ZERO-INFLATED REGRESSION MODELS FOR COUNT DATA: AN APPLICATION TO UNDER-5 DEATHS

A THESIS PAPER
SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE
MASTER OF SCIENCE
BY
MD ABDULLAH AL MAMUN
DR. MUNNI BEGUM - ADVISOR

BALL STATE UNIVERSITY
MUNCIE, INDIANA
MAY 2014
Acknowledgements

Foremost, I would like to express my sincere gratitude to my advisor Dr. Munni Begum for the continuous support of my thesis study. This work would not be possible without her patience, motivation, enthusiasm, and immense knowledge. Her guidance helped me in all the time of the study; specifically, from the set up of the research questions to writing the conclusion of the report. Besides my advisor, I would like to thank the rest of my thesis committee: Dr. Dale Umbach and Dr. Rahmatullah Imon for their encouragement, insightful comments and patience. I am thankful to all my classmates for their kind supports. Last but not the least, I would like to thank my family and friends for supporting me throughout my life.

Md Abdullah Al Mamun

April 7, 2014
Contents

Acknowledgements ................................................. ii
List of Tables ....................................................... v
List of Figures ...................................................... vi

1 Introduction ....................................................... 1

2 Zero-inflated Count Data Regression Models ............... 4
  2.1 Zero-inflated Poisson (ZIP) Regression ..................... 4
  2.2 Zero-inflated Negative Binomial (ZINB) Regression ....... 6
  2.3 Hurdle Regression Model ................................. 8
  2.4 Zero-inflated Generalized Poisson (ZIGP) regression .... 10

3 The Simulation Experiment ..................................... 13
  3.1 The simulation experiment on zero-inflated count data generated from Poisson distribution ................. 13
  3.2 The simulation experiment on zero-inflated count data generated from negative binomial distribution ........ 18

4 Under-5 deaths in Bangladesh ............................... 24
  4.1 A brief overview on BDHS 2011 survey .................... 24
  4.2 Under-5 mortality and it’s determinants .................. 24
  4.3 Exploratory analysis of under-5 death ................. 26
4.3.1 Why it is important to model zero counts and positive counts separately 26

5 An Application of Zero-inflated Models in Under-5 Deaths in Bangladesh 33

5.1 Estimation of coefficients 34

5.2 Model comparison 37

5.3 A closer look on estimates of the parameters using Hurdle model 37

6 Conclusions and discussions 40

6.1 Limitation of the study 41

Bibliography 42
List of Tables

4.1 Descriptive statistics of the variables, BDHS 2011 ........................................ 27
4.2 Frequency of number of under-5 deaths in BDHS 2011 survey .......................... 28
5.1 Estimation of coefficients using different count data models for under-5 mor-
tality in Bangladesh, BDHS 2011 ........................................................................ 36
5.2 Finding the best model for under-5 mortality in Bangladesh, BDHS 2011 .......... 37
5.3 Estimate of parameters of Hurdle model for under-5 deaths ......................... 39
List of Figures

3.1 AIC statistic of each models for different values of lambda and percentage of zeros (data generated from ZIP) ........................................... 15

3.2 Prediction of zero counts by each models in 1,000 simulation run (data generated from ZIP) .......................................................... 17

3.3 MSE of the count data models in 1,000 simulation run (data generated from ZIP) .......................................................... 18

3.4 Histogram of 1000 variates generated from Poisson and Negative binomial distribution inflated at zero ........................................... 20

3.5 AIC of each models for different values of lambda and percentage of zeros (data generated from ZINB) ........................................... 21

3.6 Prediction of zero counts by each models in 1,000 simulation run (data generated from ZINB) .......................................................... 22

3.7 MSE of the count data models in 1,000 simulation run (data generated from ZINB) .......................................................... 23

4.1 Percentage of women experiencing under-5 deaths by age and number of live births: How zero counts differ than the rests ........................................... 29

4.2 Percentage of women experiencing under-5 deaths by parents level of education: How zero counts differ than the rests ........................................... 29

4.3 Percentage of women experiencing under-5 deaths by wealth index: how zero counts differ than the rests ........................................... 30
4.4 Percentage of women experiencing under-5 deaths by number of births at teen age: how zero counts differ than the rests ........................................... 31
4.5 Percentage of women experiencing under-5 deaths by number of births at teen age: how zero counts differ than the rests ........................................... 32
Chapter 1

Introduction

In many studies count data may possess excess amount of zeros. Zero counts may not occur in the same process as other positive counts. Lambert (1992) discussed this issue and proposed zero-inflated Poisson (ZIP) model with an application of defects in manufacturing (also proposed by Greene (1994)). Zero-inflated count data may not have equality of mean and variance. In such case over-dispersion (or under-dispersion) need to be taken into account. A number of mixture and compound Poisson distribution, such as, double Poisson, Poisson log-normal, Poisson-geometric, Poisson-negative binomial, modified Poisson, Poisson-Pascal, Quasi-Poisson, mixed-Poisson, Poisson-binomial, and so on, have been suggested by many researchers to address the inequality of mean and variance of the Poisson distribution. However, Consul (1973) proposed generalized Poisson distribution which can take account of over-dispersion of Poisson distribution. It contains two parameters $\lambda_1$ and $\lambda_2$ where $\lambda_2$ can be positive, zero, or negative. If $\lambda_2$ is zero then it becomes the standard Poisson. Still there is a mean variance relationship but equality assumption becomes flexible. Later Consul and Famoye (1992) improved the model for regression analysis. An extension of generalized Poisson distribution is zero-inflated generalized Poisson (ZIGP) proposed by Famoye and Singh (2006). In presence of over-dispersion in the data negative binomial model can be preferred when Poisson mean
has a gamma distribution. A natural extension of negative binomial model to accommodate excess zeros in the data is zero-inflated negative binomial (ZINB) model discussed by Mwalili (2008). Another popular approach to model the excess zeros in count data is to use truncated models. Hurdle model developed by Cragg (1971) is an example of truncated models for count data.

In this study we considered the three zero-inflated models, ZIP, ZINB and Hurdle model, to observe whether there is any effect of proportion of zeros in the performance of the models with given overall rate of the counts. We also considered two classical count data models, Poisson and negative binomial for comparison. The principal motivation of this study was that the real life data often inflated by zero counts and may or may not be over-dispersed (or under-dispersed). We have used two data generating processes, Poisson and negative binomial. In both cases we inflated zero counts excessively (varied from 30% to 80%) for different rate of the counts. The over-dispersion accounted by the variance, that may affect the fit of the model based on the functional relationship between mean and variance. Mean-variance relationship for ZIP is linear while the relationship is quadratic for ZINB. Hurdle model has the flexibility to choose Poisson or negative binomial or geometric distribution in modeling positive counts. We tried to investigate how the model fit changes for different rate of counts with different proportion of zeros in our simulation experiment.

Numerous applications of ZI models can be found in the literature. In this study we have fitted these models on the number of under-5 deaths per woman in Bangladesh. Under-5 death is showing a declining trend over the last two decades. However, still it is very high and require interventions to lower the death toll. Number of under-5 deaths was 527,000 in 1990 and 140,000 in 2010 in Bangladesh. And the rate was 143 per 1000 live births in 1990 and 48 per 1000 live births in 2010. The major causes of under-5 deaths are early perinatal (death in first three days of life), neonatal tetanus, acute respiratory infection
(ARI), malnutrition, diarrhoea, measles, and accidents [2] in Bangladesh. To understand the disparity in under-5, we need to study the socio-economic factors. Preceding birth interval and nutritional status of the mother and the child have significant effects on under-5 deaths [22]. Also parents level of education [25], mother’s health seeking behaviour [13], availability and accessibility of the health care [1], birth order, maternal age of mother [21], and economic status of the household of the mother [25] are the factors need to consider to understand the under-5 mortality. Malnutrition is a potential source of deadly diseases which may be responsible for under-5 deaths. As the CPI (consumer price index) rises over the last decade along with nearly 7% of rate of inflation per year in Bangladesh, proportion of malnourished children age 6-23 months rises among the lowest two asset quartiles and decreases among the highest two asset quartiles. Also the overall rate of malnourished children decreased over that period [11]. In this study we have modeled the number of under-5 deaths experienced by a mother to the associated available factors suggested by the literature. In BDHS 2011 survey, nearly 80% of the mothers never experienced any under-5 deaths. That means, our response variable is zero-inflated, which motivated us to fit ZI models to the data.
Chapter 2

Zero-inflated Count Data Regression Models

In this section, we briefly outlined a non exhaustive list of commonly used regression models for zero-inflated count data. The following four regression models, zero-inflated Poisson (ZIP) regression, zero-inflated negative binomial (ZINB) regression, hurdle regression, and zero-inflated generalized Poisson (ZIGP) regression are frequently used to model zero-inflated count data.

2.1 Zero-inflated Poisson (ZIP) Regression

This model was proposed by Lambert (1992) [15] with an application to defects in a manufacturing process. In ZIP regression, the responses $Y = (Y_1, Y_2, ..., Y_n)'$ are independent. One assumption of this model is that with probability $p$ the only possible observation is 0, and with probability $(1 - p)$, a Poisson($\lambda$) random variable is observed in $Y$. Thus the occurrence of $Y_i$ follows the following distributions:
\[ Y_i = \begin{cases} 0 & \text{with probability } p_i + (1 - p_i)e^{-\lambda_i} \\ k & \text{with probability } (1 - p_i)\frac{e^{-\lambda_i\lambda_i^k}}{k!}, k = 1, 2, \ldots \end{cases} \] (2.1)

The mean and variance of ZIP distribution are \( E(Y_i) = (1 - p_i)\lambda_i \) and \( V(Y_i) = (1 - p_i)(\lambda_i + \lambda_i^2) - ((1 - p_i)\lambda_i)^2 \), respectively. Note that this distribution approaches to Poisson(\( \lambda \)) as \( p \to 0 \).

Lambert (1992) [15] described a manufacturing process for two distinct sources of zero counts in number of defective items. One obvious source for zero counts is perfect state, when it is impossible for an item to be defective, and the probability of a perfect state is \( p \). Another source is imperfect state, when a process may produce a non-defective items with probability \( (1 - p)e^{-\lambda} \) when the number of defective items follows Poisson. The transient perfect state increases the number of zeros in the data.

The Poisson mean vector \( \lambda = (\lambda_i, ..., \lambda_n) \) has the canonical link \( \log(\lambda) \) for a Poisson regression model which satisfies:

\[
\log(\lambda) = \mathbf{B}\mathbf{\beta} \\
\lambda = e^{\mathbf{B}\mathbf{\beta}}
\]

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

\[
\text{logit}(p) = \log \left( \frac{p}{1 - p} \right) = \mathbf{G}\gamma \\
p = \frac{e^{G\gamma}}{1 + e^{G\gamma}} \\
(1 - p) = \frac{1}{1 + e^{G\gamma}}
\]
Here \( B \) and \( G \) are covariate matrices. The covariate matrix \( B \) is responsible for Poisson outcome in \( Y \) and the covariate matrix \( G \) is responsible for excess zeros in \( Y \). Also, \( B \) and \( G \) can be identical, or may have some common covariates depending on the types of the study. Now we have:

\[
f(y_i) = \begin{cases} 
  p_i + (1 - p_i)e^{-\lambda_i} = \frac{e^{G_i \gamma}}{1 + e^{G_i \gamma}} + \frac{e^{-B_i \beta}}{1 + e^{G_i \gamma}} & \text{; when } y_i = 0 \\
  (1 - p_i)\frac{e^{-\lambda_i} \lambda_i^y}{y!} = \frac{e^{-B_i \beta} (e^{B_i \beta})^y}{(1 + e^{G_i \gamma})^y!} & \text{; when } y_i = 1, 2, \ldots 
\end{cases}
\]  

(2.2)

The likelihood function is:

\[
L(\gamma, \beta; y_i) = \prod \left( \left( \frac{e^{G_i \gamma}}{1 + e^{G_i \gamma}} + \frac{e^{-B_i \beta}}{1 + e^{G_i \gamma}} \right) \left( \frac{e^{-B_i \beta} (e^{B_i \beta})^y}{(1 + e^{G_i \gamma})^y!} \right) \right) 
\]  

(2.3)

And the log-likelihood function is as follows which can be used to estimate the parameter vectors \( \gamma \) and \( \beta \), as well as, \( \lambda \) and \( p \):

\[
l(\gamma, \beta; y_i) = \sum_{y_i=0} \log(e^{G_i \gamma} + e^{-B_i \beta}) - \sum_{i=1}^n \log(1 + e^{G_i \gamma}) + \sum_{y_i>0} (y_iB_i \beta - e^{B_i \beta}) - \sum_{y_i>0} \log(y_i!) 
\]  

(2.4)

The parameter estimation can be carried out by employing EM algorithm and/or Newton-Raphson algorithm.

## 2.2 Zero-inflated Negative Binomial (ZINB) Regression

We briefly discuss on ZINB regression in this subsection. For a detailed see Hilbe (2011) \[12\] and Mwalili(2008) \[18\]. The ZINB distribution is a mixture distribution, similar to ZIP distribution, where the probability \( p \) for excess zeros and with probability \( (1 - p) \) the rest of the counts followed negative binomial distribution. Note that the negative binomial distribution is a mixture of Poisson distributions, which allows the Poisson mean \( \lambda \) to be distributed as
Gamma, and in this way overdispersion is modelled. Negative binomial distribution is given by:

\[
P(Y = y) = \frac{\Gamma(y + \tau)}{y!\Gamma(\tau)} \left( \frac{\tau}{\lambda + \tau} \right)^\tau \left( \frac{\lambda}{\lambda + \tau} \right)^y, \quad y = 0, 1, \ldots; \lambda, \tau > 0 \quad (2.5)
\]

where \( \lambda = \text{E}(Y) \), \( \tau \) is a shape parameter which quantifies the amount of overdispersion, and \( Y \) is the response variable of interest. The variance of \( Y \) is \( \lambda + \lambda^2/\tau \). Clearly, the negative binomial distribution approaches a Poisson distribution when \( \tau \) tends to \( \infty \) (no overdispersion). Consequently, the ZINB distribution is given by:

\[
P(Y = y) = \begin{cases} 
  p + (1 - p)(1 + \frac{\lambda}{\tau})^{-\tau}, & y = 0 \\
  (1 - p) \frac{\Gamma(y + \tau)}{y!\Gamma(\tau)} (1 + \frac{\lambda}{\tau})^{-\tau}(1 + \frac{x_i}{\tau})^{-y}, & y = 1, 2, \ldots
\end{cases}
\]

The mean and variance of the ZINB distribution are \( \text{E}(Y) = (1 - p)\lambda \) and \( \text{V}(Y) = (1 - p)\lambda(1 - p\lambda + \lambda/\tau) \), respectively. It is to be noted that this distribution approaches the ZIP distribution and the negative binomial distribution as \( \tau \to \infty \) and \( p \to 0 \), respectively. If both \( 1/\tau \) and \( p \approx 0 \) then ZINB distribution reduces to Poisson distribution.

The ZINB regression model relates \( p \) and \( \lambda \) to covariate matrix \( \mathbf{X} \) and \( \mathbf{Z} \) with regression parameters \( \beta \) and \( \gamma \) as:

\[
\log(\lambda_i) = \mathbf{x}_i^T \beta \quad \text{and} \quad \logit(p_i) = \mathbf{z}_i^T \gamma \quad i = 1, 2, \ldots, n \quad (2.6)
\]

The ZINB log-likelihood given the observed data is:

\[
l(\beta, \gamma, \tau; \mathbf{y}, \mathbf{z}, \mathbf{x}) = \sum_{i=1}^{n} \log (1 + e^{\mathbf{x}_i^T \gamma}) - \sum_{i=1: y_i=0}^{n} \log \left( e^{\mathbf{x}_i^T \gamma} + \left( \frac{e^{x_i \beta + \tau}}{\tau} \right)^{-\tau} \right) \\
+ \sum_{i=1: y_i>0}^{n} \left( \tau \log \left( \frac{e^{x_i \beta + \tau}}{\tau} \right) + y_i \log (1 + e^{-x_i \beta}) \right) \\
+ \sum_{i=1: y_i>0}^{n} \left( \log \Gamma(\tau) + \log \Gamma(1 + y_i) - \log \Gamma(\tau + y_i) \right) \quad (2.7)
\]
Parameter estimation can be carried out by ML technique using quasi-Newton optimization method.

2.3 Hurdle Regression Model

A detail outline of hurdle model can be found in the text by Cameron and Trivedi (1998) [3] (see also Cragg (1971) [7] and Mullahy (1986) [17]. The hurdle model may be define as a two-part model where the first part is a binary outcome model, and the second part is a truncated count model. As per Cameron and Trivedi (1998, p. 123) “Such a partition permits the interpretation that positive observations arise from crossing the zero hurdle or the zero threshold. The first part models the probability that the threshold is crossed. In principle, the threshold need not be at zero; it could be any value. Further, it need not be treated as known. The zero value has special appeal because in many situations it partitions the population into subpopulations in a meaningful way.” So, a data set is split into zero and non-zero (positive) values to fit two different model with associated covariates in regression. A variety of probability distribution can be considered for zero counts, and frequent use in real-life data are binomial distribution, Poisson distribution, and negative binomial distribution. Also, for the positive count data frequent use of probability distributions are Poisson distribution, negative binomial distribution and geometric distribution. Cameron and Trivedi (1998) defined the probability functions as:

\[
Pr[y = 0] = f_1(0) \tag{2.8}
\]
\[
Pr[y = j] = \frac{1 - f_1(0)}{1 - f_2(0)} f_2(y), \quad j > 0 \tag{2.9}
\]

which collapse to the standard model, as well as, to a single model if \( f_1(\cdot) = f_2(\cdot) \)

The basic idea is that a binomial probability governs the binary outcome of whether a count variate has a zero or a positive realization. If the realization is positive, the hurdle is crossed, and the conditional distribution of the positives is governed by a truncated-at-zero
count data model. Mullahy (1986) provided the general form of hurdle count regression models, together with applications to daily consumption of various beverages.

The mean and variance of the hurdle model can be shown as:

\[
E[y|x] = Pr[y > 0|x]E_{y>0}[y > 0, x] \\
V[y|x] = Pr[y > 0|x]V_{y>0}[y > 0, x] + Pr[y = 0|x]E_{y>0}[y > 0|x]
\] (2.10) (2.11)

Cameron and Trivedi (1998) showed the probability function of the hurdle model where they considered negative binomial distribution of second kind (NB2) (that is, of quadratic variance function). Let \( \mu_{1i} = \exp(x\beta_1) \) be the NB2 mean parameter for the case of zero counts. Similarly, let \( \mu_{2i} = \mu_2(x\beta_2) \) for the positive set \( J = \{1, 2, \ldots\} \). Further define the indicator function \( I[y_i \in J] = 1 \) if \( y_i \in J \) and \( I[y_i \in J] = 0 \) if \( y_i = 0 \). From the NB distribution with a quadratic variance function, the following probabilities can be obtained:

\[
Pr[y_i = 0|x_i] = (1 + \alpha_1 \mu_{1i})^{-1/\alpha_1}; \\
1 - Pr[y_i = 0|x_i] = \sum_{y_i \in J} h(y_i|x_i) = 1 - (1 + \alpha_1 \mu_{1i})^{-1/\alpha_1}; \\
Pr[y_i|x_i, y_i > 0] = \frac{\Gamma(y_i + \alpha_2^{-1})}{\Gamma(\alpha_2^{-1})\Gamma(y_i + 1)} \left( \frac{1}{1 + \alpha_2 \mu_{2i}} \right)^{-\alpha_2^{-1}} \left( \frac{\mu_{2i}}{\mu_{2i} + \alpha_2^{-1}} \right)^{y_i}
\] (2.12) (2.13) (2.14)

The equation in (2.12) gives the probability of zero counts, while (2.13) is the probability that the threshold is crossed. Equation (2.14) is the truncated-at-zero NB2 distribution.
The log-likelihood function splits in two components for zeros and positive counts as:

\[ l_1(\beta_1, \alpha_1) = \sum_{i=1}^{n} [(1 - I[y_i \in J])] ln(Pr[y_i = 0|x_i]) \]
\[ + \sum_{i=1}^{n} I[y_i \in J] ln(1 - Pr[y_i = 0|x_i]) \]
\[ l_2(\beta_2, \alpha_2) = \sum_{i=1}^{n} I[y_i \in J] ln(Pr[y_i > 0|x_i]) \]

and, \( l(\beta_1, \beta_2, \alpha_1, \alpha_2) = l_1(\beta_1, \alpha_1) + l_2(\beta_2, \alpha_2) \)

(2.15)

Here \( l_1(\beta_1, \alpha_1) \) is the log-likelihood for the binary process that splits the observations into zeros and positives, and \( l_2(\beta_2, \alpha_2) \) is the likelihood function for the truncated negative binomial part for the positives. Cameron and Trivedi (1998, p. 125) argued that, the two mechanisms are assumed (functionally) independent, so the joint likelihood can be maximized by separately maximizing each component. The Poisson hurdle and the geometric hurdle models examined in Mullahy (1986) can be obtained from (2.12) through (2.14) by setting \( \alpha_1 = \alpha_2 = 0 \) and \( \alpha_1 = \alpha_2 = 1 \), respectively. Note that when \( \alpha_1 = 1, Pr[y_i = 0|x_i] = (1 + \mu_{1i})^{-1} \), so that if \( \mu_{1i} = exp(x_i \beta_1) \), the binary process model is a logit model.

2.4 Zero-inflated Generalized Poisson (ZIGP) regression

The generalized Poisson distribution has first been introduced by Consul and Jain (1970). ZIGP models have recently been found useful for the analysis of count data with a large amount of zeros and proposed by Famoye and Singh (2006) with an application to domestic violence data. In the case of overdispersed data with too many zeros, ZINB model may fit the data well than ZIP model. But the iterative technique to estimate the parameters of ZINB model sometimes fails to converge; Lambert (1992) also pointed out

The generalized Poisson regression model for mean \( \mu_i \) and for fixed dispersion parameter \( \alpha \) can be written as:

\[
f(y_i, \mu_i, \alpha) = \left( \frac{\mu_i}{1 + \alpha \mu_i} \right)^{y_i} \cdot \frac{(1 + \alpha y_i)^{y_i-1}}{y_i!} \cdot \exp \left[ -\frac{\mu_i(1 + \alpha y_i)}{1 + \alpha \mu_i} \right]
\] (2.16)

Here, \( y_i = 0, 1, 2, ... \) for \( i = 1, 2, ..., \mu_i = e^{x\beta}, x \) is the covariate matrix and \( \beta \) are unknown regression parameters. Mean and variance of generalized Poisson regression (GPR) model are \( E(y_i|x_i) = \mu_i \) and \( V(y_i|x_i) = \mu_i(1 + \alpha \mu_i) \), respectively. Note that when \( \alpha = 0 \) the probability model reduces to the Poisson regression model. When \( \alpha > 0 \), the GPR model in (2.16) represents count data with overdispersion and when \( \alpha < 0 \) then the model represents underdispersion.

Famoye and Singh (2006) proposed the ZIGP model as:

\[
P(Y = y_i|x_i, z_i) = \varphi_i + (1 - \varphi_i)f(y_i, \mu_i, \alpha), \quad y_i = 0
\]

\[
= (1 - \varphi_i)f(y_i, \mu_i, \alpha), \quad y_i > 0
\] (2.17)

Still \( \mu_i = e^{x\beta} \), and the new parameter \( \varphi_i \) is the probability of \( y_i = 0 \) for given \( z_i \). Here \( z_i \) are the the covariate matrix which are responsible for zeros in \( y_i \). \( x \) and \( z \) can be identical or may have some common covariates. The link function for \( \varphi \) is \( \text{logit}(\varphi) = \log\left(\frac{\varphi}{1-\varphi}\right) = z\gamma \). Here \( \gamma \) is the unknown regression parameter for \( Z \). The mean and variance of the ZIGP
model in [2.17] are given,

\[ E(y_i|x_i) = (1 - \varphi_i)\mu_i \tag{2.18} \]
\[ V(y_i|x_i) = (1 - \varphi_i)[\mu_i^2 + \mu_i(1 + \alpha\mu_i)^2] - (1 - \varphi)^2\mu_i^2 \tag{2.19} \]

Note that the model reduces to GPR model if \( \varphi = 0 \). It reduces to ZIP if \( \alpha = 0 \). The log-likelihood for ZIGP model is given by:

\[
\begin{align*}
\log l(\beta, \gamma, \alpha) &= -\sum_{i=1}^{n} \log(1 + \omega_i) + \sum_{y_i=0} \log(\omega_i + v_i) \\
&\quad + \sum_{y_i>0} \left\{ y_i \log(\xi_i) + (y_i - 1) \log(1 + \alpha y_i) - \log(y_i!) - \xi_i(1 + \alpha y_i) \right\} \tag{2.20}
\end{align*}
\]

Here, \( \xi_i = \frac{\mu_i}{1 + \alpha\mu_i}, \omega_i = e^{-\xi_i}, \omega_i = \frac{\varphi_i}{1 - \varphi_i} = e^{\gamma} \). The log-likelihood function can be used to get score function and Fisher’s information matrix to estimate the parameters \( \alpha, \beta \) and \( \gamma \) for the given covariate matrix \( \mathbf{x} \) and \( \mathbf{z} \). Numerical algorithms can be used to estimate the parameters.
Chapter 3

The Simulation Experiment

3.1 The simulation experiment on zero-inflated count data generated from Poisson distribution

Lambert(1992) discussed the ZIP regression model and likelihood functions of the model with the estimation techniques. The author performed a simulation experiment to understand the asymptotic results of the parameter estimation along with the properties of MLE’s, behavior of confidence interval and tests. The author performed the experiment for 2,000 times on finite sample of size (n) of 25, 50, and 100 with a single covariate x taken from n uniformly spaced values between 0 and 1. Lambert assigned the two sets of parameters, \( \gamma = (-1.5, 2) \) and \( \beta = (1.5, 2) \). The response variable generated \( y \) by first drawing a uniform (0, 1) random vector \( U \) of length n and then assigning \( y_i = 0 \) if \( U_i \geq p_i \), otherwise \( y_i \sim \text{Poisson}(\lambda_i) \).

We have simulated zero-inflated count data which accounted three covariates with random values for \( \beta \) and \( \gamma \). The underlying interest of doing this was to see the performance of different count data models for different parameter values. We have considered three popular zero-inflated count models, ZIP, ZINB and Hurdle, and classical count data models,
Poisson and Negative Binomial regression model. Our assumption was zero-inflated count data models will have superior performance in model fitting if proportion of zeros is large enough (say, more than 30% than the rest of the counts) regardless of the values of the parameters $\beta$ and $\gamma$. In that case, we are also interested to see if there any influence of parameter values in the performance between ZIP, ZINB and Hurdle models. We know, ZIP and ZINB should have similar performance if the data are not over-dispersed. Hurdle model is free of over-dispersion and should also have similar performance as the other two. In presence of over-dispersion in the data we are also interested to see the performance of each count data models.

We considered three covariates, two of them were continuous and the third one was binary. The two continuous covariates generated from normal distribution truncated at zero with mean 0 and 10, respectively, and with standard deviation 1 and 4, respectively. The binary covariate was generated from Bernoulli distribution with success probability 0.7. We considered random values for the parameters within a specified boundary in each trial of the simulations. We considered $\beta = ((0.01 \text{ to } 1.0), (0.5 \text{ to } 1.5), (-0.04 \text{ to } -0.02), (0.4 \text{ to } 1.0))$ and $\gamma = ((-1.3 \text{ to } -0.7), (-1.0 \text{ to } 0.01), (0.01 \text{ to } 0.02), (0.01 \text{ to } 1.5))$. In each trial it takes a random set of parameters to generate $\lambda_i$ and $p_i$ where $\lambda = exp(\beta X)$ and $\gamma = logit(p)$. The response variable $y_i$ was generated in the same way as by Lambert. We repeated the trial for 1,000 times for sample size 100. The values of $\lambda$ varies from 0.62 to 13.64 while the values of $p$ varies from 0.18 to 0.58. We have fitted ZIP, ZINB and Hurdle regression models using \texttt{zeroinfl()} and \texttt{hurdle()} functions through \texttt{pscl} package developed by Jackman (2012) [14] available from the Comprehensive R Archive Network (CRAN) at \url{http://CRAN.R-project.org/package=pscl}. The Poisson and negative binomial regression models are fitted in this study using R [20] by the \texttt{glm()} function developed by Chambers and Hastie (1992) [4] in the \texttt{stats} package and using the \texttt{glm.nb()} function in
In this study we were interested to explore how the available zero-inflated count data models behave with different sets of parameters. Figure 3.1 shows that each zero-inflated models have superior performance than the classical count data models, that is, Poisson and negative binomial models.

And it does not affected by neither $\lambda$ nor $p$ for any of the zero-inflated models. Moreover, if lambda increases but proportion of zeros remain high then zero-inflated models perform much better. Among the three zero-inflated models, their performance are indistinguishable according to Vuong test \[24\]. In fact, ZIP and ZINB have identical performance as the
dispersion was one for data generated from Poisson with inflated zeros. But ZINB is sensitive to convergence issue. However, Hurdle models also perform along with the rest of the two zero-inflated models. As ZIP and ZINB shows exactly same performance, we became interested to generate data from zero-inflated negative binomial and compare the performance of these models. We have discussed this issue in the next section.

It may not be a good choice to analyze the relative error using estimated mean vs true mean. As the main interest centered on counts rather than estimate of mean, we may consider how the models are performing in predicting the frequency of each counts, for example, number of zeros, number of one, and so on. We have mentioned earlier that, as we generated data using ZIP distribution the dispersion is very close to one. So ZIP and ZINB have almost identical results. Additionally, if we consider Poisson distribution with logit link for Hurdle model, it should also have identical estimate of parameters as ZIP model. In predicting the counts ZIP, ZINB and Hurdle demonstrated their superiority than the classical models as expected. The main difference between the zero-inflated models and the classical models were to estimate the number of zeros. When the true percentages of zeros varies from 20% to 60% and zero-inflated models almost perfectly predicted this frequency while the classical models never able to predict the true percentage of zeros better. Among the Poisson and negative binomial models, sometimes negative binomial able to predict the zeros better than Poisson, especially when the data is slightly deviated from unit dispersion (Figure 3.2). Among the five models Hurdle model have outstanding performance in predicting the zero counts. It almost perfectly predicted the frequency of zeros in each simulation experiment.
Instead of considering only zero counts we may consider each of the counts and estimate mean square error (MSE) of each models to compare their performance. If a model works better the MSE should be close to zero, if not it should be away from the zero to the right. All three zero-inflated count data models, ZIP, ZINB, and Hurdle had nearly zero MSE while Poisson and negative binomial had very large MSE (Figure 3.3). Sometimes negative binomial performed better but many times it fails.
3.2 The simulation experiment on zero-inflated count data generated from negative binomial distribution

In the previous section we have discussed whether there is any effect of different parameter values on the performance of the models. We have noticed that though we have generated data from Poisson distribution with inflated zeros, it performs equally better by ZINB and Hurdle models as ZIP model. One possible reason might be unit dispersion of the data, and ZINB acts like ZIP in such case. When over-dispersion is under consideration, that is, if the count data set exhibits more zero counts that than would be allowed by Poisson model, we may consider zero-inflated generalized Poisson model (ZIGP) (proposed by Famoye and
Singh (2006) [9]) or Hurdle model (proposed by Mullahy (1986) [17]). One difficulty with the ZIGP model is we have to have some prior knowledge which variables are responsible for over-dispersion. It may not known in advance in some studies. On the other hand, the Hurdle model can take account of over-dispersion. It is a two-component models: A truncated count component, such as Poisson, geometric or negative binomial, is employed for positive counts, and a hurdle component models zero vs. larger counts. For the latter, either a binomial model or a censored count distribution can be employed. ZINB should also be able to take the over-dispersion in presence of excess zeros into account, but it has a convergence problem in some situation.

In this section we have generated zero-inflated count data from negative binomial distribution and fit the three ZI models and two classical count data models. We are motivated to do so as data generated from ZI Poisson and ZI negative binomial are apparently identical. In Figure 3.4 we have shown two histogram of ZI count data generated from Poisson and negative binomial for sample size 1,000 with same mean (5), dispersion 1 and 10, respectively, and proportion of excess zeros 40%. It is hard to differentiate between two graphs. This motivated us to check model efficiency on data generated from negative binomial distribution with excess zeros. We have considered similar covariates as Poisson generated data. The dispersion parameter set at 2 to generate 100 samples for different $\beta$ and $\gamma$ values within specified range for 1,000 times for the simulation experiment.
We consider AIC statistic to compare efficiency of each models. Figure 3.5 shows whether increase of $\lambda$ have any impact on the AIC statistic of each models. ZINB and Hurdle models almost overlap with each other and their AIC statistic do not affected by the changes of $\lambda$. ZIP and Poisson models performed same as ZINB and Hurdle for small $\lambda$ while failed to keep the AIC statistic lower as $\lambda$ increases. AIC statistic for negative binomial model improved a lot than the earlier simulation experiment when data generated from Poisson, and became a competitive of ZINB and hurdle models. Proportion of zeros does not have impact on the performance of the models.
The AIC statistic over the 1,000 simulation run shows that ZINB always performed better than the rest of the models. It was quite surprising that ZIP models did not fit well with the data even if the data is zero-inflated while negative binomial showed huge improvements. Only a unit change to the right of the dispersion parameter distorted the performance of ZIP for any set of the the parameter values.

In the area of studies where predicting the frequency of zeros have higher interest may require additional care in choosing the model. Figure 3.6 shows that Hurdle model predicted the zero counts almost perfectly while ZIP and ZINB over and under-estimated around 5%. Though negative binomial model fitted very well in terms of AIC statistic, did not able to
estimate the frequency of zeros well. Poisson had worse performance among the rest of the four models in predicting the zero counts.

![Prediction of zero counts by each model](image)

Figure 3.6: Prediction of zero counts by each models in 1,000 simulation run (data generated from ZINB)

Instead of considering only zero counts we may consider each of the counts and estimate mean square error (MSE) of each models to compare their performance. If a model works better the MSE should be close to zero, if not it should be away from the zero to the right. Hurdle model had the least MSE while ZINB and negative binomial model have nearly same performance (Figure 3.7). Though negative binomial model did not performed very well in predicting zero, it predicted the higher count better than Poisson and ZIP models. ZIP model has slightly higher MSE than NB and Hurdle models. Poisson showed very high MSE.
Figure 3.7: MSE of the count data models in 1,000 simulation run (data generated from ZINB)
Chapter 4

Under-5 deaths in Bangladesh

4.1 A brief overview on BDHS 2011 survey

Bangladesh Demographic and Health Survey (BDHS) 2011 is the sixth survey of its kind conducted in Bangladesh. The 2011 BDHS is a nationwide sample survey of men and women of reproductive age that provides information on childhood mortality levels, fertility preferences, use of family planning methods, and maternal, child, and newborn health. The sample for the 2011 BDHS is nationally representative along with representative for each divisions (there are seven administrative divisions in Bangladesh). 17,842 women in reproductive age (12-49 years of age) and 3,997 eligible men were interviewed in BDHS 2011 survey [19].

4.2 Under-5 mortality and it’s determinants

DHS surveys conducted in Bangladesh since 1993 confirm a declining trend in under-5 mortality. Between the period 1989-1993 and 2007-2011 under-5 mortality decreased from 133 to 53 per 1,000 live births. BDHS 2011 survey explored the socioeconomic differentials of under-5 mortality. It reveals that the mothers education plays a critical role in lowering the under-5 death rate. Mothers who have passed secondary level (grade 10) of education have experienced nearly half of under-5 deaths as compared to mothers who have no education.
Similar trend for wealth quintile where under-5 death rate in highest quintile is 37 and in lowest quintile 64 per 1,000 live births. BDHS 2011 investigated causes of death of under-5 death using verbal autopsy. It reported the leading causes of death of under-5 is pneumonia (22%), followed by sepsis (15%), birth asphyxia (12%), drowning (9%), and pre-term birth (7%).

Survival of a child may depend on various factors including its personal characteristics, biological conditions, knowledge and practices of it's parents, environmental issues, strength of regional and national health systems, and many more. Mosley and Chen (1984) discussed an analytical framework to study child survival in developing countries. They emphasized on considering socioeconomic variables with medical and biological variables to understand child survival. BDHS 2011 also implemented this framework to investigate the child mortality in the last five years of the survey. BDHS 2011 survey collected biological and medical data of children who had born in last five years of the survey. As there is no standard data repository system exists in Bangladesh, mothers or care-givers are the only source of data for children. To avoid possible bias due to lack of evidence or lack of authentic memory of the mother, BDHS 2011 survey did not collected data for children who had born earlier than 5 years of the survey. Our study considered number of under-5 deaths experienced by each woman having at least one child. So our study data contains information of under-5 deaths prior than the last five years of the survey. In that case, due to completeness of the data we have to use only those variables which are available for all under-5 deaths. Biological data of a child such as causes of death, low birth weight (LLB), premature delivery, and immunization are not included in our study. Also mother’s birth specific practices, such as, antenatal care (ANC), postnatal care (PNC), place of delivery, and attendance of the delivery are not included in the study. We considered women in reproductive age (12-49 years of age) who gave at least one live birth in their lifetime. As we are modeling expected number of under-5 deaths experienced by a woman for some given socioeconomic variables, women with no live birth are not legitimate for such model.
Numerous studies confirm that socioeconomic variables, such as, *mother’s level of education, father’s level of education, household’s wealth status, mother’s age at each births, child’s sex, birth order, and preceding birth interval* may have causal association with under-5 death in developing countries.

### 4.3 Exploratory analysis of under-5 death

We have considered only those women who gave at least one live birth in their reproductive age. Table 4.1 presents the descriptive statistics of the variables.

#### 4.3.1 Why it is important to model zero counts and positive counts separately

From the policy making view point, it is important to investigate why some mothers experienced under-5 deaths of their children while some mothers never experienced such events. Medical reasons such as pneumonia, sepsis, birth asphyxia are the leading causes of under-5 deaths. These causes of death may also be influenced by socioeconomic characteristics in individual level and household level, such as, *mothers health seeking behavior, parents level of education, accessibility of health services, and strength of the health system of that region*. In this study we are considering the socioeconomic characteristics in individual level and household level to find the causal association with under-5 deaths.

In this study we considered three zero-inflated count data models and two classical count data models. Table 4.2 presents the number of under-5 deaths of 16,004 women surveyed in BDHS 2011. The data seems to be a good candidate to be fitted by zero-inflated count data models where nearly 80% of the women never experienced under-5 death of their children.
Table 4.1: Descriptive statistics of the variables, BDHS 2011

<table>
<thead>
<tr>
<th>Variables</th>
<th>Categories</th>
<th>Mean/percentages</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mothers age</td>
<td></td>
<td>31.85</td>
<td>16,004</td>
</tr>
<tr>
<td>Mothers age in five year categories</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>15-19</td>
<td>6.44</td>
<td>1031</td>
</tr>
<tr>
<td></td>
<td>20-24</td>
<td>18.60</td>
<td>2976</td>
</tr>
<tr>
<td></td>
<td>25-29</td>
<td>20.13</td>
<td>3222</td>
</tr>
<tr>
<td></td>
<td>30-34</td>
<td>16.32</td>
<td>2612</td>
</tr>
<tr>
<td></td>
<td>35-39</td>
<td>14.07</td>
<td>2252</td>
</tr>
<tr>
<td></td>
<td>40-44</td>
<td>13.19</td>
<td>2111</td>
</tr>
<tr>
<td></td>
<td>45-49</td>
<td>11.25</td>
<td>1800</td>
</tr>
<tr>
<td>Highest education level by mother (medu)</td>
<td>No education</td>
<td>27.76</td>
<td>4442</td>
</tr>
<tr>
<td></td>
<td>Primary</td>
<td>30.73</td>
<td>4918</td>
</tr>
<tr>
<td></td>
<td>Secondary</td>
<td>34.25</td>
<td>5482</td>
</tr>
<tr>
<td></td>
<td>Higher</td>
<td>7.26</td>
<td>1162</td>
</tr>
<tr>
<td>Highest education level by husband</td>
<td>No education</td>
<td>30.55</td>
<td>4889</td>
</tr>
<tr>
<td></td>
<td>Primary</td>
<td>27.32</td>
<td>4373</td>
</tr>
<tr>
<td></td>
<td>Secondary</td>
<td>28.05</td>
<td>4489</td>
</tr>
<tr>
<td></td>
<td>Higher</td>
<td>14.08</td>
<td>2253</td>
</tr>
<tr>
<td>Household’s wealth index (wealth)</td>
<td>Lowest</td>
<td>17.91</td>
<td>2867</td>
</tr>
<tr>
<td></td>
<td>Second</td>
<td>18.72</td>
<td>2996</td>
</tr>
<tr>
<td></td>
<td>Middle</td>
<td>19.19</td>
<td>3071</td>
</tr>
<tr>
<td></td>
<td>Fourth</td>
<td>20.87</td>
<td>3340</td>
</tr>
<tr>
<td></td>
<td>Highest</td>
<td>23.31</td>
<td>3730</td>
</tr>
<tr>
<td>Number of births at teen age (nbteen)</td>
<td>No</td>
<td>28.89</td>
<td>2867</td>
</tr>
<tr>
<td></td>
<td>One</td>
<td>46.36</td>
<td>2996</td>
</tr>
<tr>
<td></td>
<td>Two</td>
<td>20.12</td>
<td>3071</td>
</tr>
<tr>
<td></td>
<td>Three or more</td>
<td>4.63</td>
<td>3340</td>
</tr>
<tr>
<td>Number of births at less than 24 months of preceding birth interval (prec24)</td>
<td>No</td>
<td>28.89</td>
<td>4623</td>
</tr>
<tr>
<td></td>
<td>One</td>
<td>46.36</td>
<td>7420</td>
</tr>
<tr>
<td></td>
<td>Two</td>
<td>20.12</td>
<td>3220</td>
</tr>
<tr>
<td></td>
<td>Three or more</td>
<td>4.63</td>
<td>741</td>
</tr>
<tr>
<td>Number of births after the third (bord3)</td>
<td>No</td>
<td>71.97</td>
<td>11518</td>
</tr>
<tr>
<td></td>
<td>One</td>
<td>19.34</td>
<td>3095</td>
</tr>
<tr>
<td></td>
<td>Two</td>
<td>5.45</td>
<td>878</td>
</tr>
<tr>
<td></td>
<td>Three or more</td>
<td>3.23</td>
<td>513</td>
</tr>
</tbody>
</table>

Primary defined as passed grade 5
Secondary defined as passed grade 10
Table 4.2: Frequency of number of under-5 deaths in BDHS 2011 survey

<table>
<thead>
<tr>
<th>Number of deaths</th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>12,769</td>
<td>79.79</td>
</tr>
<tr>
<td>1</td>
<td>2,442</td>
<td>15.26</td>
</tr>
<tr>
<td>2</td>
<td>563</td>
<td>3.52</td>
</tr>
<tr>
<td>3</td>
<td>155</td>
<td>0.97</td>
</tr>
<tr>
<td>4</td>
<td>44</td>
<td>0.27</td>
</tr>
<tr>
<td>5</td>
<td>23</td>
<td>0.14</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>0.02</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>0.02</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>16,004</strong></td>
<td><strong>100.00</strong></td>
</tr>
</tbody>
</table>

Mothers current age may have effects on the degree of experiencing under-5 deaths. Women with higher age are likely to give more births, and giving more births raise the risk of experiencing under-5 deaths. Also experiencing a child’s death may influence parents to give more births. However, both variables, mother’s current age and total number of live births of a mother, may have influence on never experiencing under-5 deaths with opposite slope. Figure 4.1 might be a motivational one for employing zero-inflated count data models in such case. The trend shows irregular pattern after seven births due to smaller number of mothers available for the latter groups (right panel in figure 4.1).

Parents level of education may have effects on the degree of experiencing under-5 deaths. In BDHS 2011 survey, we have found that mothers who have more than 10 years of education (more than secondary) experienced 5% of under-5 deaths while mothers with no education (less than 5 years of education) experienced 25% of under-5 deaths. Similar pattern is found for father’s education. Also, parents who never experienced under-5 death showed similar trend with opposite slope (Figure 4.2). This finding was also motivational to employ zero-inflated count data models where factors affecting zero counts can be modeled separately.

While public health concern centered with controlling under-5 death rate, it is important to analyze why some women never experience under-5 deaths rather than which factors are responsible for lowering the under-5 deaths.
Figure 4.1: Percentage of women experiencing under-5 deaths by age and number of live births: How zero counts differ than the rests

Figure 4.2: Percentage of women experiencing under-5 deaths by parents level of education: How zero counts differ than the rests
Wealth index (also known as asset quintile) is another factor which have substantial effect on under-5 deaths \[19\]. Women who are from those households that belongs to the lowest 20% according to the wealth index had higher experience of under-5 deaths than the women who are from those households that belongs to the highest quintile. Again, the slope for the percentage of women who experience no under-5 deaths, are showing a trend with opposite slope of women who had at least one experience of under-5 deaths (Figure 4.3).

![Figure 4.3: Percentage of women experiencing under-5 deaths by wealth index: how zero counts differ than the rest](image)

Early marriage is a risk factor for under-5 deaths. The legal age of marriage for women in Bangladesh is 18 years. Nearly one third of the women are married before their 18\textsuperscript{th} birthday in 2010 in a rural area of Bangladesh [11]. Considering all women in BDHS 2011 survey, we have found that nearly 70 percent of the women gave birth by age 20 [19]. Figure 4.4 shows that women who have higher incidence of giving birth at teen age have experienced higher proportion of under-5 deaths. Also women who did not have given birth in their teen age have substantially lower proportion of under-5 deaths.
Studies suggested that optimal birth interval should be 36 months to 59 months to avoid the under-5 deaths and undernutrition for a birth \(^{22}\) in developing countries. BDHS 2011 survey reported very few number of mothers had birth interval more than 59 months. On the other hand, nearly one third of the women had at least one event of shorter preceding birth interval (less than 24 months). Figure 4.5 shows that women with shorter birth spacing is causally associated with number of under-5 deaths. Also shorter birth spacing associated with ‘no under-5 deaths’ with negative slope in terms of percentages of women.

Figure 4.4: Percentage of women experiencing under-5 deaths by number of births at teen age: how zero counts differ than the rests
Figure 4.5: Percentage of women experiencing under-5 deaths by number of births at teen age: how zero counts differ than the rests
Chapter 5

An Application of Zero-inflated Models in Under-5 Deaths in Bangladesh

In this chapter we are going to discuss an application of zero-inflated count data models. Nearly 20% mothers in Bangladesh had at least one event of under-5 death in their reproductive age (source BHDS 2011 [19]). BDHS 2011 is a cross sectional study on around 18,000 women of reproductive age (15-49 years of age). It is a certain event with probability one of zero under-5 deaths for those mother who never gave a live birth. So for this study we have considered only those women who gave at least one birth. 16,004 of such women were available. We consider number of under-5 deaths of each mother as a response variable. The risk factors for under-5 deaths are as follows: mother’s level of education (medu), number of births during teen age (nbteen), number of births with preceding birth interval less than 24 months (prec24), number of births with birth order more than three (bord3), and household’s wealth index (wealth) for that mother. We could not use some important variables as recommended in many studies to understand under-5 deaths, such as, father’s level of education, residence of the mother, mother’s age, mother’s age at first birth, and total number of births;
either to avoid multicollinearity or they were not significant in any models. Some important variables which might be categorized as biological variables of the mother, such as, mother’s BMI (Body Mass Index as a measure of nutritional status) in each pregnancy period, any complication during each pregnancy, place of birth, birth attended by, number of ante-natal care (ANC) taken, number of post-natal care (PNC) taken, and whether the birth is a multiple births; are not considered as they were not available in the BDHS 2011 survey for all mothers over the reproductive age. Some important child specific biological variables also not considered in this study as we considered mothers as our response subject. Exploratory analysis in Chapter 3 motivated us to consider the selected variables to be included in the models. Also the exploratory analysis show some evidence to consider zero-inflated count data models to model under-5 deaths per women with associated factors.

5.1 Estimation of coefficients

We have fitted ZIP, ZINB and Hurdle regression models on the under-5 mortality data as around 80% of the mothers in BDHS 2011 survey have never experienced any under-5 deaths. The data seems to be a good candidate for the ZI models. All the three ZI models also can be modeled with different link functions, namely logit, probit, cloglog (complementary log log), and cauchit (Cauchy) to account the zero counts. For this study we have considered logit link. The link functions neither improved the model nor have practical inclination with this kind of study. For positive counts we have considered Poisson and negative binomial distribution for ZIP and ZINB models, respectively. Hurdle model has a bit more flexibility. We have discussed it earlier that it is a two component model and zero counts can be model separately with the mentioned link functions while the positive counts can be modeled using Poisson, negative binomial or geometric distribution. To determine which distribution works better for Hurdle model, we performed Wald test (also log-likelihood ratio test, though both produce same conclusion) with different distribution for positive counts and zero counts in
Hurdle model. We also had to determine which variables can take account of excess zeros. As we have shown in Chapter 4 all of the variables selected for modeling have influence in controlling the event of under-5 death of a mother while each of these variables have opposite slope in number of such events of a mother. Such exploration motivated us to consider all the five variables for excess zeros and also for positive counts. After doing so, we have found that ‘wealth’ does not have significant effect on positive counts while it has somewhat significant effects on zero counts only for Hurdle models. Wald test also suggested no significant improvement on the models if we use ‘wealth’ variable in the model for positive counts. We kept the wealth variable for ZIP and ZINB also for comparison purpose of the models. We have used the \texttt{pscl} package to fit these models in \texttt{R}. We also fitted classical count data models, Poisson and negative binomial to compare our results. Table 5.1 shows the parameter estimates of each models with associated standard error.

Table 5.1 shows that mothers level of education is a protective factors for number of under-5 deaths of a mother. Number of births at teen age, number of births with less than 24 months of the preceding birth, and number of births after the third are the risk factors for number of under-5 births of a mother. All the ZI models and classical models showed a consensus on that with slightly different estimate of the coefficients but with same directions. Notice that ZIP and ZINB with logit link have almost similar estimates of the parameters. It might be due to the fact that there is nearly unit dispersion in the data. On the other hand, the Hurdle model with negative binomial distribution for positive counts and logit link for the zero counts have the lowest AIC among the models. Estimates of the parameters for zero counts is only available for ZI models. Mother’s level of education and wealth index, (the higher the better), are the significant covariates for a mother to have zero under-5 deaths. While number of births at the teen age, shorter than 24 months of preceding birth interval, and number of births after the third, are the responsible covariates for a mother of not possessing zero under-5 births. Notice that, Hurdle model have different sign for the
Table 5.1: Estimation of coefficients using different count data models for under-5 mortality in Bangladesh, BDHS 2011

<table>
<thead>
<tr>
<th>Models:</th>
<th>ZIP</th>
<th>ZINB</th>
<th>Hurdle</th>
<th>Poisson</th>
<th>Negative Binomial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covariates for positive counts (Intercept)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>−1.21***</td>
<td>−1.21***</td>
<td>−1.66***</td>
<td>−1.69***</td>
<td>−1.76***</td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.12)</td>
<td>(0.04)</td>
<td>(0.05)</td>
</tr>
<tr>
<td>medu1</td>
<td>−0.06</td>
<td>−0.06</td>
<td>−0.11</td>
<td>−0.16***</td>
<td>−0.17***</td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.07)</td>
<td>(0.04)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>medu2</td>
<td>−0.19**</td>
<td>−0.19**</td>
<td>−0.26*</td>
<td>−0.65***</td>
<td>−0.61***</td>
</tr>
<tr>
<td></td>
<td>(0.06)</td>
<td>(0.06)</td>
<td>(0.12)</td>
<td>(0.05)</td>
<td>(0.05)</td>
</tr>
<tr>
<td>medu3</td>
<td>−0.29</td>
<td>−0.29</td>
<td>0.00</td>
<td>−1.01***</td>
<td>−0.94***</td>
</tr>
<tr>
<td></td>
<td>(0.21)</td>
<td>(0.21)</td>
<td>(0.35)</td>
<td>(0.12)</td>
<td>(0.13)</td>
</tr>
<tr>
<td>wealth2</td>
<td>−0.01</td>
<td>−0.03</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wealth3</td>
<td>−0.10*</td>
<td>−0.09</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wealth4</td>
<td>−0.08</td>
<td>−0.10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wealth5</td>
<td>−0.17**</td>
<td>−0.18**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nbteen</td>
<td>0.11***</td>
<td>0.11***</td>
<td>0.11***</td>
<td>0.18***</td>
<td>0.18***</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.03)</td>
<td>(0.02)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>prec24</td>
<td>0.12***</td>
<td>0.12***</td>
<td>0.16***</td>
<td>0.15***</td>
<td>0.21***</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.03)</td>
<td>(0.02)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>bord3</td>
<td>0.22***</td>
<td>0.22***</td>
<td>0.27***</td>
<td>0.30***</td>
<td>0.32***</td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.02)</td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>Log(theta)</td>
<td>8.35</td>
<td>1.80***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(7.66)</td>
<td>(0.39)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Covariates for zero counts (Intercept) |  |  |  |  |  |
| | 0.39* | 0.39* | −0.29 | | |
| | (0.17) | (0.17) | (0.69) | | |
| medu1 | 0.73*** | 0.73*** | −1.39*** | | |
| | (0.16) | (0.16) | (0.41) | | |
| medu2 | 1.04*** | 1.04*** | −2.15*** | | |
| | (0.17) | (0.17) | (0.52) | | |
| medu3 | 0.90* | 0.90* | −1.99*** | | |
| | (0.35) | (0.35) | (0.53) | | |
| wealth2 | 0.04 | 0.04 | −0.26 | | |
| | (0.17) | (0.17) | (0.30) | | |
| wealth3 | 0.22 | 0.22 | −0.54 | | |
| | (0.17) | (0.17) | (0.30) | | |
| wealth4 | 0.28 | 0.28 | −0.68* | | |
| | (0.17) | (0.17) | (0.30) | | |
| wealth5 | 0.29 | 0.29 | −0.68* | | |
| | (0.18) | (0.18) | (0.31) | | |
| nbteen | −0.20*** | −0.20*** | 0.78*** | | |
| | (0.08) | (0.08) | (0.17) | | |
| prec24 | −2.55*** | −2.55*** | 4.23*** | | |
| | (0.32) | (0.32) | (1.12) | | |
| bord3 | −3.08*** | −3.07*** | 5.03*** | | |
| | (0.53) | (0.53) | (1.40) | | |
| Log(theta) | | | | | −3.21*** |
| | | | | | (0.28) |

| AIC | 16565.11 | 16567.14 | 16514.35 | 17424.38 | 17338.48 |
| Log Likelihood | −8264.56 | −8264.57 | −8233.18 | −8701.19 | −8657.24 |
| Num. obs. | 16004 | 16004 | 16004 | 16004 | 16004 |
| BIC | 17508.87 | 17439.65 | 17399.12 | 17399.12 | 17399.12 |
| Deviance | 10529.52 | 9421.87 | | | |

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$
estimates parameters for zero counts than the ZIP and ZINB. This is because of the way the models are defined. For the hurdle model, the zero hurdle component describes the probability of observing a positive count whereas, for the ZIP and ZINB models, the zero-inflation component predicts the probability of observing a zero count from the point mass component.

5.2 Model comparison

To compare the performance of each models we may use Voung test as the models are non-nested (Vuong (1989) [24]). Table 5.2 shows that Hurdle model is the superior one than the rests. Also ZIP and ZINB has almost similar performance and Voung test inferred that these two models does not have significant difference. It might be a reason of statistical insignificance of the estimate of dispersion parameter, log(theta).

Table 5.2: Finding the best model for under-5 mortality in Bangladesh, BDHS 2011

<table>
<thead>
<tr>
<th></th>
<th>ZIP</th>
<th>ZINB</th>
<th>Hurdle</th>
<th>Poisson</th>
<th>NB</th>
<th>Best model</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZIP</td>
<td>0.075</td>
<td>0.048*</td>
<td>0.000***</td>
<td>0.000***</td>
<td>Hurdle</td>
<td></td>
</tr>
<tr>
<td>ZINB</td>
<td>0.048*</td>
<td>0.000***</td>
<td>0.000***</td>
<td>Hurdle</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hurdle</td>
<td>0.000***</td>
<td>0.000***</td>
<td>Hurdle</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poisson</td>
<td>0.000***</td>
<td>0.000***</td>
<td>Hurdle</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NB</td>
<td>0.000***</td>
<td>0.000***</td>
<td>NB</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

***p < 0.001, **p < 0.01, *p < 0.05

5.3 A closer look on estimates of the parameters using Hurdle model

To find the best model we have considered the significant covariates in the model. But we also need to account the practical implication of a covariate. Like, exploratory analysis confirms that ‘wealth index’ may have important role in controlling under-5 deaths. Also
when we have used only ‘wealth index’ in the model, it shows significant in each of the
models. But when we considered it in presence of other covariates in the model, it became
insignificant. However, ‘wealth index’ never being exclusively significant in the ZI models,
except Hurdle model in zero counts part. Wald test also shows no improvement of the
models if we consider ‘wealth index’ in the model. As example, following is R output of
fitting the Hurdle model to the data with and without of wealth index. Under the null
hypothesis, both models are equivalent, the Wald test failed to reject the null hypothesis.

```r
> m_hrdl0 = hurdle(died5 ~ factor(medu)+factor(wealth)+nbteen+
+ prec24+bord3 | factor(medu)+factor(wealth)+nbteen+prec24+
+ bord3, data = data, dist="negbin", link="logit", zero.dist="negbin")
> m_hrdl = hurdle(died5 ~ factor(medu)+nbteen+prec24+bord3 | factor(medu)+
+ factor(wealth)+nbteen+prec24+
+ bord3, data = data, dist="negbin", link="logit", zero.dist="negbin")
> lrtest(m_hrdl0, m_hrdl)
Likelihood ratio test
Model 1: died5 ~ factor(medu) + factor(wealth) + nbteen + prec24 + bord3 | factor(medu) + factor(wealth) + nbteen + prec24 + bord3
Model 2: died5 ~ factor(medu) + nbteen + prec24 + bord3 | factor(medu) + factor(wealth) + nbteen + prec24 + bord3
  #Df LogLik Df Chisq Pr(>Chisq)
1 24 -8233.2
2 20 -8237.0 -4 7.7049 0.103
```

Notice that a mother is less likely to have a positive counts of under-5 deaths of her
children with higher level of education. The odds ratio is 0.137 ($e^{-1.99}$), that is, a mother
who have more than 10 years of education is 0.137 times less likely to experience under-5
deaths in comparison to a mother who have no education. Mothers belong to higher wealth
index also less likely to experience any under-5 death of her children and the odds ratio is
0.509 where the reference category is lowest wealth index. Mothers who gave birth in their
teen age are highly likely to experience under-5 deaths. Also note that the model showed
huge weight on shorter preceding birth interval (<24 months) and birth order more than
three. Both of these variables have huge impact on experiencing under-5 births according to
the model. On the other hand, women who have experienced under-5 deaths are more likely
Table 5.3: Estimate of parameters of Hurdle model for under-5 deaths

|                          | Estimate | Std. Error | z value | Pr(>|z|) |
|--------------------------|----------|------------|---------|----------|
| **Count model coefficients (truncated negbin with log link)** |          |            |         |          |
| (Intercept)              | −1.64    | 0.11       | −15.06  | 2e−16    |
| medu1                    | −0.09    | 0.07       | −1.33   | 0.18     |
| medu2                    | −0.28    | 0.11       | −2.44   | 0.0146   |
| medu3                    | −0.11    | 0.34       | −0.33   | 0.7432   |
| nbteen                   | 0.11     | 0.03       | 3.97    | 7.08e−05 |
| prec24                   | 0.16     | 0.03       | 5.77    | 7.90e−09 |
| bord3                    | 0.27     | 0.02       | 12.21   | <2e−16   |
| log(theta)               | 1.78     | 0.39       | 4.61    | 3.95e−06 |
| **Zero hurdle model coefficients (censored negative binomial with log link)** |          |            |         |          |
| (Intercept)              | −0.29    | 0.69       | −0.42   | 0.67     |
| medu1                    | −1.40    | 0.41       | −3.39   | 0.000693 |
| medu2                    | −2.15    | 0.52       | −4.12   | 3.73e−05 |
| medu3                    | −1.99    | 0.53       | −3.73   | 0.000190 |
| wealth2                  | −0.26    | 0.30       | −0.88   | 0.378    |
| wealth3                  | −0.54    | 0.30       | −1.80   | 0.072009 |
| wealth4                  | −0.69    | 0.30       | −2.25   | 0.024571 |
| wealth5                  | −0.67    | 0.31       | −2.20   | 0.027571 |
| nbteen                   | 0.78     | 0.17       | 4.58    | 4.58e−06 |
| prec24                   | 4.23     | 1.12       | 3.77    | 0.000164 |
| bord3                    | 5.03     | 1.34       | 3.60    | 0.000318 |
| Log(theta)               | −3.21    | 0.28       | −11.59  | <2e−16   |

| AIC                      | 16514.06 |
| Log Likelihood           | −8237.03 |
| Num. obs.                | 16004    |

***p < 0.001, **p < 0.01, *p < 0.05

to belong to the uneducated group with higher number of such experience. The incidence rate ratio (IRR) for mother who have 10 years of education is 0.896 ($e^{-0.11}$) times less likely than the mother who have no education. Giving birth at the teen age, with ‘shorter preceding birth interval’ and ‘birth order higher than three’ have significant influence on number of under-5 deaths of a mother with IRR 2.181, 1.172 and 1.305 respectively.
Chapter 6

Conclusions and discussions

From our simulation experiment we see that the zero inflated models, ZIP, ZINB and Hurdle, are consistent over the changes of the model parameters. But specification of the correct model is very important. Sometimes overdispersion of a data may not be significant if the percentage of zeros is too high (might be 80% or more) and in such case ZIP and ZINB have nearly identical estimate of the parameters. But ZIP does not fit the data well, if there is over-dispersion with moderate percentage of zeros. Hurdle model has a higher flexibility to fit a model with mixture of distribution for zeros and positive counts. And it performs in a competitive way with ZIP and ZINB.

Number of under-5 deaths of the children of a mother is a good candidate to apply zero inflated count data models. We have found that the ZI models have better performance than the Poisson and Negative binomial models in terms of AIC statistic. Although under-5 mortality shows a decreasing trend over time in Bangladesh, still it is high compared to developed countries. Higher enrollment of women in education may have some effects in lowering the rate. But birth at teen age, and shorter birth interval is still very high (see BDHS 2011 report [19]) in Bangladesh. It might be interesting to notice that wealth status of a mother is not a significant covariate to determine the number of under-5 deaths.
Various studies showed higher importance on the wealth status, but in the presence of other covariates it may become inferior, as we have found. In addition, it may have nice implication in practical life. It might be harder to eradicate poverty in short period of time, but it might be possible to control early marriage, advocate longer birth spacing, and have less number of children.

6.1 Limitation of the study

Under-5 deaths may depend on child specific variables also. Studies showed that under-weight births, place of delivery, complications during birth, and diseases including pneumonia, sepsis, and birth asphyxia with non-disease cause of death, drowning, are the responsible factors for under-5 death in Bangladesh. We were not able to consider those variables as our study set up focused on mother related factors only. Many deaths may not be explained by the mother related factors if it is solely related with child specific factors. In my future study, I will find an appropriate model to incorporate both mother specific and child specific factors to explain under-5 deaths.
Bibliography


