RESEARCH ARTICLE | FEBRUARY 23 2024

Zero-inflated regression models for measuring accident

Nurani Hartatik ✉; Joewono Prasetijo; Yudi Dwi Prasetyo; Khilda Nistrina; Atqiya Muslihati

AIP Conf. Proc. 2838, 070013 (2024)
https://doi.org/10.1063/5.0180308

Zero-Inflated Regression Models for Measuring Accident

Nurani Hartatik[1,2, a)], Joewono Prasetijo[2, b)], Yudi Dwi Prasetyo[3, c)], Khilda Nistrina[4, d)] and Atqiya Muslihati[5, e)]

[1]*Department of Civil Engineering, Universitas 17 Agustus, Surabaya, Indonesia*
[2]*Department of Transportation Engineering Technology (STTS), Universiti Tun Hussein Onn Malaysia, Johor, Malaysia*
[3]*Directorate General for The National Road Implementation East Java – Bali of Ministry of Public Works Indonesia, Sidoarjo, Indonesia*
[4]*Faculty of Information Technology, Universitas Bale Bandung, Indonesia*
[5]*Faculty of Applied Sciences and Technology, UTHM, Johor, Malaysia*

[a)]*Corresponding author: nuranihartatik@untag-sby.ac.id*
[b)] *joewono@uthm.edu.my*
[c)] *yudiprasetyo875@gmail.com*
[d)] *khildanistrina@unibba.ac.id;*
[e)] *atqiyamus313@gmail.com*

Abstract. Worldwide, hundreds of thousands of deaths and thousands more are injured every year in traffic accidents around the world. This is owing to an increase in road traffic throughout time, as well as a wide range of traffic compositions. Nowadays, road accidents have become a major concern, and analyzing accidents data has become an important concern for analysts. Therefore, analysis of accidents data requires a lot of attention because accident data is very complex. The road accidents process results in various frequency calculations, for example, deaths and injured number, and/or involved cars in the accidents. However, the probability distribution governing the occurrence of this count may be different. In addition to the problem of excess zeros, lack of data is a common occurrence in results of traffic accidents. Thus, this study discusses the use of the zero-inflated model in analyzing traffic accidents and the variables used by the researchers by reviewing the literature related to the use of zero-inflated models in accident cases. To find a better zero-inflated model that can be used to calculate accident data and to identify the variables that are commonly used to calculate traffic accidents. The result showed that the models that are more widely used by researchers to calculate traffic accidents are commonly known as (ZINB) the zero-inflated negative binomial model and (ZIP) the zero-inflated Poisson model. Both model types have been used since they are approaches for resolving the problem of overdispersion.

INTRODUCTION

The outcomes of interest in health, psychological, economic, and social studies are typically very infrequent behaviors and phenomena. When accounting for the occurrence of particular behavioral events, such as the number of school absences, unhealthy, hospitalization, cigarettes smoked, and automobile accidents number, research studies frequently use a lot of zero data. Vehicle accidents are a problem that requires serious handling. For this reason, the study that needs to be done is to analyze the existing traffic accident data. Data analysis was conducted to identify accident-prone areas (black spots).

Factors such as human (driver), environment, and good roads equipped with complete supporting facilities have a lower accident rate compared to roads that have fewer supporting facilities. Traffic accidents are influenced by infrastructure, road geometric conditions, and several other factors. Therefore, the Handling of accidents must be

Proceedings of the 8th International Conference on Engineering, Technology, and Industrial Applications 2021 (8th ICETIA 2021)
AIP Conf. Proc. 2838, 070013-1–070013-7; https://doi.org/10.1063/5.0180308
Published by AIP Publishing. 978-0-7354-4864-3/$30.00

carried out based on the causal factors. Count data is a sort of data in which the values are usually non-negative, have a bottom bound of zero, and have a lot of zeros and overdispersion (i.e., greater-than-expected variability). One of the analytical methods that can be used to calculate zero is using regression analysis.

The relationship between the dependent and independent variables can be modeled using regression analysis. The Poisson regression model is one of the regression models that can be used to examine the relationship between the dependent variable in discrete data and the independent variable is continuous, discrete, or mixed data. There are various assumptions in the Poisson regression model. The variance of the dependent variable must be equal to the mean, which is one of the assumptions that must be met (equidispersion) (1).

Various assumptions must be met in the Poisson regression model, including the response variable having a Poisson distribution, no multicollinearity between predictor variables, and equidispersion. The mean and variance of the same response variable are referred to as equidispersion. The mean of the dependent variable is the same as the mean of the independent variable (equidispersion) (2). In practice, these assumptions are not always met, such as when the variance value exceeds the average value, a condition known as overdispersion. An excess of zeros, or a scenario in which the proportion of zero values in the response variable data is greater than other values, can cause overdispersion of count data (3).

The estimated Zero Inflated Poisson (ZIP) regression model can be employed in more than zero situations. Excess zeros, according to the theory, are generated by a different process from computed values and can be modmodeleddependently. As a result, the ZIP model is divided into two parts: a Poisson calculation model and a logit model for predicting zero excess. This ZIP model is highly recommended in some instances, according to (4), while the Zero Inflated Poisson model is more successful for many zero-outcomes, according to (5). Like The observed data can be described as a modified count model with two parts: the first portion yields the sum of the excess zeros and zeros from the Poisson distribution, while the second part yields the postpositive-summ the truncated Poisson distribution (6). The Zero Inflated Poisson (ZIP) regression model is not suited for modeling the data in conditions of excess zero overdispersion, and the model that is formed will yield a skewed estimate. This has led to the creation of statistical models that use the Zero Inflated Negative Binomial (ZINB) regression model to tackle the difficulties of excess zero overdispersion.

The Zero Inflated Model (ZI) includes multiple models in addition to the ZIP, including the Zero-Inflated Poisson (ZIP), the Zero Inflated Negative Binomial (ZINB), the Zero Inflated Binomial (ZIB), and the Zero Inflated Generalized Poisson (ZIGP) (7). In comparison to other common regression models, the Zero Inflated (ZI) model was cited more than 1000 times as a title, abstract, or keyword in all articles. As a result, the ZI model is regarded a more advanced model that is required to appropriately calculate the statistic's excess zero (8).

In general, the ZI model can be thought of as a mixed distribution of two components, including arithmetic distributions such as Poisson, binomial, negative, or geometric binomials, and distributions of degenerates at zero. This ZI regression model differs from the others in terms of the nature of the arithmetic distribution used for the probability mass function. So far, the ZI model that is widely used is the ZIP regression model proposed by (4). Therefore, this paper is a review to compile and compare some developed models for measurinaccidentsnt in different variable amodelsdel from the previous researchers. The goal of this research is to develop a better zero-inflated model for calculating accident data and to identify the variables that are often utilized in traffic accident calculations.

DISCUSSIONS

Zero Inflated Model in Accidents

The method used in this research is a literature review, which is a study conducted to analyze selected literature from several sources so that it generates new conclusions and ideas. The journals used in this study are journals that discuss topics with keywords, namely the calculation of traffic accidents number using the ZI model. Frequently, accident data is tabulated as categorical or calculated data. The most prevalent strategies for examining the cause-and-effect relationship behind the occurrence of count events have typically been Poisson regression and negative binomial regression models (9).

The model offers both benefits and drawbacks. As a result, accident data analysis necessitates a great deal of attention due to the complexity of the data. The road accidents process results in various counting frequencies, for example, the number of injured, deaths, and people or cars involved in the accidents. The probability distribution governing the occurrence of this calculation may be different. One of the models that can be used is the zero-inflated

model. Several researchers have used this model for analysis. The variables utilized to produce traffic accident statistics are listed in **Table 1** as a result of journal analysis.

TABLE 1. accidents analysis variable using zero-inflated model

No	Author	Variable
1.	Chimba, *et.al*, (2014) (10)	➢ Geometry ➢ Number of accidents
2.	Ayati and Abbasi, (2014) (11)	➢ Geometry ➢ Load
3.	Prasetijo and Musa, (2016) (5)	➢ Type of vehicle
4.	Sapuan, *et.al* (2016) (12)	➢ Type of accidents ➢ Geometry
5.	Rizaldi, *et.al* (2017) (13)	➢ Type of accidents ➢ Geometry ➢ Type of vehicle ➢ Number of accidents
6.	Weihong and Zhenzou (2018) (14)	➢ Type of accidents ➢ Geometry
7.	Fisa, *et.al* (2019) (15)	➢ Type of accidents ➢ Geometry
8.	Elvik, *et.al* (2019) (16)	➢ Type of accidents ➢ Geometry
9.	Raihan, *et.al* (2019) (17)	➢ Type of accidents ➢ Geometry ➢ Number of accidents
10.	Azeze, *et.al* (2020) (18)	➢ Gender ➢ Age of driver ➢ Driving under fatigue ➢ Does not give priority ➢ Type of accidents ➢ Geometry ➢ Number of accidents
11.	Xujia, *et.al* (2020) (19)	➢ Type of vehicle ➢ Gender ➢ Age of driver ➢ Geometry ➢ Load
12	Worku and Tesfaw, (2020) (20)	➢ Type of vehicle ➢ Gender ➢ Age of driver ➢ Driving under fatigue ➢ Does not give priority ➢ Type of accidents ➢ Geometry ➢ Load ➢ Number of accidents
13	Yaghoubi, *et,al* (2020) (21)	➢ Type of vehicle ➢ Geometry ➢ Number of accidents
14	Borgoni, *et.al* (2021) (22)	➢ Type of vehicle ➢ Geometry
15	Redon, *et.al* (2021) (23)	➢ Geometry ➢ Number of accidents

Shankar et al. (1997) used an improved zero probability process to estimate the frequency of accidents (24). Shankar et al. (2003) also used the zero-inflated Poisson (ZIP) model to account for excess zero in the outcome variable in accidents involving pedestrians and motorized traffic (25). Using Poisson regression, we can determine the most relevant accidents factors and the type of vehicle that has the most accidents. Lukusa and Phoa (2009) utilize the ZIP model framework to overcome missing data in calculating traffic accidents data (26). In addition, some researchers also use two models to account for excess zero, such as Prasetijo and Musa, (2016), Worku and Tesfaw, (2020), and Yaghoubi, *et, al* (2020) (5)(20)(21). They combined the zero-inflated Poisson model (ZIP) with the zero-inflated negative binomial model (ZINB). Because the zero-inflated negative binomial provides a method for solving the overdispersion problem, both models are used. The Poisson regression has a problem in that when there are many zeros for the dependent variable when the accuracy of the Poisson regression is reduced. To solve this problem, zero-inflated Poisson regression (ZIP) has been introduced. After doing a comparison, Worku and Tesfaw, (2020) stated that the zero-inflated Poisson (ZIP) is the most fitted model for the road traffic accident dataset (20). Furthermore, Ayati and Abbasi, (2014), Weihong and Zhenzou (2018), Fisa, *et.al* (2019), and Azeze, *et.al* (2020) calculate accident data using the Poisson, Negative Binomial, and Zero-inflated Poisson (ZIP) and Zero-inflated Negative binomial (ZINB) models (11)(14)(15)(18). They demonstrated the superiority of Zero-inflated regression models over traditional Poisson and Negative binomial models. After doing a comparison between these models, according to Weihong and Zhenzou (2018), ZINB fits the data well as compared to other models (15).

Then, Sapuan, *et.al* (2016) utilizes Poisson and negative binomial distributions to calculate accident data (12). Rizaldi, *et.al* (2017), Elvik, *et.al* (2019), Raihan, *et.al* (2019), and Redon, *et.al* (2021) used the zero-inflated negative binomial model to calculate the predicted number of accidents (13)(16)(17)(23). In contrast to Borgoni, *et.al* (2021), who only uses the zero-inflated Poisson model in analyzing accident data (22). The zero-inflated model could correctly calculate the excess zero in the statistic produce a biased estimate in the condition of excess zero and equidisperssion that the basic model could not estimate. From **Table 1**, it can be concluded that the variables that are often used by researchers to analyze traffic accidents using the Zero Inflated model are: road geometry, number of accidents, types of accidents, payload, type of vehicle, gender, age of the driver, driving under fatigue, not giving priority. The purpose of the selection of variables is to select significant factors and determine the level of influence of traffic accidents factors. All researchers are working to minimize traffic accidents number and optimize overall road safety.

Poisson Regression (PR) Model

For an independent sample of n pairs of observations (y_i,x_i), $i \in 1, 2, \ldots, n$, where y_i denoted "the number of events that occurred" and x_i is the value of explanatory variables for the i^{th} subject (4). Assume $y_i \sim$ Poisson (μ_i), $I = 1, 2\ldots,n$ the probability density function of Poisson accidents variables, Y_i, is given by

$$P(y_i|\mu_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, y_i = 0,1,2,\ldots \quad (1)$$

Where $\mu > 0$, represents the expected number of occurrences in a fixed time. The variance and mean for the Poisson regression model iaregiven as follows:

$$E(y_i) = Var(y_i) = \mu_i \quad (1.1).$$

Overdispersion

When the variance of the count response variable exceeds the mean, $Var[y_i] > E[y_i], i = 1,2,3,\ldots,n$ a feature of overdispersion will occur (27). As a result, overdispersion trouble occurs, and the Poisson maximum likelihood estimator received may be wrong (28).

$H_o: \delta = 0$ (There is no overdispersion implies equidispersed), versus

$H_o: \delta > 0$ (Overdispersion exists in the dataset)

Negative Binomial Regression (NBR) Model

The random variable $Y_i, i = 1,2,3, \dots, n$ is a negatively binomial distribution count with parameters μ and δ the probability density function is expressed as follows (4):

$$f(y_i : \mu_i \delta) = \frac{r(y_i{}^{1}/_{\delta}}{r({}^{1}/_{\delta})y_i!}(1 + \delta\mu_i)^{-{}^{1}/_{\delta}}(1 + \frac{1}{\delta\mu_i})^{-yi}, y_i = 0,1,2, \dots \quad (2)$$

With mean and variance, respectively, given by

$$E(Y_i) = \mu_i = \exp(x_i^T \beta), and\ Var(Y_i) = \mu_i(1 + \delta\mu_i) \quad (2.1)$$

The term δ (read as delta) is called the dispersion parameter. If the dispersion parameter closes to null (δ $\rightarrow$0). Then the NBR model reduces to the classical Poisson regression model.

Zero Inflated Poisson (ZIP) Model

The ZIP distribution is assumed for two different underlying states. Firstly stateω_i, produces for only zeros, while to a standard Poisson count with mean $\mu_i$and hereafter a chance of extra zeros. In general, the first state is called structural zeros and the others state from the Poisson model is called sampling zeros (4). This two-state process gives the following probability mass function:

$$P(Y_i = y_i) = \begin{cases} \omega_i + (1 - \omega_i)^{e^{\mu_i}}, y_i = 0 \\ (1 - \omega_i)\frac{e^{-\mu_i}\mu_i{}^{y_i}}{y_i!}, y_i = 1,2,3, \dots \end{cases} \quad (3)$$

Where and μ_i. The parameter depends on the covariatesX_i and Z_i, respectively. The mean and the variance of the ZIP regression model, respectively are:

$$E(y_i) = \mu_i(1 - \omega_i) and\ Var(y_i) = \mu_i(1 - \omega_i)(1 + \omega_i\mu_i)$$

Zero Inflated Negative Binomial (ZINB) Model

The ZINB model is designed to demonstrate variables in the excess zeros and overdispersion. The researcher (4) demonstrated, the ZINB model given an appropriate design for the overdispersed response variable as compared and the ZIP model

$$P(Y_i = y_i) = \begin{cases} \omega_i + (1 - \omega_i)(1 + \delta_{\mu_i})^{\frac{i}{\delta_{\delta i}}}, y_i = 1 \\ (1 - \omega_i)\frac{r(y_i + {}^{i}/_{\delta})}{y_i! r({}^{1}/_{\delta})}(1 + \delta_{\mu_i})^{-{}^{1}/_{\delta}}\left(1 + \frac{1}{\delta_{\mu_i}}\right)^{-y_i}, y_i > 0 \end{cases} \quad (4)$$

Where$\delta > 0$, is a dispersion parameter.

The variance and the mean of the ZINB model are:

$$E(y_i) = \mu_i(1 - \omega_i), Var(y_i) = \mu_i(1 - \omega_i)(1 + \omega_i\mu_i + \delta_{\mu_i}) \quad (4.1)$$

The parameters depend on covariates X_i and Z_i, respectively. Then, the model is as follow:

$$\log(\mu_i) = x_i^T \beta, \log\left(\frac{\omega_i}{1-\omega_i}\right) z_i^T \gamma \dots \dots \dots \quad (4.2)$$

CONCLUSION

Poisson distribution is commonly used to model calculated data and assumes that the variance and distribution's mean are equal. The most of calculated data is obtained whose variance is greater than the average. Usually, this occurs due to the heterogeneity of the unobserved data. Another cause of excess variability is the occurrence of extra zero counts. In situations like this, the zero-inflated model is very suitable for use in data calculations. In analyzing an accident, variables are also needed. It can be concluded that the variables that are often used by researchers to analyze traffic accidents using the Zero Inflated model are: road geometry, number of accidents, types of accidents, payload, type of vehicle, gender, age of the drivers, driving under fatigue, not giving priority. The purpose of the

selection of variables is to select significant factors and determine the level of influence of traffic accidents factors. All researchers are working to minimize traffic accidents number and optimize overall road safety. Then, from the analysis of several kinds of literature, it can be concluded that the models that are more widely used by researchers to calculate traffic accidents are the zero-inflated Poisson model and the zero-inflated negative binomial model. Both models type have been used because the zero-inflated Poisson and the zero-inflated negative binomial are methods that can solve over-dispersion problem.

REFERENCES

1. Kusuma W, Komalasari D, Hadijati M. Model Regresi Zero Inflated Poisson Pada Data Overdispersion. J Mat. 2013;1–15.
2. Myers R., Montgemery D., Vining G., Robinson T. Generalized Linier Models with Applications in Engineering and The Sciences. Boston: Se. New Jersey: John Wiley and Sons; 2010.
3. Purnama DI. Perbandingan regresi zero inflated poisson (ZIP), regresi zero inflated negative binomial (ZINB) dan regresi hurdle negative binomial (HNB) untuk memodelkan data konsumsi rokok harian pendudukan dewasa di Indonesia. J Mat Stat komputasi. 2021;357–69.
4. Labert D. Zero-inflated poisson regression, with an application to defects in manufacturing," Technometrics. Technometrics. 1992;1–14.
5. Prasetijo J, Zahidah Musa W. Modeling Zero- Inflated Regression of Road Accidents at Johor Federal Road F001. MATEC Web Conf. 2016;1–7.
6. Sakthivel KM, Rajitha CS. Estimation of Zero Infolated Paramenter in Zero Inflated Poisson Model. J Math trends Technol. 2018;1–6.
7. Heilbron D. Zero-altered and other regression models for count data with added zeros. Biometrial. 1994;531–47.
8. T ML, Lee S-M, Li C-S. Review of ero-Inflated Models with Missing Data. Artik Curr Res Biostat. 2017;1–13.
9. Cameron A, Trivedi PK. regression analysis of cound data second edition. USA: Cambridge University Press; 2012.
10. Chimba D, Sando T, Kwigizile V, Kutela B. Modeling school bus accidents using zero inflated model. J Transp Stat. 2014;22–35.
11. Ayati E, Abbasi E. Modeling accidents on mshhad urban highways. Open J Saf Sci Technol. 2014;22–35.
12. Sapuan MS, Razali AM, Zamzuri ZH, Ibrahim K. Simulation on poisson and negative binomial models of count road accidents modeling. AIP Conf Proc. 2016;1–6.
13. Rizaldi A, Dixit V, Pande A, Junirman RA. Predicting casualty-accidents count by highway design standards compliance. Int J Transp Sci Technol. 2017;174–83.
14. Weihong M, Zhenzhou Y. Analysis and comparison of Traffic Accidents Regresion Prediction Model. Int Conf Electromechanical Control Technol Transp. 2018;364–9.
15. Fisa R, Nakazwe C, Michelo C, Musonda P. Modelling deaths associated with road traffic accidents and other factors on great north road in zambia between the years 2010 and 2016 using poisson model. Open Public Health J. 2019;1–10.
16. Elvik R, Sagberg F, Langeland PA. An analysis of factors influencing accidents on road bridges in norway. Accid Anal Prev. 2019;1–6.
17. Raihan MA, Alluri P, Wensong W, Albert G. Estimation of bicycle accident modification factors (CMFs) on urban facilities using zero inflated negative binomial models. Accid Anal Prev. 2019;303–13.
18. Azeze M, Seyoum A, Tesfa E, Debusho LK. Predictors of Human Death by Road Traffic Accidents in Bahir Dar City, North Western Ethiopia: A count Data Analysis Regression Model. Int J Theor Appl Math. 2020;95–104.
19. Xujia G, Xuedong Y, Ma L, Liu X. Modeling the service-route-based accident frequency by a spatiotemporal-random-effect zero-inflated negative binomial model: an empirical analysis for bus-involved accidents. Accid Anal Prev. 2020;1–9.
20. Worku G, Tesfaw D. The application of count regression models on traffic acidents in case of Addis Ababa, Ethiopia. Abyssinia J Sci Technol. 2020;26–33.
21. M. Yaghoubi M, Rassafi AA, Emrani N. Investigating the effect of road characteristics on fatal accident count and accident severity: case study Birjand-Qayen royte. AUT J Civ Eng. 2020;1–6.

22. Borgoni R, Gilardi A, Zappa D. Assessing the risk of car accidents in road network. Soc Indic Res. 2021;429–47.
23. Alvaro BR, Mateu J, Montes F. Modeling accidents risk at the road level through zero-inflated negative binomial models: A case study of multiple road networks. Spat Stat. 2021;1–9.
24. Shankar V, Milton J, Mannering F. Modeling accidents frequencies as zero altered probability processes: an empirical inquiry. Accid Anal Prev. 1997;829–37.
25. N. Shankar V, F. Ulfarsson G, M. Pendyala R, B. Nebergall M. Modeling accidents involving pedestrians and motorized traffic. Saf Sci. 2003;627–40.
26. Lukusa MT, Phoa FKH. A horvitz-type estimation on incomplete traffic accidents data analyzed via a zero-inflated poisson model. Accid Anal Prev. 2019;1–9.
27. Dean, C, F Lawless J. Tests for detecting overdispersion in poisson regression models. J Am Stat Assoc. 1989;467–72.
28. K Trivedi P, A Cameron C. Regression analysis of count data. Cambridge. UK: Cambridge University Press; 2013.