Introducing Hurdle_T as a Count Regression Model with High Potency in Solving Problems

Ojekudo, Nathaniel Akpofure and Akpan, Nsikan Paul

Department of Computer Science, Ignatius Ajuru University of Education, Port Harcourt, Nigeria.

Department of Mathematics and Statistics, University of Port Harcourt, Nigeria.

Authors’ contributions

This work was carried out in collaboration between both authors. Author ONA designed the study, performed the statistical analysis, wrote the protocol and wrote the first draft of the manuscript. Author ANP managed the analyses of the study, managed the literature searches. Both authors read and approved the final manuscript.

Article Information

DOI: 10.9734/ARJOM/2020/v16i530191

Editores:
(1) Dr. Krasimir Yankov Yordzhev, South-West University, Bulgaria.

Reviewers:
(1) Colins Ojwang’ Odhiambo, Strathmore University, Kenya.
(2) Zlatin Zlatev, Trakia University, Bulgaria.
(3) Janilson Pinheiro de Assis, Federal Rural University of the Semi-arid Region, Brazil.

Complete Peer review History: http://www.sdiarticle4.com/review-history/55699

Received: 10 February 2020
Accepted: 15 April 2020
Published: 08 May 2020

Abstract

Count data regression models exhibit different strengths and weaknesses in their bids to solving problems. The study considers six count models namely Poisson Regression Model (PRM), Negative Regression Model (NBRM), Zero Inflated Poisson (ZIP), Zero Inflated Negative Binomial (ZINB), Zero Truncated Poisson (ZTP) and Zero Truncated Negative Binomial (ZTNB) and an additional model called hurdle_T. These models are used to analyze two health data sets. The data on male breast cancer reveals that male breast cancer cuts across all age brackets or categories but it is more prevalent between the ages of 50 and 60. The PRM yields a better result than the NBRM in the case of cancer data as shown by the information criteria. The analysis of the second data, which is on doctor’s visit reveals that ZINB yields a better result than the other five models, followed by NBRM, then the ZTNB before their Poisson counterparts. The hurdle_T model shows the propensity of each coefficient as reflected by the positive count in the Tobit (Binary) model. The study also shows that at 65 years and above, gender has significant effect on doctor’s visit. In particular, females, more than males attract more doctors’ visit in the said age range. Government policies should provide more funds in the health sector to accommodate cancer cases in terms of the provision of awareness, studies/research and infrastructural development.

*Corresponding author: E-mail: nsipaulakpan@gmail.com;
Males should be encouraged to visit clinics especially in their late forties and above for breast cancer related checkup. At age 65 and above, doctors visit to patients are frequent, especially to females. Policy of government in the health sector should accommodate a favourable adjustment in the budget to take care of doctors’ visit.

Keywords: Count data regression; Poisson regression model; negative binomial model and Hurdle_T model.

Abbreviations (Variable Descriptions)

| Docvis | # Doctor’s Visit |
|--------|------------------|
| Private | =1 if Private Supplementary Insurance |
| Medicaid | =1 if Medical public Insurance |
| Age | Age |
| Age2 | Age Squared |
| Educyr | Years of Education |
| Actlim | =1 if Activity Limitation |
| Phylim | =1 if Physical Limitation |
| Totchr | # Chronic Condition |
| Female | =1 if female |

1 Introduction

Count is synonymous to existence of an event. For an occurrence to be counted, it means that it has to be in existence or had existed. Every event or occurrence can be expressed in a mathematical form, made up of dependent and independent variables. In many economic contents, the dependent or response variable of interest (y) is seen as a non negative integer or count which we wish to analyse in terms of a set of covariates (x). Unlike the classical regression model, the response variable is discrete, with a distribution that places probability mass at non negative integer values only. Regression models for counts, like other limited or discrete variable models such as the logit and probit are nonlinear with various properties and special features intimately connected to discreteness and nonlinearity.

A count regression involves the relationship between a set of covariates and the number of events of interest in a fixed time interval. Count data arises as outcomes of an underlying count process in continuous time. Thus, misleading outcomes may occur from data gotten from the abundance of structured and non-structured zeros [1]. [2] in summary, stated that the approaches of differentiating zero sources and modeling them based on distinct categories or structural zeros under the framework of compositional data analysis have difficulties, and challenges. In Econometric models and its application, analysis is focused on the role of regressors X introduced by specifying $\lambda = \exp(X'\beta)$ , where the parameter vector $\beta$ may be estimated by maximum likelihood.

Various researches, especially in the case of [1] in revisiting zero-inflated and hurdle models, Poisson and Negative binomial count models as the best performing model, conducted simulations by varying samples size and levels of abundance zero. The research showed that the negative binomial model as the best model when fitted with both structured and non-structured zeros under various settings.

This research focuses on the role of Hurdle_T model (use of tobit as binary model) in cases of Zero Inflated Counts and Truncated models as against other form of binary models like the probit and logit. The probits and logits functions are symmetric around zero and are mostly used in microeconomics. The hurdles regression models used the logit and probit as binary model but this study prefers the tobit. The tobit is
preferred because of its inclination to solving censored normal regression models and it fits in properly in censoring forms below zero.

The count model exist as outcome of two processes, one generating the zeros together with below zeros and the other generating the positive values. The model relaxes the assumption that the zeros together with below zeros and the positive come from the same data generation process. The model combines a binary model to predict zeros together with below zeros with a zero truncated Poisson or a zero truncated negative binomial model to predict the non zero counts. The suppressing regressors for the hurdle_T model is

\[
f(y) = \begin{cases} 
  f_1(1) & \text{if } y = 0 \\
  \frac{1 - f_1(0)}{1 - f_2(0)} f_2(y) & \text{if } y \geq 1 
\end{cases}
\] 

(1.1)

Where \( Pr(y = 0) = f_1(0) \) and \( Pr(y > 0) = 1 - f_1(0) \) and the positive counts come from the truncated density.

\[
f_2(y|y > 0) = f_2(y) / \{1 - f_2(0)\}
\]

which is multiplied by \( x' \beta + \epsilon \) to ensure the probabilities sum to 1.

This model has the interpretation that it reflects a two stage decision making process, each part being a model of one decision. The two parts of the model function independently.

**1.1 Binary model**

Basically, three binary models falls into this discussion of model formation though only two have previously been used. The three models are the logit, probit and tobit. Each of them has their area of strength but they can conveniently fit the purpose of this discussion. The logit and probit functions are symmetric around zero and are widely used in microeconomics.

We can give binary models a latent interpretation by linking it with the linear regression model and also providing the basis for extension to some multinormal models. We differentiate between the observed binary outcome \( Y \) and an underlying continuous unobservable or latent variable \( Y^* \) that satisfies the simple-indexed model.

\[
Y^* = \begin{cases} 
1 & \text{IF } Y^* > 0 \\
0 & \text{IF } Y^* \leq 0 
\end{cases}
\] 

(1.2)

From these equations, we have

\[
Pr(y = 1) = Pr(x' \beta + u > 0) = Pr(-u < x' \beta) = f(x' \beta)
\]

Where \( f() \) is the Cdf of \( -u \). This yields the probit model if \( u \) is standard normally distributed and the logit model if \( u \) is logistically distributed.
2 Literature Review

Count data models have had a wide applicability and acceptability in health economics and health sciences. A number of separate application areas can be distinguished. For instance, using individual level patient records from a controlled experiment, [3] and [4] estimate the effect of a drug treatment on the number of epileptic seizures over a given period. [5] used survey data from the 1979 National Health Interview Survey in estimating the effect of smoking on the number of days of respiratory illness during a two-week recall period. [6] study the number of emergency room visits for respiratory diseases, again using count data regression models and others in [7,8,9,10,11,12].

Six count models are considered in this work. We start with the Poisson Regression model; the Negative Binomial 1 (NB1) and (NB2) 2 Regression models which are the foundation for other count models. We talk about Zero Truncated Poisson (ZTP) and Zero Truncated Negative Binomial (ZTNB) models. Combining zero truncated Poisson with a binary model we develop the hurdle_T regression model. Finally, we consider the zero inflated Poisson and zero related negative binomial.

2.1 Poisson distribution

Looking at Poisson distribution, we assume X to be a random variable with a discrete distribution that is defined over a non-negative variable or count. X as having a Poisson distribution with parameter λ written as X-Poisson λ iff

\[ \Pr[Y=y] = \frac{e^{-\lambda} \lambda^y}{y!} \]  

The Poisson distribution has expected mean value as

\[ \operatorname{E}[y] = \lambda \]  
\[ \operatorname{Var}[y] = \lambda \]

This shows the equality of mean and variance property of the Poisson distribution.

2.2 Poisson Regression Model (PRM)

The Poisson Regression Model (PRM) extends the Poisson distribution by allowing each observation to have a different value of \( \lambda_i \). The PRM assumes that the observed count for distribution i is drawn from a Poisson distribution with mean \( \lambda_i \), where \( \lambda_i \) is estimated from observed features [13]. This is sometimes referred to as incorporating unobserved heterogeneity and leads to the structural equation.

\[ \lambda_i = \operatorname{E}(y_i|x_i) = \exp(x_i \beta) \]  

The PRM in simpler terms actually relates the probability function of a dependent variable y to a vector of independent variable x. let K be the number of regressors, X is then a column vector of dimension (KX_i) and y is the number of observations in the sample.

"We assume the following 3 assumptions:
1. $f(y|\lambda) = \frac{e^{-\lambda} \lambda^y}{y!} \quad y = 0, 1, 3$

where $f(y|\lambda)$ is the conditional probability function $y$ given $\lambda$ and it must be $\lambda > 0$.

2. $\lambda = \exp(x'\beta)$

where $\beta$ is a (Kx1) vector of parameters and $X$ is a (Kx1) vector of regressor including a constant.

3. Observed pairs $(y_i, x_i), i = 1 \ldots n$ are independently distributed.

Combining assumption 1 and 2 we have

$$F(y|x) = \frac{\exp(-\exp(x'\beta))\exp(yx'\beta)}{y!} \quad y = 0, 1, 2, 3 \ldots$$

The Poisson distribution has only one parameter that simultaneously determines the conditional mean and variable. The Poisson regression model as defined by the assumption above implies an exponential means function (log linear).

$$E(y|x) = \hat{\lambda} = \exp(x'\beta)$$

and an exponential conditional variance function

$$\text{var}(y|x) = \exp(x'\beta)$$.  

The fact that conditional mean and conditional variance are equal in the Poisson regression model we call the factor as equidispersion.

### 2.3 Negative binomial regression models

The negative binomial model distribution is the most commonly used alternative to the Poisson model. The negative binomial model has the interpretation of a Poisson mixture model that accounts in a specific way for the randomness of the Poisson parameter [13]. The negative binomial regression models the number of occurrences (count) of an event when the event has extra Poisson variation (overdispersion).

"The Poisson regression models is

$$y_i \sim \text{Poisson}(\hat{\lambda}_i)$$

where

$$\hat{\lambda}_i = \exp(x_i\beta + \text{offsets}_j)"$$

For observed counts $y_i$ with covariance $x_i$ for jth observation. One derivation of the negative binomial mean-dispersion model is that individual units follow a Poisson regression model but there is an omitted variable $v_i$ and $v_j$ follows a gamma distribution with mean and variance different.
\[ y_i \sim \text{Poisson} \left( \lambda_i \right) \]

where

\[ \lambda_i = \exp(x_i \beta + \text{offset} + v_j) \]

and

\[ \lambda_j = \text{Gamma}(I/\alpha, \alpha \lambda_j). \]

The probability function of the negative binomial model is given as

\[
\begin{align*}
    f(y|x, \lambda) &= \frac{\Gamma(\alpha + y)}{\Gamma(\alpha)\Gamma(y + 1)} \left( \frac{\alpha}{\lambda + \alpha} \right)^\alpha \left( \frac{\lambda}{\lambda + \alpha} \right)^y \\
    \text{with } E(y|\alpha, \lambda) &= \lambda \quad \Gamma() \text{ is the gamma function}
\end{align*}
\]

and

\[ \text{var}(y|\alpha, \lambda) = \lambda + \alpha^{-1} \lambda^2 \]

The regression model is completed by setting

\[ \lambda_i = \exp(x^T \beta). \] An alternative common specification uses

\[ \alpha = \alpha^{-1} \lambda_i \]

resulting in a negative binomial model with linear variance function.

To make the step to the negative binomial regression model, the parameters \( \alpha \) and \( \lambda \) are specified in terms of exogenous variables.

The NB regression model addresses the failure of the PRM by adding a parameter, \( \alpha \) that reflects unobserved heterogeneity among observations. For example, with 3 independent variable, PRM is

\[
\begin{align*}
    \lambda_i &= \exp\left( \beta_0 + \beta_{1x_1} + \beta_{2x_2} + \beta_{3x_3} + \Sigma \right) \\
    &= \exp\left( \beta_0 + \beta_{1x_1} + \beta_{2x_2} + \beta_{3x_3} \right) \exp(\Sigma) \\
    &= \exp(\beta_0 + \beta_{1x_1} + \beta_{2x_2} + \beta_{3x_3} \delta_i) \\
\end{align*}
\]

Equation 2 above follows basic algebra which the last step simply defines \( \delta \equiv \exp(\Sigma) \).

To identify the model, we assume that \( E(\delta) = 1 \) which corresponds to the assumption \( E(\Sigma) = U \) in the LRM. With this assumption, it is easy to show that

\[ E(\lambda) = \lambda E(\delta) = \lambda. \]

Thus, the PRM and the NB RM have the same mean structure. That means that if the assumptions of the NB RM are correct, the expected rate for a given level of the independent variables will be the same in both models.
However, the standard errors in the PRM will be biased downward resulting in spuriously large z-values and spuriously small P-values. The distribution of observations given both the values of the $X$'s and $\delta$ is still Poisson in the NB RM. That is,

$$\text{Pr}(y_i|x_i,\delta_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}$$  \hspace{1cm} (2.3.2)

Because $\delta$ is unknown we cannot compute $\text{Pr}(y|x)$. this limitation is resolved by assuming that $\delta$ is drawn from a gamma distribution. Then, we can compute $\text{Pr}(y/x)$ as a weighted combination of $\text{Pr}(y/x, \delta)$ for all values of $\delta$, where the weights are determined by $\text{Pr}(\delta)$. The mathematics for $\text{Pr}(y/x, \delta)$ are complex but it leads to NB distribution.

$$\text{Pr}(y/x) = \frac{\Gamma(y + \alpha^{-1})}{y!\Gamma(\alpha^{-1})} \left( \frac{\alpha^{-1}}{\alpha^{-1} + \hat{\lambda}} \right)^{\alpha^{-1}} \left( \frac{\hat{\lambda}}{\alpha^{-1} + \hat{\lambda}} \right)^{y}$$  \hspace{1cm} (2.3.3)

### 2.4 Negative Binomial II (NB II)

Counts are usually overdispersed in real world situation. Meaning that the conditional variance is larger than
the conditional mean. The NB RM addresses the problem by allowing overdispersion through the $\alpha$
parameter. The most commonly used function is

$$\text{Var}(y_i|x) = \lambda_i + \alpha \beta_i^2$$

This is referred to as the NB II model. The NB II is obtained from $\alpha = \alpha^2$ and $\lambda = \exp(x^T \beta)$. In this case, the
conditional expectation function is $E(y_i|x) = \exp(x^T \beta)$

while the conditional variance is given by

$$\text{Var}(y_i|x) = \exp(x^T \beta) + \alpha^2[\exp(x^T \beta)]^2$$  \hspace{1cm} (2.4)

The conditional variance is always greater than the conditional mean.

The NB I mode is obtained by letting $\alpha$ vary across individuals such that $\alpha = \alpha^2 \exp(x^T \beta)$ and $\lambda = \exp(x^T \beta)$.

The parameterization produces a variance that is linear function of the mean

$$\text{Var}(y_i|x) = ((1 + \sigma^2)\exp(x^T \beta))$$  \hspace{1cm} (2.4.1)

Another way of differentiating the NB I and II is in terms of dispersion function $\mathcal{O}(y|x)$.

For NB I Model, $\mathcal{O} = (1 + \sigma^2)$ while for NB II $\mathcal{O} = 1 + \sigma^2\exp(x^T \beta)$

Thus, the probability function for NB II can be written as

$$F(y/=) = \frac{\Gamma(\sigma^{-2} + y)}{\Gamma(\sigma^{-2})\Gamma(y + 1)} \left( \frac{\sigma^{-2}}{\exp(x^T \beta) + \sigma^{-2}} \right)^{\sigma^{-2}} \left( \frac{\exp(x^T \beta)}{\exp(x^T \beta) + \sigma^{-2}} \right)^{y}$$  \hspace{1cm} (2.4.2)
For $\sigma^2 \to 0$. This model converges on the PRM.

### 2.5 Models for truncated counts

A major problem occurs when observation are made with possible exclusion of some observations because of their inactiveness which is popularly known as zero. For instance, talking about lecturers in this department, there may be the presence of some lecturers who by virtue of an agreement were not taking courses. It does not foreclose them from being lecturers even though they are not taking courses. But for course taking, they are not listed.

Zero truncated count models are designed for data in which observations with an outcome of zero have been excluded from the sample.

The zero-truncated Poisson model begins with the Poisson regression model

$$
Pr(y_i|x) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}
$$

where $\lambda_i = \exp(\chi_i^\top \beta)$. For a given set of $\chi$’s, the probability of observing a zero is

$$
Pr(y = 0|\chi) = \exp(-\lambda)
$$

So the probability of a non-zero count is $Pr(y>0|\chi) = 1-\exp(-\lambda)$.

Since our counts should be truncated at zero, we compute the probability for each positive outcome given that the outcome is greater than zero. In observing the law of conditional probability,

$$
Pr(A|B) = \frac{Pr(A \text{ and } B)}{Pr(B)}
$$

the probability of observing a specific value of $y$ is given that we know the count is not zero is

$$
Pr(y_i|y_i>0, \chi_i) = \frac{Pr(y_i|\chi_i)}{Pr(y_i>0|\chi_i)} = \frac{Pr(y_i, \chi_i)}{1-\exp(-\lambda_i)} \quad \text{for } y > 0
$$

This formula simply increases each unconditional probability by the factor $(1-\exp(-\lambda))^\text{-1}$ forcing the probability mass of the truncated distribution to sum to 1.

There are two types of counts associated with this. First, we can compute the expectation without truncation as we may want to estimate the expected number of lecturers taking courses in the entire dependent, not just everybody. In the zero truncated Poisson, the expected rate is the same as the standard PRM

$$
E(y_i|\chi_i) = \exp(\chi_i^\top \beta)
$$

Secondly, we may be interest in the expected rate given that the count is positive, written as

$$
E(y_i|y>0, \chi_i)
$$

For example, among the lecturers, how many courses are they taking each? The conditional rate must be longer than the unconditional rate since we are excluding zeros. As with the conditional probability, the rate increases proportionally to a probability of appositive count.
\[ E(y_i|y_i>0, x_i) = \frac{\Gamma(y_i + \alpha^{-1})}{y_i!\Gamma(\alpha^{-1})} \left( \frac{\alpha^{-1}}{\alpha^{-1} + \lambda_i} \right)^{\alpha^{-1}} \left( \frac{\lambda_i}{\alpha^{-1} + \lambda_i} \right)^{\frac{y_i}{\alpha^{-1}}} \]  
(2.5.2)

So that \( \Pr(y=0|x) = (1 + \alpha \lambda)^{-\alpha} \) and

\[ \Pr(y>0|x) = 1-(1 + \alpha \lambda)^{-\frac{2}{\alpha}} \]

Accordingly, the conditional probability in the zero truncated Negative binomial is

\[ \Pr(y_i>0, x_i) = \frac{1}{1-(1 + \alpha \lambda)^{-\frac{2}{\alpha}}} \]

and the conditional mean equals

\[ E(y_i|y_i>0, x_i) = \frac{\lambda_i}{1-(1 + \alpha \lambda)^{-\frac{2}{\alpha}}} \]

(2.5.3)

It should be noted that the adverse effect of overdispersion are worse with truncated models. When the sample is not truncated, using the PRM in the presence of overdispersion does not bias the estimation but the zero truncated not only bias the estimate but also make it inconsistent.

3 Methodology

The hurdles regression models used the logit and probit as a binary model but this study intends to use the tobit, christened as Hurdle_T model. The tobit is taken because of its inclination to solving censored normal regression models and it fits in properly in censoring form below zero, where the latent variable is linear in regressors with additive error that is normally distributed and homoskedastic.\(^7\) The tobit model begins with

\[ Y^* = x^\prime \beta + \varepsilon \]

where \( \varepsilon \) is the error term and \( \varepsilon \sim N (0, \sigma^2) \), with a variance \( \sigma^2 \), constants across observations for latent variables. The observed variable \( Y \) is related to the latent variable \( Y^* \) through the observation rule

\[ Y = \begin{cases} Y^* & \text{if } Y^* > L \\ L & \text{if } Y^* \leq L \end{cases} \]

The probability of an observation being censored is

\[ \Pr(Y^* \leq L) = \Pr(x^\prime \beta + \varepsilon \leq L) = \phi \left\{ \left( L - x^\prime \beta \right) / \sigma \right\} \]

where \( \phi () \) is the Standard normal cumulative distribution functions.

The truncated mean or expected value of \( Y \) for the uncensored observation is

\[ E(y / x, y > L) = x^\prime \beta + \sigma \phi \left\{ (x^\prime \beta - L) / \sigma \right\} \phi \left\{ (L - x^\prime \beta) / \sigma \right\} \]

(3.1)
Where $\phi()$ is the standard normal density.

For the case of left censored date with the censoring point $\gamma(L = \gamma)$, the density function can be written as

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} (y_i - x_i^T\beta)^2 \right\} \left[ \phi\left( \frac{(\gamma - x_i^T\beta)}{\sigma} \right) \right]^{1-d_i}$$

(3.1.1)

Where $\left[ \phi\left( \frac{(\gamma - x_i^T\beta)}{\sigma} \right) \right]^{1-d_i}$ reflects the contribution to the likelihood of the censored observations. If $d=1$ denotes that the censoring indicator for the outcome that the observation is not censored and when $d=0$ indicate a censored observation.

3.1 Choice of binary model and the build up of the Hurdle_T model

From the brief discussion on binary model, one settles for the tobit model, owing to its versatility on modeling both observed and unobserved outcomes.

Let $Y$ be a count outcome that range from zero to some maximum value, suppose that zero counts are generated by a binary process; where we use tobit to model the binary outcome $Y=0$ versus $Y>0$, the positive outcomes are generated by either the Zero Truncated Poisson (ZTP) or Zero Truncated Negative Binomial (ZTNB). In this problem, zero is then seen as a hurdle that one gets past before reaching positive counts. Since they are separate equations, we allow the process to predict zero and also determine positive counts. Where $E(y_i|y_i>0, x_i)$ is defined for ZTP and ZTNB.

3.2 Zero inflated count models

There are two main reasons why zeros are of particular interest in count data models. First, empirically, their fraction is often too high to be compatible with a standard underlying count data model. Secondly, theoretically, zeros often reflect corner solution outcomes in economic choice model.

The Negative Binomial Regression Models (NBRM) improves upon the under prediction of zeros in the Poisson Regression Models (PRM) by increasing the conditional variance without changing the conditional mean. The hurdle_T model addresses the under prediction of zeros by using two equations.

The zero-inflated count model changes the mean structure to allow zeros to be generated by two distinct processes of scientific productivity.

The zero inflated model assumes that there are two latent groups. An individuals in the Always Zero group (Group A) has an outcome of 0 with a probability of 1 whereas an individual in the Not Always Zero group (Group –A) might have a zero count but there is a non-zero probability that she has a positive count.

3.2.1 Counts for group a (Always Zero)

This group has a binary outcome and can be modeled using the tobit

$$\varphi_i = \Pr(A_i = 1|z_i) = F(z_i\gamma)$$

where $\varphi$ is the probability of being in Group A for individual i. The Z-variables are referred to as inflation variables because they inflate the number of 0s.
If two variables affect the probability of an individual being in Group A, we model this with a logit equation

$$
\varphi_i = \frac{\exp(\gamma_0 + \gamma_1 z_1 + \gamma_2 z_2)}{1 + \exp(\gamma_0 + \gamma_1 z_1 + \gamma_2 z_2)}
= \frac{1}{1 + \exp\{-\gamma_0 + \gamma_1 z_1 + \gamma_2 z_2\}}
$$

(3.1.2)

$$
z_1, z_2 \to \infty \quad \varphi_i \to 1
$$

The presence of observed variable indicating group membership would make would standardize the model and of course, we call it binary regression model but because group membership is a latent variable, we cannot know whether an individual is in Group A or Group –A.

### 3.2.2 Counts for Group –A

This group are not always zero group. The probability of each count is determined by either a Poisson or a negative binomial regression. For the zero-inflated Poisson (ZIP) model, we have

$$
\Pr(y_i | x_i, A_i = 0) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}
$$

(3.1.3)

Or for the zero-inflated negative binomial (ZINB) model as

$$
\Pr(y_i | x_i, A_i = 0) = \frac{\Gamma(y_i + \alpha^{-1})}{y_i! \Gamma(\alpha^{-1})} \left( \frac{\alpha^{-1}}{\alpha^{-1} + \lambda_i} \right)^{\alpha^{-1}} \left( \frac{\lambda_i}{\alpha^{-1} + \lambda_i} \right)^{\lambda_i}
$$

(3.1.4)

In both equation,

$$
\lambda_i = \exp(x_i \beta)
$$

If we knew which observations were in Group –A, these equations would define the PRM and the NBRM. But the equations apply only in those observations in Group –A and we do not have an observed variable indicating group membership.

### Table 1. Comparing coefficient across models. Dependent variable=docvis, with p-values in parentheses

| Regressor | PRM | NBRM | ZIP | ZINB | ZIP | ZTP | ZTNB | TOBIT/ZTP (HURDLE_T) | TOBIT/ZTNB (HURDLE_T) |
|-----------|-----|------|-----|------|-----|-----|------|---------------------|----------------------|
| Private   | 0.13393 | 0.15795 | 0.0866 | 0.1178 | 0.0865 | 0.1020 | 1.17769/0.08652 | 1.17769/0.1020 |
|           | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.003) | (0.000) / 0.000 | (0.000) / 0.0033 |
| Medicaid  | 0.0925 | 0.0932 | 0.0826 | 0.1040 | 0.0829 | 0.0957 | 0.61944/0.08288 | 0.61944/0.0957 |
|           | (0.000) | (0.040) | (0.000) | (0.019) | (0.000) | (0.041) | (0.101) / 0.000 | (0.101) / 0.041 |
| Age       | 0.2923 | 0.2295 | 0.2487 | 0.2802 | 0.2487 | 0.2705 | 2.25125/0.24866 | 2.25125/0.2705 |
|           | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) / 0.000 | (0.000) / 0.000 |
| Age2      | -0.00193 | -0.00195 | -0.0016 | -0.0019 | -0.0017 | -0.0018 | -0.0149/-0.0017 | -0.0149/-0.0018 |
|           | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) / 0.000 | (0.000) / 0.000 |
4 Explanation of Results

PRM: The coefficients for PRM are all significant as the p-values, each equals 0.000. The overall fit is significant, since $\chi^2$ statistics is significant (p-value = 0.000).

NBRM: All the coefficients are significant except that of females (p-value=0.380). The overall fit is significant (p-value=0.000).

ZIP: All coefficients are significant (p-value=0.000) Overall fit is significant (p-value=0.000).

ZINB: All coefficients are significant as p<0.05. The overall regression is significant (p-value=0.000).

ZTP: All coefficients are significant (p=0.000). Overall fit is significant (p-value=0.000).

ZTNB: All coefficients are significant (p<0.05). Overall fit is significant (p-value=0.000).

HURDLE T (TOBIT/ZTP): All coefficient are significant except Medicaid (0.101) and Female (0.211) in the TOBIT. Overall fit is significant (p-value=0.000).

HURDLE T (TOBIT/ZTNB): All coefficient are significant except Medicaid (0.101) and Female (0.211) in the TOBIT and Medicaid (0.041) and Female (0.035). Overall fit is significant (p-value=0.000).

4.1 Discussion on the Hurdle_T model.

A positive signed coefficients in the tobit model means that the corresponding regressor increases the probability of a positive observation. In the second part a positive coefficient means that conditional on a positive count, the corresponding variable increases the value of the count.

Thus, private, medical, age, education, acllim, totchr, all increases the probability of the positive counts in the positive observation. The result in the tobit also shows that except medical and females, all other
The positive coefficient in the ZTP and ZTNB also increases the corresponding variable value corresponding to them.

5 Conclusion

Count Data Regression models exhibit different strength and weaknesses in their bids to solving problems. This study concentrated on making a right choice of count data model as it analyzes the issue of male breast cancer and the need to develop a good health policy on the need for doctor’s visitation to patients who by the nature of their condition, may not be able to have access to good health facility. The call for doctor’s visitation ranges from physical limitation, activity limitation and for the peculiar case of our country where there is no health insurance policy, financial limitation.

Eight count data regression models were used to analyze the issue of breast cancer amongst male and the issue of doctor’s visitation. These are the PRM, NBRM, ZIP, ZINB, ZTP ZTNB and the hurdle_T model (Tobit/ZTP and Tobit/ZTNB). This study reveals that Male Breast cancer cuts across all age brackets or category but it is more for ages between 50 and 60 and so, as one approaches age 50, regular check ups are of a necessity. The strength of count data models are also affected by the nature of data available for analysis. A poor data gathering affects and limits the number of count data models that can be used, as in the case of breast cancer in Nigeria where only two models could be used; the PRM and the NBRM. The use of the information criterion revealed that the PRM is a better model than the NBRM. The second health related case of doctor’s visitation permitted the use of the other models as it created room for inflation and truncation of zeros. The analysis revealed that the models with negative binomial tendencies are preferred to the ones with Poisson. In all, the ZINB yielded a better result than all, followed by the NBRM, then the ZTNB before their Poisson counterparts. The study also revealed that out of about 3677 people visited as shown by the data, the female folks of 2210 need more doctor’s visit at their reap age than the male with a population of 1467. It is therefore recommended that governmental policies should be tailored towards providing awareness to male of 50 to 60 year age bracket on the possibility of risk of breast cancer and provision of adequate funding to cater for doctor’s visitation to females of old age (65 years and above) to reduce pains of old age to the elderly.

Competing Interests

Authors have declared that no competing interests exist.

References

[1] Nekesa F, Odhiambo C, Chaba L. Comparative assessment of zero-inflated models with application to HIV exposed infants data. Open Journal of Statistics. 2019;9(6):664-685. DOI: 10.4236/ojs.2019.96043

[2] Hu Mei-Chen, Martina Pavlicova, Edward V. Nunes. Zero-inflated and hurdle models of count data with extra zeros: Examples from an HIV-risk reduction intervention trial. The American Journal of Drug and Alcohol Abuse. 2011;37(5):341.

[3] Diggle P, Liang KY, Zeger SL. Analysis of longitudinal data. Oxford University Press; 1995.

[4] Chib S, Greenberg E, Winckelmann R. Markov chain Monte Carlo simulation methods in econometrics. Econometric Theory. 1996;12:409-431.

[5] Mullahy J, Portney PK. Air pollution, cigarette smoking and the production of respiratory health. Journal of Health Economics. 1990;9:193-205.
[6] Jergensen B, Lundbye-Christensen S, Xue-Kun SP, Sun I. A state space models for multivariate longitudinal count data. Biometrika. 1999;86:169-181.

[7] Cameron AC, Trivedi PK. Microeconometrics methods and application. Cambridge University Press; 2009.

[8] Faddy MJ. Extended Poisson modelling and analysis of count data. Biometric Journal. 1997;39:431-440.

[9] Gurmu S, Rilstone P, Stern S. Semiparametric estimation of count regression models. Journal of Econometrics. 1998;88(1):123-150.

[10] Haab TC, McConnell KE. Count data models and the problem of zeros in recreation demand analysis. American Journal of Agricultural Economics. 1996;78:80-102.

[11] Hausman JA, Hall BH, Griliches Z. Econometric models for count data with an application to the patents - R & D relationship. Econometrics. 1984;52:909-938.

[12] Nagata C, Kawakami N, Shimizu H. Trends in the incidence rate and risk factors for breast cancer in Japan. Breast Cancer Res Treat. 1997;44:75-82.

[13] Winkelmann R. Econometric analysis of count data. Springer Verlay Berlin Heidelberg; 2008.

[14] Amir H, Makwaya CK, Moshiro C, Kwesigabo G. Carcinoma of the male breast: A sexually transmitted disease? East African Medical Journal. 1996;73:187-189.

[15] Nggada H. Yawe, Abdulazeez J, Khalil MA. Breast cancer burden in Maiduguri, North Eastern Nigeria. The Breast Journal. 2008;14(3):284-286.

© 2020 Akpofure and Paul; This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
 http://www.sdiarticle4.com/review-history/55699