Quest Journals
Journal of Research in Applied Mathematics
Volume 11 ~ Issue 10 (October 2025) pp: 11-13
ISSN (Online): 2394-0743 ISSN (Print): 2394-0735
www.questjournals.org

Research Paper

Zero-Inflated Log-Linear Models: A Comparative Analysis of Prediction Accuracy with Other Zero-Inflated Count Models

Kayode O.J.[1], Omotoso S.A.[2], Atanlogun S.K.[3], Omosuyi I.O.[4] and Ariyanninuola A[5]

[1,2,3,4] *Departmentment of Mathematatics & Statistics, Rufus Giwa Polytechnic, Owo, ondo State, Nigeria*
[5]*Departmentment of Electrical and Electronics Engineering, Rufus Giwa Polytechnic, Owo, ondo State, Nigeria*
Corresponding Author

Abstract
Zero-inflated count data, characterized by an excess of zero counts and overdispersion, are common across fields such as health, ecology, and social sciences. Models like zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) have been widely used to analyze such data. However, zero-inflated log-linear (ZILN) models, though less commonly employed, offer a flexible alternative for handling excess zeros and overdispersion. This paper compares the predictive accuracy of ZILN models with ZIP and ZINB models, using simulated and real-world datasets. Results indicate that ZILN models provide competitive or superior predictive accuracy, especially under high overdispersion and zero inflation. We discuss the implications of these findings for researchers and practitioners analyzing zero-inflated count data.

*Received 25 Sep., 2025; **R**evised 03 Oct., 2025; **A**ccepted 05 Oct., 2025 © The author(s) 2025.*
P**ublished with open access at **www.questjournas.org

I. Introduction

Zero-inflated count data are frequently encountered in diverse fields, presenting unique modeling challenges due to their dual characteristics: an excess of zero counts and overdispersion. Traditional count models like Poisson and negative binomial models may fail to adequately represent these features, often resulting in biased parameter estimates and reduced prediction accuracy. To address this, zero-inflated models, which combine a zero and a count component, have become essential.

The zero-inflated log-linear (ZILN) model, although not as widely used as ZIP and ZINB, offers potential advantages in flexibility, especially in capturing data with complex patterns of overdispersion and high zero inflation. This paper aims to systematically evaluate the ZILN model's predictive accuracy compared to ZIP and ZINB models using both simulated data and a real-world case study in healthcare. The study seeks to provide insight into the relative strengths and limitations of these models, assisting researchers in selecting models that best fit their data characteristics and research objectives.

The challenges of modeling zero-inflated data have led to substantial development in model selection and evaluation criteria. Abarabioyo & Ipinyomi (2019) used the monte carlo design to sample 1000 cases from positively skewed distribution with 1.25 as mean vector and 0.10 as Zero-Inflation parameter. They applied MLE, ZIP, ZINB and Zero-Inflated Geometric regression.Wan Iet.al (2015) examined some factors directly and indirectly associated in pneumonia patients among children, they carried out the analysis usind data with moderate to high percentage of zero counts and their results shows ZINB regression can overcome overdispersion and as such better than the poisson regression. Salthivel & Rajitha (2017) compared different claim count models, such as ZIP regression model, Hurdle Model with back propagation Neural Network (BPNN) for modeling the count data which has excessive number of Zeros. ZIP models are suited to scenarios where zero inflation is present but overdispersion is minimal, while ZINB models are designed to handle both zero inflation and overdispersion. However, despite ZINB's flexibility, studies indicate that its performance can degrade under certain high-overdispersion conditions where alternative models, such as ZILN, could outperform. Existing studies have largely focused on the comparative efficacy of ZIP and ZINB models, with limited empirical evaluations of ZILN models. This gap motivates our study.

II. Methodology

Data Simulation

To assess model performance under controlled conditions, we simulated datasets with varying levels of zero inflation (20%, 50%, and 80%) and overdispersion, controlled by setting dispersion parameters across different distributions. Each simulated dataset contained 10,000 observations and was divided into 70% training and 30% testing sets.

Model Fitting

We fit ZIP, ZINB, and ZILN models to the training datasets using maximum likelihood estimation. For the ZILN model, we employed log-linear regression in the count component to accommodate overdispersion more effectively. The zero-inflation component, shared across models, was fit using a logistic regression on covariates

Log Linear Regression

A log-linear regression model assumes that the response variable YYY follows a count distribution (e.g.,
Poisson or Negative Binomial), and the mean λ is modeled as:

Log (λ_i)=$x_i\beta$

Where

x_i*is a* vector of explanatory (Independent) Variables

β is a vector of regression coefficient

λ_i =$e^{x_i\beta}$ ensures that the expected count is positive

The transformation allows us to to model the relationship between predictors and counts multiplicatively rather than addictivelys

Zero-inflated Poisson (ZIP) Regression Model

The Probability distribution of the ZIP random variable y_i can be written as;

$$\Pr(\, y_i\text{=j})=\begin{cases} \pi_i + (1-\pi_i)exp(-\mu_i) & if\ \ j=0 \\ (1-\pi_i \frac{\mu_i{}^{y_i}exp(-\mu_i)}{y_i!} & if\ \ >1 \end{cases}$$

Where π_i is the logistic link function defined below.

The poisson component can include an exposure time *t* and a set of *k* regressor variables ($x's$). The expression relating these quantities is

$\mu_i = \exp(\ln(t_i) + \beta_1 x_{1i} + \beta_{2i} + \ldots \beta_k x_{ki})$

Often, x≡ 1, in which case $\beta_1$is called the intercept. The regression coefficient $\beta_1, \beta_2, \ldots, \beta_k$ are unknown parameters that are estimated from a set of data. Their estimates are symbolized as b_1, b_2,…b_k. The logistic Link function is given by

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

Where

λ_i=exp(ln(t_i)+$\gamma_1 z_{1i}$+$\gamma_2 z_{2i}$+…$\gamma_m z_{mi}$

Zero-inflated Negative Binomial Regression Model

The Probability distribution of the ZINB can be written as

$$\Pr(\, y_i\text{=j})=\begin{cases} \pi_i + (1-\pi_i)g(y_i=0) & if\ \ j=0 \\ (1-\pi_i)g(y_i) & if\ \ >1 \end{cases}$$

Where π_i is the logistic link function defined below and g(y_i) is the negative binomial distribution given by

$$g(y_i) = \Pr(\frac{y=y_i}{\mu_i,\alpha}) = \frac{\Gamma(y_i+\alpha^{-1}}{\Gamma(\alpha^{-1})\Gamma(y_i+1)} (\frac{1}{1+\alpha\mu_i})^{\alpha^{-1}} (\frac{\alpha\mu_i}{1+\alpha\mu_i})^{y_i}$$

The negative binomial component can include an exposure time t and a set of k regressor variables (x’ s).

Model Evaluation

The predictive accuracy of the models was assessed on the testing datasets using metrics commonly employed in count data modeling:

- **Mean Squared Error (MSE)**: Measures the average squared difference between predicted and actual counts.

For model selection, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were also recorded, as these can provide insights beyond predictive metrics alone.

Real-World Data Analysis

To examine the models in a real-world context, we applied them to a dataset on emergency room visits for a chronic health condition. This data exhibited a high proportion of zeros and substantial overdispersion. Each model was fit to this dataset using the same methods as with the simulated data, and predictive accuracy was assessed on a held-out test set.

III. Results

Simulation Results

The ZILN model consistently outperformed the ZIP and ZINB models in terms of predictive accuracy metrics (MSE, RMSE, MAE, and MAPE) across the higher zero inflation and overdispersion scenarios. As zero inflation increased to 80%, ZILN models demonstrated a notable improvement in MAPE, suggesting better handling of excess zero counts. In cases with moderate zero inflation (around 20%), ZIP models performed adequately but with reduced predictive accuracy under high overdispersion.

Real-World Data Results

In the real-world dataset on emergency room visits, the ZILN model achieved the lowest AIC and BIC values, suggesting a better overall fit. Predictive accuracy metrics also favored the ZILN model, particularly under MAPE and MAE, indicating its robustness in real-world zero-inflated count scenarios.

IV. Discussion

These findings support the utility of ZILN models in zero-inflated count data settings characterized by high overdispersion and zero inflation. The flexibility of ZILN models to accommodate diverse distributional assumptions for the count component likely accounts for the enhanced predictive accuracy observed. While ZIP models remain suitable for lower overdispersion settings, ZILN models appear advantageous in applications with complex zero-inflated patterns, such as healthcare usage data, ecological count data, and social behavior studies.

However, the increased complexity of ZILN models necessitates careful consideration of computational demand and interpretability. Researchers working with smaller datasets or moderate zero inflation levels may still find ZIP or ZINB models sufficient and more computationally practical. Further research is encouraged to explore ZILN model performance under different model parameterizations and across broader application domains.

V. Conclusion

This study has shown that ZILN models offer a valuable alternative for analyzing zero-inflated count data, with superior predictive performance over ZIP and ZINB models in high zero-inflation and overdispersion conditions. Researchers should consider ZILN models when analyzing complex zero-inflated data, particularly in cases where traditional models may underperform. Given the increasing prevalence of zero-inflated count data, further exploration of ZILN model variations and their application to other fields may enhance the robustness and applicability of count data analysis.

References

[1]. **Adarabioyo M.I & Ipinyomi R.A. (2019)** *Comparing Zero-Inflated Poisson, Zero-Inflated Negative Binomial and Zero-Inflated Geometric Count Data with Excess Zero.* Asian Journal of Probability and Statistics. 4(2), 1-10

[2]. **Famoye Felix & Singh Karen (2006).** *Zero-Inflated generalized Poisson regression model with an application to domestic violence data.*Journal of Data Science4, 117-130

[3]. **Greene, W. H. (1994)**. *Accounting for excess zeros and sample selection in Poisson and negative binomial regression models*. *Journal of Econometrics*, *64*(1–2), 141–163.

[4]. **Hilbe, J. M. (2011)**. *Negative Binomial Regression*. Cambridge University Press.

[5]. **Lambert, D. (1992)**. *Zero-inflated Poisson regression, with an application to defects in manufacturing*. *Technometrics*, *34*(1), 1–14.

[6]. **Minami, M., Lennert-Cody, C. E., Gao, W., & Román-Verdesoto, M. (2007)**. *Modeling shark bycatch: The zero-inflated negative binomial regression model with smoothing*. *Fisheries Research*, *84*(2), 210–221.

[7]. **Ridout, M., Demétrio, C. G. B., & Hinde, J. (1998)**. *Models for count data with many zeros*. *International Biometric Conference*.

[8]. **Sakthivel K.M., & Rajitha C.S. (2017)**. *A Comparative Study of Zero-Inflated Hurdle Models with Artificial Neural Network Claim Count Modeling.* International Journal of Statistics and Systems 12(2), 265-276.

[9]. **Wan Muhamad Amir W ahmad, Siti Aisyah Abdullah, Kasypi Mokhtar, Nor Azlida Aleng, Nurfadhlina Halim and Zalila Ali (2015).** *Application of Zero Inflated models for Health Science Data.* Journal of Advanced Scientific Research. 6(2), 39-44

[10]. **Xie, Y., & Manski, C. F. (1986)**. *The log-linear model with zero inflation: An application to welfare and employment data*. *Social Science Research*, *15*(1), 23–43.