Optimizing Healthcare Analytics: A Zero-Inflated Poisson Approach to Pediatric Emergency Room Visits

Nadashree Bose[1] and Hemlata Joshi[2]

[1,2]Department of Statistics and Data Science, CHRIST(Deemed to be University), Bengaluru, India
[1,2]Corresponding Authors: nadashree.bose01@gmail.com; hemlata.joshi28@gmail.com

Abstract

In various fields, the modeling of count data holds significant importance. The Poisson regression model is a commonly utilized tool for this purpose. However, this model assumes that the data has uniform dispersion, a condition often not met in real-world observations. The nature of overdispersion can vary depending on the specific context. When the overdispersion is primarily due to an excessive number of zero counts, the Zero-inflated Poisson regression model becomes a more suitable choice for modeling count data. The paper commences by offering a summary of the theoretical foundations of both Poisson regression and Zero-inflated Poisson regression. To evaluate their performance, use the Mean-Squared error (MSE) as a comparative metric. Next, apply these models to analyze the frequency of hospital emergency room visits by children between 10-18 years of age. The overdispersion of the visit count in our dataset is mostly caused by the excessive occurrence of zero counts. The findings demonstrate that the Zero-inflated Poisson regression model outperforms the standard Poisson regression model in terms of MSE.

Keywords : Poisson model, Zero-inflated Poisson model, Count data, Excess of zeroes, Mean-Squared error, Emergency room visit

1 Introduction

A generalized linear regression model (or GLM) formulated by [1] is made up of three components:

1. A random component, that defines the conditional distribution of the dependent variable Y_i, within the exponential family, such as the Binomial, Gamma, Poisson, Gaussian, or inverse-Gaussian families of distributions, given the values of the independent variables in the model.
2. A linear predictor - which is a linear function of explanatory variables,

$$\psi_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_j X_{ij} \tag{1}$$

3. Link function that can be inverted $g(\cdot)$, which establishes a connection between the mean of the response variable, $\theta_i = E(Y_i)$, to the linear predictor:

$$g(\theta_i) = \psi_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_j X_{ij} \tag{2}$$

The assumptions of a generalized linear model are as follows:

1. The response variable and the explanatory variables are not assumed to have a linear relationship in a generalized linear model. Instead, it presumes a linear relationship between the transformed expected response variable, as specified by the link function, and the independent variables.
2. There should be no multicollinearity among the explanatory variables.
3. Auto-correlation among the residuals should be absent.

Analyzing count data is a central concern in public health research, typically addressed through Poisson regression models. Poisson regression model is a type of generalised linear model where the response variable represents counts and it follows poisson distribution. These models assume that the mean and variance are equal, but this assumption is frequently violated in practical situations due to the turning up of numerous zeros or overdispersion. Overdispersion mainly occurs when the variance of the response variable exceeds its mean. When using the standard Poisson regression model in such cases, the variance of estimated coefficients tends to be underestimated.

To tackle this issue, this study adopts the ZIP model as the preferred count model. The ZIP model by [2] proves more suitable than the standard Poisson regression when dealing with count data exhibiting overdispersion or when there's an unexplained excess of zero counts. This model has the advantage of categorizing the population into two groups: one with zero outcomes and the other with non-zero outcomes. In the context of ZIP regression, the zero outcomes are modeled by including a proportion (π) of additional zeroes and a proportion $(1-\pi)$ of zeroes generated from a Poisson distribution.

This paper applies the Poisson regression model and the ZIP regression model to a dataset in the public health sector with an emphasis on the applications of these models in that field. Coefficient estimates in both scenarios is calculated and then the outcomes are compared using the MSE criterion. Our primary focus is to shed light on how these two methods differ in handling count data with an abundance of zero values.

The rest of the paper is organized as follows: In Section [2], the methodology is outlined, defining the Poisson Regression Model and the ZIP Regression Model, along with their respective maximum likelihood estimation procedures. We also introduce the MSE function in this section. Section [3] provides a practical application of these models in the context of public health, including a detailed description of the dataset, data analysis, and the presentation of results and significant findings. In Section [4], the paper is concluded, with an emphasis on the significance of applying such regression models in the specified domain. Finally, in Section [5], potential future research directions are explored, and suggestions for further investigation are offered.

2 Methodology

A generalized linear model (GLM) in statistics is an adaptable generalization of conventional linear regression. The GLM generalizes linear regression by enabling each measurement's variance to be a function of its predicted value and by allowing the linear model to be connected to the response variable via a link function.

In order to combine many other statistical models, such as linear regression, logistic regression, and Poisson regression, [1] created generalized linear models. They suggested an iteratively reweighted least squares approach for maximum likelihood estimation (MLE) of the model parameters.

Each result of the dependent variables Y is considered in a GLM to be produced from a specific distribution in the exponential family that includes the normal, binomial, Poisson, and gamma distributions. The independent variables, X, influence the distribution's mean, θ , in the following ways:

$$\mathrm{E}(Y|X) = \theta = g^{-1}(X\beta) \tag{3}$$

where $\mathrm{E}(Y|X)$ is the expected value of Y conditional on X; $X\beta$ is the linear predictor, a linear combination of unknown parameters β; g is the link function.

This approach assumes that the variance V is normally a function of the mean given by:

$$\mathrm{Var}(Y|X) = \mathrm{V}(g^{-1}(X\beta)) \tag{4}$$

Usually, Bayesian, maximum quasi-likelihood, or maximum likelihood estimation techniques are used to estimate the unknown parameters, β.

2.1 Poisson Regression

Poisson regression is a common method for modeling data for counts. It operates ψ the assumption that the response variable follows a Poisson distribution, and the expected value's logarithm can be expressed as a linear combination of unknown parameters. Let y_i represent the response variable. It is assumed that y_i adheres to a Poisson distribution with a mean θ_i, determined by a function of covariates x_i. Consequently, the Poisson regression takes the form:

$$P(y_i) = \frac{e^{-\theta_i}\theta_i^{y_i}}{y_i!}; y_i = 0, 1, 2, \ldots, \theta_i > 0 \tag{5}$$

with mean and variance, $E(y_i) = V(y_i) = \theta_i$. Here, θ_i is the conditional mean, defined as $\theta_i = E(y_i|x_i) = e^{x_i^T\gamma}$. The vector $x_i = (xi_1, xi_2, \ldots, xi_k)$ comprises the covariates, while $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_k)$ represents the vector of parameters which are unknown. The value of k specifies the dimension of the covariate vector within the model.

In order to estimate the Poisson regression's parameters, maximum likelihood approaches can be employed. Assuming that the observations $(y_i|x_i)$ are independent, the log-likelihood function is used to estimate the parameter vector γ is formulated as:

$$\ln L(\gamma) = \sum_{i=1}^{n} \left[y_i x_i^T \gamma - e^{x_i^T\gamma} - \ln(y_i!) \right] \tag{6}$$

By Fisher Scoring method the Poisson maximum likelihood estimator of γ is obtained as:

$$\hat{\gamma}_{PMLE} = (X'\hat{W}X)^{-1}X'\hat{W}\hat{z} \tag{7}$$

where, $\hat{W} = diag[\hat{\theta}_i]$ and $\hat{z}$ is a vector, with the i^{th} element given by $\hat{z}_i = \log(\hat{\theta}_i) + \frac{y_i - \hat{\theta}_i}{\theta_i}$.

An essential characteristic of the Poisson distribution is that the conditional variance of y_i equals λ_i. This property is known as equidispersion. When the variance of y_i exceeds λ_i, it indicates the presence of overdispersion.

2.2 Zero-inflated Poisson Regression

For modeling zero-inflation in count data, the zero-inflated Poisson (ZIP) model was proposed by [2]. For count data containing extra zeros, the ZIP model can be thought of as a mixing model. There are two categories of zeros that can be identified in the count data for the ZIP model. The first one, known as structural zeros, belongs to a non-susceptible group. The structural zero happens with a π chance. Random zeros are the second type of zeros in the count data for the ZIP model that come from a vulnerable group. The random zero has a Poisson distribution and occurs with probability $(1 - \pi)$. ZIP regression serves as a suitable model for count data marked by overdispersion and an abundance of zero values. This distribution blends characteristics of both the Poisson and logit distributions. The permissible values for the variable Y encompass non-negative integers, namely 0, 1, 2, 3, and so forth.

Each observation has two probable instances. In scenario 1, there is no count. However, if instance 2 occurs, a Poisson model is used to create counts (including zeros). Let's say that case 1 happens with probability π and case 2 happens with probability $(1 - \pi)$. Consequently, the following is a textual representation of the probability distribution of the Zero-inflated Poisson (ZIP) random variable y_i,

$$P(y_i = j) = \begin{cases} \pi_i + (1-\pi_i)e^{-\theta_i} & \text{if } j = 0 \\ (1-\pi_i)\frac{\theta_i^{y_i} e^{-\theta_i}}{y_i!} & \text{if } j > 0 \end{cases} \tag{8}$$

where π_i is described as the logistic link function below.

The Poisson component can include the exposure duration t, and the k regressor variables (the x's). These quantities are expressed as follows:

$$\theta_i = \exp(\ln(t_i) + \gamma_1 x_{1i} + \gamma_2 x_{2i} + \ldots + \gamma_k x_{ki}) \tag{9}$$

In many cases, $x_1 \equiv 1$, in which case γ_1 is referred to as the intercept. The unknown parameters that make up the regression coefficients $\gamma_1, \gamma_2, \ldots, \gamma_k$ are calculated from a set of data. Their estimates are represented by $b_1, b_2, \ldots, b_k$. This logistic link function π_i is provided by-

$$\pi_i = \frac{\lambda_i}{1 + \lambda_i} \tag{10}$$

where,

$$\lambda_i = \exp\left(\ln(t_i) + \tau_1 z_{1_i} + \tau_2 z_{2_i} + \ldots + \tau_m z_{m_i}\right)$$

The logistic component consists of a collection of m regressor variables (the z's) and an exposure duration t. The most common method of estimating the coefficients of ZIPRM is to use the maximum likelihood method. The likelihood function's logarithm is -

$$\mathcal{L}(\gamma; \tau) = \mathcal{L}_1 + \mathcal{L}_2 - \mathcal{L}_3 \tag{11}$$

where,

$$\mathcal{L}_1 = \sum_{i:y_i=0} \ln\left[\lambda_i + e^{-\theta_i}\right]$$

$$\mathcal{L}_2 = \sum_{i:y_i>0} \left(y_i \ln(\theta_i) - \theta_i - \ln(y_i!)\right)$$

$$\mathcal{L}_3 = \sum_{i=1}^{n} \ln(1+\lambda_i)$$

The slope of $\mathcal{L}$ is given by:

$$\frac{\partial \mathcal{L}}{\partial \gamma_r} = \sum_{i:y_i=0} \left(\frac{-x_{ir}\theta_i}{\lambda_i \exp(\theta_i)+1}\right) + \sum_{i:y_i>0} (y_i - \theta_i)\, x_{ir}, \text{ for } r = 1,2,\ldots,k \tag{12}$$

$$\frac{\partial \mathcal{L}}{\partial \tau_r} = \sum_{i:y_i=0} \left(\frac{z_{ir}\lambda_i \exp(\theta_i)}{\lambda_i \exp(\theta_i)+1}\right) - \sum_{i=1}^{n} \frac{\lambda_i}{1+\lambda_i} z_{ir}, \text{ for } r = 1,2,\ldots,m \tag{13}$$

The second derivatives are:

$$\frac{\partial^2 \mathcal{L}}{\partial \gamma_r \partial \gamma_s} = \sum_{i:y_i=0} \left(\frac{x_{ir}x_{is}\theta_i((\theta_i-1)\lambda_i \exp(\theta_i)-1)}{(\lambda_i \exp(\theta_i)+1)^2}\right) - \sum_{i:y_i>0} (\theta_i x_{ir} x_{is}), \text{ for } r,s = 1,2,\ldots,k \tag{14}$$

$$\frac{\partial^2 \mathcal{L}}{\partial \tau_r \partial \tau_s} = \sum_{i:y_i=0} \left(\frac{z_{ir}z_{is}\lambda_i \exp(\theta_i)}{(\lambda_i \exp(\theta_i)+1)^2}\right) - \sum_{i=1}^{n} \frac{z_{ir}z_{is}\lambda_i}{(\lambda_i+1)^2}, \text{ for } r,s = 1,2,\ldots,m \tag{15}$$

$$\frac{\partial^2 \mathcal{L}}{\partial \gamma_r \partial \tau_s} = \sum_{i:y_i=0} \frac{x_{ir}z_{is}\lambda_i\theta_i \exp(\theta_i)}{(\lambda_i \exp(\theta_i)+1)^2}, \text{ for } r = 1,2,\ldots,k; s = 1,2,\ldots,m \tag{16}$$

Using the simplex approach, this complex likelihood function is estimated. The final step of the estimated coefficients is defined as -

$$\hat{\gamma}_{ZIPMLE} = (X^T \hat{W} X)^{-1} X^T \hat{W} \hat{v} \tag{17}$$

where, W = diag$[\hat{\theta}_i]$ and $\hat{v}$ is a vector where the i^{th} element equals to $\hat{v}_i = ln(\hat{\theta}_i) + \frac{y_i - \hat{\theta}_i}{\hat{\theta}_i}$

The general form of Mean-Squared error(MSE) for the Poisson MLE and ZIP MLE is defined as:

$$MSE(\hat{\gamma}_{PMLE}) = MSE(\hat{\gamma}_{ZIPMLE}) = tr[(X'\hat{W}X)^{-1}] = \sum_{k=1}^{p} \frac{1}{\sigma_k} \tag{18}$$

where, σ_k is the eigen value of the $(X^T \hat{W} X)$ matrix.

3 Empirical Application

With the aid of actual data, both models' applications are demonstrated in this section. The drawbacks of the Poisson Regression Model when the regressand variable has a large number of zero counts are then discussed.

The Emergency Room Dataset sourced from the NHIS 2022 (U.S.) database is examined in this study, with the focal point being the count of emergency room visits made by children between the age of 10-18 years within the preceding 12 months. This dataset encompasses a total of 13 variables, with 12 covariates alongside the target variable, all of which are comprehensively detailed in Table 3.1. The data consists of 495 samples of children from the age group of 10-18 years.

Table 3.1: Variable description for the Emergency Room Dataset

Variable Name	**Variable Type**	**Variable Description**
Visits	Discrete	Emergency Room Visits by Children below age 18 (past 12 months)
Sex	Binary	Gender of the Child (1 if Male, 2 if Female)
Age	Discrete	Age of the Child (between 10-18 years)
PHstat	Categorical	General Health Status (1 if Excellent, 2 if Very Good, 3 if Good, 4 if Fair, 5 if Poor)
Height	Discrete	Height of the Child (in Inches)
Weight	Discrete	Weight of the Child (in Pounds and without shoes)
BMI	Categorical	Body Mass Index of the Child (1 if Underweight, 2 if Healthy Weight, 3 if Overweight, 4 if Obese, 9 if Unknown)
Medication	Categorical	Child needed Prescription Medication but did not get it due to cost (past 12 months) (1 if Yes, 2 if No, 8 if Not Ascertained)
Traffic	Categorical	Traffic causes Safety Issues in the Neighborhood (1 if Yes, 2 if No, 8 if Not Ascertained, 9 if Don't Know)
Behaviour	Categorical	Difficulty controlling Behaviour of the Child (1 if No Difficulty, 2 if Some Difficulty, 3 if a lot of Difficulty, 4 if Can't do it at all)
Dental	Categorical	Time since last Dental Exam/Cleaning (0 if Never, 1 if Within the past year, 2 if Within the last 2 years, 3 if Within the last 3 years, 4 if Within the last 5 years, 5 if Within the last 10 years, 8 if Not Ascertained, 9 if Don't Know)
Asthma	Binary	Asthma Emergency Room Visit by the Child (past 12 months)
Diabetes	Categorical	Child ever had Diabetes or not (1 if Yes, 2 if No, 9 if Don't Know)

The pertinent question here is how frequently a child has visited a hospital emergency room for concerns about their health over the course of the previous 12 months. 61% of them are male, and 39% are female. The minimum and maximum heights are determined to be 43 and 96 inches, respectively. The range of weights is 50 pounds at the minimum and 996 pounds at the maximum. Between the ages of 10 and 18, hospital emergency room visits by children have a mean of 0.3575 and a standard deviation of 0.9548. It is evident that the variance is significantly larger than the mean. This shows that the data is overdispersed. Also, in Figure 1 the distribution of emergency room visits is shown, where the excess of zeros is clearly visible. A Poisson regression model is typically not the best model to fit the data in these scenarios.

The 12 independent variables are encoded and the regression is first performed with the new 39 variables. The Shapiro-Wilk test is performed and the p-value is obtained, which is nearly equal to zero. This shows that the residuals are not normally distributed. Hence, a GLM is a better fit here. Among the 39 variables, 9 variables are chosen based on their significance. Those variables are, "medication.1", "medication.2", "dental.1", "dental.2", "dental.3", "dental.4", "dental.8", "asthma.1" and "diabetes.1". The significance of these variables are assessed with the help of Lagrange Multiplier Test or the Likelihood Ratio Test, which gives a p-value of 0.9975 (>0.05). The same variables are taken and fitted into both Poisson regression model and ZIP regression model and the summary is shown in Table 3.2 and 3.3. All the variables for the Poisson regression model and the count part of the ZIP regression model are significant based on the p-value, as it is less than 0.05.

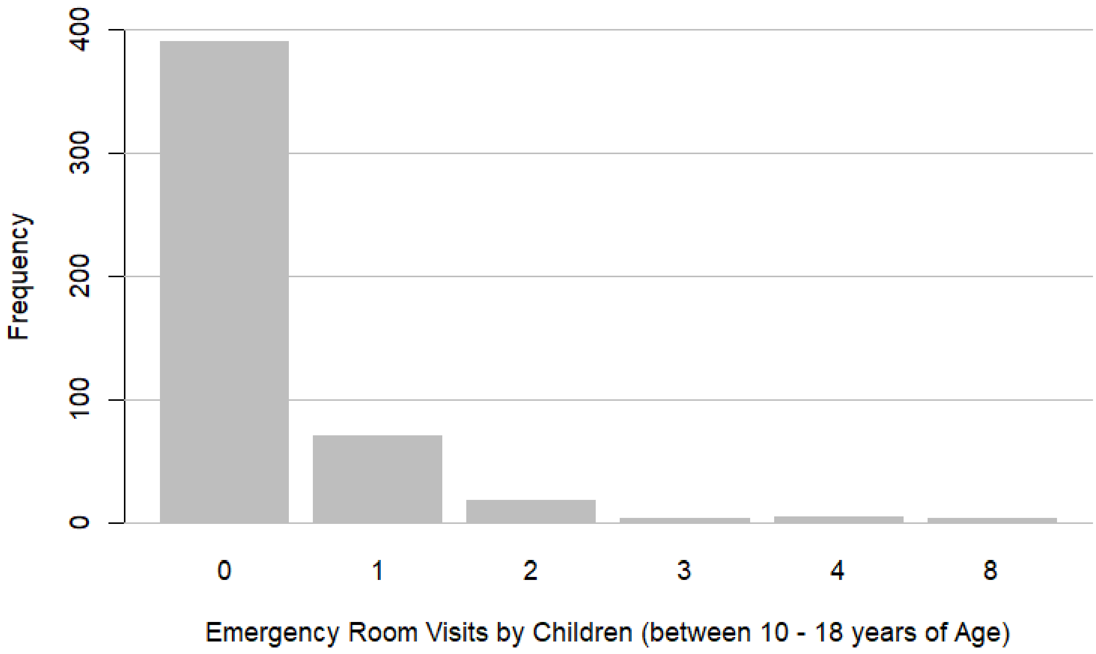

Figure 1: Frequency Distribution of Emergency Room Visits by Children(between 10-18 years of age)

Table 3.2: Estimation of Poisson Regression Model

	Estimate	**Std.Err.**	**z value**	**Pr(>|z|)**
(Intercept)	3.539637	0.621023	5.699685	<2e-16
medication.1	-3.32571	0.528658	-6.29087	<2e-16
medication.2	-3.53964	0.368335	-9.60983	<2e-16
dental.1	-1.4602	0.510558	-2.86	0.004236
dental.2	-1.89712	0.645497	-2.93901	0.003293
dental.3	-1.86835	0.766793	-2.43658	0.014827
dental.4	-2.203	0.878014	-2.50907	0.012105
dental.8	-1.4602	0.65371	-2.23371	0.025502
asthma.1	1.86142	0.205768	9.04621	<2e-16
diabetes.1	2.376486	0.458989	5.177655	<2e-16

Table 3.3: Estimation of ZIP Regression Model

	Pr(Y>0)				**Pr(Y=0)**			
	Estimate	**Std.Err.**	**z-value**	**Pr(>|z|)**	**Estimate**	**Std.Err.**	**z-value**	**Pr(>|z|)**
(Intercept)	3.8255	0.6567	5.8254	<2e-16	-24.851	301404.2	-0.0001	0.9999
medication.1	-2.8801	0.5229	-5.5079	<2e-16	-1.6684	626101.5	<2e-16	0.9999
medication.2	-2.4592	0.3996	-6.1534	<2e-16	25.9229	301404.2	0.0001	0.9999
dental.1	-1.746	0.5534	-3.1551	0.0016	-0.381	1.1923	-0.3196	0.7493
dental.2	-3.2634	0.662	-4.9299	<2e-16	-16.8143	3714.999	-0.0045	0.9964
dental.3	-2.5735	1.1341	-2.2692	0.0233	-1.4576	2.9148	-0.5001	0.617
dental.4	-2.1969	0.9143	-2.4029	0.0163	43.6785	548883.5	0.0001	0.9999
dental.8	-1.746	0.6877	-2.539	0.0111	0.5307	320948.8	<2e-16	0.9999
asthma.1	0.8306	0.2537	3.2742	0.0011	-60.4554	548886.5	-0.0001	0.9999
diabetes.1	1.296	0.4845	2.6751	0.0075	-17.1966	2714.247	-0.0063	0.9949

The performance of the models are evaluated in terms of Mean-Squared Error(MSE) and Akaike Information Criteria(AIC) value. The results are shown in Table 3.4.

Table 3.4: Performance Metrics of Poisson and ZIP regression Models

	MSE	***AIC***
Poisson	5.592646	677.3876
ZIP	0.387683	661.0843

The findings presented in Table 3.4 unequivocally demonstrate that the ZIP Regression Model outperforms the Poisson Regression Model, as evidenced by its lower MSE. Furthermore, the AIC for the ZIP Regression Model is also notably lower, underlining its superior fit to the data. Hence, in the presence of overdispersion in the count dependent variable, a ZIP regression model better explains the overabundance of zeros than the general Poisson regression model.
The count part of the zero inflated poisson regression model is given by:

$$\begin{aligned}\text{Poisson with log link } \ln(\theta) =& log(3.8255) - 2.8801(medication.1) - 2.4592(medication.2)\\ &- 1.7460(dental.1) - 3.2634(dental.2) - 2.5735(dental.3)\\ &- 2.1969(dental.4) - 1.7460(dental.8) + 0.8306(asthma.1)\\ &+ 1.2960(diabetes.1)\end{aligned} \tag{19}$$

Asthma.1 and diabetes.1 have positive signs, indicating that a positive change in these factors induces an increase in the number of visits to the emergency room by children of age 10-18 years. The number of positive signs the same with the Poisson regression model. However, there are differences in the percentage changes. The percentage change in medication.2 for the Poisson regression model is -97%, and for the ZIP regression model is -91%, indicating those who could afford medication did not have to visit the emergency room as often than those who haven't. Similarly, in asthma.1 and diabetes.1 the percentage of change for Poisson regression model and ZIP regression model are 543%, 976%, and 129%, 265%, respectively. It indicates those who had visited the emergency room of a hospital due to asthma in the past 12 months have higher chance of getting admitted than those who did not go, and those who ever had diabetes had a greater chance of visiting the emergency room against those who did not ever have diabetes. Odds ratio can be regarded while interpreting the parameters of the zero outcomes model. The odds ration in dental.2 is almost 100%, indicating those who did not have any dental clinic visit within last 2 years, have 100% higher probability of visiting the emergency room than the others.

By means of all these findings, it is simple to conclude that the ZIP regression model surpasses the poisson regression model in terms of accurately predicting the quantity of visitors to a hospital's emergency room. This analysis underscores a key insight, in discrete data scenarios where an overabundance of zero counts is observed in the dependent variable, the ZIP Regression Model consistently outperforms the Poisson Regression Model.

4 Conclusions

In conclusion, this research highlights the critical significance of count data modeling in various fields, particularly in the realm of public health. While the Poisson regression model serves as a common tool for this purpose, it often assumes uniform dispersion, a limitation frequently encountered in real-world datasets. This study demonstrates that the ZIP regression model emerges as a superior choice when dealing with overdispersion caused by an excess of zero counts. Through a meticulous analysis of hospital emergency room visit data for children between 10-18 years of age, this article have substantiated that the ZIP Regression model consistently outperforms the traditional Poisson Regression model, as evidenced by the lower MSE. These findings underscore the necessity of adopting ZIP Regression models in public health research, where datasets with excess zero counts are prevalent. Adopting this model can result in more precise and informative studies, which will ultimately improve healthcare decision-making and resource allocation in the public health sector.

5 Future Scope

The article's findings suggest several compelling future research directions in count data modeling, especially in the context of public health. These include applying the ZIP regression model to more complex healthcare datasets, refining the model itself, exploring predictive modeling for healthcare service utilization, analyzing spatial and temporal patterns, and investigating clinical applications and health policy evaluations. Additionally, integrating machine learning techniques into count data modeling offers exciting prospects for enhancing our understanding of healthcare utilization and improving evidence-based decision-making in public health. These diverse avenues promise to expand the field's horizons and address critical healthcare challenges.

References

[1] Nelder,J.A., Wedderburn,R.W.M.: Generalized Linear Models. Journal of Royal Statistical Society. 135(3), 370-384 (1972)

[2] Lambert,D.: Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. Technometrics. 34(1), 1–14 (1992)

[3] Nanjundan,G., Naika,T.R.: Asymptotic Comparison of Method of Moments Estimators (MME) and Maximum Likelihood Estimators (MLE) of Parameters in Zero-Inflated Poisson Model. Applied Mathematics. 3(6), 610-616 (2012)

[4] Mouatassim,Y., Ezzahid,E.H.: Poisson regression and Zero-inflated Poisson regression: Application to private health insurance data. European Actuarial Journal. 2(2), 187-204 (2012)

[5] Eggers,J.: On Statistical Methods for Zero-Inflated Models. Behavior Research Methods. 49(5), 1805-1825 (2017)

[6] Agresti,A.: An Introduction to Categorical Data Analysis. Wiley, New York (1996)

[7] Centers for Disease Control and Prevention. National Health Interview Survey. U.S. Department of Health & Human Services. https://www.cdc.gov/nchs/nhis/2022nhis.htm (2023). Accessed 22 Sept 2023