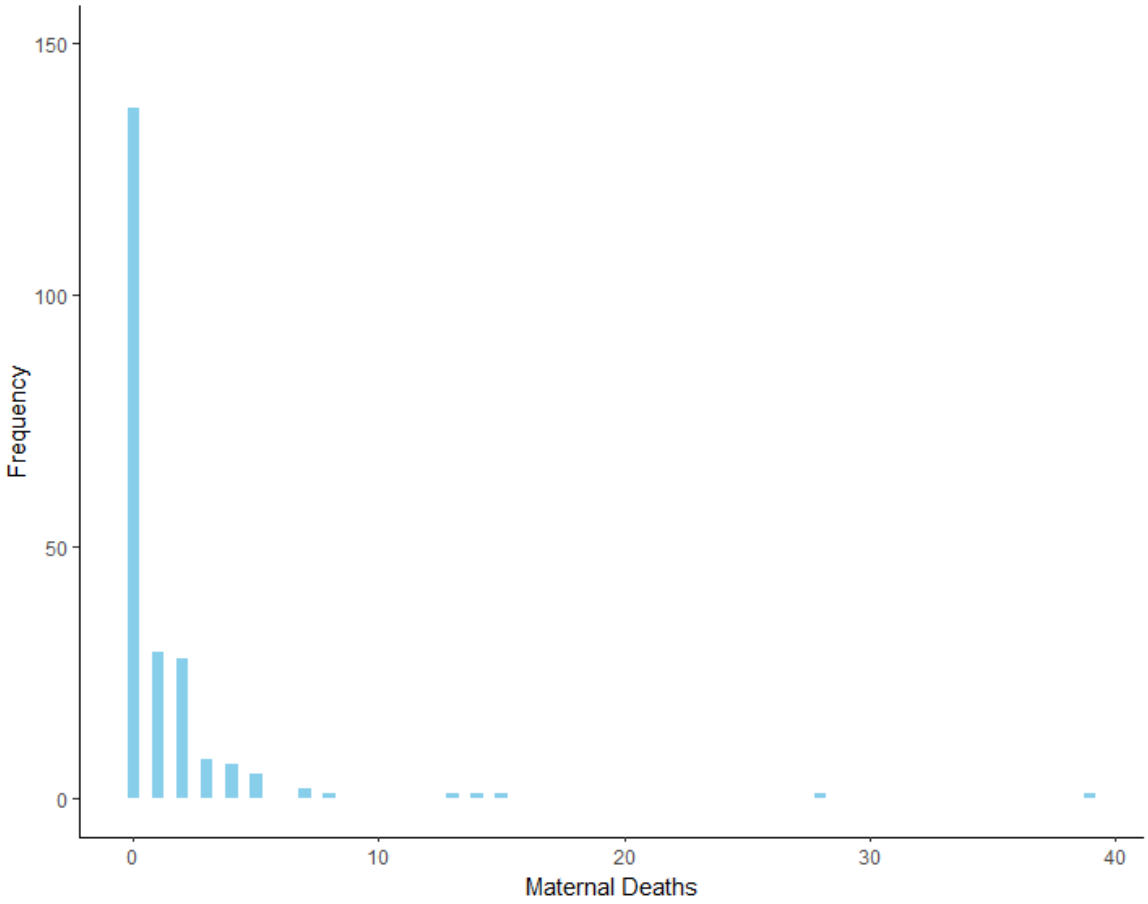


Figure 4.9: Maternal Death Counts

FGMComplicatons (r = 0.9742), RupturedUterus (r = 0.9744), Sepsis (r = 0.9740), and Stillbirth (r = 0.9764).

4.2.2 Maternal Death Models

Prior to formulating the count models, we compute the weights for the binary Hurdle-REWLR models. The weights penalize the misclassification made by setting a higher class weight to the positive counts while reducing weight for the zero counts.

The weights are calculated as outlined by Maalouf and Siddiqi (2014):

Table 4.2: Average Count of Obstetric Conditions reported in facilities with and without reported maternal deaths

Factor	No Maternal Deaths	Maternal Deaths
BreechDelivery	0.3	2.7
CS	15.0	174.7
LiveBirths	44.7	560.8
EarlyTeenPreg	2.3	22.9
LateTeenPreg	12.7	60.0
NormalDeliveries	36.5	404.2
ANC4Visits	42.7	272.8
Uterotonics3stg	96.2	1139.7
Carbatosin	9.5	112.9
Oxytocin	75.0	889.0
AntHaemorrage	7.5	94.8
Eclampsia	2.2	28.5
ObstructedLabour	0.7	10.4
PostHaemorrage	3.4	42.4
FGMComplicatons	1.5	18.8
RupturedUterus	1.7	21.8
Sepsis	0.4	6.1
Stillbirth	0.0	0.6

$$w_1 = \frac{\tau}{\bar{y}}; \quad w_0 = \frac{(1-\tau)}{(1-\bar{y})} \tag{4.1}$$

We have 293 deaths for the 53792 live births from our sample data. The latest report by the Kenya Ministry of Health on Health and Health-related SDGs revealed the latest deaths per live birth ratio reported in Nairobi as 97.

$$\bar{Y} = \frac{293}{53792} = 0.0054 \qquad \tau = \frac{97}{100000} = 0.00097$$

$$w_1 = \frac{0.00097}{0.0054} = 0.1796 \qquad w_0 = \frac{(1-0.00097)}{(1-0054)} = 1.0045$$

We use correlation analysis to select the predictors to use for the analysis. Predictors with high correlations are more linearly dependent and thus have the same effect on the dependent variables.

The factors that influence observing a maternal death in the facility are attending at least 4 ANC visits, antepartum haemorrhage, and receiving uterotonics during or immediately after birth. Specific effects of these factors on the various models are presented in Table 4.3. Upon observing a maternal death, the determinants of the actual number of maternal deaths that a facility could report are the occurrence of macerated stillbirth, attending of at least 4 ANC visits, adolescent pregnancies, antepartum hemorrhage, breech deliveries, postpartum hemorrhage, receiving Carbatosin and giving birth by cesarean section. The coefficients of the count model component is presented in Table 4.4.

Table 4.3: Binary Component Coefficients

	PH.BINARY	NBH.BINARY	PHRE.BINARY	NBHRE.BINARY
(Intercept)	-3.9561920	-3.9561920	8.2592952	8.2591187
Stillbirth	21.4971965	21.4971965	-0.0242153	-0.0529712
ANC4Visits	-0.0021931	-0.0021931	0.0260967	0.0261088
LateTeenPreg	-0.0072837	-0.0072837	0.0091975	0.0091929
AntHaemorrage	1.2112676	1.2112676	-4.6352784	-4.6355084
PostHaemorrage	-0.1450389	-0.1450389	0.2713996	0.2708596
BreechDelivery	0.1438486	0.1438486	-0.0180133	-0.0183037
Carbatosin	-0.0103415	-0.0103415	-0.1092431	-0.1002919
Uterotonics3stg	-0.0785952	-0.0785952	0.3119054	0.3110578

Table 4.4: Count Component Coefficients

	PH.COUNT	NBH.COUNT	PHRE.COUNT	NBHRE.COUNT
(Intercept)	-0.0475	-0.0475	0.6303	-0.0444
Stillbirth	0.9126	0.9127	-0.0051	0.9126
ANC4Visits	0.0011	0.0011	0.0386	0.0011
LateTeenPreg	0.0023	0.0023	0.0006	0.0023
AntHaemorrage	-0.0307	-0.0307	0.0018	-0.0304
PostHaemorrage	0.0302	0.0302	NA	0.0305
BreechDelivery	0.0317	0.0317	NA	0.0318
Carbatosin	0.2087	0.2087	NA	0.2131
Uterotonics3stg	-0.0196	-0.01961	NA	-0.0201

Table 4.5 shows the resulting AICs following the fit of Poisson, Negative Binomial, Poisson Hurdle, NB Hurdle, Poisson Hurdle REWLR, NB Hurdle REWLR models to the maternal mortality data. NB Hurdle REWLR produced the lowest AIC, indicating a better fit than the other count models.

Table 4.5: AIC for Maternal Mortality Models

Model	AIC
Poisson	469.6684
PH	335.9051
PH-RE	370.6200
NB	473.5588
NB-H	337.9054
NBH-RE	284.1434

ROC and the corresponding AUC values were obtained as shown in Figure 4.10. The Hurdle models employed logistic regression algorithm for classification in the binary component while the Hurdle-REWLR used REWLR algorithm. Both models scored highly on AUC with values close to 1, an indication of good model performance. The classification algorithm introduced by the study's models emerged the better performing algorithm, with higher AUC scores.

It was also of interest to the study how the Hurdle-REWLR predicted zero counts compared to their counterpart standard Hurdle models. From Table 4.6, we observe that Poisson Hurdle and NB Hurdle models accurately predicted the observed number of zero counts in the sample data. The predicted zero counts from the sample data were slightly less than that observed in the sample data.

Table 4.6: Observed and Expected Zero Counts

Oberved	Poisson	PH	PH-RE	NB	NB-H	NBH-RE
137	122	137	102	126	137	102

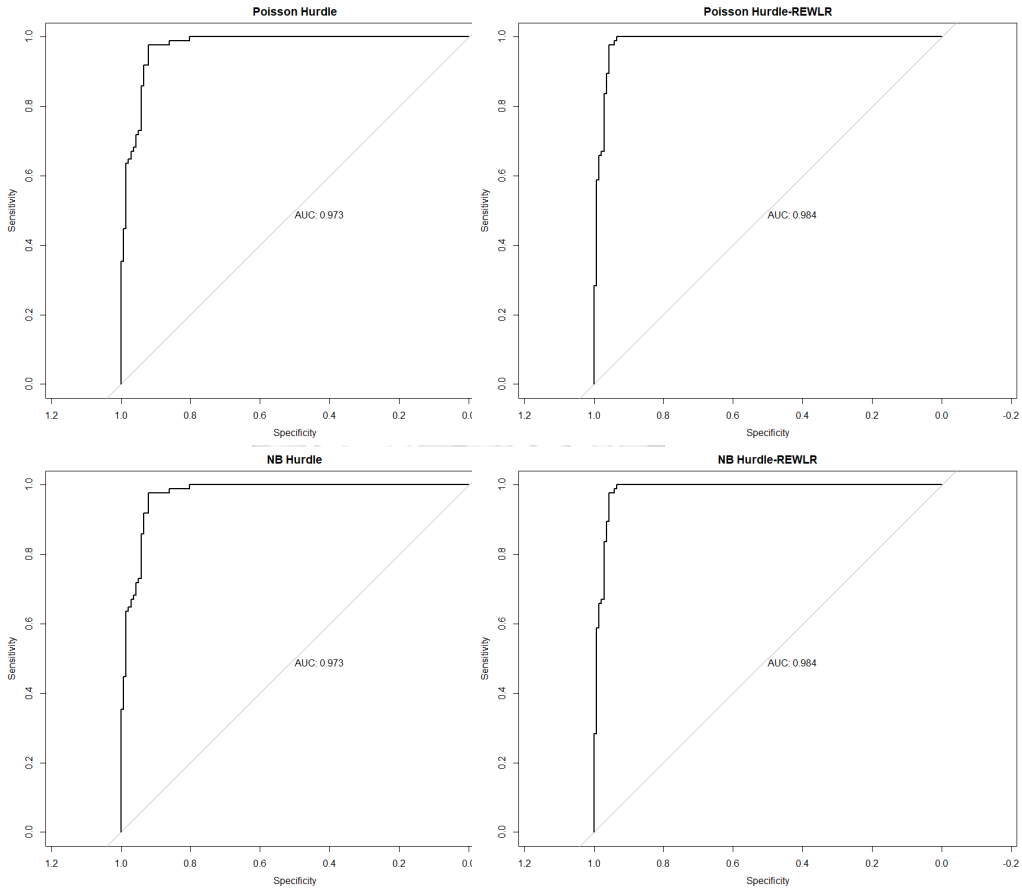


Figure 4.10: ROC-AUC for the various models

Chapter 5

Discussion, Conclusion and Recommendation

5.1 Introduction

This section presents an interpretation of the study's research findings in relation to findings made by previous researchers on the same topic. We further draw conclusions and make recommendations based on our study outputs.

5.2 Discussion

Cases of rare events in count data where the proportion of zero counts is significantly less than that of the natural numbers have been shown to influence binary estimations in zero-inflated count models. Theoretically, more extreme rare events are expected to impose extreme bias towards the majority group, i.e., zero counts. Because of this hypothesis, the current study conducted extensive simulation and analysis with varying proportions of zeros and sample sizes to evaluate the performance of the Hurdle-REWLR models in the various simulation conditions. The study also evaluated the performance of the study models alongside the standard hurdle models when fit on maternal mortality data to determine the factors which influence maternal mortality in Nairobi, Kenya.

The analysis to determine factors which influence maternal deaths in Nairobi resulted in NB Hurdle-REWLR outperforming the other models in terms of the Akaike Information Criterion. The Hurdle-REWLR models adjusted for the population estimates by introducing

weights and regularizing the coefficients. The predicted zero counts from the sample data emerged to be slightly less than observed. The number estimated by the Hurdle-REWLR models could be expected from a sample that accurately represents the Nairobi population. The introduction of the weights makes the Hurdle-REWLR models ideal for estimations and inference.

In both the Hurdle and hurdle-REWLR models, specific demographic and obstetric factors significantly affected the response variable. Age, described by the number of pregnant adolescents, and attendance of at least 4 ANC visits are some of the demographic factors shown by past literature to influence maternal deaths. In addition, childbirth-related conditions of postpartum haemorrhage, treated by uterotonics, antepartum haemorrhage, breech delivery and Macerated stillbirth were also discovered to influence the number of maternal deaths. Haemorrhage is among the obstetric factors which were highlighted by (Organization et al., 2019) to be some of the causes of maternal death globally. These findings were also in line with Nyaboga (2009) research which outlined the influential factors and causes of maternal mortality in Kenya's national referral hospital, KNH. Some of the factors identified in their research which the current study has outlined, include age, ANC attendance, and Postpartum haemorrhage.

Simulation analysis findings revealed NB Hurdle-REWLR to produce the lowest AIC value compared to the other models. The percentage difference in AICs between NB Hurdle-REWLR and the other misspecified models increased as the zeros in the data increased. The Poisson Hurdle-REWLR model outperformed the NB Hurdle in Poisson Hurdle REWLR simulated data but was inferior in NB Hurdle simulated data. It could not account for the extra dispersion introduced in the NB simulated data. The Hurdle-REWLR models, through their binary component, also outperformed the standard Hurdle models, in terms of ROC-AUC statistics. The classification algorithm used for the Hurdle-REWLR performed better at classifying imbalanced data.

The selection of NB Hurdle REWLR as the ideal model over the standard NB Hurdle model was influenced by the degree of zero inflation in the simulation analysis. The two models gave almost similar results when the proportions of zeros and non-zeros in the scenarios

where the data were not significantly different. For instance, in the NB Hurdle simulated data, the model performed better than the NB Hurdle REWLR model at 50% zero inflation but was outperformed for the subsequent degrees of zero inflation of 60%, 75% and 90%. This outcome conformed to the basic concept of the Hurdle-REWLR model. As outlined by Maalouf and Siddiqi (2014), REWLR is modified from logistic regression with the aim of unbiased prediction in rare events with imbalanced data. If the proportions of zero and non-zero counts are balanced, REWLR is not expected to outperform logistic regression.

Despite their foundations on similar concepts, the performance of the Poisson Hurdle REWLR was inferior to that of the NB Hurdle REWLR model. In the simulation analysis, Poisson Hurdle REWLR outperformed its counterpart in data generated by its distribution, and the Poisson Hurdle simulated data only by small units of percentage change in AIC. In the other simulation scenarios, the NB Hurdle REWLR model claimed superiority by quite huge margins of the percentage change in AIC. In addition, NB Hurdle REWLR was the best performing model for the fit on maternal mortality data. Such results have been witnessed in various performance comparison studies including Fenta et al. (2020) and Mamun (2014). It is common for the NB model to outperform its Poisson counterpart when there is some dispersion in the data.

In the evaluation to assess how the Hurdle-REWLR predicted zero counts compared to the Hurdle models, the hurdle models predicted the exact number of zeros available in the sample data. The binary component of the Hurdle models uses logistic regression to predict the zero counts. The prediction accuracy can thus be attributed to the bias towards the majority class. Rahim et al. (2019) assessed the performance of SMOTE logistic as a classifier in rare events data and revealed a similar outcome where SMOTE logistic regression approach was more accurate compared to the logistic regression model but was outperformed by the latter in test prediction accuracy.

The Negative Binomial hurdle REWLR model was selected based on the Akaike information criterion. The model was then fit to the maternal mortality data. The covariate factors that were significantly associated with maternal deaths at the binary level include attendance of at least 4 ANC visits, antepartum haemorrhage, and receiving uterotonics during or immediately

after birth. Upon observing maternal death within a facility, the covariate factors influencing the number of maternal deaths reported are Macerated stillbirth, attendance of at least 4 ANC visits, adolescent pregnancies, antepartum haemorrhage, breech deliveries, postpartum haemorrhage, receiving Carbatosin and giving birth by cesarean section.

The Hurdle-REWLR model has an advantage over the Hurdle models because of their ability to introduce weights hence producing more accurate estimates that can be used for inference of population parameters. When the zero-inflated sample accurately represents the population, choosing between these two groups of models could be based on Akaike Information Criterion.

5.3 Conclusion

The main aim of this study was to create Poisson and NB Hurdle-REWLR models for zero-inflated data and evaluate their performance in comparison to the standard Hurdle models. The Hurdle-REWLR in their binary component accounted for an imbalance between majority and minority proportions. That was the differentiating factor between the two models.

The proposed study models were then applied to simulated and maternal mortality data, where NB Hurdle-REWLR outperformed the other models. The difference in AIC based performance between the NB Hurdle REWLR model and the other models increased with an increase in the degree of zero inflation. The ideal model performed better in cases of class imbalance. The study findings also highlighted a case of biased classification. The binary component of the Hurdle model, using logistic regression, classified all the observed zero counts in the maternal mortality as zeroes. Despite the prediction being an exact fit, the NB Hurdle model was inferior in AIC measures. NB Hurdle REWLR was thus selected as the ideal model in rare event cases where class imbalance exists.

The study further outlined factors influencing maternal deaths in Nairobi: adolescent pregnancy, attendance of at least 4 ANC visits, postpartum haemorrhage, antepartum haemor-

rhage, breech delivery, and Macerated stillbirth. Most of these factors have been identified as determinants or causes of maternal deaths literature reviewed by this study.

Findings from this research are expected to provide reliable estimates of the number of maternal deaths in Nairobi, Kenya. Without the risk of overfitting zero counts, researchers will be able to realize the actual maternal mortality ratio and the factors associated with zero maternal death counts. The research results will assist in supporting existing policies and developing new programs and interventions to reduce the number of deaths due to childbirth and maternity.

5.4 Recommendation

5.4.1 Recommendation for further research

One area for further research is the implementation of the Hurdle-REWLR models on normally distributed covariates. The covariates of the current study data consisted of count data, majority being zero-inflated just as the dependent variable; this limited the covariate effect on the dependent variable.

5.4.2 Policy recommendation

This study recommends that the proposed interventions be implemented to halt any avoidable deaths of women during and immediately after childbirth. These interventions, such as the implementation of BEmONC or CEmONC has yet to be rolled out in all healthcare facilities. Actualizing this would go a long way in preventing maternal deaths due to obstetric conditions. Maternal deaths due to demographic and social factors such as adolescent pregnancies and attendance of ANC visits can be countered by educating the public on all the associated risks of these practices or lack-off.

References

Ali, E. (2020). Zero-inflated poisson regression model for a new class of flexible link functions: A case study on healthcare utilization.

Arefaynie, M., Kefale, B., Yalew, M., Adane, B., Dewau, R., and Damtie, Y. (2022). Number of antenatal care utilization and associated factors among pregnant women in ethiopia: zero-inflated poisson regression of 2019 intermediate ethiopian demography health survey. *Reproductive Health*, 19(1):1–10.

Aryuyuen, S., Bodhisuwan, W., and Supapakorn, T. (2014). Zero inflated negative binomial-generalized exponential distribution and its applications. *Songklanakarin Journal of Science and Technology*, 36(4):483–491.

Chaudhari, M., Hubbard, R., Reid, R. J., Inge, R., Newton, K. M., Spangler, L., and Barlow, W. E. (2012). Evaluating components of dental care utilization among adults with diabetes and matched controls via hurdle models. *BMC oral health*, 12(1):1–12.

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*. University of Minnesota.

Diop, A., Deme, E. H., and Diop, A. (2021). Zero-inflated generalized extreme value regression model for binary data and application in health study. *arXiv preprint arXiv:2105.00482*.

Fenta, S. M. and Fenta, H. M. (2020). Risk factors of child mortality in ethiopia: application of multilevel two-part model. *PLoS One*, 15(8):e0237640.

Fenta, S. M., Fenta, H. M., and Ayenew, G. M. (2020). The best statistical model to estimate predictors of under-five mortality in ethiopia. *Journal of Big Data*, 7(1):1–14.

Fitriani, R., Chrisdiana, L. N., and Efendi, A. (2019). Simulation on the zero inflated negative binomial (zinb) to model overdispersed, poisson distributed data. In *IOP Conference Series: Materials Science and Engineering*, volume 546, page 052025. IOP Publishing.

Greene, W. H. (1994). Accounting for excess zeros and sample selection in poisson and negative binomial regression models.

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.

Hilbe, J. M. (2014). *Modeling count data*. Cambridge University Press.

Hutchinson, M. K. and Holtman, M. C. (2005). Analysis of count data using poisson regression. *Research in nursing & health*, 28(5):408–418.

Jabeur, S. B. (2017). Bankruptcy prediction using partial least squares logistic regression. *Journal of Retailing and Consumer Services*, 36:197–202.

Kibika, S. A. (2020). *The Zero Inflated Negative Binomial-Shanker distribution and its application to HIV exposed infant data*. PhD thesis, Strathmore University.

King, G. (1989). Event count models for international relations: Generalizations and applications. *International Studies Quarterly*, 33(2):123–147.

King, G. and Zeng, L. (2001). Logistic regression in rare events data. *Political analysis*, 9(2):137–163.

Lambert, D. (1992). Zero-inflated poisson with an regression, in manufacturing to defects application. *Technometrics*, 34(1):14.

Loquiha, O., Hens, N., Chavane, L., Temmerman, M., and Aerts, M. (2013). Modeling heterogeneity for count data: A study of maternal mortality in health facilities in mozambique. *Biometrical Journal*, 55(5):647–660.

Maalouf, M., Homouz, D., and Trafalis, T. B. (2018). Logistic regression in large rare events and imbalanced data: A performance comparison of prior correction and weighting methods. *Computational Intelligence*, 34(1):161–174.

Maalouf, M. and Siddiqi, M. (2014). Weighted logistic regression for large-scale imbalanced and rare events data. *Knowledge-Based Systems*, 59:142–148.

Mamun, M. A. A. (2014). Zero-inflated regression models for count data: an application to under-5 deaths.

Mason, C., Twomey, J., Wright, D., and Whitman, L. (2018). Predicting engineering student attrition risk using a probabilistic neural network and comparing results with a backpropagation neural network and logistic regression. *Research in Higher Education*, 59(3):382–400.

McDowell, A. (2003). From the help desk: hurdle models. *The Stata Journal*, 3(2):178–184.

Miller, J. M. (2007). *Comparing Poisson, Hurdle, and ZIP model fit under varying degrees of skew and zero-inflation*. PhD thesis, University of Florida.

Min, Y. and Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical modelling*, 5(1):1–19.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of econometrics*, 33(3):341–365.

Mwangi, A., Nangami, M., Tabu, J., Ayuku, D., Were, E., and Fabian, E. (2019). A system approach to improving maternal and child health care delivery in kenyan communities and primary care facilities: baseline survey on maternal health. *African Health Sciences*, 19(2):1841–1848.

Neelon, B., Chang, H. H., Ling, Q., and Hastings, N. S. (2016). Spatiotemporal hurdle models for zero-inflated count data: exploring trends in emergency department visits. *Statistical methods in medical research*, 25(6):2558–2576.

Nekesa, F. V. (2019). *Distributions of zero-inflated models with application to HIV exposed infants*. PhD thesis, Strathmore University.

Nusinovici, S., Tham, Y. C., Yan, M. Y. C., Ting, D. S. W., Li, J., Sabanayagam, C., Wong, T. Y., and Cheng, C.-Y. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of clinical epidemiology*, 122:56–69.

Nyaboga, E. O. (2009). *Maternal mortality at Kenyatta National'hospital (Nairobi, Kenya) 2000-2008*. PhD thesis.

Odhiambo, C. and Kinoti, F. (2019). Evaluation and comparison of patterns of maternal complications using generalized linear models of count data time series. *International Journal of Statistics in Medical Research*, 8:32–39.

Organization, W. H. et al. (2019). Trends in maternal mortality 2000 to 2017: estimates by who, unicef, unfpa, world bank group and the united nations population division.

Rahim, A. H. A., Rashid, N. A., Nayan, A., and Ahmad, A.-R. (2019). Smote approach to imbalanced dataset in logistic regression analysis. In *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)*, pages 429–433. Springer.

Rose, C. E., Martin, S. W., Wannemuehler, K. A., and Plikaytis, B. D. (2006). On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. *Journal of biopharmaceutical statistics*, 16(4):463–481.

Smith, H., Ameh, C., Godia, P., Maua, J., Bartilol, K., Amoth, P., Mathai, M., and van den Broek, N. (2017). Implementing maternal death surveillance and response in kenya: incremental progress and lessons learned. *Global Health: Science and Practice*, 5(3):345–354.

Tecla, S. J., Franklin, B., David, A., and Jackson, T. K. (2017). Assessing facility readiness to offer basic emergency obstetrics and neonatal care (bemonc) services in health care facilities of west pokot county, kenya. *J Clin Simul Res*, 7:25–39.

Tolles, J. and Meurer, W. J. (2016). Logistic regression: relating patient characteristics to outcomes. *Jama*, 316(5):533–534.

Wang, H. (2020). Logistic regression for massive data with rare events. In *International Conference on Machine Learning*, pages 9829–9836. PMLR.

Yego, F., D'este, C., Byles, J., Williams, J. S., and Nyongesa, P. (2014). Risk factors for maternal mortality in a tertiary hospital in kenya: a case control study. *BMC pregnancy and childbirth*, 14(1):1–9.

Zare, N., Haem, E., Lankarani, K. B., Heydari, S. T., and Barooti, E. (2013). Breast cancer risk factors in a defined population: weighted logistic regression approach for rare events. *Journal of breast cancer*, 16(2):214–219.

Zhen, Z., Shao, L., and Zhang, L. (2018). Spatial hurdle models for predicting the number of children with lead poisoning. *International journal of environmental research and public health*, 15(9):1792.

Ziraba, A. K., Madise, N., Mills, S., Kyobutungi, C., and Ezeh, A. (2009). Maternal mortality in the informal settlements of nairobi city: what do we know? *Reproductive health*, 6(1):1–8.

Appendix A

R CODES

The R code used for simulations and model fitting in Chapter 4.

A.1 Libraries

```
library(ggplot2)
library(sandwich)
library(msm)
library(dplyr)
library(tidyr)
library(vcd)
library(countreg)
library(pscl)
library(VGAM)
library(rewlr)
library(kableExtra)
library(readxl)
library(gridExtra)
library(glmnet)
library(plotrix)
library(ZIM)
library(tidyverse)
```

A.2 Simulations and Analysis

```
#Generate data from Poisson Hurdle distribution.
#Repeat for Poisson Hurdle-REWLR, NB Hurdle, and NB Hurdle-REWLR.

# Probability of 0 = 1-p

#Set the seed for reproducible results
set.seed(2345)
#Assigned weights
w1 = 3.5
w0 = 1

#Zero-altered poisson random number generator function
zero.aic.func <- function(n, pi, zero.prop) {
rhpois <- function(n=n, mu, zprob){
ifelse(rbinom(n, 1, zprob) == 1, 0, rpois(n, mu))
}
Y <- rhpois(n, mu = 1.3, zprob = pi) #Poisson Hurdle
X <- rpois(n, 3.5)
dsname <- data.frame(Y, X)

#Poisson Regression
model1 <- glm(Y ~ X, family="poisson", data=dsname)
aic1 <- summary(model1)$aic

#Poisson Hurdle Regression
model2 <- hurdle(Y ~ X, data=dsname, dist = "poisson", link="logit")
aic2 <- AIC(model2)
```

```
#REWLR-Hurdle Poisson Regression
model3.a <- vglm(Y[Y > 0] ~ X[Y > 0], family = pospoisson(), data=dsname)
model3.b <- rewlr(I(Y > 0) ~ X, weights0 = w0, weights1 = w1, data=dsname)
aic.val3.a <- AICvlm(model3.a)
aic.val3.b <- model3.b$aic

#Negative Binomial Regression
model4 <- glm.nb(Y ~ X, data=dsname)
aic4 <- summary(model4)$aic

#Negative Binomial Hurdle Regression
model5 <- hurdle(Y ~ X, dist = "negbin")
aic5 <- AIC(model5)

#REWLR-Hurdle Negative Binomial Regression
model6.a <- vglm(Y[Y > 0] ~ X[Y > 0], family = posnegbinomial(), data=dsname)
model6.b <- rewlr(I(Y > 0) ~ X, weights0 = w0, weights1 = w1, data=dsname)
aic.val6.a <- AICvlm(model6.a)
aic.val6.b <- model6.b$aic
aic6 <- aic.val6.a + aic.val6.b

#AIC Values based on arious zero-proportions
aic.values <- data.frame(AIC=rbind(aic1, aic2, aic3, aic4, aic5, aic6),
Model=c("Poisson", "PH", "PH-RE","NB", "NB-H", "NBH-RE"),
Zero.Proportion=c(rep(pi, 6)), row.names = NULL)

#Figure: Performance based on AIC values for sets of models, sample size, zero%
plot <- ggplot(aic.values, aes(x=Model, y=AIC))+
geom_bar(stat = "identity", fill="azure3")+
geom_text(aes(label=round(AIC, digits=0)), vjust=1.6, color="black", size=3.5)+
```

```
scale_x_discrete(limits=c("Poisson", "NB", "PH", "NB-H", "PH-RE", "NBH-RE"))+
labs(y = "AIC Value", x = "Count Model")+
ggtitle(zero.prop)+
theme_bw()

return(plot)
}

plot1 <- zero.aic.func(200, 0.50, "n=200; 50% zero-inflation")
plot2 <- zero.aic.func(200, 0.40, "n=200; 60% zero-inflation")
plot3 <- zero.aic.func(200, 0.25, "n=200; 75% zero-inflation")
plot4 <- zero.aic.func(200, 0.10, "n=200; 90% zero-inflation")

png(file="D:/MSc/Thesis/Thesis-template-20220215T185005Z-001/
Thesis-template/Figs/PH200.png", width=600, height=350)
grid.arrange(
plot1, plot2, plot3, plot4,
ncol=2, nrow = 2)
dev.off()

plot5 <- zero.aic.func(1000, 0.50, "n=1000; 50% zero-inflation")
plot6 <- zero.aic.func(1000, 0.40, "n=1000; 60% zero-inflation")
plot7 <- zero.aic.func(1000, 0.25, "n=1000; 75% zero-inflation")
plot8 <- zero.aic.func(1000, 0.10, "n=1000; 90% zero-inflation")

png(file="D:/MSc/Thesis/Thesis-template-20220215T185005Z-001/
Thesis-template/Figs/PH500.png", width=600, height=350)
grid.arrange(
plot5, plot6, plot7, plot8,
ncol=2, nrow = 2)
```

```
dev.off()

plot9  <- zero.aic.func(500, 0.50, "n=500; 50% zero-inflation")
plot10 <- zero.aic.func(500, 0.40, "n=500; 60% zero-inflation")
plot11 <- zero.aic.func(500, 0.25, "n=500; 75% zero-inflation")
plot12 <- zero.aic.func(500, 0.10, "n=500; 90% zero-inflation")

png(file="D:/MSc/Thesis/Thesis-template-20220215T185005Z-001/
Thesis-template/Figs/PH1000.png", width=600, height=350)
grid.arrange(
plot9, plot10, plot11, plot12,
ncol=2, nrow = 2)
dev.off()

#Compute Percentage change in AIC
aic.chg.func <- function(n, pi) {

# Zero-altered poisson random number generator
rhpois <- function(n=n, mu, zprob){
ifelse(rbinom(n, 1, zprob) == 1, 0, rpois(n, mu))
}

Y <- rhpois(n, mu = 1.3, zprob = pi) #Poisson Hurdle
X <- rpois(n, 3.5) # Independent variable X
dsname <- data.frame(Y, X)

#Poisson Regression
model1 <- glm(Y ~ X, family="poisson", data=dsname)
aic1 <- summary(model1)$aic
```

```
#Poisson Hurdle Regression
model2 <- hurdle(Y ~ X, data=dsname, dist = "poisson", link="logit")
aic2 <- AIC(model2)

## REWLR-Hurdle Poisson Regression
# Error due to vglm: https://bookdown.org/fxpalacio/bookdown_curso/GLM.html
model3.a <- vglm(Y[Y > 0] ~ X[Y > 0], family = pospoisson(), data=dsname)
model3.b <- rewlr(I(Y > 0) ~ X, weights0 = w0, weights1 = w1, data=dsname)
aic.val3.a <- AICvlm(model3.a)
aic.val3.b <- model3.b$aic
aic3 <- aic.val3.a + aic.val3.b

#Negative Binomial Regression
model4 <- glm.nb(Y ~ X, data=dsname)
aic4 <- summary(model4)$aic

#Negative Binomial Hurdle Regression
model5 <- hurdle(Y ~ X, dist = "negbin")
aic5 <- AIC(model5)

## REWLR-Hurdle Negative Binomial Regression
model6.a <- vglm(Y[Y > 0] ~ X[Y > 0], family = posnegbinomial(), data=dsname)
model6.b <- rewlr(I(Y > 0) ~ X, weights0 = w0, weights1 = w1, data=dsname)

aic.val6.a <- AICvlm(model6.a)
aic.val6.b <- model6.b$aic
aic6 <- aic.val6.a + aic.val6.b

#AIC Values based on various zero-proportions
aic.values.1 <- data.frame(ref="Poisson Hurdle (PH)", sample.size = n,
```

```
Zero.Proportion=1-pi, AIC=cbind(aic2, aic3, aic5, aic6))
colnames(aic.values.1) <- c("Reference","Sample size", "Zero Proportion",
"PH", "PHRE", "NBH", "NBHRE")
aic.values.1$PH <- round(aic.values.1$PH, 0)
aic.values.1$PHRE <- paste(round(aic.values.1$PHRE, 0), '(',
round(((aic.values.1$PHRE - aic.values.1$PH)/aic.values.1$PHRE)*100, 2),'%)')
aic.values.1$NBH <- paste(round(aic.values.1$NBH, 0), '(',
round(((aic.values.1$NBH - aic.values.1$PH)/aic.values.1$NBH)*100, 2),'%)')
aic.values.1$NBHRE <- paste(round(aic.values.1$NBHRE, 0), '(',
round(((aic.values.1$NBHRE - aic.values.1$PH)/aic.values.1$NBHRE)*100, 2),'%)')

return(aic.values.1)
}

ph <- rbind(aic.chg.func(200, 0.50),aic.chg.func(200, 0.40),
aic.chg.func(200, 0.25),aic.chg.func(200, 0.10),
aic.chg.func(1000, 0.50),aic.chg.func(1000, 0.40),
aic.chg.func(1000, 0.25),zero.aic.func(1000, 0.10))


#Step 1: Generate data from Poisson Hurdle-REWLR distribution
aic.chg.func <- function(n, pi, zero.prop) {

# Zero-altered poisson random number generator
rhpois <- function(n=n, mu, zprob){
ifelse(rbinom(n, 1, zprob) == 1, 0, rpois(n, mu))
}

Y <- rhpois(n, mu = 1.3, zprob = pi^w1) #Poisson Hurdle-RE
X <- runif(n, -1, 1) # Independent variable X
```

```
dsname <- data.frame(Y, X)

#Poisson Regression
model1 <- glm(Y ~ X, family="poisson", data=dsname)
aic1 <- summary(model1)$aic

#Poisson Hurdle Regression
model2 <- hurdle(Y ~ X, data=dsname, dist = "poisson", link="logit")
aic2 <- AIC(model2)

## REWLR-Hurdle Poisson Regression
model3.a <- vglm(Y[Y > 0] ~ X[Y > 0], family = pospoisson(), data=dsname)
model3.b <- rewlr(I(Y > 0) ~ X, weights0 = w0, weights1 = w1, data=dsname)
aic.val3.a <- AICvlm(model3.a)
aic.val3.b <- model3.b$aic
aic3 <- aic.val3.a + aic.val3.b

#Negative Binomial Regression
model4 <- glm.nb(Y ~ X, data=dsname)
aic4 <- summary(model4)$aic

#Negative Binomial Hurdle Regression
model5 <- hurdle(Y ~ X, dist = "negbin")
aic5 <- AIC(model5)

## REWLR-Hurdle Negative Binomial Regression
model6.a <- vglm(Y[Y > 0] ~ X[Y > 0], family = posnegbinomial(), data=dsname)
model6.b <- rewlr(I(Y > 0) ~ X, weights0 = w0, weights1 = w1, data=dsname)

aic.val6.a <- AICvlm(model6.a)
```

```
aic.val6.b <- model6.b$aic
aic6 <- aic.val6.a + aic.val6.b

#AIC Values based on various zero-proportions
aic.values.1 <- data.frame(ref="Poisson Hurdle - REWLR (PHRE)",
sample.size = n, Zero.Proportion=1-pi, AIC=cbind(aic2, aic3, aic5, aic6))
colnames(aic.values.1) <- c("Reference","Sample size", "Zero Proportion",
 "PH", "PHRE", "NBH", "NBHRE")
aic.values.1$PHRE <- round(aic.values.1$PHRE, 0)
aic.values.1$PH <- paste(round(aic.values.1$PH, 0), '(',
round(((aic.values.1$PH - aic.values.1$PHRE)/aic.values.1$PH)*100, 2),'%)')
aic.values.1$NBH <- paste(round(aic.values.1$NBH, 0), '(',
round(((aic.values.1$NBH - aic.values.1$PHRE)/aic.values.1$NBH)*100, 2),')')
aic.values.1$NBHRE <- paste(round(aic.values.1$NBHRE, 0), '(',
round(((aic.values.1$NBHRE - aic.values.1$PHRE)/aic.values.1$NBHRE)*100, 2),'%)')

return(aic.values.1)
}

phre <- rbind(aic.chg.func(200, 0.50),aic.chg.func(200, 0.40),
aic.chg.func(200, 0.25),aic.chg.func(200, 0.10),
aic.chg.func(1000, 0.50),aic.chg.func(1000, 0.40),
aic.chg.func(1000, 0.25),zero.aic.func(1000, 0.10))


#Step 1: Generate data from NB Hurdle distribution
aic.chg.func <- function(n, pi) {

# Zero-altered negative binomial random number generator
rhnbinom <- function(n=n, mu, size=0.5, zprob){
```

```
ifelse(rbinom(n, 1, zprob) == 1, 0, rnbinom(n, size = 0.5, mu = mu))
}

Y <- rhnbinom(n, mu = 1.3, size = 3, zprob = pi) #NB Hurdle
X <- runif(n, -1, 1) # Independent variable X
dsname <- data.frame(Y, X)

#Poisson Regression
model1 <- glm(Y ~ X, family="poisson", data=dsname)
aic1 <- summary(model1)$aic

#Poisson Hurdle Regression
model2 <- hurdle(Y ~ X, data=dsname, dist = "poisson", link="logit")
aic2 <- AIC(model2)

## REWLR-Hurdle Poisson Regression
model3.a <- vglm(Y[Y > 0] ~ X[Y > 0], family = pospoisson(), data=dsname)
model3.b <- rewlr(I(Y > 0) ~ X, weights0 = w0, weights1 = w1, data=dsname)
#aic.val3.a <- (-2*logLik.vlm(model3.a))+(2*3)
aic.val3.a <- AICvlm(model3.a)
aic.val3.b <- model3.b$aic
aic3 <- aic.val3.a + aic.val3.b

#Negative Binomial Regression
model4 <- glm.nb(Y ~ X, data=dsname)
aic4 <- summary(model4)$aic

#Negative Binomial Hurdle Regression
model5 <- hurdle(Y ~ X, dist = "negbin")
aic5 <- AIC(model5)
```

```
## REWLR-Hurdle Negative Binomial Regression
model6.a <- vglm(Y[Y > 0] ~ X[Y > 0], family = posnegbinomial(), data=dsname)
model6.b <- rewlr(I(Y > 0) ~ X, weights0 = w0, weights1 = w1, data=dsname)

#aic.val6.a <- (-2*logLik.vlm(model3.a))+(2*3)
aic.val6.a <- AICvlm(model6.a)
aic.val6.b <- model6.b$aic
aic6 <- aic.val6.a + aic.val6.b


#AIC Values based on various zero-proportions
aic.values.1 <- data.frame(ref="NB Hurdle (NBH)", sample.size = n,
Zero.Proportion=1-pi, AIC=cbind(aic2, aic3, aic5, aic6))
colnames(aic.values.1) <- c("Reference","Sample size", "Zero Proportion",
 "PH", "PHRE", "NBH", "NBHRE")
aic.values.1$NBH <- round(aic.values.1$NBH, 0)
aic.values.1$PH <- paste(round(aic.values.1$PH, 0), '(',
round(((aic.values.1$PH - aic.values.1$NBH)/aic.values.1$PH)*100, 2),'%)')
aic.values.1$PHRE <- paste(round(aic.values.1$PHRE, 0), '(',
round(((aic.values.1$PHRE - aic.values.1$NBH)/aic.values.1$PHRE)*100, 2),'%)')
aic.values.1$NBHRE <- paste(round(aic.values.1$NBHRE, 0), '(',
round(((aic.values.1$NBHRE - aic.values.1$NBH)/aic.values.1$NBHRE)*100, 2),'%)')

return(aic.values.1)
}

nbh <- rbind(aic.chg.func(200, 0.50),aic.chg.func(200, 0.40),
aic.chg.func(200, 0.25),aic.chg.func(200, 0.10),
aic.chg.func(1000, 0.50),aic.chg.func(1000, 0.40),
```

```
aic.chg.func(1000, 0.25),zero.aic.func(1000, 0.10))


#Step 1: Generate data from NB Hurdle REWLR distribution
aic.chg.func <- function(n, pi, zero.prop) {

# Zero-altered negative binomial random number generator
rhnbinom <- function(n=n, mu, size=0.5, zprob){
ifelse(rbinom(n, 1, zprob) == 1, 0, rnbinom(n, size = 0.5, mu = mu))
}

Y <- rhnbinom(n, mu = 1.3, size = 3, zprob = pi^w1) #NB Hurdle-RE
X <- runif(n, -1, 1) # Independent variable X
dsname <- data.frame(Y, X)

#Poisson Regression
model1 <- glm(Y ~ X, family="poisson", data=dsname)
aic1 <- summary(model1)$aic

#Poisson Hurdle Regression
model2 <- hurdle(Y ~ X, data=dsname, dist = "poisson", link="logit")
aic2 <- AIC(model2)

## REWLR-Hurdle Poisson Regression
model3.a <- vglm(Y[Y > 0] ~ X[Y > 0], family = pospoisson(), data=dsname)
model3.b <- rewlr(I(Y > 0) ~ X, weights0 = w0, weights1 = w1, data=dsname)
#aic.val3.a <- (-2*logLik.vlm(model3.a))+(2*3)
aic.val3.a <- AICvlm(model3.a)
aic.val3.b <- model3.b$aic
aic3 <- aic.val3.a + aic.val3.b
```

```
#Negative Binomial Regression
model4 <- glm.nb(Y ~ X, data=dsname)
aic4 <- summary(model4)$aic

#Negative Binomial Hurdle Regression
model5 <- hurdle(Y ~ X, dist = "negbin")
aic5 <- AIC(model5)

## REWLR-Hurdle Negative Binomial Regression
model6.a <- vglm(Y[Y > 0] ~ X[Y > 0], family = posnegbinomial(), data=dsname)
model6.b <- rewlr(I(Y > 0) ~ X, weights0 = w0, weights1 = w1, data=dsname)
#
#aic.val6.a <- (-2*logLik.vlm(model3.a))+(2*3)
aic.val6.a <- AICvlm(model6.a)
aic.val6.b <- model6.b$aic
aic6 <- aic.val6.a + aic.val6.b


#AIC Values based on various zero-proportions
aic.values.1 <- data.frame(ref="NB Hurdle - REWLR (NBHRE)", sample.size = n,
Zero.Proportion=1-pi, AIC=cbind(aic2, aic3, aic5, aic6))
colnames(aic.values.1) <- c("Reference","Sample size", "Zero Proportion",
"PH", "PHRE", "NBH", "NBHRE")
aic.values.1$NBHRE <- round(aic.values.1$NBHRE, 0)
aic.values.1$PH <- paste(round(aic.values.1$PH, 0), '(',
round(((aic.values.1$PH - aic.values.1$NBHRE)/aic.values.1$PH)*100, 2),'%)')
aic.values.1$NBH <- paste(round(aic.values.1$NBH, 0), '(',
round(((aic.values.1$NBH - aic.values.1$NBHRE)/aic.values.1$NBH)*100, 2),'%)')
aic.values.1$PHRE <- paste(round(aic.values.1$PHRE, 0), '(',
```

```
round(((aic.values.1$PHRE - aic.values.1$NBHRE)/aic.values.1$PHRE)*100, 2),'%)')

return(aic.values.1)
}

nbhre <- rbind(aic.chg.func(200, 0.50),aic.chg.func(200, 0.40),
aic.chg.func(200, 0.25),aic.chg.func(200, 0.10),
aic.chg.func(1000, 0.50),aic.chg.func(1000, 0.40),
aic.chg.func(1000, 0.25),zero.aic.func(1000, 0.10))


#Combine all
allaic <- rbind(ph, phre, nbh, nbhre)

allaic %>% knitr::kable(format='latex') %>%
kable_classic_2(full_width = F, html_font = "Cambria")
```

A.3 Analysis on Maternal Mortality Data

A.3.1 Exploratory Data Analysis

```
#Read in Data
maternal <- read.csv("D:/MSc/Thesis/Analysis/Maternal Mortality Data.csv")
maternal1 <- maternal[, -1]

#Rename Columns
maternal2 <- rename(maternal1, MaternalDeaths=Maternal.Deaths,
AssistedDeliveries=assisted.Vaginal.deliveries, BreechDelivery=breach.delivery,
```

```
CS=caesarian.sections, LiveBirths=live.birth,
EarlyTeenPreg=no.adolesc..10.14.years..pregn.at.1st.anc.visit,
LateTeenPreg=no.adolesc..15.19.years..preg..at.1st.anc.visit,
NormalDeliveries=normal.deliveries,
ANC4Visits=anc.4.visits,
Uterotonics3stg=Number.of.women.giving.birth.who.received.
uterotonics.in.the.third.stage.of.labor..or.immediately.after.birth.,
Carbatosin=Mothers.given.uterotonics.within.1.minute..Carbatosin.,
Oxytocin=Mothers.given.uterotonics.within.1.minute..Oxytocin.,
Eclampsia=Eclampsia, AntHaemorrage=Ante.partum.Haemorrage
PostHaemorrage=Post.Partum.Haemorrage,
ObstructedLabour=Obstructed.Labour,
RupturedUterus=Ruptured.Uterus, Sepsis=Sepsis,
FGMComplicatons=Mothers.with.delivery.complications.
associated.with.FGM, Stillbirth=Macerated.still.Birth)

#EDA
#Central Tendency
mean(maternal2$MaternalDeaths)
std.error(maternal2$MaternalDeaths)
#Spread
sd(maternal2$MaternalDeaths)

data.frame(table(maternal2$MaternalDeaths)) %>%
kbl() %>%
kable_classic_2(full_width = F, html_font = "Cambria")

#Histogram
plot2 <- ggplot(maternal2, aes(x=MaternalDeaths)) +
geom_histogram(binwidth=1, fill="skyblue")+
```

```
labs(x = "Maternal Deaths", y = "Frequency")+
ylim(0, 150)+
theme_classic()
plot2

#Correlation Analysis
cor(maternal2)

#Averaging the factors
sum1 <- as.data.frame(t(maternal2 %>%
group_by(MaternalDeaths.bin) %>%
summarise_all(mean))) %>%
mutate(across(where(is.numeric), round, 1))

i <- which(str_detect(row.names(sum1), "^Maternal"))
sum2 <- sum1%>% slice(-i)

colnames(sum2) <- c("No Maternal Deaths","Maternal Deaths")

sum2 %>% slice(2:n()) %>%
knitr::kable(format='latex') %>%
kable_classic_2(full_width = F, html_font = "Cambria")
```

A.3.2 Count Models

```
# Sample data = 293 deaths per 53792 live births;
#Y-bar = 293/53792 = 0.0054
# Population data = 342 deaths per 100000 live births;
Tau = 1600/100000 = 0.016
```

```
#Source: https://www.health.go.ke/wp-content/uploads/2022/01/
Kenya-SDG-Progress-Report_-April21.pdf

w1 = 0.00163/0.0054
w0 = (1 - 0.00163)/(1 - 0.0054)

#Poisson Hurdle Regression
fit2 <- hurdle(MaternalDeaths ~ Stillbirth+ANC4Visits+LateTeenPreg+
AntHaemorrage+PostHaemorrage+BreechDelivery+Carbatosin+Uterotonics3stg,
dist = "poisson", link="logit", data=maternal2)
f.exp2 <- round(sum(predict(fit2, type = "prob")[,1]),0)
f.aic2 <- AIC(fit2)
summary(fit2)

## REWLR-Hurdle Poisson Regression
fit3.a <- vglm(MaternalDeaths ~ AssistedDeliveries+BreechDelivery
+CS+EarlyTeenPreg, family = pospoisson(), data=Maternal2.gt0)
fit3.b <- rewlr(MaternalDeaths.bin ~ Stillbirth+ANC4Visits+
LateTeenPreg+AntHaemorrage+PostHaemorrage+BreechDelivery+Carbatosin+
Uterotonics3stg, weights0 = w0, weights1 = w1, data=maternal2)
f.exp3 <- round(sum(1-predict.rewlr(fit3.b)))
f.aic.val3.a <- AICvlm(fit3.a)
f.aic.val3.b <- fit3.b$aic
f.aic3 <- f.aic.val3.a + f.aic.val3.b
summary(fit3.a)
summary.rewlr(fit3.b)


#Negative Binomial Hurdle Regression
fit5 <- hurdle(MaternalDeaths ~ Stillbirth+ANC4Visits+LateTeenPreg+
```

```
AntHaemorrage+PostHaemorrage+BreechDelivery+Carbatosin+
Uterotonics3stg, data=maternal2, dist = "negbin")
f.exp5 <- sum(predict(fit5, type = "prob")[,1])
f.aic5 <- AIC(fit5)
summary(fit5)

## REWLR-Hurdle Negative Binomial Regression
fit6.a <- vglm(MaternalDeaths ~ Stillbirth+ANC4Visits+LateTeenPreg+
AntHaemorrage+PostHaemorrage+BreechDelivery+Carbatosin+
Uterotonics3stg, family = posnegbinomial(), data=Maternal2.gt0)
fit6.b <- rewlr(MaternalDeaths.bin ~ Stillbirth+ANC4Visits+
LateTeenPreg+AntHaemorrage+PostHaemorrage+BreechDelivery+Carbatosin
+Uterotonics3stg, weights0 = w0, weights1 = w1, data=maternal2)
f.exp6 <- round(sum(1-predict.rewlr(fit6.b)))
f.aic.val6.a <- AICvlm(fit6.a)
f.aic.val6.b <- fit6.b$aic
f.aic6 <- f.aic.val6.a + f.aic.val6.b
summary(fit6.a)
summary.rewlr(fit6.b)


################  Area under Curve  ################
library(pROC)
#Poisson Hurdle
fit2.lm <- lm(MaternalDeaths ~ Stillbirth+ANC4Visits+LateTeenPreg+AntHaemorrage+Pos
auc1 <- roc(maternal2$MaternalDeaths.bin ~ fit2.lm$fitted, plot=TRUE, print.auc=TRU

#NB Hurdle
fit5.lm <- lm(MaternalDeaths ~ Stillbirth+ANC4Visits+LateTeenPreg+AntHaemorrage+Pos
auc3 <- roc(maternal2$MaternalDeaths.bin ~ fit5.lm$fitted, plot=TRUE, print.auc=TRU
```

```
#Poisson Hurdle - REWLR
auc2 <- roc(maternal2$MaternalDeaths.bin ~ summary.rewlr(fit3.b)$fitted, plot=TRUE,
#NB Hurdle - REWLR
auc4 <- roc(maternal2$MaternalDeaths.bin ~ summary.rewlr(fit6.b)$fitted, plot=TRUE,

#Comparison of coefficients between models
#Binary part
coef.bin <- data.frame(PH.BINARY =
summary(fit2)$coefficients$zero[,1],
NBH.BINARY = summary(fit5)$coefficients$zero[,1],
PHRE.BINARY = summary.rewlr(fit3.b)$B,
NBHRE.BINARY = summary.rewlr(fit6.b)$B)
#Count part
coef.cnt <- data.frame(cbind(PH.COUNT =
summary(fit2)$coefficients$count[,1],
NBH.COUNT = summary(fit5)$coefficients$count[,1],
PHRE.COUNT = summary(fit3.a)@coef3[, 1],
NBHRE.COUNT = summary(fit6.a)@coef3[-2, 1]))

coef.cnt$PH.COUNT[10] = "NA"
coef.cnt$NBHRE.COUNT[10] = "NA"
coef.cnt$PHRE.COUNT[c(6,7,8,9,10)] = "NA"

coef.bin %>% knitr::kable(format='latex') %>%
kable_classic_2(full_width = F, html_font = "Cambria")

coef.cnt %>% knitr::kable(format='latex') %>%
kable_classic_2(full_width = F, html_font = "Cambria")
```

```
#AIC Values based on various zero-proportions
aic.values <- data.frame(Model=c("Poisson", "PH", "PH-RE","NB",
"NB-H", "NBH-RE"), AIC=rbind(f.aic1, f.aic2, f.aic3,
f.aic4, f.aic5, f.aic6), row.names = NULL)
aic.values %>% knitr::kable(format='latex') %>%
kable_classic_2(full_width = F, html_font = "Cambria")

#Zero counts
zero.counts <- data.frame(cbind(observed,f.exp1,f.exp2,f.exp3,
f.exp4,f.exp5,f.exp6))
colnames(zero.counts) <- c("Oberved","Poisson", "PH", "PH-RE","NB",
"NB-H", "NBH-RE")

zero.counts %>% knitr::kable(format='latex') %>%
kable_classic_2(full_width = F, html_font = "Cambria")
```

Appendix B

Turnitin Report

Curiginal

Document Information

Analyzed document	Sharon Okello Thesis.pdf (D138655736)
Submitted	2022-05-31T13:02:00.0000000
Submitted by	
Submitter email	Awuor.Okello@strathmore.edu
Similarity	1%
Analysis address	library.strath@analysis.urkund.com

Sources included in the report

SA	**ST404A3.pdf** Document ST404A3.pdf (D27768399)	1
W	URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4493133/ Fetched: 2020-04-16T00:11:45.3870000	5
SA	**Assignment 3.pdf** Document Assignment 3.pdf (D27768595)	1
SA	**assignment3_1414386.pdf** Document assignment3_1414386.pdf (D27768445)	1
SA	**Handin 1.pdf** Document Handin 1.pdf (D108014019)	1

Entire Document

Improving Performance of Hurdle Models using Rare-Event Weighted Logistic Regression: Application to Maternal Mortality Data Sharon Awuor Okello Submitted in partial fulfilment of the requirements for the Degree of Master of Science in Statistical Sciences of Strathmore University Institute of Mathematical Sciences Strathmore University Nairobi, Kenya May 31, 2022 This thesis is available for Library use through open access on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Declaration I declare that this work has not been previously submitted and approved for 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 proposal itself. © No part of this thesis may be reproduced without the permission of the author and Strathmore University. Name: ..... Sharon Awuor Okello ..... Signature: ..... Date: ..... May 31, 2022 ..... Approval The thesis of Sharon Awuor Okello was reviewed and approved by the following: Dr. Collins Ojwang' Odhiambo Supervisor, Institute of Mathematical Sciences, Strathmore University. Dr. Evans Otieno Omondi Supervisor, Institute of Mathematical Sciences, Strathmore University. Dr. Godfrey Madigu Dean, Institute of Mathematical Sciences, Strathmore University. Dr. Bernard Shibwabo Director, Office of Graduate Studies, Strathmore University. ii

Abstract Hurdle models, which are commonly used alongside zero-inflated models to analyze dis- persed zero-inflated count data, employ a logit link function to predict whether an observation takes a positive count or a zero count based on a set of covariates. However, the logit model tends to be biased toward the majority zero class in cases involving rare events, and may underestimate the positive counts when their proportion is significantly smaller than that of the zero counts. This research aimed to develop and assess the performance of hurdle models incorporating rare-event weighted logistic regression and their applications to maternal mortality data. Poisson and Negative Binomial (NB) Hurdle Rare Event Weighted Logistic Regression (REWLR) model estimates were developed and fit on various simulation conditions and maternal mortality data for performance evaluation using AIC measures. The Negative Binomial Hurdle REWLR emerged to be the best performing among all the evaluated models due to the ability to handle dispersion and adjust for class imbalance. The research findings will provide reliable estimates of the maternal mortality ratio in Nairobi without the risk of over-fitting zero counts. iii

List of figures Figure 1.1: MMR Trends between 2000 - 2017: Source Organization et al. (2019) 4 Figure 4.1: AICs from Models fit on Poisson Hurdle simulated data, n = 200 . . 24 Figure 4.2: AICs from Models fit on Poisson Hurdle simulated data, n = 1000 . . 25 Figure 4.3: AICs from Models fit on Poisson Hurdle-RE simulated data, n = 200 25 Figure 4.4: AICs from Models fit on Poisson Hurdle-RE simulated data, n = 1000 26 Figure 4.5: AICs from Models fit on NB Hurdle simulated data, n = 200 . . . . . 26 Figure 4.6: AICs from Models fit on NB Hurdle simulated data, n = 1000 . . . . 27 Figure 4.7: AICs from Models fit on NB Hurdle-RE simulated data, n = 200 . . 27 Figure 4.8: AICs from Models fit on NB Hurdle-RE simulated data, n = 1000 . . 28 Figure 4.9: Maternal Death Counts . . . . . 29 vii

List of tables Table 3.1: Variable Definition . . . . . 20 Table 4.1: AIC (Percentage Change in AIC) for Misspecified and Actual Models 23 Table 4.2: Average Count of Obstetric Conditions reported in facilities with and without reported maternal deaths . . . . . 30 Table 4.3: Binary Component Coefficients . . . . . 31 Table 4.4: Count Component Coefficients . . . . . 31 Table 4.5: AIC for Maternal Mortality Models . . . . . 32 Table 4.6: Observed and Expected Zero Counts . . . . . 32 viii

List of abbreviations OLS Ordinary Least Squares WHO World Health Organization MDSR Maternal Death Surveillance and Response MMR Maternal Mortality Ratio SDG Sustainable Development Goals KNH Kenyatta National Hospital ANC Antenatal Care REWLR Rare-Event Weighted Logistic Re- gression GLM Generalized Linear Model MNCH Maternal, Newborn and Child Health ix

Acknowledgement I want to express my sincere gratitude to my academic supervisors, Dr Collins Odhiambo and Dr Evans Omondi, for the valuable advice and support throughout the research period and guidance during the thesis write-up. I am also grateful to the Strathmore Institute of Mathematical Sciences and the faculty who have imparted knowledge and offered support throughout this academic program. x

Dedication This thesis is dedicated to God for giving me the gift of life, knowledge and perseverance. To my mum Prisca Atieno Muga for her love and support. To my beloved son Haris Hawi Nyangi for whom I am motivated to be the best version of myself. xi

Chapter 1 Introduction 1.1 Background to the study Count data are generated by enumeration processes that produce discrete non-negative numbers. Due to the heteroskedastic and skewed nature of these data, the standard OLS models are not suitable for parameter estimations (Hutchinson and Holtman, 2005). Count models provide a better fit. Poisson and Negative Binomial regression are the most commonly used models for count data estimations. The Poisson model, considered the standard count model, assumes that the sample variance and sample mean are equal, a condition referred to as equidispersion. However, this is seldom the case. In practice, the sample variance is often either greater than (overdispersed) or less than the mean (under-dispersed). Poisson models provide a poor fit for such data. Overdispersion in count data may arise due to several reasons, including the presence of excess zero counts in the data (Hilbe, 2011). The negative binomial model offers a better fit for overdispersed data but may also suffer overdispersion limitations. Overdispersion in a negative binomial model could occur when the observed model variance is greater than NB's expected variance, at times, due to more zeros than the model can accommodate (Hilbe, 2014). Zero-inflated mixture models are better suitable for modelling count data with more zeros than can be accounted for by the regular count models. These models - Zero-inflated Poisson (ZIP), Zero-inflated Negative Binomial (ZINB), Poisson Hurdle (PH) and Negative Binomial Hurdle (NBH) - propose separate data-generating processes for zero and positive counts. Zero-inflated models introduced by Lambert (1992) and Greene (1994) propose a mixture distribution, where data is generated from Bernoulli and Poisson or Negative Binomial 1

processes. The most outstanding feature of the zero-inflated models, as explained by (Rose et al., 2006) is the assumption of the existence of an at-risk group that can never experience an event (structural zeros), and an at-risk group that may still not experience the event (sampling zeros). Zeros can thus be estimated using a mixture of a binary distribution - which estimates the probability of structural zeros, and a count model - which estimates all counts, including zeros. The general structure of the zero-inflated model is given by: P(Y i = y i) = ? ? ? ? ? ? ? ? ? π i +(1−π i)p(y i ;λ i |y i = 0) y i = 0 (1−π i) p(y i ;λ i) y i < 0; (1.1) Where π i is the probability of being a structural zero, p(y i ;μ) is the probability mass function of the count model, and p(y i = 0;μ) is the probability mass function of a count model for the count zero. A typical example that would motivate the application of zero-inflated models is the case of modelling the weekly number of cigarettes smoked by a group of people, following, say, a policy implementation that aims to reduce cigarette smoking. Participants who respond that they have smoked 'zero' cigarettes may either be non-smokers who cannot have any value other than zero (structural zeros) or smokers who have reduced their weekly consumption to zero (sampling zeros). Hurdle models, also called zero-altered models, provide an alternative means of modelling zero-inflated data. The distinctive feature between these models and the zero-inflated models is that hurdle models assume the existence of a

single structural source of zeros. The general concept of the hurdle models is that a binomial probability model determines whether a count response variable has a zero or a positive number. If the response variable returns a positive value, the 'hurdle' is crossed,

a maternal mortality setting, as in this study, the number of zeros reported is often excessive. From a perspective of the total number of live births, maternal deaths can be viewed as a 2

rare event. Based on WHO recommendation, data collected on maternal death through the Maternal Death Surveillance and Response (MDSR) systems include zero-reporting where weekly statistics are submitted even if no death has occurred (Smith et al., 2017). These are reported as 'zero' deaths. However, the zero deaths reported aren't distinguishable as from a structural or sampling source. One can't divide the population of women giving birth into a risk and a not-at-risk group and be certain the not-at-risk group will only report zero cases of death. Because of this, the current research focuses on Hurdle models for estimation of maternal deaths. The logistic regression model is vital in the formulations of zero-inflated mixture models. In Hurdle models, either logistic or probit regression models are used to estimate the probability of obtaining a positive count (Hilbe, 2014). However, logistic regression models show limitations when predicting probabilities in imbalanced classes, e.g., prediction of zero versus positive counts for the binary component of Hurdle models. In the case of maternal deaths reported where the zero class may always be significantly larger than the non-zero class, logistic regression tends to underestimate the probability of crossing the 'hurdle'. 1.2 Maternal Mortality in Kenya Maternal death is the death of a woman while pregnant or within 42 days of pregnancy termination, irrespective of the duration and site of the pregnancy, from any cause related to or aggravated by the pregnancy or its management but not from accidental or incidental causes (Organization et al., 2019). According to 2017 WHO estimates, MMR declined by 38% globally between 2000 and 2017 from 342 deaths to 211 deaths per 100,000 live births. Kenya reported an impressive 52% reduction in MMR from 708 to 342 for the same period (Organization et al., 2019). This significant reduction in maternal death cases can be attributed to policies and initiatives implemented by the Kenyan government, such as Free Maternity Program, Beyond Zero, Linda Mama Campaign, among others. Despite all the strides towards reducing the number of maternal deaths, MMR is still high in Kenya. Reducing the number of maternal deaths remains a national priority. 3

The third SDG of the UN launched in 2015 aims for global MMR reduction to 70 or less, or at most 140, by 2030. With the current pace of progress, Kenya may fall short of this target despite the programs and initiatives in place. More research is needed to guide new initiatives and support existing policies. Figure 1.1: MMR Trends between 2000 - 2017: Source Organization et al. (2019) Past studies on maternal mortality have aimed at establishing incidences, analyzing trends, identifying factors that may influence maternal deaths or specific causes of death with the goal of reducing cases of maternal deaths. One of those studies by Nyaboga (2009) described the trends, magnitude, contributing factors and causes of maternal mortality in Kenya's national referral hospital, KNH. His research identified Age, Parity, Place of Delivery, Contraceptive use, ANC attendance, and Socioeconomic status as the influential factors for maternal mortality in the national referral hospital. The specific maternal death causes were outlined in his paper as HIV, Abortion complications, Eclampsia, Sepsis and Postpartum haemorrhage. Recommendations based on the research were in line with implementing BEmONC or CEmONC interventions in all healthcare facilities. Emergency Obstetric and Newborn Care (EmONC) describes a set of interventions that treat leading causes of perinatal and maternal mortality (Tecla et al., 2017). Basic EmONC (BEmONC) services include: administration 4

of antibiotics to counter sepsis, anticonvulsants for hypertension disorders, uterotonic for postpartum haemorrhage, Manual placenta removal, Assisted vaginal delivery, retained products of conception extraction and neonatal resuscitation. Comprehensive EmONC (CEmONC) services include all BEmONC components in addition to Caesarean section surgical capability and blood transfusions (Odhiambo and Kinoti, 2019). Research into maternal mortality has also involved developing and comparing models to determine the best-suited model for estimating and predicting maternal deaths. 1.3 Statement of the Problem In cases of rare events, zero inflation in count data may extend significantly beyond 51% of the total observations. This results in a case of class imbalance where the proportion of zero counts is greater than that of positive counts. In such cases, the standard logistic regression used to estimate the probability of non-zero counts for Hurdle models may not be optimal. Hence the need for models with better predictive ability. The current research incorporates Rare-Event Weighted Logistic Regression (REWLR) in our Hurdle models' binary component to improve the

model performance. 1.4 Objective of the study 1.4.1 General Objective The main objective of this study is to assess the performance of hurdle models incorporating rare-event weighted logistic regression and their applications to maternal mortality data. 1.4.2 Specific Objectives • To assess the performances of Hurdle-REWLR models for various proportions of zero-inflation. 5

• To investigate the performance of Hurdle-REWLR models with the application of maternal mortality data. 1.4.3 Research Questions i. Does incorporating REWLR in Hurdle models improve the performance of the models? ii. Does the degree of class imbalance between zero and non-zero classes influence model performance? iii. Which obstetric factors influence maternal deaths in Nairobi? 1.5 Justification The goal of any statistical study is often to predict an outcome of interest or provide inference about the same. Practical inferences about a population require the sample statistics and estimates to be generalizable to a broader population, hence the need for models best suited for the population from which a sample is drawn. Statisticians and researchers have adjusted distributions and modified models to find the distribution that best explains their data and models that offer the best fit. Subsequently, count data models were extended to accommodate the excess zeros that arise in situations where the event of interest is rare. Formulations of these zero-modified count models involved nesting logistic regression to estimate the probabilities of a structural zero or a zero count for zero-inflated and hurdle models, respectively. GLM literature suggests the logit function is symmetric, so the response curve approaches zero and one at the same rate. This feature makes logistic regression inefficient due to the risk of underestimating the probability of a rare event. Maternal deaths in Kenya are already under-reported due to the inefficient data collection systems and the high number of women who give birth outside healthcare facilities. Employ- ing models that may underestimate the already under-reported maternal death cases would 6

harm maternal health policies, as it would offer a false sign of relief. Hence the need for count estimation models adjusted to deal with rare events. 1.6 Significance of the Study This study proposes a model extension to improve the predictive ability and thus the perfor- mance of zero altered models. The model, applied to maternal mortality data, would ensure accurate estimation of the death cases and eliminate any false relief caused by over-estimation of the zero-death cases. 7

Chapter 2 Literature review 2.1 Introduction Variations of two statistical approaches have been used in modelling count data characterized by excess zeros in the outcome variable. This chapter provides an overview of these statistical approaches, the concepts behind them, and their applications in past research. We also review the logistic regression model and its limitations in estimating probabilities of rare events. 2.2 Models 2.2.1 Hurdle Models Mullahy (1986) developed the Poisson hurdle model to handle zero-inflated count data in cases where sampling and structural zeros were not distinguishable. His proposed 2-part model analyzed zero counts separately from positive counts. He applied the model to study peoples' daily consumption of beverages based on certain socio-demographic factors. The study results revealed that the hurdle model allowed for more flexibility in model specification than the basic model. The models proposed in this research could also account for both under-dispersion and over-dispersion. King (1989) separately developed hurdle models in an application to a political science study. His research aimed to develop an approach that models the onset of war separately from its escalation. The model was developed following Mullahy (1986) theory of data generation mechanism, where certain factors determine whether a country goes or does not go into war, 8

and once a country crosses the hurdle, factors such as alliances will determine the number of wars with which the country will be involved. This model proved to be an improvement of Mullahy's hurdle model. Rose et al. (2006) applied the Poisson and Negative Binomial hurdle models in estimating the number of adverse events reported for each subject following a vaccination injection. They assumed a single source of zeros (sampling) because their study design made it such that all subjects were at risk of experiencing at least one adverse event. This assumption favoured hurdle over zero-inflated models. The goodness of fit statistics for ZINB and NBH were indistinguishable. A quasi-experimental study by Chaudhari et al. (2012) utilized hurdle models in the estimation of the total dental utilization using data obtained from dental claims. The model allowed them to decompose the hurdle likelihood function to allow for individual estimation of the probability of dental care, type of dental care and level of utilization. The likelihood decomposing feature gave the Negative Binomial Hurdle model the edge over the other models. Hurdle models have also been widely applied in mortality estimation studies. Fenta and Fenta (2020) determined NBH model over ZIP, ZINB and PH for estimating risk factors of child mortality in Ethiopia. In a different study, NBH emerged to be the best statistical model for estimating predictors of under-five mortality in Ethiopia. The hurdle model was also selected as the best fitting model in the Mamun (2014) study to estimate under-five deaths. Both pieces of research involved comparing the Hurdle models to the zero-inflated models and, in some cases, the standard count models. Besides the application of hurdle models in the various research fields, researchers have also developed modified versions of the models to provide better for their data. One of the hurdle model extensions was by Min and Agresti (2005), to accommodate correlated data. The authors modified the Hurdle model to include a random effect for their research to estimate the number of episodes of side effects recorded at

each visit and compared two treatments. Fitting the random effects hurdle models proved less complex than fitting a zero-inflated random-effects model. In addition, the model provided more straightforward interpretations. 9

A two-part model meant the two parts could be fitted and estimated separately, hence reducing complexity. 2.2.2 Zero-inflated Models Lambert (1992) introduced a zero-inflated Poisson model in an application to model defects in manufacturing. Her study aimed to propose a new model, ZIP, for handling zero inflation in count data, where both the binary and count components of the model could depend on covariates. This was motivated by the poor predictive ability of the Poisson, Generalized Poisson and Negative Binomial model in estimating defects for a count response variable which consisted of 81% zero defects. Formulation of the proposed model involved multiply- ing the probability of a structural zero by probability for the counts. The analysis showed ZIP to outperform both Poisson and NB models. It was also the preferred model for the engineers in the manufacturing study that motivated Lambert's research; separating zeros into structural and sampling was logical given the research design. A study limitation noted by Lambert (1992) was that the model was not as easy to fit because it was not known which of the zeros were perfect (structural). Greene (1994) developed ZINB by extending ZIP; they also explored performance differences between ZINB, Poisson, Negative Binomial, and ZIP using credit card approval data. The outcome of interest was the number of major derogatory reports; out of a sample of 1023, 89.4% had zero reports. The study's objective was to compare the models' performances for zero-inflated data, which showed evidence of dispersion due to heterogeneity. Study results indicated that the NB model outperformed the Poisson model, ZIP model also outperformed the Poisson model but had a slightly worse fit than NB. This was an implication that the negative binomial ZIP model was necessary to accommodate two sources of overdispersion. Since their introduction, zero-inflated models have been continually modified and applied in many fields. Aryuyuen et al. (2014) developed the ZINB - Generalized Exponential distribution to provide a better fit for heavy-tailed over-dispersed zero-inflated data. He assessed the new model's performance compared to ZIP and ZINB, applied on simulated data 10

and actual data for hospital stays by senior US residents. The resultant model proved to be a better fit than the ZIP and ZINB distributions. Kibika (2020) developed the ZINB - Shanker distribution by combining Zero-inflated Negative Binomial and Shanker distribution. The goal of developing the new model was to allow greater flexibility by increasing randomness in the ZINB probability distribution function. The model was used to model HIV cases among infants exposed to HIV through breastfeeding, etc. Overall fit tests revealed ZINB to offer the best fit. ZINB-Shanker distribution proved competitive for larger sample sizes. Diop et al. (2021) proposed a modification to ZIP which involved the use of the quantile function of the Generalized Extreme Value (GEV) distribution as a link function for zero- inflated data with rare events. The approach was proposed to curb the drawbacks of logistic regression when dealing with imbalanced data, where the probability of a rare event is underestimated. Ali (2020) did a comparison study between ZIP, ZIP-GEV, ZIP-clog log and ZIP-probit. The analysis results revealed the Zero-inflated Poisson with a GEV link function to be the best performing model. Zero-inflated models have also been widely utilized in maternal health studies. Arefaynie et al. (2022) used ZIP regression in a study to determine the number of antenatal care and associated factors in Ethiopia. Fitriani et al. (2019) and Loquiha et al. (2013) used ZINB regression to model maternal mortality in Malang and Mozambique. 2.2.3 Logistic Regression Model The logistic regression model is the most commonly used statistical model for classifying binary data. It estimates the probability of a binary outcome, independently or dependent on a set of predictors. It has been broadly utilized in many fields including healthcare Yego et al. (2014), epidemiology Tolles and Meurer (2016), education Mason et al. (2018), economics Jabeur (2017), etc. Hurdle models also employ logistic regression to assign the probability that governs whether a count takes on a zero or a positive value. Models based on various link functions, including logit, probit, log-log, clog-log, have been proposed for the binary response estimation, but logistic regression remains the most popular Desjardins (2013). Its 11

convenient interpretation and implementation make it an ideal method for modelling binary response variable Ali (2020). Logistic regression, however, has drawbacks when applied to the classification of imbalanced binary events. Research has highlighted this limiting feature of the logistic regression and proposed solutions to account for class imbalance in binary data. Rahim et al. (2019) applied Synthetic Minority Over-sampling Technique (SMOTE) sampling to Logistic regression, intending to improve its classification accuracy in bankruptcy detection. The study results showed that the SMOTE logistic regression outperformed the standard logistic regression with imbalanced data. In his study, Wang (2020) investigated the sampling-based interventions for imbalanced binary classes. The two approaches considered were undersampling the majority class or oversampling the minority class. The results reveal that undersampling the majority class did not always penalize the estimations, and oversampling the minority class did not consistently reduce estimation efficiency. King and Zeng (2001) proposed a different approach for dealing with imbalanced binary classes, which involved applying weights and prior correction in the estimation of probabili- ties and regression coefficients. Their study results showed that the models implementing the recommended corrections outperformed the existing standard methods. However, the study's

recommended approach turned out to be over-correcting bias in Maximum Likelihood Estimations. Maalouf and Siddiqi (2014) developed the Rare Event Weighted Logistic Regression (REWLR) for classifying large imbalanced data with a rare event. The proposed algorithm applied weights and regularization terms to achieve better predictive accuracy, counter over-fitting and reduce bias and variance. Weighted Logistic Regression Approach for Rare Events was used in a study by Zare et al. (2013) to determine risk factors for female breast cancer where the choice of REWLR over Logistic Regression was influenced by the rarity of events of interest in their research. In a comparison study, REWLR proved to perform better than other algorithms, including the Truncated-Regularized Iteratively Re-weighted Least Squares algorithm and Truncated-regularized Prior Correction Maalouf et al. (2018). The authors rec- 12

ommended the application of appropriate corrections and adjustments to Logistic Regression when data is imbalanced. 2.3 Our Research From the reviewed research, the question that remains to be explored is how a modified logistic regression would affect the performance of hurdle models. To the best of the author's knowledge, no study has attempted to improve the predictive performance of hurdle models' binary component by accommodating class imbalance. Based on the problem we have described so far, the objective of the current research is to develop and assess the performance of hurdle models nested with rare-event weighted logistic regression when applied to maternal mortality data. 2.4 Conclusion This chapter highlighted past research involving zero-inflated data, which elucidated the decision behind the choice model for the current research and the need for more robust classification models. The researchers concurred that decision about whether to apply hurdle models or zero-inflated models in modelling count data with excessive zeros should be guided by the beliefs about the data-generating mechanism of the zeros Min and Agresti (2005); Miller (2007); Desjardins (2013). Rose et al. (2006) proposed hurdle models be considered if there is a chance of zero deflation in the data. The previous research work also supported the need for better-performing classification techniques in the hurdle model's binary component for data with rare events. 13

Chapter 3 Methodology 3.1 Introduction This chapter details the development of the modified hurdle model which is achieved by incorporating REWLR for binary component estimations. The model is applied to simulated data to assess model performance with various proportions of zero counts and a real dataset to assess factors that influence maternal mortality in Nairobi. 3.2 Research Design This study employs a descriptive research design to assess the effects of select obstetric and demographic factors on the number of maternal deaths in Nairobi. It uses a secondary source of data from publicly available information. The maternal mortality data was pulled from JPHES, a portal of District Health Information Software (DHIS2), that streamlines health data reporting. The data contains the number of maternal deaths and other obstetric and demographic factors recorded in MNCH facilities in Nairobi between October 2021 and January 2022. The study also introduces a modified Hurdle model that is based on Mullahy (1986)'s Hurdle models, and Maalouf and Siddiqi (2014)'s Rare-Events Weighted Logistic Regression model. The general structure of a hurdle model as proposed by Mullahy (1986) is given by: 14

P(Y i = y i) = ? ? ? ? ? ? ? ? ? (1– p i) y i = 0 (p i) p(y i ;λ i) 1–p(y i ;λ i |y i =0) y i < 0; (3.1) This is the two-part model which uses a logistic regression model to estimate p i and a zero-truncated count model for the estimation of the zero-truncated count model. logit(p i) = x 1i β 1 and log(λ i) = x 2i β 2 (3.2) We obtain the zero-truncated model by excluding the probability that y i = 0 from the count distribution, which is achieved by dividing the probability mass function of the count model by 1 minus the probability of a zero count i.e., p(y i ;λ i) 1– p(y i ;λ i |y i = 0) . (3.3) The probability p i of a positive count in hurdle models is typically modeled using a logistic regression model, presented as: p i = e Xβ 1+e Xβ = 1 1+e –Xβ (3.4) where β is the vector of coefficients, and X is a vector of predictors. We use MLE to find the parameter estimates of the hurdle model; this is obtained by separately maximizing the log-likelihood functions of the binary and the zero-truncated distributions. The log-likelihood function of the Hurdle model using a logistic regression model for the binary component is given by: 15

ℓ(β 1 ,β 2) = ln n ∏ i=1 (p i) y i (1– p i) (1–y i) × p(y i ;μ i) 1– p(y i ;μ i |y i = 0) = n ∑ i=1 y i ln p i +(1–y i)ln(1– p i)+ln p(y i ;μ i) 1– p(y i ;μ i |y i = 0) = n ∑ i=1 ln1– p i + n ∑ i=1 y i ln p i 1 – p i + n ∑ i=1 ln p(y i ;μ i) 1– p(y i ;μ i |y i = 0) = n ∑ i=1 ln1– p i + n ∑ i=1 y i (xβ)+ n ∑ i=1 ln p(y i ;μ i) 1– p(y i ;μ i |y i = 0) = n ∑ i=1 – ln1+e xβ + n ∑ i=1 y i (xβ)+ n ∑ i=1 ln p(y i ;μ i) 1– p(y i ;μ i |y i = 0) (3.5) The maximum likelihood estimate for the binary component is the mean of the y variable from the n draws, i.e. p i = 1 n n ∑ i=1 y i (3.6) There is no closed form solution to obtain the maximum likelihood estimates for the zero- truncated component. MLE are therefore obtained by using IRLS method of Newton-Raphson algorithm to solve the score equations. 3.3 Hurdle-REWLR Model The proposed model overcomes logistic regression, and hence hurdle models', weakness in the case of imbalanced data by adopting regularization, weighting, and bias correction on logistic regression's log likelihood function. The log-likelihood function of the REWLR model introduced by Maalouf and Siddiqi (2014) is given by: ℓ(β) = ln n ∏ i=1 (p i) w 1 y i (1– p i) w 0 (1–y i) – λ 2 ‖β‖ 2 = –w 0 n ∑ i=1 ln1+e xβ +(w 1 – w 0) n ∑ i=1 y i xβ – λ 2 ‖β‖ 2 (3.7) 16

where: i. w s are the weights applied to counter imbalance in the data, which penalize the misclassification made by setting a higher class weight to the minority class (positive counts) while reducing weight for the majority class (zeros). w 1 = τ ¯y ; w 0 = (1–τ) (1– ¯y) (3.8) (a) τ is the proportion of (non-zero) events in the population (b) ¯y is the proportion of (non-zero) events in the sample; ii. λ 2 ‖β‖ 2 is a regularization term that introduces a penalty for large values of β hence avoids overfitting. The log-likelihood of the binary logistic component and the zero-truncated Poisson or NB component are estimated separately and then combined for model fit assessments. Neither the binary nor the zero-truncated components have closed form solutions for max- imum likelihood estimation. MLEs are thus obtained by using IRLS method of Newton- Raphson algorithm to solve REWLR and zero truncated Poisson or zero truncated NB score equations. 3.3.1 Poisson Hurdle-REWLR Model The Probability Mass Function of the Poisson Hurdle-REWLR Model is given by: P(Y i = y i) = ? ? ? (1– p i) y i = 0 (p i) e –λ i λ y i i (1–e –λ i)y i ! y i < 0 (3.9) 17

Model estimates are obtained by maximizing the MLE function of the Poisson Hurdle- REWLR distribution: ℓ(β 1 ,β 2) = ln n ∏ i=1 (p i) w 1 y i (1– p i) w 0 (1–y i) – λ 2 ‖β‖ 2 + e –λ i λ y i i 1 –e –λ i y i ! ! = –w 0 n ∑ i=1 ln1+e xβ +(w 1 – w 0) n ∑ i=1 y i xβ – λ 2 ‖β‖ 2 + n ∑ i=1 –λ +y i lnλ i – ln(1–e –λ i)–ln(y i !) (3.10) Since both components have no closed form solutions, MLEs are thus obtained by using IRLS method of Newton-Raphson algorithm. 3.3.2 Negative Binomial Hurdle-REWLR Model The Probability Mass Function of the Negative Binomial Hurdle-REWLR Model is given by: P(Y i = y i) = ? ? ? ? ? (1– p i) y i = 0 p i 1– k μ i +k k Γ(y i +k) y i !Γ(k) μ i μ i +k y i k μ i +k k y i < 0; (3.11) where the dispersion parameter k is given by 1 α . Model estimates are obtained by maximizing the MLE function of the Negative Binomial Hurdle-REWLR distribution: 18

ℓ(β 1 ,β 2) = ln n ∏ i=1 (p i) w 1 y i (1– p i) w 0 (1–y i) – λ 2 ‖β‖ 2 + 1 1– 1 1+αμ i α –1 Γ y i + α –1 y i !Γ(α –1) αμ i 1 +αμ i y i 1 1 +αμ i α –1 = –w 0 n ∑ i=1 ln1+e xβ +(w 1 – w 0) n ∑ i=1 y i xβ – λ 2 ‖β‖ 2 + n ∑ i=1 lnΓ y i + α –1 –lnΓ(α –1)–lny i !– y i + a –1 ln(1+αμ i) +y i lnαμ i – ln h 1–(1+αμ i) –a –1 i (3.12) Since both components have no closed form solutions, MLEs are thus obtained by using IRLS method of Newton-Raphson algorithm. 3.4 Simulations Data is simulated under PH and NBH distributions. Simulations are performed using a combination of sample size and proportion of zeros observed. The following experimental conditions are applied for the simulation study: • Zero inflation - 50%, 60%, 75%, and 90%. • Sample size - 200, 1000 • The dispersion parameter value is set at 3 Both the binary and zero-truncated components incorporate a count predictor variable x simulated from a Poisson distribution x ~ Poisson(λ). The simulated covariate mimics the type of covariates for the real maternal mortality data, e.g., number of pregnant women attending at least 4 ANC visits, number of assisted Vaginal deliveries, etc. Data is generated in R using the rpois(), rbinom(), rhpois(), rhnbinom() functions from the stats, actuar and countreg packages. 19

Analyses for the simulated and real data is performed in R using hurdle() from the pscl package, glm from stats package and vglm() from VGAM package. 3.5 Maternal Mortality data The data contains information on obstetric outcomes, including maternal deaths for public and private facilities in Nairobi that offer MNCH services. It covers the duration of October 2021 to January 2022, containing records for 222 MNCH facilities. Data is available for at least one facility in all the 17 sub-counties in Nairobi: Westlands, Dagoretti North, Dagoretti South, Langata, Kibra, Roysambu, Kasarani, Ruaraka, Embakasi South, Embakasi North, Embakasi Central, Embakasi East, Embakasi West, Makadara, Kamukunji, Starehe and Mathare. Nairobi is a cosmopolitan county and Kenya's capital city hence the data offers a good representation of the Kenyan population. Table 3.1: Variable Definition Variable Description MaternalDeaths Number of Maternal deaths in MNCH Nairobi facilities between October 2021 - January 2022 AssistedDeliveries Number of women who had assisted vaginal deliveries BreechDelivery Number of women who had breech delivery CS Number of women who gave birth by caesarian sections LiveBirths Number of live births EarlyTeenPreg Number of adolescents (10-14 years) pregnant at 1st ANC visit LateTeenPreg Number of adolescents (15-19 years) pregnant at 1st ANC visit NormalDeliveries Number of women who had normal deliveries ANC4Visits Number of women who have attended at least 4 ANC visits Uterotonics3stg Number of women giving birth who received uterotonics in the third stage of (or immediately after birth) Continued on next page 20

Table 3.1 – continued from previous page Variable Definition Carbatosin Number of Mothers given uterotonics within 1 minute (Carbatosin) Oxytocin Number of Mothers given uterotonics within 1 minute (Oxytocin) AntHaemorrage Number of women who had Ante partum Haemorrage PostHaemorrage Number of women who had Post Partum Haemorrage ObstructedLabour Number of women who had Obstructed Labour Eclampsia Number of women who had Eclampsia RupturedUterus Number of women who had Ruptured Uterus Sepsis Number of women who had sepsis FGMComplicatons Number of Mothers with delivery complications associated with FGM Stillbirth Number of women who had Macerated stillbirth The response variable for this research is the number of maternal deaths reported in MNCH Nairobi facilities between October 2021 - January 2022. The predictors consist of obstetric factors, maternal complications, and demographic factors that previous literature suggested influence maternal deaths. 3.6 Model selection

The study uses Akaike information criterion (AIC) to compare the model fit between the modified hurdle models and the standard Hurdle

models. AIC is computed as: AIC = –2log(L)+2K; where L is the likelihood, and K is the number of parameters in the model.

AIC evaluates how well a model fits the data from which it was generated. The best-fitting model yields the lowest AIC values. 21

Chapter 4 Results and Interpretation 4.1 Simulation In this section, we present the theoretical results of the study models through simulation analyses performed on the Poisson Hurdle-REWLR, NB Hurdle-REWLR and Poisson Hurdle, and NB Hurdle models. We simulated data from four different distributions: Poisson Hurdle, Poisson Hurdle-REWLR, Negative Binomial Hurdle-REWLR, and Negative Binomial Hurdle-REWLR. For each model, we generated 200 and 1000 random samples with varying zero proportions from the true model, and then all the models were fitted to the simulated datasets. A predictor variable was simulated from the Poisson distribution, with a constant mean of 3 across all simulation conditions. The dispersion parameter k was used with a pre-stipulated value of 3. We applied AIC to compare the true and misspecified models in terms of the percentage of the differences in the AICs for the misspecified models and true model (%ΔAIC). Where ΔAIC = AIC(Misspecified model)–AIC(True model). First, we evaluated the performance of HNB, HNB-REWLR and HP-REWLR models when the data are simulated from a Hurdle Poisson model. The Hurdle-REWLR models reported the lowest AIC values among all the other models. The percentage differences in the AICs between the misspecified models and the Poisson Hurdle models increased as the proportions of zero in the data increased. This was the trend for both the small (n=200) and large (n=1000) sample sizes. 22

Table 4.1: AIC (Percentage Change in AIC) for Misspecified and Actual Models Reference n Zeros PH PHRE NBH NBHRE PH 200 0.50 418 402 (-3.98) 420 (0.43) 403 (-3.63) PH 200 0.60 512 479 (-6.88) 514 (0.3) 480 (-6.64) PH 200 0.75 519 478 (-8.68) 521 (0.37) 480 (-8.16) PH 200 0.90 582 527 (-10.43) 584 (0.42) 529 (-9.92) PH 1000 0.50 2252 2136 (-5.44) 2254 (0.09) 2140 (-5.23) PH 1000 0.60 2473 2302 (-7.43) 2474 (0.04) 2303 (-7.36) PH 1000 0.75 2730 2513 (-8.65) 2732 (0.07) 2515 (-8.56) PH 1000 0.90 2855 2592 (-10.17) 2857 (0.07) 2593 (-10.1) PHRE 200 0.50 578 (9.53) 523 580 (9.84) 525 (0.36) PHRE 200 0.60 599 (9.13) 544 600 (9.29) 545 (0.26) PHRE 200 0.75 562 (10.03) 506 564 (10.35) 509 (0.67) PHRE 200 0.90 587 (9.61) 531 589 (9.92) 534 (0.5) PHRE 1000 0.50 2807 (9.61) 2537 2809 (9.68) 2542 (0.18) PHRE 1000 0.60 2952 (9.4) 2675 2954 (9.44) 2676 (0.06) PHRE 1000 0.75 2976 (9.32) 2699 2978 (9.38) 2701 (0.07) PHRE 1000 0.90 2921 (9.5) 2644 2923 (9.54) 2645 (0.04) NBH 200 0.50 374 (7.56) 380 (8.98) 346 352 (1.61) NBH 200 0.60 499 (5.84) 488 (3.76) 470 460 (-2.27) NBH 200 0.75 562 (7.87) 540 (4.11) 518 496 (-4.47) NBH 200 0.90 649 (6.26) 617 (1.39) 608 576 (-5.57) NBH 1000 0.50 2054 (5.9) 2069 (6.56) 1933 1948 (0.76) NBH 1000 0.60 2161 (5.13) 2144 (4.37) 2050 2033 (-0.85) NBH 1000 0.75 2864 (8.81) 2767 (5.61) 2612 2514 (-3.88) NBH 1000 0.90 3128 (9.65) 2984 (5.29) 2826 2682 (-5.38) NBHRE 200 0.50 614 (12.66) 588 (8.78) 562 (4.58) 536 NBHRE 200 0.60 685 (11.4) 649 (6.45) 643 (5.58) 607 NBHRE 200 0.75 679 (14.28) 640 (9.12) 621 (6.26) 582 NBHRE 200 0.90 680 (12.82) 645 (8.02) 629 (5.69) 593 NBHRE 1000 0.50 3337 (12.58) 3180 (8.27) 3074 (5.11) 2917 NBHRE 1000 0.60 3044 (11.13) 2896 (6.59) 2853 (5.18) 2705 NBHRE 1000 0.75 3426 (14.11) 3245 (9.3) 3125 (5.81) 2943 NBHRE 1000 0.90 3279 (14.71) 3111 (10.08) 2966 (5.69) 2797 23

Analysis on NB Hurdle data resulted in NB Hurdle REWLR outperforming NB Hurdle at 60%, 75% and 90% zero inflation for both small and large sample sizes. NB Hurdle outperformed the other models only when the data had a 50% zero inflation. Data generated by the study model distributions, Poisson Hurdle-REWLR and NB Hurdle- REWLR, performed best on the true models, based on AIC statistics. In the Poisson Hurdle REWLR generated data, the lowest AIC values were recorded by the same model, for all the simulation conditions. In a small sample size, the model performed best in data with 75% zero inflation while in large sample size, Poisson Hurdle-REWLR performance is best in 60% inflation data. The least percentage change in AIC was achieved by the NB Hurdle REWLR model. Models fit on the NB Hurdle-REWLR simulated data achieved the lowest AIC. The least percentage change in AIC was achieved by the NB Hurdle model. Overall, NB Hurdle REWLR outperformed the Poisson Hurdle, Poisson Hurdle REWLR and NB Hurdle models. Plots of the resulting AIC values are presented in figure 4.1 and figure 4.2 for Poisson Hurdle simulated data, figure 4.3 figure 4.4 for Poisson Hurdle REWLR simulated data, figure 4.5 and figure 4.6 for NB Hurdle simulated data, figure 4.7 and figure 4.7 for NB Hurdle REWLR simulated data. Figure 4.1: AICs from Models fit on Poisson Hurdle simulated data, n = 200 24

Figure 4.2: AICs from Models fit on Poisson Hurdle simulated data, n = 1000 Figure 4.3: AICs from Models fit on Poisson Hurdle-RE simulated data, n = 200 25

Figure 4.4: AICs from Models fit on Poisson Hurdle-RE simulated data, n = 1000 Figure 4.5: AICs from Models fit on NB Hurdle simulated data, n = 200 26

Figure 4.6: AICs from Models fit on NB Hurdle simulated data, n = 1000 Figure 4.7: AICs from Models fit on NB Hurdle-RE simulated data, n = 200 27

Figure 4.8: AICs from Models fit on NB Hurdle-RE simulated data, n = 1000 4.2 Application to Maternal Deaths Data 4.2.1 Descriptive Statistics The study sample data reported 293 maternal deaths of the 53792 recorded live births. The sample variance of 3.758 exceeds the sample mean of 1.32, indicating overdispersed data. The data also exhibits zero inflation as 61.71% of the dependent variable counts are zero. Table 4.2 below exhibits the average counts for some of the obstetric factors used as covariates in this study. The facilities which reported maternal deaths had higher average counts of all the conditions. For instance, the number of Mothers given uterotonics was higher in the group that experienced maternal deaths. Stillbirth occurrence was also primarily associated with maternal death. Correlation analysis between the maternal deaths and the predictors revealed that Maternal deaths was highly correlated with BreechDelivery (r = 0.7342), Uterotonics3stg (r = 0.9615), Oxytocin (r = 0.9615), Carbatosin (r = 0.9615), AntHaemorrage (r = 0.9743), Eclampsia (r = 0.9742), ObstructedLabour (r = 0.9741), PostHaemorrage (r = 0.9741), 28

Figure 4.9: Maternal Death Counts FGMComplicatons (r = 0.9742), RupturedUterus (r = 0.9744), Sepsis (r = 0.9740), and Stillbirth (r = 0.9764). 4.2.2 Maternal Death Models Prior to formulating the count models, we compute the weights for the binary Hurdle-REWLR models. The weights penalize the misclassification made by setting a higher class weight to the positive counts while reducing weight for the zero counts. The weights are calculated as outlined by Maalouf and Siddiqi (2014): 29

Table 4.2: Average Count of Obstetric Conditions reported in facilities with and without reported maternal deaths Factor No Maternal Deaths Maternal Deaths BreechDelivery 0.3 2.7 CS 15.0 174.7 LiveBirths 44.7 560.8 EarlyTeenPreg 2.3 22.9 LateTeenPreg 12.7 60.0 NormalDeliveries 36.5 404.2 ANC4Visits 42.7 272.8 Uterotonics3stg 96.2 1139.7 Carbatosin 9.5 112.9 Oxytocin 75.0 889.0 AntHaemorrage 7.5 94.8 Eclampsia 2.2 28.5 ObstructedLabour 0.7 10.4 PostHaemorrage 3.4 42.4 FGMComplicatons 1.5 18.8 RupturedUterus 1.7 21.8 Sepsis 0.4 6.1 Stillbirth 0.0 0.6 w 1 = τ ¯y ; w 0 = (1–τ) (1– ¯y) (4.1) We have 293 deaths for the 53792 live births from our sample data. The latest report by the Kenya Ministry of Health on Health and Health-related SDGs revealed the latest deaths per live birth ratio reported in Nairobi as 97. ¯ Y = 293 53792 = 0.0054 τ = 97 100000 = 0.00097 w 1 = 0.00097 0.0054 = 0.1796 w 0 = (1–0.00097) (1–0054) = 1.0045 30

We use correlation analysis to select the predictors to use for the analysis. Predictors with high correlations are more linearly dependent and thus have the same effect on the dependent variables. The factors that influence observing a maternal death in the facility are attending at least 4 ANC visits, antepartum haemorrhage, and receiving uterotonics during or immediately after birth. Specific effects of these factors on the various models are presented in Table 4.3. Upon observing a maternal death, the determinants of the actual number of maternal deaths that a facility could report are the occurrence of macerated stillbirth, attending of at least 4 ANC visits, adolescent pregnancies, antepartum hemorrhage, breech deliveries, postpartum hemorrhage, receiving Carbatosin and giving birth by cesarean section. The coefficients of the count model component is presented in Table 4.4. Table 4.3: Binary Component Coefficients PH.BINARY NBH.BINARY PHRE.BINARY NBHRE.BINARY (Intercept) -3.9561920 -3.9561920 8.2592952 8.2591187 Stillbirth 21.4971965 21.4971965 -0.0242153 -0.0529712 ANC4Visits -0.0021931 -0.0021931 0.0260967 0.0261088 LateTeenPreg -0.0072837 -0.0072837 0.0091975 0.0091929 AntHaemorrage 1.2112676 1.2112676 -4.6352784 -4.6355084 PostHaemorrage -0.1450389 -0.1450389 0.2713996 0.2708596 BreechDelivery 0.1438486 0.1438486 -0.0180133 -0.0183037 Carbatosin -0.0103415 -0.0103415 -0.1092431 -0.1002919 Uterotonics3stg -0.0785952 -0.0785952 0.3119054 0.3110578 Table 4.4: Count Component Coefficients PH.COUNT NBH.COUNT PHRE.COUNT NBHRE.COUNT (Intercept) -0.0475 -0.0475 0.6303 -0.0444 Stillbirth 0.9126 0.9127 -0.0051 0.9126 ANC4Visits 0.0011 0.0011 0.0386 0.0011 LateTeenPreg 0.0023 0.0023 0.0006 0.0023 AntHaemorrage -0.0307 -0.0307 0.0018 -0.0304 PostHaemorrage 0.0302 0.0302 NA 0.0305 BreechDelivery 0.0317 0.0317 NA 0.0318 Carbatosin 0.2087 0.2087 NA 0.2131 Uterotonics3stg -0.0196 -0.01961 NA -0.0201 31

Table 4.5 shows the resulting AICs following the fit of Poisson, Negative Binomial, Poisson Hurdle, NB Hurdle, Poisson Hurdle REWLR, NB Hurdle REWLR models to the maternal mortality data. NB Hurdle REWLR produced the lowest AIC, indicating a better fit than the other count models. Table 4.5: AIC for Maternal Mortality Models Model AIC Poisson 469.6684 PH 335.9051 PH-RE 370.6200 NB 473.5588 NB-H 337.9054 NBH-RE 284.1434 It was also of interest to the

study how the Hurdle-REWLR predicted zero counts compared to their counterpart standard Hurdle models. From Table 4.6, we observe that Poisson Hurdle and NB Hurdle models accurately predicted the observed number of zero counts in the sample data. The predicted zero counts from the sample data were slightly less than that observed in the sample data. Table 4.6: Observed and Expected Zero Counts Oberved Poisson PH PH-RE NB NB-H NBH-RE 137 122 137 102 126 137 102 32

Chapter 5 Discussion, Conclusion and Recommendation 5.1 Introduction This section presents an interpretation of the study's research findings in relation to findings made by previous researchers on the same topic. We further draw conclusions and make recommendations based on our study outputs. 5.2 Discussion Cases of rare events in count data where the proportion of zero counts is significantly less than that of the natural numbers have been shown to influence binary estimations in zero-inflated count models. Theoretically, more extreme rare events are expected to impose extreme bias towards the majority group, i.e., zero counts. Because of this hypothesis, the current study conducted extensive simulation and analysis with varying proportions of zeros and sample sizes to evaluate the performance of the Hurdle-REWLR models in the various simulation conditions. The study also evaluated the performance of the study models alongside the standard hurdle models when fit on maternal mortality data to determine the factors which influence maternal mortality in Nairobi, Kenya. The analysis to determine factors which influence maternal deaths in Nairobi resulted in NB Hurdle-REWLR outperforming the other models in terms of the Akaike Information Criterion. The Hurdle-REWLR models adjusted for the population estimates by introducing 33

weights and regularizing the coefficients. The predicted zero counts from the sample data emerged to be slightly less than observed. The number estimated by the Hurdle-REWLR models could be expected from a sample that accurately represents the Nairobi population. The introduction of the weights makes the Hurdle-REWLR models ideal for estimations and inference. In both the Hurdle and hurdle-REWLR models, specific demographic and obstetric factors significantly affected the response variable. Age, described by the number of pregnant ado- lescents, and attendance of at least 4 ANC visits are some of the demographic factors shown by past literature to influence maternal deaths. In addition, childbirth-related conditions of postpartum haemorrhage, treated by uterotonics, antepartum haemorrhage, breech delivery and Macerated stillbirth were also discovered to influence the number of maternal deaths. Haemorrhage is among the obstetric factors which were highlighted by (Organization et al., 2019) to be some of the causes of maternal death globally. These findings were also in line with Nyaboga (2009) research which outlined the influential factors and causes of maternal mortality in Kenya's national referral hospital, KNH. Some of the factors identified in their research which the current study has outlined, include age, ANC attendance, and Postpartum haemorrhage. Simulation analysis findings revealed NB Hurdle-REWLR to produce the lowest AIC value compared to the other models. The percentage difference in AICs between NB Hurdle- REWLR and the other misspecified models increased as the zeros in the data increased. The Poisson Hurdle-REWLR model outperformed the NB Hurdle in Poisson Hurdle REWLR simulated data but was inferior in NB Hurdle simulated data. It could not account for the extra dispersion introduced in the NB simulated data. The selection of NB Hurdle REWLR as the ideal model over the standard NB Hurdle model was influenced by the degree of zero inflation in the simulation analysis. The two models gave almost similar results when the proportions of zeros and non-zeros in the scenarios where the data were not significantly different. For instance, in the NB Hurdle simulated data, the model performed better than the NB Hurdle REWLR model at 50% zero inflation but was outperformed for the subsequent degrees of zero inflation of 60%, 75% and 90%. 34

This outcome conformed to the basic concept of the Hurdle-REWLR model. As outlined by Maalouf and Siddiqi (2014), REWLR is modified from logistic regression with the aim of unbiased prediction in rare events with imbalanced data. If the proportions of zero and non-zero counts are balanced, REWLR is not expected to outperform logistic regression. Despite their foundations on similar concepts, the performance of the Poisson Hurdle REWLR was inferior to that of the NB Hurdle REWLR model. In the simulation analysis, Poisson Hurdle REWLR outperformed its counterpart in data generated by its distribution, and the Poisson Hurdle simulated data only by small units of percentage change in AIC. In the other simulation scenarios, the NB Hurdle REWLR model claimed superiority by quite huge margins of the percentage change in AIC. In addition, NB Hurdle REWLR was the best performing model for the fit on maternal mortality data. Such results have been witnessed in various performance comparison studies including Fenta et al. (2020) and Mamun (2014). It is common for the NB model to outperform its Poisson counterpart when there is some dispersion in the data. In the evaluation to assess how the Hurdle-REWLR predicted zero counts compared to the Hurdle models, the hurdle models predicted the exact number of zeros available in the sample data. The binary component of the Hurdle models uses logistic regression to predict the zero counts. The prediction accuracy can thus be attributed to the bias towards the majority class. Rahim et al. (2019) assessed the performance of SMOTE logistic as a classifier in rare events data and revealed a similar outcome where SMOTE logistic regression approach was more accurate compared to the logistic regression model but was

outperformed by the latter in test prediction accuracy. The negative binomial hurdle REWLR model was selected based on the Akaike information criterion. The model was then fit to the maternal mortality data. The covariate factors that were significantly associated with maternal deaths at the binary level include attendance of at least 4 ANC visits, antepartum haemorrhage, and receiving uterotonics during or immediately after birth. Upon observing maternal death within a facility, the covariate factors influencing the number of maternal deaths reported are Macerated stillbirth, attendance of at least 4 35

ANC visits, adolescent pregnancies, antepartum haemorrhage, breech deliveries, postpartum haemorrhage, receiving Carbatosin and giving birth by cesarean section. The Hurdle-REWLR model has an advantage over the Hurdle models because of their ability to introduce weights hence producing more accurate estimates that can be used for inference of population parameters. When the zero-inflated sample accurately represents the population, choosing between these two groups of models could be based on Akaike Information Criterion. 5.3 Conclusion The main aim of this study was to create Poisson and NB Hurdle-REWLR models for zero- inflated data and evaluate their performance in comparison to the standard Hurdle models. The Hurdle-REWLR in their binary component accounted for an imbalance between majority and minority proportions. That was the differentiating factor between the two models. The proposed study models were then applied to simulated and maternal mortality data, where NB Hurdle-REWLR outperformed the other models. The difference in AIC based performance between the NB Hurdle REWLR model and the other models increased with an increase in the degree of zero inflation. The ideal model performed better in cases of class imbalance. The study findings also highlighted a case of biased classification. The binary component of the Hurdle model, using logistic regression, classified all the observed zero counts in the maternal mortality as zeroes. Despite the prediction being an exact fit, the NB Hurdle model was inferior in AIC measures. NB Hurdle REWLR was thus selected as the ideal model in rare event cases where class imbalance exists. The study further outlined factors influencing maternal deaths in Nairobi: adolescent pregnancy, attendance of at least 4 ANC visits, postpartum haemorrhage, antepartum haemor- rhage, breech delivery, and Macerated stillbirth. Most of these factors have been identified as determinants or causes of maternal deaths literature reviewed by this study. 36

Findings from this research are expected to provide reliable estimates of the number of maternal deaths in Nairobi, Kenya. Without the risk of overfitting zero counts, researchers will be able to realize the actual maternal mortality ratio and the factors associated with zero maternal death counts. The research results will assist in supporting existing policies and developing new programs and interventions to reduce the number of deaths due to childbirth and maternity. 5.4 Recommendation 5.4.1 Recommendation for further research One area for further research is the implementation of the Hurdle-REWLR models on normally distributed covariates. The covariates of the current study data consisted of count data, majority being zero-inflated just as the dependent variable; this limited the covariate effect on the dependent variable. 5.4.2 Policy recommendation This study recommends that the proposed interventions be implemented to halt any avoidable deaths of women during and immediately after childbirth. These interventions, such as the implementation of BEmONC or CEmONC has yet to be rolled out in all healthcare facilities. Actualizing this would go a long way in preventing maternal deaths due to obstetric conditions. Maternal deaths due to demographic and social factors such as adolescent pregnancies and attendance of ANC visits can be countered by educating the public on all the associated risks of these practices or lack-off. 37

References Ali, E. (2020). Zero-inflated poisson regression model for a new class of flexible link functions: A case study on healthcare utilization. Arefaynie, M., Kefale, B., Yalew, M., Adane, B., Dewau, R., and Damtie, Y. (2022). Number of antenatal care utilization and associated factors among pregnant women in ethiopia: zero-inflated poisson regression of 2019 intermediate ethiopian demography health survey. Reproductive Health, 19(1):1–10. Aryuyuen, S., Bodhisuwan, W., and Supapakorn, T. (2014). Zero inflated negative binomial- generalized exponential distribution and its applications. Songklanakarin Journal of Science and Technology, 36(4):483–491. Chaudhari, M., Hubbard, R., Reid, R. J., Inge, R., Newton, K. M., Spangler, L., and Barlow, W. E. (2012). Evaluating components of dental care utilization among adults with diabetes and matched controls via hurdle models. BMC oral health, 12(1):1–12. Desjardins, C. D. (2013).

Diop, A., Deme, E. H., and Diop, A. (2021). Zero-inflated generalized extreme value regression model for binary data and application in health study. arXiv preprint arXiv:2105.00482. Fenta, S. M. and Fenta, H. M. (2020). Risk factors of child

mortality in ethiopia: application of multilevel two-part model. PLoS One, 15(8):e0237640. Fenta, S. M., Fenta, H. M., and Ayenew, G. M. (2020). The best statistical model to estimate predictors of under-five mortality in ethiopia. Journal of Big Data, 7(1):1–14. Fitriani, R., Chrisdiana, L. N., and Efendi, A. (2019). Simulation on the zero inflated negative binomial (zinb) to model overdispersed, poisson distributed data. In IOP Conference Series: Materials Science and Engineering, volume 546, page 052025. IOP Publishing.

Greene, W. H. (1994). Accounting for excess zeros and sample selection in poisson and negative binomial regression models.

Hilbe, J. M. (2011). Negative binomial regression. Cambridge University Press. Hilbe, J. M. (2014). Modeling count data. Cambridge University Press. Hutchinson, M. K. and Holtman, M. C. (2005). Analysis of count data using poisson regression. Research in nursing & health, 28(5):408–418. Jabeur, S. B. (2017). Bankruptcy prediction using partial least squares logistic regression. Journal of Retailing and Consumer Services, 36:197–202. Kibika, S. A. (2020). The Zero Inflated Negative Binomial-Shanker distribution and its application to HIV exposed infant data. PhD thesis, Strathmore University. 38

King, G. (1989). Event count models for international relations: Generalizations and applications. International Studies Quarterly, 33(2):123–147. King, G. and Zeng, L. (2001). Logistic regression in rare events data. Political analysis, 9(2):137–163. Lambert, D. (1992). Zero-inflated poisson with an regression, in manufacturing to defects application. Technometrics, 34(1):14. Loquiha, O., Hens, N., Chavane, L., Temmerman, M., and Aerts, M. (2013). Modeling het- erogeneity for count data: A study of maternal mortality in health facilities in mozambique. Biometrical Journal, 55(5):647–660. Maalouf, M., Homouz, D., and Trafalis, T. B. (2018). Logistic regression in large rare events and imbalanced data: A performance comparison of prior correction and weighting methods. Computational Intelligence, 34(1):161–174. Maalouf, M. and Siddiqi, M. (2014). Weighted logistic regression for large-scale imbalanced and rare events data. Knowledge-Based Systems, 59:142–148. Mamun, M. A. A. (2014). Zero-inflated regression models for count data: an application to under-5 deaths. Mason, C., Twomey, J., Wright, D., and Whitman, L. (2018). Predicting engineering student attrition risk using a probabilistic neural network and comparing results with a backpropagation neural network and logistic regression. Research in Higher Education, 59(3):382–400. McDowell, A. (2003). From the help desk: hurdle models. The Stata Journal, 3(2):178–184. Miller, J. M. (2007).

Comparing Poisson, Hurdle, and ZIP model fit under varying degrees of skew and zero-inflation. PhD thesis, University of Florida.

Min, Y. and Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data.

Statistical modelling, 5(1):1–19.

Mullahy, J. (1986). Specification and testing of some modified count data models. Journal of econometrics, 33(3):341–365.

Neelon, B., Chang, H. H., Ling, Q., and Hastings, N. S. (2016). Spatiotemporal hurdle models for zero-inflated count data: exploring trends in emergency department visits. Statistical methods in medical research, 25(6):2558–2576. Nekesa, F. V. (2019). Distributions of zero-inflated models with application to HIV exposed infants. PhD thesis, Strathmore University. Nusinovici, S., Tham, Y. C., Yan, M. Y. C., Ting, D. S. W., Li, J., Sabanayagam, C., Wong, T. Y., and Cheng, C.-Y. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. Journal of clinical epidemiology, 122:56–69. Nyaboga, E. O. (2009). Maternal mortality at Kenyatta National'hospital (Nairobi, Kenya) 2000-2008. PhD thesis. 39

Odhiambo, C. and Kinoti, F. (2019). Evaluation and comparison of patterns of maternal complications using generalized linear models of count data time series. International Journal of Statistics in Medical Research, 8:32–39. Organization, W. H. et al. (2019). Trends in maternal mortality 2000 to 2017: estimates by who, unicef, unfpa, world bank group and the united nations population division. Rahim, A. H. A., Rashid, N. A., Nayan, A., and Ahmad, A.-R. (2019). Smote approach to imbalanced dataset in logistic regression analysis. In Proceedings of the Third Interna- tional Conference on Computing, Mathematics and Statistics (iCMS2017), pages 429–433. Springer. Rose, C. E., Martin, S. W., Wannemuehler, K. A., and Plikaytis, B. D. (2006).

Journal of biopharmaceutical statistics, 16(4):463–481. Smith, H., Ameh, C., Godia, P., Maua, J., Bartilol, K., Amoth, P., Mathai, M., and van den Broek, N. (2017). Implementing maternal death surveillance and response in kenya: incremental progress and lessons learned. Global Health: Science and Practice, 5(3):345– 354. Tecla, S. J., Franklin, B., David, A., and Jackson, T. K. (2017). Assessing facility readiness to offer basic emergency obstetrics and neonatal care (bemonc) services in health care facilities of west pokot county, kenya. J Clin Simul Res, 7:25–39. Tolles, J. and Meurer, W. J. (2016). Logistic regression: relating patient characteristics to outcomes. Jama, 316(5):533–534. Wang, H. (2020). Logistic regression for massive data with rare events. In International Conference on Machine Learning, pages 9829–9836. PMLR. Yego, F., D'este, C., Byles, J., Williams, J. S., and Nyongesa, P. (2014). Risk factors for maternal mortality in a tertiary hospital in kenya: a case control study. BMC pregnancy and childbirth, 14(1):1–9. Zare, N., Haem, E., Lankarani, K. B., Heydari, S. T., and Barooti, E. (2013). Breast cancer risk factors in a defined population: weighted logistic regression approach for rare events. Journal of breast cancer, 16(2):214–219. Zhen, Z., Shao, L., and Zhang, L. (2018). Spatial hurdle models for predicting the number of children with lead poisoning. International journal of environmental research and public health, 15(9):1792. Ziraba, A. K., Madise, N., Mills, S., Kyobutungi, C., and Ezeh, A. (2009). Maternal mortality in the informal settlements of nairobi city: what do we know? Reproductive health, 6(1):1–8. 40

Appendix A R CODES The R code used for simulations and model fitting in Chapter 4. A.1 Libraries

gridExtra) library(glmnet) library(plotrix) library(ZIM) library(tidyverse) 41

A.2 Simulations and Analysis #Generate data from Poisson Hurdle distribution. #Repeat for Poisson Hurdle-REWLR, NB Hurdle, and NB Hurdle-REWLR. # Probability of 0 = 1-p #Set the seed for reproducible results set.seed(2345) #Assigned weights w1 = 3.5 w0 = 1 #Zero-altered poisson random number generator function zero.aic.func >- function(n, pi, zero.prop) { rhpois >- function(n=n, mu, zprob){ ifelse(rbinom(n, 1, zprob) == 1, 0, rpois(n, mu)) } Y >- rhpois(n, mu = 1.3, zprob = pi) #Poisson Hurdle X >- rpois(n, 3.5) dsname >- data.frame(Y, X) #Poisson Regression model1 >- glm(Y ~ X, family="poisson", data=dsname) aic1 >- summary(model1)$aic #Poisson Hurdle Regression model2 >- hurdle(Y ~ X, data=dsname, dist = "poisson", link="logit") aic2 >- AIC(model2) 42

#REWLR-Hurdle Poisson Regression model3.a >- vglm(Y[Y < 0] ~ X[Y < 0], family = pospoisson(), data=dsname) model3.b >- rewlr(I(Y < 0) ~ X, weights0 = w0, weights1 = w1, data=dsname) aic.val3.a >- AICvlm(model3.a) aic.val3.b >- model3.b$aic #Negative Binomial Regression model4 >- glm.nb(Y ~ X, data=dsname) aic4 >- summary(model4)$aic #Negative Binomial Hurdle Regression model5 >- hurdle(Y ~ X, dist = "negbin") aic5 >- AIC(model5) #REWLR-Hurdle Negative Binomial Regression model6.a >- vglm(Y[Y < 0] ~ X[Y < 0], family = posnegbinomial(), data=dsname) model6.b >- rewlr(I(Y < 0) ~ X, weights0 = w0, weights1 = w1, data=dsname) aic.val6.a >- AICvlm(model6.a) aic.val6.b >- model6.b$aic aic6 >- aic.val6.a + aic.val6.b #AIC Values based on arious zero-proportions aic.values >- data.frame(AIC=rbind(aic1, aic2, aic3, aic4, aic5, aic6), Model=c("Poisson", "PH", "PH-RE","NB", "NB-H", "NBH-RE"), Zero.Proportion=c(rep(pi, 6)), row.names = NULL) #Figure: Performance based on AIC values for sets of models, sample size, zero% plot >- ggplot(aic.values, aes(x=Model, y=AIC))+ geom_bar(stat = "identity", fill="azure3")+ geom_text(aes(label=round(AIC, digits=0)), vjust=1.6, color="black", size=3.5)+ 43

scale_x_discrete(limits=c("Poisson", "NB", "PH", "NB-H", "PH-RE", "NBH-RE"))+ labs(y = "AIC Value", x = "Count Model")+ ggtitle(zero.prop)+ theme_bw() return(plot) } plot1 >- zero.aic.func(200, 0.50, "n=200; 50% zero-inflation") plot2 >- zero.aic.func(200, 0.40, "n=200; 60% zero-inflation") plot3 >- zero.aic.func(200, 0.25, "n=200; 75% zero-inflation") plot4 >- zero.aic.func(200, 0.10, "n=200; 90% zero-inflation") png(file="D:/MSc/Thesis/Thesis-template-20220215T185005Z-001/ Thesis-template/Figs/PH200.png", width=600, height=350) grid.arrange(plot1, plot2, plot3, plot4, ncol=2, nrow = 2) dev.off() plot5 >- zero.aic.func(1000, 0.50, "n=1000; 50% zero-inflation") plot6 >- zero.aic.func(1000, 0.40, "n=1000; 60% zero-inflation") plot7 >- zero.aic.func(1000, 0.25, "n=1000; 75% zero-inflation") plot8 >- zero.aic.func(1000, 0.10, "n=1000; 90% zero-inflation") png(file="D:/MSc/Thesis/Thesis-template-20220215T185005Z-001/ Thesis-template/Figs/PH500.png", width=600, height=350) grid.arrange(plot5, plot6, plot7, plot8, ncol=2, nrow = 2) 44

dev.off() plot9 >- zero.aic.func(500, 0.50, "n=500; 50% zero-inflation") plot10 >- zero.aic.func(500, 0.40, "n=500; 60% zero-inflation") plot11 >- zero.aic.func(500, 0.25, "n=500; 75% zero-inflation") plot12 >- zero.aic.func(500, 0.10, "n=500; 90% zero-inflation") png(file="D:/MSc/Thesis/Thesis-template-20220215T185005Z-001/ Thesis-template/Figs/PH1000.png", width=600, height=350) grid.arrange(plot9, plot10, plot11, plot12, ncol=2, nrow = 2) dev.off() #Compute Percentage change in AIC aic.chg.func >- function(n, pi) { # Zero-altered poisson random number generator rhpois >- function(n=n, mu, zprob){ ifelse(rbinom(n, 1, zprob) == 1, 0, rpois(n, mu)) } Y >- rhpois(n, mu = 1.3, zprob = pi) #Poisson Hurdle X >- rpois(n, 3.5) # Independent variable X dsname >- data.frame(Y, X) #Poisson Regression model1 >- glm(Y ~ X, family="poisson", data=dsname) aic1 >- summary(model1)$aic 45

#Poisson Hurdle Regression model2 >- hurdle(Y ~ X, data=dsname, dist = "poisson", link="logit") aic2 >- AIC(model2) ## REWLR-Hurdle Poisson Regression # Error due to vglm: https://bookdown.org/fxpalacio/bookdown_curso/GLM.html model3.a >- vglm(Y[Y < 0] ~ X[Y < 0], family = pospoisson(), data=dsname) model3.b >- rewlr(I(Y < 0) ~ X, weights0 = w0, weights1 = w1, data=dsname) aic.val3.a >- AICvlm(model3.a) aic.val3.b >- model3.b$aic aic3 >- aic.val3.a + aic.val3.b #Negative Binomial Regression model4 >- glm.nb(Y ~ X, data=dsname) aic4 >- summary(model4)$aic #Negative Binomial Hurdle Regression model5 >- hurdle(Y ~ X, dist = "negbin") aic5 >- AIC(model5) ## REWLR-Hurdle Negative Binomial Regression model6.a >- vglm(Y[Y < 0] ~ X[Y < 0], family = posnegbinomial(), data=dsname) model6.b >- rewlr(I(Y < 0) ~ X, weights0 = w0, weights1 = w1, data=dsname) aic.val6.a >- AICvlm(model6.a) aic.val6.b >- model6.b$aic aic6 >- aic.val6.a + aic.val6.b #AIC Values based on various zero-proportions aic.values.1 >- data.frame(ref="Poisson Hurdle (PH)", sample.size = n, 46

Zero.Proportion=1-pi, AIC=cbind(aic2, aic3, aic5, aic6)) colnames(aic.values.1) >- c("Reference","Sample size", "Zero Proportion", "PH", "PHRE", "NBH", "NBHRE") aic.values.1$PH >- round(aic.values.1$PH, 0) aic.values.1$PHRE >- paste(round(aic.values.1$PHRE, 0), '(', round(((aic.values.1$PHRE - aic.values.1$PH)/aic.values.1$PHRE)*100, 2),'%)') aic.values.1$NBH >- paste(round(aic.values.1$NBH, 0), '(', round(((aic.values.1$NBH - aic.values.1$PH)/aic.values.1$NBH)*100, 2),'%)') aic.values.1$NBHRE >- paste(round(aic.values.1$NBHRE, 0), '(', round(((aic.values.1$NBHRE - aic.values.1$PH)/aic.values.1$NBHRE)*100, 2),'%)') return(aic.values.1) } ph >- rbind(aic.chg.func(200, 0.50),aic.chg.func(200, 0.40), aic.chg.func(200, 0.25),aic.chg.func(200, 0.10), aic.chg.func(1000, 0.50),aic.chg.func(1000, 0.40), aic.chg.func(1000, 0.25),zero.aic.func(1000, 0.10)) #Step 1: Generate data from Poisson Hurdle-REWLR distribution aic.chg.func >- function(n, pi, zero.prop) { # Zero-altered poisson random number generator rhpois >- function(n=n, mu, zprob){ ifelse(rbinom(n, 1, zprob) == 1, 0, rpois(n, mu)) } Y >- rhpois(n, mu = 1.3, zprob = pi^w1) #Poisson Hurdle-RE X >- runif(n, -1, 1) # Independent variable X 47

dsname >- data.frame(Y, X) #Poisson Regression model1 >- glm(Y ~ X, family="poisson", data=dsname) aic1 >- summary(model1)$aic #Poisson Hurdle Regression model2 >- hurdle(Y ~ X, data=dsname, dist = "poisson", link="logit") aic2 >- AIC(model2) ## REWLR-Hurdle Poisson Regression model3.a >- vglm(Y[Y < 0] ~ X[Y < 0], family = pospoisson(), data=dsname) model3.b >- rewlr(I(Y < 0) ~ X, weights0 = w0, weights1 = w1, data=dsname) aic.val3.a >- AICvlm(model3.a) aic.val3.b >- model3.b$aic aic3 >- aic.val3.a + aic.val3.b #Negative Binomial Regression model4 >- glm.nb(Y ~ X, data=dsname) aic4 >- summary(model4)$aic #Negative Binomial Hurdle Regression model5 >- hurdle(Y ~ X, dist = "negbin") aic5 >- AIC(model5) ## REWLR-Hurdle Negative Binomial Regression model6.a >- vglm(Y[Y < 0] ~ X[Y < 0], family = posnegbinomial(), data=dsname) model6.b >- rewlr(I(Y < 0) ~ X, weights0 = w0, weights1 = w1, data=dsname) aic.val6.a >- AICvlm(model6.a) 48

aic.val6.b >- model6.b$aic aic6 >- aic.val6.a + aic.val6.b #AIC Values based on various zero-proportions aic.values.1 >- data.frame(ref="Poisson Hurdle - REWLR (PHRE)", sample.size = n, Zero.Proportion=1-pi, AIC=cbind(aic2, aic3, aic5, aic6)) colnames(aic.values.1) >- c("Reference","Sample size", "Zero Proportion", "PH", "PHRE", "NBH", "NBHRE") aic.values.1$PHRE >- round(aic.values.1$PHRE, 0) aic.values.1$PH >- paste(round(aic.values.1$PH, 0), '(', round(((aic.values.1$PH - aic.values.1$PHRE)/aic.values.1$PH)*100, 2),'%)') aic.values.1$NBH >- paste(round(aic.values.1$NBH, 0), '(',

round(((aic.values.1$NBH - aic.values.1$PHRE)/aic.values.1$NBH)*100, 2),')') aic.values.1$NBHRE >- paste(round(aic.values.1$NBHRE, 0), '(', round(((aic.values.1$NBHRE - aic.values.1$PHRE)/aic.values.1$NBHRE)*100, 2),'%)') return(aic.values.1) } phre >- rbind(aic.chg.func(200, 0.50),aic.chg.func(200, 0.40), aic.chg.func(200, 0.25),aic.chg.func(200, 0.10), aic.chg.func(1000, 0.50),aic.chg.func(1000, 0.40), aic.chg.func(1000, 0.25),zero.aic.func(1000, 0.10)) #Step 1: Generate data from NB Hurdle distribution aic.chg.func >- function(n, pi) { # Zero-altered negative binomial random number generator rhnbinom >- function(n=n, mu, size=0.5, zprob){ 49

ifelse(rbinom(n, 1, zprob) == 1, 0, rnbinom(n, size = 0.5, mu = mu)) } Y >- rhnbinom(n, mu = 1.3, size = 3, zprob = pi) #NB Hurdle X >- runif(n, -1, 1) # Independent variable X dsname >- data.frame(Y, X) #Poisson Regression model1 >- glm(Y ~ X, family="poisson", data=dsname) aic1 >- summary(model1)$aic #Poisson Hurdle Regression model2 >- hurdle(Y ~ X, data=dsname, dist = "poisson", link="logit") aic2 >- AIC(model2) ## REWLR-Hurdle Poisson Regression model3.a >- vglm(Y[Y < 0] ~ X[Y < 0], family = pospoisson(), data=dsname) model3.b >- rewlr(I(Y < 0) ~ X, weights0 = w0, weights1 = w1, data=dsname) #aic.val3.a >- (-2*logLik.vlm(model3.a))+(2*3) aic.val3.a >- AICvlm(model3.a) aic.val3.b >- model3.b$aic aic3 >- aic.val3.a + aic.val3.b #Negative Binomial Regression model4 >- glm.nb(Y ~ X, data=dsname) aic4 >- summary(model4)$aic #Negative Binomial Hurdle Regression model5 >- hurdle(Y ~ X, dist = "negbin") aic5 >- AIC(model5) 50

REWLR-Hurdle Negative Binomial Regression model6.a >- vglm(Y[Y < 0] ~ X[Y < 0], family = posnegbinomial(), data=dsname) model6.b >- rewlr(I(Y < 0) ~ X, weights0 = w0, weights1 = w1, data=dsname) #aic.val6.a >- (-2*logLik.vlm(model3.a))+(2*3) aic.val6.a >- AICvlm(model6.a) aic.val6.b >- model6.b$aic aic6 >- aic.val6.a + aic.val6.b #AIC Values based on various zero-proportions aic.values.1 >- data.frame(ref="NB Hurdle (NBH)", sample.size = n, Zero.Proportion=1-pi, AIC=cbind(aic2, aic3, aic5, aic6)) colnames(aic.values.1) >- c("Reference","Sample size", "Zero Proportion", "PH", "PHRE", "NBH", "NBHRE") aic.values.1$NBH >- round(aic.values.1$NBH, 0) aic.values.1$PH >- paste(round(aic.values.1$PH, 0), '(', round(((aic.values.1$PH - aic.values.1$NBH)/aic.values.1$PH)*100, 2),'%)') aic.values.1$PHRE >- paste(round(aic.values.1$PHRE, 0), '(', round(((aic.values.1$PHRE - aic.values.1$NBH)/aic.values.1$PHRE)*100, 2),'%)') aic.values.1$NBHRE >- paste(round(aic.values.1$NBHRE, 0), '(', round(((aic.values.1$NBHRE - aic.values.1$NBH)/aic.values.1$NBHRE)*100, 2),'%)') return(aic.values.1) } nbh >- rbind(aic.chg.func(200, 0.50),aic.chg.func(200, 0.40), aic.chg.func(200, 0.25),aic.chg.func(200, 0.10), aic.chg.func(1000, 0.50),aic.chg.func(1000, 0.40), 51

aic.chg.func(1000, 0.25),zero.aic.func(1000, 0.10)) #Step 1: Generate data from NB Hurdle REWLR distribution aic.chg.func >- function(n, pi, zero.prop) { # Zero-altered negative binomial random number generator rhnbinom >- function(n=n, mu, size=0.5, zprob){ ifelse(rbinom(n, 1, zprob) == 1, 0, rnbinom(n, size = 0.5, mu = mu)) } Y >- rhnbinom(n, mu = 1.3, size = 3, zprob = pi^w1) #NB Hurdle-RE X >- runif(n, -1, 1) # Independent variable X dsname >- data.frame(Y, X) #Poisson Regression model1 >- glm(Y ~ X, family="poisson", data=dsname) aic1 >- summary(model1)$aic #Poisson Hurdle Regression model2 >- hurdle(Y ~ X, data=dsname, dist = "poisson", link="logit") aic2 >- AIC(model2) ## REWLR-Hurdle Poisson Regression model3.a >- vglm(Y[Y < 0] ~ X[Y < 0], family = pospoisson(), data=dsname) model3.b >- rewlr(I(Y < 0) ~ X, weights0 = w0, weights1 = w1, data=dsname) #aic.val3.a >- (-2*logLik.vlm(model3.a))+(2*3) aic.val3.a >- AICvlm(model3.a) aic.val3.b >- model3.b$aic aic3 >- aic.val3.a + aic.val3.b 52

#Negative Binomial Regression model4 >- glm.nb(Y ~ X, data=dsname) aic4 >- summary(model4)$aic #Negative Binomial Hurdle Regression model5 >- hurdle(Y ~ X, dist = "negbin") aic5 >- AIC(model5) ## REWLR-Hurdle Negative Binomial Regression model6.a >- vglm(Y[Y < 0] ~ X[Y < 0], family = posnegbinomial(), data=dsname) model6.b >- rewlr(I(Y < 0) ~ X, weights0 = w0, weights1 = w1, data=dsname) # #aic.val6.a >- (-2*logLik.vlm(model3.a))+(2*3) aic.val6.a >- AICvlm(model6.a) aic.val6.b >- model6.b$aic aic6 >- aic.val6.a + aic.val6.b #AIC Values based on various zero-proportions aic.values.1 >- data.frame(ref="NB Hurdle - REWLR (NBHRE)", sample.size = n, Zero.Proportion=1-pi, AIC=cbind(aic2, aic3, aic5, aic6)) colnames(aic.values.1) >- c("Reference","Sample size", "Zero Proportion", "PH", "PHRE", "NBH", "NBHRE") aic.values.1$NBHRE >- round(aic.values.1$NBHRE, 0) aic.values.1$PH >- paste(round(aic.values.1$PH, 0), '(', round(((aic.values.1$PH - aic.values.1$NBHRE)/aic.values.1$PH)*100, 2),'%)') aic.values.1$NBH >- paste(round(aic.values.1$NBH, 0), '(', round(((aic.values.1$NBH - aic.values.1$NBHRE)/aic.values.1$NBH)*100, 2),'%)') aic.values.1$PHRE >- paste(round(aic.values.1$PHRE, 0), '(', 53

round(((aic.values.1$PHRE - aic.values.1$NBHRE)/aic.values.1$PHRE)*100, 2),'%)') return(aic.values.1) } nbhre >- rbind(aic.chg.func(200, 0.50),aic.chg.func(200, 0.40), aic.chg.func(200, 0.25),aic.chg.func(200, 0.10), aic.chg.func(1000, 0.50),aic.chg.func(1000, 0.40), aic.chg.func(1000, 0.25),zero.aic.func(1000, 0.10)) #Combine all allaic >- rbind(ph, phre, nbh, nbhre) allaic %<% knitr::kable(format='latex') %<% kable_classic_2(full_width = F, html_font = "Cambria") A.3 Analysis on Maternal Mortality Data A.3.1 Exploratory Data Analysis #Read in Data maternal >-

read.csv("D:/MSc/Thesis/Analysis/Maternal Mortality Data.csv") maternal1 >- maternal[, -1] #Rename Columns maternal2 >- rename(maternal1, MaternalDeaths=Maternal.Deaths, AssistedDeliveries=assisted.Vaginal.deliveries, BreechDelivery=breach.delivery, 54

CS=caesarian.sections, LiveBirths=live.birth, EarlyTeenPreg=no.adolesc..10.14.years..pregn.at.1st.anc.visit, LateTeenPreg=no.adolesc..15.19.years..preg..at.1st.anc.visit, NormalDeliveries=normal.deliveries, ANC4Visits=anc.4.visits, Uterotonics3stg=Number.of.women.giving.birth.who.received. uterotonics.in.the.third.stage.of.labor..or.immediately.after.birth., Carbatosin=Mothers.given.uterotonics.within.1.minute..Carbatosin., Oxytocin=Mothers.given.uterotonics.within.1.minute..Oxytocin., Eclampsia=Eclampsia, AntHaemorrage=Ante.partum.Haemorrage PostHaemorrage=Post.Partum.Haemorrage, ObstructedLabour=Obstructed.Labour, RupturedUterus=Ruptured.Uterus, Sepsis=Sepsis, FGMComplicatons=Mothers.with.delivery.complications. associated.with.FGM, Stillbirth=Macerated.still.Birth) #EDA #Central Tendency mean(maternal2$MaternalDeaths) std.error(maternal2$MaternalDeaths) #Spread sd(maternal2$MaternalDeaths) data.frame(table(maternal2$MaternalDeaths)) %<% kbl() %<% kable_classic_2(full_width = F, html_font = "Cambria") #Histogram plot2 >- ggplot(maternal2, aes(x=MaternalDeaths)) + geom_histogram(binwidth=1, fill="skyblue")+ 55

labs(x = "Maternal Deaths", y = "Frequency")+ ylim(0, 150)+ theme_classic() plot2 #Correlation Analysis cor(maternal2) #Averaging the factors sum1 >- as.data.frame(t(maternal2 %<% group_by(MaternalDeaths.bin) %<% summarise_all(mean))) %<% mutate(across(where(is.numeric), round, 1)) i >- which(str_detect(row.names(sum1), "^Maternal")) sum2 >- sum1%<% slice(-i) colnames(sum2) >- c("No Maternal Deaths","Maternal Deaths") sum2 %<% slice(2:n()) %<% knitr::kable(format='latex') %<% kable_classic_2(full_width = F, html_font = "Cambria") A.3.2 Count Models # Sample data = 293 deaths per 53792 live births; #Y-bar = 293/53792 = 0.0054 # Population data = 342 deaths per 100000 live births; Tau = 1600/100000 = 0.016 56

#Source: https://www.health.go.ke/wp-content/uploads/2022/01/ Kenya-SDG-Progress-Report_-April21.pdf w1 = 0.00163/0.0054 w0 = (1 - 0.00163)/(1 - 0.0054) #Poisson Hurdle Regression fit2 >- hurdle(MaternalDeaths ~ Stillbirth+ANC4Visits+LateTeenPreg+ AntHaemorrage+PostHaemorrage+BreechDelivery+Carbatosin+Uterotonics3stg, dist = "poisson", link="logit", data=maternal2) f.exp2 >- round(sum(predict(fit2, type = "prob")[,1]),0) f.aic2 >- AIC(fit2) summary(fit2) ## REWLR-Hurdle Poisson Regression fit3.a >- vglm(MaternalDeaths ~ AssistedDeliveries+BreechDelivery +CS+EarlyTeenPreg, family = pospoisson(), data=Maternal2.gt0) fit3.b >- rewlr(MaternalDeaths.bin ~ Stillbirth+ANC4Visits+ LateTeenPreg+AntHaemorrage+PostHaemorrage+BreechDelivery+Carbatosin+ Uterotonics3stg, weights0 = w0, weights1 = w1, data=maternal2) f.exp3 >- round(sum(1-predict.rewlr(fit3.b))) f.aic.val3.a >- AICvlm(fit3.a) f.aic.val3.b >- fit3.b$aic f.aic3 >- f.aic.val3.a + f.aic.val3.b summary(fit3.a) summary.rewlr(fit3.b) #Negative Binomial Hurdle Regression fit5 >- hurdle(MaternalDeaths ~ Stillbirth+ANC4Visits+LateTeenPreg+ 57

AntHaemorrage+PostHaemorrage+BreechDelivery+Carbatosin+ Uterotonics3stg, data=maternal2, dist = "negbin") f.exp5 >- sum(predict(fit5, type = "prob")[,1]) f.aic5 >- AIC(fit5) summary(fit5) ## REWLR-Hurdle Negative Binomial Regression fit6.a >- vglm(MaternalDeaths ~ Stillbirth+ANC4Visits+LateTeenPreg+ AntHaemorrage+PostHaemorrage+BreechDelivery+Carbatosin+ Uterotonics3stg, family = posnegbinomial(), data=Maternal2.gt0) fit6.b >- rewlr(MaternalDeaths.bin ~ Stillbirth+ANC4Visits+ LateTeenPreg+AntHaemorrage+PostHaemorrage+BreechDelivery+Carbatosin +Uterotonics3stg, weights0 = w0, weights1 = w1, data=maternal2) f.exp6 >- round(sum(1-predict.rewlr(fit6.b))) f.aic.val6.a >- AICvlm(fit6.a) f.aic.val6.b >- fit6.b$aic f.aic6 >- f.aic.val6.a + f.aic.val6.b summary(fit6.a) summary.rewlr(fit6.b) #Comparison of coefficients between models #Binary part coef.bin >- data.frame(PH.BINARY = summary(fit2)$coefficients$zero[,1], NBH.BINARY = summary(fit5)$coefficients$zero[,1], PHRE.BINARY = summary.rewlr(fit3.b)$B, NBHRE.BINARY = summary.rewlr(fit6.b)$B) #Count part coef.cnt >- data.frame(cbind(PH.COUNT = summary(fit2)$coefficients$count[,1], 58

NBH.COUNT = summary(fit5)$coefficients$count[,1], PHRE.COUNT = summary(fit3.a)@coef3[, 1], NBHRE.COUNT = summary(fit6.a)@coef3[-2, 1])) coef.cnt$PH.COUNT[10] = "NA" coef.cnt$NBHRE.COUNT[10] = "NA" coef.cnt$PHRE.COUNT[c(6,7,8,9,10)] = "NA" coef.bin %<% knitr::kable(format='latex') %<% kable_classic_2(full_width = F, html_font = "Cambria") coef.cnt %<% knitr::kable(format='latex') %<% kable_classic_2(full_width = F, html_font = "Cambria") #AIC Values based on various zero-proportions aic.values >- data.frame(Model=c("Poisson", "PH", "PH-RE","NB", "NB-H", "NBH-RE"), AIC=rbind(f.aic1, f.aic2, f.aic3, f.aic4, f.aic5, f.aic6), row.names = NULL) aic.values %<% knitr::kable(format='latex') %<% kable_classic_2(full_width = F, html_font = "Cambria") #Zero counts zero.counts >- data.frame(cbind(observed,f.exp1,f.exp2,f.exp3, f.exp4,f.exp5,f.exp6)) colnames(zero.counts) >- c("Oberved","Poisson",

"PH", "PH-RE","NB", "NB-H", "NBH-RE") zero.counts %<% knitr::kable(format='latex') %<% kable_classic_2(full_width = F, html_font = "Cambria") 59

Curiginal

Hit and source - focused comparison, Side by Side

Submitted text As student entered the text in the submitted document.

Matching text As the text appears in the source.

1/9	SUBMITTED TEXT	13 WORDS	75%	MATCHING TEXT	13 WORDS

and a zero-truncated model determines the magnitude of the positive counts (Mullahy, 1986). In

SA ST404A3.pdf (D27768399)

2/9	SUBMITTED TEXT	20 WORDS	75%	MATCHING TEXT	20 WORDS

models. AIC is computed as: AIC = −2log(L)+2K; where L is the likelihood, and K is the number of parameters in the model.

models. The AIC is computed using the formula AIC = −2log(L) + 2q, where L is the likelihood and q is the number of parameters in the model.

W https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4493133/

5/9	SUBMITTED TEXT	21 WORDS	95%	MATCHING TEXT	21 WORDS

Evaluating the performance of two competing models of school suspension under simulation-the zero-inflated negative binomial and the negative binomial hurdle. University of Minnesota.

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,

W https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4493133/

7/9	SUBMITTED TEXT	17 WORDS	100%	MATCHING TEXT	17 WORDS

Greene, W. H. (1994). Accounting for excess zeros and sample selection in poisson and negative binomial regression models.

SA Assignment 3.pdf (D27768595)

3/9	SUBMITTED TEXT	18 WORDS	100%	MATCHING TEXT	18 WORDS

Comparing Poisson, Hurdle, and ZIP model fit under varying degrees of skew and zero-inflation. PhD thesis, University of Florida.

Comparing Poisson, hurdle and ZIP model fit under varying degrees of skew and zero-inflation. PhD Thesis, University of Florida,

W https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4493133/

Curiginal

4/9	SUBMITTED TEXT	16 WORDS	100%	MATCHING TEXT	16 WORDS

Min, Y. and Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data.

Min Y and Agresti A. (2005) Random effect models for repeated measures of zero-inflated count data.

W https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4493133/

8/9	SUBMITTED TEXT	15 WORDS	100%	MATCHING TEXT	15 WORDS

Mullahy, J. (1986). Specification and testing of some modified count data models. Journal of econometrics, 33(3):341–365.

SA assignment3_1414386.pdf (D27768445)

6/9	SUBMITTED TEXT	15 WORDS	100%	MATCHING TEXT	15 WORDS

On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data.

On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data.

W https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4493133/

9/9	SUBMITTED TEXT	1 WORDS	40%	MATCHING TEXT	1 WORDS

library(ggplot2) library(sandwich) library(msm) library(dplyr) library(tidyr) library(vcd) library(countreg) library(pscl) library(VGAM) library(rewlr) library(kableExtra) library(readxl) library(

library(readxl) library(Rcmdr) library(Rmisc) library(dplyr) library(tidyr) library(ggplot2) library(gridExtra) library(kableExtra) library(xtable) library(multcomp) library(psych) library(effects) library(

SA Handin 1.pdf (D108014019)

Appendix C

Ethics Review Approval

Strathmore UNIVERSITY

24th May 2022

Ms Okello Sharon,
awuor.okello@strathmore.edu

Dear Ms Okello,

RE: Improving Performance of Hurdle Models using Rare-Event Weighted Logistic Regression: Application to Maternal Mortality Data.

This is to inform you that SU-IERC has reviewed and **approved** your above **SU Masters'** research proposal. Your application reference number is **SU-IERC1339/22.** The approval period is **24th May 2022 to 23rd May 2023.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-IERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-IERC within 48 hours of notification
iv. Any changes, anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-IERC within 48 hours
v. Clearance for export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days upon completion of the study to SU-IERC.

Prior to commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

for: **Dr Ben Ngoye,**
Secretary; SU-IERC

Cc: Prof Fred Were,
Chairperson; SU-IERC