2023 ■ Vol. 19 ■ no. 1

Regression models for count data with excess zeros: A comparison using survey data

Adhin Bhaskar[a], K. Thennarasu[b], Mariamma Philip[b] & T. S. Jaisoorya[c]

[a]Department of Statistics, ICMR – National Institute for Research in Tuberculosis , Chennai, India
[b]Department of Biostatistics, NIMHANS, Bengalulru, India
[c]Department of Psychiatry, NIMHANS, Bengalulru, India

Abstract ■ Presence of excess zeros and the distributions are major concern in modeling count data. Zero inflated and hurdle models are regression techniques which can handle zero inflated count data. This study compares various count regression models for survey data observed with excess zeros. The data for the study is obtained from a survey conducted to assess the harms attributable to drinkers among children. Poisson, negative binomial and their zero inflated and hurdle versions were compared by fitting them to two count response variables, number of physical and number of psychological harms. The models were compared using fit indices, residual analysis and predicted values. The robustness of the models were also compared using simulated data sets. Results indicated that the Poisson regression was less robust to deviations from the distributional assumptions. The negative binomial regression and hurdle regression model were found to be suitable to model the number of physical and number of psychological harms respectively. The results showed that excess zeros in count data does not imply zero inflation. The zero inflated or hurdle models are suitable for zero inflated data. The selection between the zero inflated and hurdle models should be based on the assumed cause of zeros.

Keywords ■ count data; Poisson; negative binomial; zero inflation; hurdle regression. **Tools** ■ R.

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

✉ adhinb6001@gmail.com

10.20982/tqmp.19.1.p001

Introduction

Zero inflation is a persistent phenomenon in count data. It occurs when the zeros observed in a dataset are beyond the range of basic count that can be handled by regression models such as Poisson or negative binomial regression. It is also considered as a source of overdispersion (Cameron & Trivedi, 1998; Martin et al., 2005), defined as the conditional variance of the response variable exceeding conditional mean. The Poisson regression model assumes that the conditional mean and conditional variance are equal (known as equidispersion) as this regression model is based on Poisson distribution. Negative binomial regression model is a suitable alternative to Poison regression model when the data is overdispersed. However the negative binomial regression model may not be able to handle the overdispersion occurring due to zero inflation. In zero inflated situations, the zeros can be classified as true zeros and sampling zeros based on their origin (Zorn, 1996). The true zeros are generated from perfect state or risk free stage where the counts are always zero. The sampling zeros or false zeros are generated from an imperfect state or an at-risk stage. The count can take any non-negative value including zero in an imperfect state or an at-risk stage. For example, the answer to the question "how many cigarettes you smoked last week?" is always zero for non-smokers, which is as true (structural) zero. However, the response of smokers can be either a positive count or zero. These particular zeros are generated from an at-risk origin, hence called as sampling zeros. Sampling zeros are produced usually due to design error, survey error, observer error, etc. (Loeys et al., 2012). However, in practice, it is quite difficult to differentiate between the two kinds of zeros as the cause of structural zeros is usually unobservable (He et al., 2014).

When data is generated from these two processes, the models based on a single distribution (such as Poisson or negative binomial models) may not be able to capture the excess zero or result in a good fit. The zero inflated data should be modeled using a model which incorporates the data generation process as well as the excess incidence of zeros. Zero inflated models and hurdle models are commonly used in situations of zero inflation. These two models assume that the data is generated from two separate distributions conditional on the independent variables. This study is trying to assess the suitability of various count regression models for count response variable with excess zeros when it related with a set of predictor variables.

Materials and methods

Zero inflated regression models

Zero inflated models were introduced by Lambert (1992) as a way to capture the excess zeros in the number of defects in manufacturing process. The zero inflated models use a mixture modeling approach, where the data generation process is governed by a count distribution and a degenerate Bernoulli distribution with mass at zero. The zero inflated models permit the occurrence of both structural zeros and sampling zeros in the data. In zero inflated models all structural zeros are generated from a Bernoulli process with probability π_i, $1 - \pi_i$ being the probability of transition to a count process from which the count values are generated. The count process is governed by a count distribution. The transition to a count process does not insure a positive realization, because the count process generates positive counts as well as zeros. The zeros generated by the count process are sampling zeros.

In Zero inflated Poisson (ZIP) regression model, a type of zero inflated model, the count process is governed by Poisson distribution (Lambert, 1992). The probability mass function (pmf) of the ZIP distribution can be written as a mixture of Poisson and degenerate Bernoulli distribution at zero as follows,

$$p\left(y_i\right) = \begin{cases} \pi_i + \left(1-\pi_i\right) e^{-\lambda_i} & \text{if } y_i = 0 \\ \left(1-\pi_i\right) \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} & \text{if } y_i = 1,\ 2,\ 3, \ldots \end{cases} \qquad (1)$$

where π_i is the probability of structural zero and λ_i is the mean function for the i^{th} individual from Poisson distribution. In ZIP regression model, two regression models are fitted simultaneously to relate π_i and λ_i with the independent variables with logit and log link function respectively. i. e.,

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = z_i^{\intercal}\gamma$$

and

$$\ln(\lambda_i) = x_i^{\intercal}\beta$$

where z_i is the set of independent variables of the binary logistic regression part with parameters γ and x_i is set the independent variables of the Poisson regression part with parameters β. Hence, the ZIP model allows structural zeros, sampling zeros and positive counts to be dependent on the independent variables. It is not necessary that the same set of independent variables to be used in both parts of the mixture model. More precisely, z_i and x_i can be two independent set of predictor for the logistic and Poisson regression parts of the ZIP model.

Sometimes, same predictors can have effect in the binary part as well as the count part. If the higher value of a predictor reduces the proportion of structural zero and increases the mean in the count part, or vice versa, the variable is said to have a consonant effect on the response variable. The consonant effect implies that the predictor works in the same direction in increasing or decreasing the overall mean of the response variable (Xu et al., 2015; Mills, 2013). In such cases, the coefficients of the predictor in the binary and Poisson regression parts will have opposite signs. On the other hand, if the variable has an opposite effect in the two parts of the models in determining the overall mean, it is called as a dissonant effect, in which case, the coefficients of the variable will have the same sign in both parts (logistic and count) of the zero inflated model (Xu et al., 2015; Mills, 2013). The ZIP model may not give satisfactory fit even after accounting for the zero inflation, if the data is having greater variability than the fitted model can estimate. The ZIP model can be extended to a Zero Inflated Negative Binomial (ZINB) model in which the count part is governed by a negative binomial regression. This refinement allows the ZINB model to permit overdispersion induced by zero inflation as well as unobserved heterogeneity (Greene, 1994). Major reasons for overdispersion are dependency between the events and unobserved heterogeneity which may be defined as the effect of omitted or unmeasured independent variables in a Poisson regression.

The probability mass function of ZINB distribution can be formulated by replacing the Poisson probability mass function in equation (1) with the probability mass function of a negative binomial distribution. The probability mass function of ZINB distribution is as follows (top of next page):

$$p\left(y_i\right) = \begin{cases} \pi_i + \left(1-\pi_i\right) \left(\frac{\theta^{-1}}{\theta^{-1}+\lambda_i}\right)^{\theta^{-1}} & \text{if } y_i = 0 \\ \left(1-\pi_i\right) \frac{\Gamma\left(y_i + \theta^{-1}\right)}{\Gamma(y_i+1)\Gamma(\theta^{-1})} \left(\frac{\theta^{-1}}{\theta^{-1}+\lambda_i}\right)^{\theta^{-1}} \left(\frac{\lambda_i}{\theta^{-1}+\lambda_i}\right)^{y_i} & \text{if } y_i = 1,\ 2,\ 3, \ldots. \end{cases} \qquad (2)$$

where θ is known as the dispersion parameter as variance increases with increase in θ (Mean = λ_i , variance = $\lambda_i + \theta\lambda_i^2$; negative binomial model). The ZINB model relates π_i and λ_i with the independent variables via logit and log link functions respectively as in ZIP model.

Hurdle regression models

Hurdle regression models are another kind of modified regression approach to accommodate the excess incidence of zeros in the data. The hurdle regression models do not discriminate zeros as structural and sampling zeros. The fundamental idea behind the formulation of hurdle regression model is that a binomial probability distribution governs the binary outcome of whether the count variate has a zero or positive realization (Mullahy, 1986). If the realization is positive, 'the hurdle' is crossed and the non-negative counts are generated from a zero truncated count distribution. Hence, all zeros are generated from a binary regression model.

The pmf of the binary process can be written as,

$$p\left(y_i\right) = \begin{cases} \pi_i & \text{if } y_i = 0 \\ 1 - \pi_i & \text{if } y_i = 1,\ 2,\ 3, \ldots \end{cases} \tag{3}$$

where π_i is the probability of zero for the ith individual. The zero truncated count model has the form,

$$p\left(y_i\right) = \begin{cases} f\left(y_i\right) & \text{if } y_i = 1,\ 2,\ 3, \ldots \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Combining equations (3) and (4), the generic form of probability mass function of the hurdle distribution can be formulated as,

$$p\left(y_i\right) = \begin{cases} \pi_i & ; y_i = 0 \\ \left(1 - \pi_i\right) f\left(y_i\right) & ; y_i = 1,\ 2,\ 3, \ldots \end{cases} \tag{5}$$

The hurdle Poisson (HP) regression is a form of hurdle regression, in which the positive count part is modeled by a zero truncated Poisson regression. The probability mass function of the distribution governing the HP regression has the form,

$$p\left(y_i\right) = \begin{cases} \pi_i & \text{if } y_i = 0 \\ \left(1 - \pi_i\right) \frac{e^{-\lambda_i}\lambda_i^{y_i}}{\left(1 - e^{-\lambda_i}\right) y_i!} & \text{if } y_i = 1,\ 2,\ 3, \ldots \end{cases} \tag{6}$$

where $\pi_i = \frac{\exp\left(z_i^\intercal \gamma\right)}{1 + \exp\left(z_i^\intercal \gamma\right)}$ and $\lambda_i = \exp\left(x_i^\intercal \beta\right)$ is the mean of the Poisson regression.

Unlike zero inflated models, the two parts of the hurdle models are assumed functionally independent (Cameron & Trivedi, 1998). Hence, the estimates of the models can be obtained by maximizing the log likelihoods separately (Mullahy, 1986; McDowell, 2003). The data may exhibit overdispersion derived from sources such as unobserved heterogeneity other than zero inflation. The HP model is inflexible to accommodate this extra Poisson variation arising in the positive count part. In such situations, the HP model can be extended to a Hurdle Negative Binomial (HNB) model similar to extending ZIP model to ZINB model.

Test for zero inflation

Wilson and Einbeck (2019) proposed a test for the presence of zero inflation based on the observed zeros in the count data. The test assumes that zeros are generated from Bernoulli process with varying probability. The test takes the random variable, total number of observed zeros, N_0 as the test statistic which follows a Poisson-binomial distribution. The test is performed after fitting a Poisson or negative binomial model and estimating the probability of zero using the predicted mean for each subject (π_i). The probability mass function of Poisson–binomial distribution is as follows,

$$p\left(N_0 = k\right) = \left\{ \prod_{i=1}^{n} \left(1 - \pi_i\right) \right\} \times \sum_{i_1 < i_2 < \ldots < i_k} w_{i1} w_{i2} \ldots w_{ik} \tag{7}$$

where $w_i = \frac{\pi_i}{1 - \pi_i}$; i = 1, 2, 3, . . . , n and the summation is taken over all possible distinct combination of $i_1, i_2, \ldots, i_k$ from 1, 2, 3, . . . , n. The test is performed by assuming H_0: Absence of zero inflation under the fitted model against H_1: The data is zero inflated.

Randomized Quantile Residuals

Residual analysis is a foremost method used for diagnosis of the model fit. Dunn and Smyth (1996) proposed Randomized Quantile Residuals (RQR) with a special attention on discrete variables. The RQR produce non-overlapping continuous residuals for discrete variables. Hence,RQR are ideal for assessing the residual distribution of count regression models. The RQR are obtained by inverting the fitted distribution function and calculating the corresponding standard normal quantile for each observation (McElduff, 2012). For continuous distribution, the distribution function $F(y_i \,|\, (\hat{\theta}_i)$ follows uniform distribution between 0 and

1. Hence, the RQR can be defined as (Dunn & Smyth, 1996)

$$q_i = \Phi^{-1}(F(y_i \,|\, \widehat{\theta}_i)) \tag{8}$$

where Φ^{-1} is the quantile function of a standard normal distribution. For discrete variable the RQR is,

$$q_i = \Phi^{-1}\left(u_i\right) \tag{9}$$

where u_i is random value from a uniform distribution with interval $[F(y_i - 1 \,|\, \hat{\theta}_i), F(y_i \,|\, \hat{\theta}_i]$. The RQR follow a standard normal distribution if the model is correctly specified (Dunn & Smyth, 1996). Hence, the test of normality of RQR can be used as a technique to assess the goodness of fit of the model. As randomness is involved in the calculation of RQR, in the present study, the residuals are calculated 1000 times for each model and the average of the mean, SD and p value of normality test are computed (Feng et al., 2020).

Data

To illustrate the aforementioned regression models, a survey conducted among school children is used. The data of 6412 students, from 73 high school and higher secondary schools were collected from the survey conducted in the district of Ernakulum, Kerala during 2014-2015. The information regarding, socio economic status, harms experienced due to others' drinking, substance use, psychological distress, and ADHD (Attention-deficit/hyperactivity disorder) were collected. Questions were asked to the students on various harms experienced due to others' drinking in the period of last one year. In the questionnaire, 11 questions were on experiencing psychological harms and four were related to physical harms. The sum of the questions related to psychological harms and physical harms constituted the response variables, number of psychological harms (ranging from 0 to 11) and number of physical harms (ranging from 0 to 4). Ethical clearance from institute ethical committee was obtained for this survey from the Government Medical College, Ernakulum, Kerala, India.

Results

Table 1 provides the descriptive statistics of the collected information. The participants of the study consisted of 49.41% ($n = 3168$) males and 50.59% ($n = 3244$) females with an average age of 15.24 (SD = 1.73) years. About, 43.68% ($n = 2801$) reported to have psychological harms and 9.75% ($n = 625$) had experienced physical harms attributable to drinkers. The count variables number of psychological harms (mean = 1.18, SD = 1.79) and number of physical harms (mean = 0.12 , SD = 0.43) were considered as the response variable for comparing various count regression models. In the preliminary stage, each variable listed in Table 1 fitted using an unadjusted Poisson regression for the total number of harms. Those variables which were not significant with $p > 0.2$ were discarded from the further analysis. The discarded variables were religion and nature of school.

The results of Poisson regression and negative binomial regression are presented in the Table 2. In order to decide on the need of negative binomial regression over Poisson regression, an auxiliary regression based test for equidispersion (which assumes that the conditional mean and variance are equal, $E(y_i \,|\, x_i) = Var(y_i \,|\, x_i)$) was carried for two count response variables after fitting Poisson regression (Cameron & Trivedi, 1998). The test performed assuming overdispersion ($E(y_i \,|\, x_i) < Var(y_i \,|\, x_i)$) under the alternative hypothesis was significant for both Poisson models fitted to number of psychological harms ($t = 24.976$, $p < 0.001$) and psychical harms ($t = 6.133$, $p < 0.001$) that supporting the presence of overdispersion in the data (Table 2). Due to overdispersion in the data, the standard errors of Poisson model were underestimated and they were much smaller than that of the negative binomial fit. Estimated standard errors of negative binomial regression fitted to number of and psychological harms were almost double than that of the Poisson model fit. However, the underestimation of standard errors was much less for the Poisson model fitted to number of physical harms. Due to the smaller standard errors the Poisson regression identified part time job ($p = 0.004$) as an additional significant risk factor for number of psychological harms and schools located in urban area ($p = 0.032$) as an additional risk factor of number of physical harms.

Since, both response variables were observed with high proportion of zeros, the Poisson and negative binomial fits were tested for possible zero inflation. The significance of the test for zero inflation pointed out that the number of psychological harms was zero inflated under Poisson ($t = 3611$, $p < 0.001$) and negative binomial model ($t = 3611$, $p < 0.001$) fits. The number of physical harms was zero inflated under Poisson fit ($t = 5787$, $p < 0.001$) but, the test was not significant for the negative binomial model fit ($t = 5787$, $p = 0.449$). In order to account for the excess zeros observed in the data, zero inflated and hurdle models were employed and the results are presented in Table 3 and Table 4.

When ZIP and ZINB models were fitted to number of psychological harms, count regression parts of models exhibited trend similar to that of Poisson and negative binomial model fits. The estimated standard errors of the coefficients in the count part of ZIP model were smaller than that of the ZINB fit, indicating overdispersion in the count part. Moreover, the estimated dispersion parameter, theta of ZINB fit was also significant ($\theta = 1.645$, $p =< 0.001$)

Table 1 ■ Socio demographic profile of the respondents

Variable	Category	n (%)	Mean (SD)
Gender	Male	3168 (49.41)	
	Female	3244 (50.59)	
Age (years)			15.24 (1.73)
Religion	Hindu	3217 (50.17)	
	Christian	2007 (31.30)	
	Muslim	1175 (18.33)	
	Others	13 (0.20)	
Residence	City	548 (8.55)	
	Town	688 (10.73)	
	Village	5176 (80.72)	
Family category	Above Poverty Line	4555 (71.04)	
	Below Poverty Line	1857 (28.96)	
Family structure	Both parents	5946 (92.73)	
	Single parent	312 (4.87)	
	Living with relative/others	154 (2.40)	
Nature of school	Government	3895 (60.75)	
	Government aided	2440 (38.05)	
	Private	77 (1.20)	
Location of school	Rural	4726 (73.71)	
	Semi urban	1292 (20.15)	
	Urban	394 (6.14)	
Part time job	Yes	473 (7.38)	
	No	5939 (92.62)	
Stay at hostel	Yes	104 (1.62)	
	No	6308 (98.38)	
Substance use	Yes	616 (9.61)	
	No	5796 (90.39)	
ADHD score			27.50 (8.41)
Psychological distress score			15.32 (6.47)
Number of Psychological harms			1.18 (1.79)
Number of physical harms			0.12 (0.43)
Experiencing psychological harm	Yes	2801 (43.68)	
	No	3611 (56.31)	
Experiencing physical harm	Yes	625(9.75)	
	No	5787 (90.25)	

confirmed overdispersion (Table 3). As seen in Poisson fit, the count part of ZIP model overestimated the significance of the predictors. Even in the presence of overdispersion, the estimated coefficients of ZIP model were consistent with the coefficients of the ZINB except for certain variables. The estimated coefficients and standard errors in the binary part of ZIP and ZINB models were quite similar. On the other hand, for number of physical harms, the estimated theta of ZINB model was found not significant ($\theta = 1.916$, $p = 0.086$), indicated the absence of overdispersion which resulted in ZINB model to give similar estimates as that of ZIP model.

The HP and HNB models exhibited trends similar to that of ZIP and ZINB models for all response variables. The overdispersion in the truncated count part resulted in HP model to underestimate the standard errors and thereby overestimated significance of the predictors in the truncated count parts. The HNB model for the number of physical harms indicated absence of overdispersion ($\theta = 0.985$, $p = 0.304$) in zero truncated part (Table 4). For all response variables, the values in the zero hurdle part of HP fit were exactly same as that of the corresponding HNB fit as the parameters were estimated by separate maximization of likelihoods.

When comparing the zero inflated and hurdle models, though, the interpretations are slightly different, the estimates and standard errors of the count part were similar for number of psychological harms, i. e., the estimates and standard errors of ZIP and HP models as well as ZINB and HNB models were similar. Though, the standard errors were similar, estimates of regression coefficients in the binary part of the zero inflated and hurdle models had opposite signs. For example, the estimate corresponding to age in the binary part of the ZIP was -0.12 (SE = 0.02) whereas, it was 0.13 (SE = 0.02) for HP model when they were fitted to number of psychological harms. The difference in sign was due to the difference in definition of the models. In the `pscl` and `countreg` packages of R software, the binary part of the hurdle model predicts the probability of a positive count, whereas, the binary part of zero inflated models

Table 2 ■ Poisson and negative binomial regression models fitted to number of psychological and physical harms

		Psychological harms		Physical harms	
Variable		Poisson b(SE)	NB b(SE)	Poisson b(SE)	NB b(SE)
Gender	Male	0.53 (0.03)*	0.60 (0.04)*	1.28 (0.10)*	1.28 (0.10)*
Age (years)		0.09 (0.01)*	0.09 (0.01)*	0.06 (0.02)*	0.06 (0.03)*
Family structure	Living with others/relative	-0.004 (0.08)	0.01 (0.13)	-0.15 (0.23)	-0.21 (0.28)
(Reference: Both parents)	Single parent	0.08 (0.05)	0.16 (0.09)	0.27 (0.13)*	0.35 (0.16)*
Family category	Below Poverty Line	0.04 (0.03)	0.07 (0.04)	0.13 (0.08)	0.15 (0.09)
Residence	City	0.06 (0.05)	0.04 (0.08)	0.09 (0.14)	0.10 (0.16)
(Reference: Village)	Town	0.04 (0.04)	0.05 (0.07)	0.29 (0.11)*	0.32 (0.13)*
Location of school	Urban	0.20 (0.05)*	0.26 (0.09)*	0.32 (0.15)*	0.34 (0.18)
(Reference: Rural)	Semi Urban	0.02 (0.03)	0.02 (0.05)	-0.11 (0.11)	-0.12 (0.12)
Part time job	Yes	0.10 (0.04)*	0.11 (0.07)	0.44 (0.09)*	0.45 (0.11)*
Stay at hostel	Yes	0.15 (0.10)	0.18 (0.16)	0.41 (0.27)	0.36 (0.33)
Substance use	Yes	0.54 (0.03)*	0.57 (0.06)*	0.82 (0.08)*	0.85 (0.10)*
ADHD score		0.02 (0.001)*	0.02 0.003)*	0.01 (0.004)*	0.02 (0.01)*
Psychological distress score		0.04 (0.002)*	0.05 (0.003)*	0.04 (0.01)*	0.05 (0.01)*
Theta (overdispersion)			0.79 (0.03)		0.74 (0.11)
Test for overdispersion		T = 24.976, p = 0.001		T = 6.133, p = 0.001	
Test for zero inflation		T = 3611, p = 0.001	T = 3611, p = 0.001	T =5787, p = 0.001	T= 5787, p = 0.449

Note. * significant at 5% level of significance

predicts the probability of structural zeros. Hence, the interpretations are entirely different for the models. In ZIP fit, as the age increases by one unit, the odd of structural zeros reduced by 11% ($\exp(-0.12) = 0.89$) whereas in HP fit, as the age increased by one unit the odds of experiencing harms increased by 14% ($\exp(0.13) = 1.14$).

The absolute goodness of fit of the models were assessed by analyzing the distribution of the Randomized Quantile Residuals (RQR). The RQR follows a standard normal distribution for model fitting the data well. The comparison of models with RQR , log likelihood and Bayesian Information Criteria (BIC) are summarized in Table 5. Among the models fitted to number of psychological harms, ZINB and HNB models had standard normal RQR, indicating good fit of the models (Table 5). With respect to the values of log likelihood and BIC (lower the value better fit), ZINB (log likelihood = -8575, BIC = 17422) and HNB (log likelihood = -8577, BIC = 17425) models had the best fit for number psychological harms. The RQR of all count models except Poisson were standard normal for physical harms. Among the models fitted to number of physical harms, the BIC penalized the complex models for estimating more parameters and chose negative binomial model as the best fitting model (BIC = 4552). Similarly, BIC of ZIP (BIC = 4596) and HP (BIC = 4594) models were smaller than ZINB (BIC = 4603) and HNB (BIC = 4600) models. Capability of models in capturing the zeros were assessed by computing the sum of the predicted probabilities of zeros. Zero inflated and hurdle models showed good prediction of zeros. The Poisson model under predicted the zeros always. The negative binomial model fitted to physical harms, exhibited competing predictive capability with the zero inflated and hurdle models. The negative binomial model was able to predict the zeros even when the dependent variable contained 90.25% zeros.

All participants in the study were expected to be under the risk of experiencing harms from drinkers. Hence, all observed zeros were assumed to have a sampling origin. In this regard, zero inflated models were avoided for establishing the relationship as they assume structural origin for zeros. Taking in to account of the values of fit indices, residual analysis and the predicted value, HNB model can be considered as the most appropriate model for modeling number of psychological harms. Negative binomial regression can be a suitable model to assess the factors predicting the number of physical harms.

Simulation study

Apart from the real data analysis a Monte Carlo based simulation study was also carried out for zero inflated data to see the robustness of models to violation of assumptions, sample size, misspecification of distribution and various other conditions. Zero inflated count data with two origins were generated from four distributions viz., ZIP, HP, ZINB and HNB. The structural zeros were generated from a process governed by a logistic function with parameters $\gamma_1 = 0.47$ and $\gamma_2 = -2.1$ for x_1 and x_2 respectively which are normally distributed. For the count part, the coefficients of x_1 and x_2 were set as $\beta_1 = -0.12$ and $\beta_2 = 1.1$ respectively. A normally (e_1) distributed and a gamma (e_2) distributed noise variables were also added in the generation of the binary and count parts respectively. The zero inflated response variable was generated for various level of zeros, dispersion, mean and sample sizes.

The logistic function for the binary part is defined as

$$p\left(\text{structural zeros}\right) = \frac{\exp\left(\gamma_0 + \gamma_1 \times x_1 + \gamma \times x_2 + e_1\right)}{1 + \exp\left(\gamma_0 + \gamma_1 \times x_1 + \gamma_2 \times x_2 + e_1\right)}.$$

The proportions of the structural zeros were changed by adjusting the values of the intercept (γ_0) for 20 level in the logistic function. The mean function for the count part was defined as $\exp(\beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + e_2)$. The mean of the count part was changed by adjusting the intercept, β_0 for four levels. Apart from varying the mean, dispersion in the ZINB and HNB random variables were also made to vary by fixing four values for theta. The simulation for each combination was repeated 2000 times and the models were fitted to each simulated data.

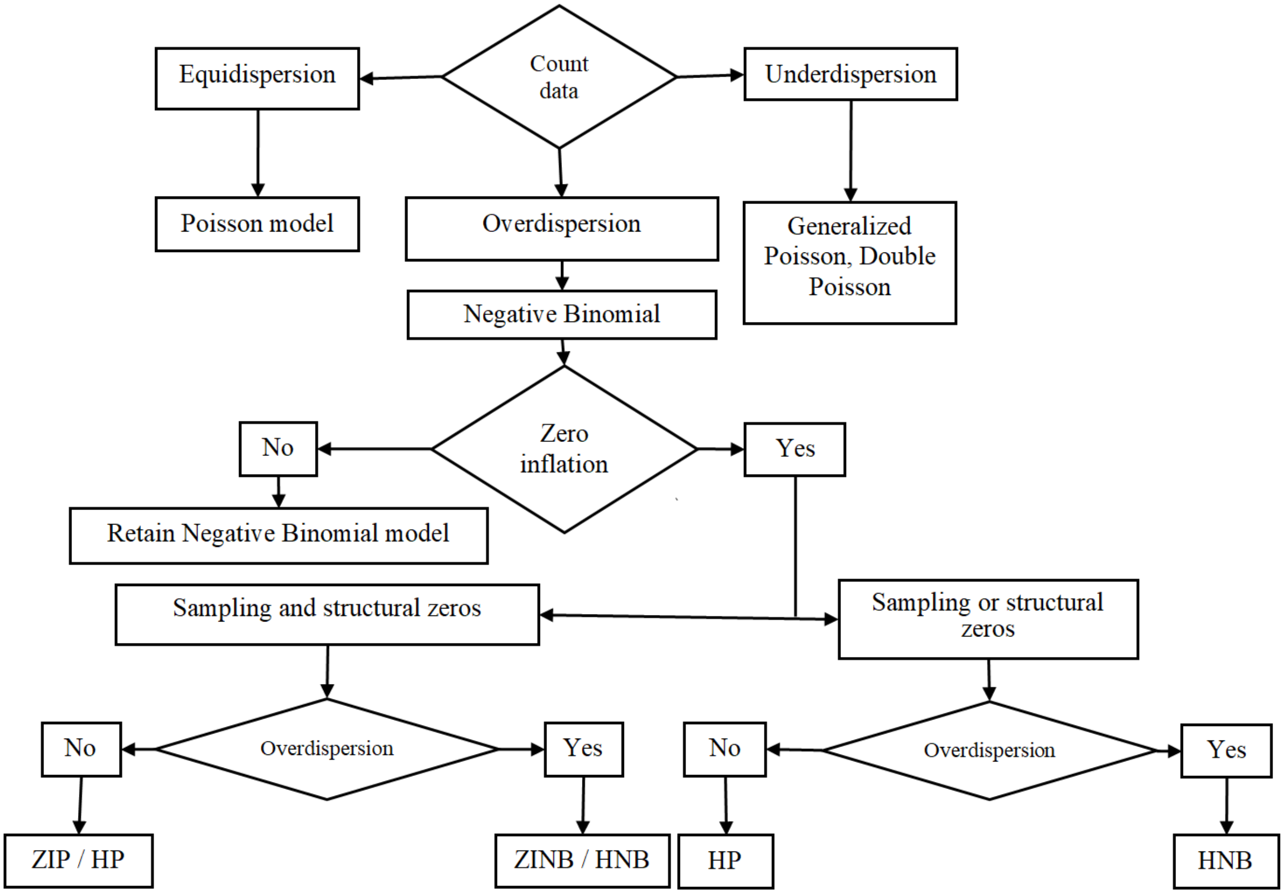
The estimates from the regression models in comparison with the parameters are plotted and given as Supplementary material I. The Poisson regression was less robust to the misspecification of assumptions. The estimates were well approximated to the parameters even in the presence of overdispersion. However, deviation from equidispersion due to excess incidence of zeros had impact on the estimation of parameters. Due to the separate origin, excess zeros and overdispersion, the estimated coefficients of Poisson and NB models were different from the parameters. These models underestimated β_1 and overestimated β_2. The level of bias in estimation increased with increase in proportion structural zeros. Though, negative binomial regression is a single distribution based model, it was capable to model count data with excess incidence of zeros, especially when the mean was low. The HNB and ZINB models estimated coefficients similar to the parameters. However, the coefficients of ZINB models were found to be sensitive to extremely larger percentage of structural zeros when the mean of the response variable was high. Among the zero inflated and hurdle models, the hurdle models were more robust to small sample size and lower proportions of zeros. The HNB model was found to be more robust to extreme situations with faster improvement in estimation with the increase in sample size over ZINB fit and it performed well for data with lower percentage of zeros as well as higher percentages of zeros. The differences in the estimated coefficients of the hurdle and zero inflated models showed that origin of the data has a role in estimating the parameters accurately.

Based on the analysis of real and simulated data sets a self-explanatory practical guideline is presented in Figure 1 for the analysis of count data. The data can be first fit with a Poisson regression followed by testing its equidispersion assumption. A test for overdispersion can be used for checking the assumption (Cameron & Trivedi, 1998). If the data is found to be overdispersed negative binomial regression can be applied. Though underdispersion is very rare to occur in count data, Double Poisson or generalized Poisson regression can be used if the Poisson regression model failed due to underdispersion. The negative binomial regression has better fit than Poisson regression for zero inflated data. Hence, if the data is observed with excess proportion of zeros, we have to go for zero inflated or hurdle regression model only if the negative binomial model fit fails. The need of zero inflated or hurdle negative binomial regression can be assessed by performing a test for zero inflation in negative binomial model fit (Wilson & Einbeck, 2019). If the test is rejected either a zero inflated or hurdle regression model can be used based on the type of zero assumed in the data. If the data is assumed to have either structural or sampling zeros hurdle regression can be fitted. The HNB model can be selected over HP (or ZINB over ZIP) based on the dispersion in the data. If the dispersion parameter estimated along with the regression coefficient in the HNB (or ZINB) is significant, we can assume overdispersion in the count regression part and select HNB model over HP model (or ZINB over HP). The models can also be chosen based on the BIC values. The model with lower BIC can be chosen as the best fitting model.

Discussion

Though, the application of Poisson regression is very limited for count data, the elaborated findings of this study recommends to start the count data analysis by fitting Poisson model followed by testing for equidispersion assumption. If the data is overdispersed, Poisson regression fails to capture the variance in the data that results in underestimating the standard errors of the estimates. The under-

Figure 1 ■ An empirical guideline for the analysis of count data



estimation of standard errors resulted in Poisson regression to identify more variables as significant for number of psychological and physical harms compared to negative binomial fit. If the data is underdispersed, generalized Poisson or double Poisson regression models can be used. However, the point to be noted here is that, underdispersion is rarely found in biomedical research (Coxe et al., 2009).

Overdispersion can occur due to either excess incidence of zeros or excess variation in the counts or as a combination of both. With the results of this study, it can be safely stated that, the negative binomial regression is a good alternative to Poisson regression, if the data is overdispersed due to the variation in counts. Though, the number of physical harms was ranging only from 0 to 4, excess proportion of zeros (90.25%) caused the Poisson regression to violate the equidispersion assumption. Hence, it was observed that excess zeros alone can cause overdispersion in count data. The results of the analysis showed that negative binomial regression has good capturing ability of zeros if the mean is small. Hence, before deciding on zero inflated or hurdle models it is always better to fit a negative binomial regression and assess the fit.

The reasons of zero inflation are broadly classified as bias in data collection and structural zeros due to the underline physical reason (Wilson & Einbeck, 2019). The selection between hurdle and zero inflated models for zero inflated data are topic of argument still. Rose et al. (2006) suggested to use zero inflated models when the origin of zeros are unsure and hurdle model when all the subjects are under risk. On the other hand, (Hu et al., 2011) argued that in hurdle models, all zeros are generated from a structural origin and the positive counts are generated from a sampling origin. However we recommend if the data is zero inflated, the selection between hurdle and zero inflated models can be decided based on the origin of zeros assumed. If the data contains only structural or sampling zeros, hurdle regression models are most suited as zero inflated models assumes both types of zeros in the data. Whereas if the data contains both structural and sampling zeros, either of zero inflated and hurdle regression models can be used. According to Hüls et al. (2017), the hurdle models can handle all type of zeros identically by definition and they never discriminate them based on their origin. The zero inflated models has the advantage of being able to differentiate zeros as structural and sampling and the hurdle models never discriminate zeros based on the origin. Nevertheless, hurdle models has the advantage of being more robust to small sample sizes. If the interest is to fit the model accounting the origin of zeros, zero inflated models can be used for larger sample sizes. Selection between the Poisson and NB versions of zero inflated and hurdle models can be decided based on the level of dispersion in the positive count part.

Conclusion

Although the Poisson regression model is the basic model used for count data, its application is very limited. Apart from the variation in the positive counts, a larger proportion of zeros alone can deviate a Poisson regression model from the equidispersion assumption. A negative binomial regression is more robust to zero inflated data; consequently, a higher proportion of zeros does not warrant the use of hurdle or zero inflated models. The need for a zero inflated or hurdle regression model should be assessed by checking the goodness of fit of the basic count regression models. The choice between zero inflated and hurdle models should be based on the assumed data generation process.

References

Cameron, A. C., & Trivedi, P. K. (1998). *Regression analysis of count data*. Cambridge University Press.

Coxe, S., West, S. G., & Aiken, L. S. (2009). The analysis of count data: A gentle introduction to poisson regression and its alternatives. *Journal of Personality Assessment*, *91*(2), 121–136. doi: 10.1080/00223890802634175.

Dunn, P. K., & Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and graphical statistics*, *5*(3), 236–244.

Feng, C., Li, L., & Sadeghpour, A. (2020). A comparison of residual diagnosis tools for diagnosing regression models for count data. *BMC Medical Research Methodology*, *20*(1), 175–199. doi: 10.1186/s12874-020-01055-2.

Greene, W. H. (1994). *Accounting for excess zeros and sample selection in poisson and negative binomial regression models* (tech. rep.). NYU Work Pap No EC-94-10.

He, H., Tang, W., Wang, W., & Crits-Christoph, P. (2014). Structural zeroes and zero-inflated models. *Shanghai Archives of Psychiatry*, *26*(4), 236–242. doi: 10.3969/j.issn.1002-0829.2014.04.008.

Hu, M.-C., Pavlicova, M., & Nunes, E. V. (2011). 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*, *37*(5), 367–375. doi: 10.3109/00952990.2011.597280.

Hüls, A., Frömke, C., Ickstadt, K., Hille, K., Hering, J., von Münchhausen, C., Hartmann, M., & Kreienbrock, L. (2017). Antibiotic resistances in livestock: A comparative approach to identify an appropriate regression model for count data. *Frontiers in Veterinary Science*, *4*, 71–99. doi: 10.3389/fvets.2017.00071.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics: A Journal of Statistics for the Physical, Chemical, and Engineering Sciences*, *34*(1), 1–99. doi: 10.2307/1269547.

Loeys, T., Moerkerke, B., De Smet, O., & Buysse, A. (2012). The analysis of zero-inflated count data: Beyond zero-inflated poisson regression: Zero-inflated poisson regression. *The British Journal of Mathematical and Statistical Psychology*, *65*(1), 163–180. doi: 10.1111/j.2044-8317.2011.02031.x.

Martin, T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, S. J., Tyre, A. J., & Possingham, H. P. (2005). Zero tolerance ecology: Improving ecological inference by modelling the source of zero observations: Modelling excess zeros in ecology. *Ecology Letters*, *8*(11), 1235–1246. doi: 10.1111/j.1461-0248.2005.00826.x.

McDowell, A. (2003). From the help desk: Hurdle models. *The Stata Journal*, *3*(2), 178–184. doi: 10.1177/1536867x0300300207.

McElduff, F. C. (2012). *Models for discrete epidemiological and clinical data* [Doctoral dissertation, UCL (University College London)].

Mills, E. D. (2013). Adjusting for covariates in zero-inflated gamma; zero-inflated log-normal models for semicontinuous data. doi: 10.17077/etd.7v3bafbd.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, *33*(3), 341–365. doi: 10.1016/0304-4076(86)90002-3.

Rose, C. E., Martin, S. W., Wannemuehler, K. A., & Plikaytis, B. D. (2006). On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. *Journal of Biopharmaceutical Statistics*, *16*(4), 463–481. doi: 10.1080/10543400600719384.

Wilson, P., & Einbeck, J. (2019). A new and intuitive test for zero modification. *Statistical Modelling*, *19*(4), 341–361. doi: 10.1177/1471082x18762277.

Xu, L., Paterson, A. D., Turpin, W., & Xu, W. (2015). Assessment and selection of competing models for zero-inflated microbiome data. *PloS One*, *10*(7), e01296061–99. doi: 10.1371/journal.pone.0129606.

Zorn, C. J. W. (1996). Evaluating zero-inflated and hurdle poisson specifications. *Midwest Political Science Association*, *18*(20), 1–16.

Open practices

The *Open Material* badge was earned because supplementary material(s) are available on the journal's web site.

Citation

Bhaskar, A., Thennarasu, K., Philip, M., & Jaisoorya, T. S. (2023). Regression models for count data with excess zeros: A comparison using survey data. *The Quantitative Methods for Psychology*, *19*(1), 1–13. doi: 10.20982/tqmp.19.1.p001.

Received: 16/07/2022 ∼ Accepted: 08/11/2022

Tables 3 to 5 follows.

Table 3 ■ Zero Inflated regression models fitted to Psychological and Physical harms

		Psychological harms				Physical harms			
		ZIP model (b(SE))		ZINB model (b(SE))		ZIP model (b(SE))		ZINB model (b(SE))	
Variable		Count part	Binary part	Count part	Binary part	Count part	Binary part	Count part	Binary part
Gender	Male	0.23(0.03)*	-0.71(0.07)*	0.26(0.04)*	-0.73(0.08)*	0.85(0.23)*	-0.65(0.31)*	0.88(0.24)*	-0.63(0.33)
Age (years)		0.03(0.01)*	-0.12(0.02)*	0.04(0.01)*	-0.13(0.02)*	-0.04(0.04)	-0.15(0.06)*	-0.04(0.04)	-0.15(0.07)*
Family structure	Living with other/relative	0.03(0.08)	0.05(0.23)	0.03(0.11)	0.03(0.26)	-0.66(0.40)	-1.19(0.93)	-0.68(0.39)	-1.38(1.02)
(Reference: Both parents)	Single parent	0.07(0.06)	-0.08(0.15)	0.09(0.07)	-0.09(0.17)	-0.05(0.23)	-0.71(0.47)	-0.06(0.24)	-0.83(0.54)
Family category	Below Poverty Line	0.02(0.03)	-0.08(0.07)	0.02(0.04)	-0.09(0.08)	-0.08(0.13)	-0.38(0.22)	-0.08(0.13)	-0.42(0.25)
Residence	City	0.11(0.05)*	0.13(0.13)	0.11(0.06)	0.16(0.15)	-0.12(0.25)	-0.38(0.44)	-0.10(0.24)	-0.40(0.46)
(Reference: Village)	Town	0.01(0.05)	-0.02(0.12)	0.02(0.06)	-0.04(0.13)	0.04(0.19)	-0.47(0.32)	0.06(0.19)	-0.49(0.34)
Location of school	Urban	-0.01(0.06)	-0.57(0.16)*	0.01(0.07)	-0.61(0.19)*	0.23(0.27)	-0.21(0.47)	0.27(0.27)	-0.15(0.51)
(Reference: Rural)	Semi Urban	0.01(0.04)	-0.01(0.09)	0.01(0.05)	-0.01(0.10)	0.18(0.18)	0.51(0.28)	0.16(0.19)	0.53(0.30)
Part time Job	Yes	0.13(0.03)*	-0.02(0.13)	0.13(0.05)*	-0.01(0.15)	0.27(0.15)	-0.48(0.29)	0.30(0.15)	-0.45(0.33)
Stay at Hostel	Yes	0.01(0.11)	-0.28(0.26)	0.04(0.13)	-0.24(0.30)	0.90(0.44)*	0.77(0.71)	0.96(0.45)*	0.96(0.77)
Substance use	Yes	0.09(0.03)*	-2.09(0.20)*	0.10(0.04)*	-2.86(0.52)	0.08(0.14)	-1.68(0.36)*	0.13(0.16)	-1.80(0.46)*
ADHD score		0.01(0.002)*	-0.03(0.01)*	0.01(0.002)*	-0.03(0.01)*	-0.004(0.01)	-0.03(0.02)*	-0.003(0.01)	-0.04(0.02)*
Psychological distress score		0.03(0.002)*	-0.05(0.01)*	0.03(0.002)*	-0.05(0.01)*	0.02(0.01)*	-0.04(0.02)*	0.02(0.01)*	-0.05(0.02)*
Theta (overdispersion)				1.65(1.13)		-	1.92(0.12)*		

Note. *: significant at 5% level of significance.

Table 4 ■ Hurdle regression models fitted to Psychological and Physical harms

		Psychological harms				Physical harms			
		HP model (b(SE))		HNB model (b(SE))		HP model (b(SE))		HNB model (b(SE))	
Variable		Count part	Binary part	Count part	Binary part	Count part	Binary part	Count part	Binary part
Gender	Male	0.23(0.03)*	0.76(0.06) *	0.26(0.04) *	0.76(0.06) *	0.74(0.23)*	1.27(0.11) *	0.77(0.25) *	1.27(0.11) *
Age (years)		0.03(0.01) *	0.13(0.02) *	0.03(0.01) *	0.13(0.02) *	-0.08(0.04)	0.09(0.03) *	-0.08(0.05)	0.09(0.03) *
Family structure	Living with other/relative	0.04(0.08)	-0.04(0.20)	0.05(0.11)	-0.04(0.20)	-0.33(0.54)	-0.15(0.31)	-0.39(0.60)	-0.15(0.31)
(Reference: Both parents)	Single parent	0.07(0.06)	0.10(0.13)	0.1(0.07)	0.10(0.13)	0.16(0.25)	0.36(0.18)	0.16(0.28)	0.36(0.18)
Family category	Below Poverty Line	0.02(0.03)	0.08(0.06)	0.02(0.04)	0.08(0.06)	-0.04(0.15)	0.19(0.10)	-0.06(0.17)	0.19(0.10)
Residence	City	0.11(0.05) *	-0.05(0.12)	0.11(0.07)	-0.05(0.12)	-0.08(0.28)	0.14(0.18)	-0.11(0.31)	0.14(0.18)
(Reference: Village)	Town	0.02(0.05)	0.02(0.10)	0.03(0.06)	0.02(0.10)	0.03(0.22)	0.36(0.15)	0.04(0.24)	0.36(0.15) *
Location of school	Urban	-0.01(0.06)	0.49(0.14) *	0.01(0.08)	0.49(0.14) *	0.10(0.30)	0.39(0.20) *	0.11(0.34)	0.39(0.20) *
(Reference: Rural)	Semi Urban	0.02(0.04)	0.01(0.08)	0.02(0.05)	0.01(0.08)	0.25(0.20)	-0.22(0.14)	0.27(0.22)	-0.22(0.14)
Part time Job	Yes	0.13(0.04) *	0.07(0.11)	0.14(0.05) *	0.07(0.11)	0.04(0.17)	0.63(0.13) *	0.04(0.19)	0.63(0.13) *
Stay at Hostel	Yes	0.002(0.11)	0.28(0.23)	0.01(0.14)	0.28(0.23)	0.56(0.54)	0.40(0.37)	0.60(0.61)	0.40(0.37)
Substance use	Yes	0.09(0.03) *	1.69(0.12) *	0.09(0.04) *	1.69(0.12) *	0.004(0.15)	1.13(0.11) *	-0.01(0.17)	1.13(0.11) *
ADHD score		0.01(0.002) *	0.03(0.004) *	0.01(0.002) *	0.03(0.004) *	0.01(0.01)	0.01(0.01) *	0.01(0.01)	0.01(0.01) *
Psychological distress score		0.03(0.002) *	0.06(0.01) *	0.03(0.002) *	0.06(0.01) *	0.03(0.01) *	0.05(0.01) *	0.03(0.01) *	0.05(0.01) *
Theta (overdispersion)				1.62(0.13) *		-		0.99(0.96)	

Note. *: significant at 5% level of significance.

Table 5 ■ Comparison of model fit indices for harms to others data

	Psychological harms					Physical harms				
	Log likelihood	BIC	Predicted zeros n (%) (Observed zeros-3611)	RQR †		Log likelihood	BIC	Predicted zeros n (%) (Observed zeros-5787)	RQR †	
Model				Mean (SD)	p value ‡				Mean (SD)	p value ‡
Poisson	-10038.3	20208	2439 (38.64)	-0.1(1.31)	0.001	-2257.7	4647	5706 (90.40)	-0.01(1.01)	0.006
NB	-8892.7	17926	3439 (54.48)	-0.01(1.01)	0.029	-2206.1	4552	5784 (91.64)	-0.01(1)	0.352
ZIP	-8641	17545	3613 (57.24)	0(1.03)	0.002	-2166	4596	5789 (91.71)	0.01(1)	0.509
ZINB	-8575	17422	3616 (57.29)	0(1)	0.178	-2166	4603	5790 (91.73)	0.01(1.01)	0.498
HP	-8644	17550	3611 (57.21)	0(1.03)	0.002	-2165	4594	5787 (91.68)	-0.01(1)	0.509
HNB	-8577	17425	3611 (57.21)	0(1)	0.136	-2164	4600	5787 (91.68)	0(1.01)	0.512

Note. †: The reported mean (SD) and p value are averaged of 1000 RQR, ‡: Kolmogorov Smirnov test for normality