P-ISSN 2586-9000
E-ISSN 2586-9027
Homepage : *https://tci-thaijo.org/index.php/SciTechAsia*
Science & Technology Asia
Vol.28 No.4 October - December 2023
Page: [54-66]

Original research article

Comparison of the Effectiveness of Regression Models for the Number of Road Accident Injuries

Wikanda Phaphan[1,2], Nutnaree Sangnuch[1], Janjira Piladaeng[3,*]

[1] *Department of Applied Statistics, Faculty of Applied Science,*
King Mongkut's University of Technology North Bangkok, Bangkok 10800, Thailand
[2] *Research Group in Statistical Learning and Inference,*
King Mongkut's University of Technology North Bangkok, Bangkok 10800, Thailand
[3] *Department of Mathematics, Faculty of Science, Burapha University, Chon Buri 20131, Thailand*

Received 5 June 2023; Received in revised form 18 October 2023
Accepted 3 November 2023; Available online 27 December 2023

ABSTRACT

This article aimed to compare the effectiveness of regression models using data on the number of road traffic injuries from the Injury Surveillance System of the Thai Department of Disease Control between 2018 and 2022. The regression models used in this study for the count data include the Poisson, negative binomial, zero-inflated Poisson, zero-inflated negative binomial, and Conway-Maxwell-Poisson models. These were compared to find a suitable regression model to predict the number of road traffic injuries. The results show that the negative binomial regression model provides an appropriate regression equation for predicting the number of road traffic injuries because it gives the lowest Akaike information criterion (AIC). Moreover, this model can still be used as a preliminary tool for predicting the number of road accident injuries since it does not rely on many independent variables.

Keywords: Count data; Injury; Maximum likelihood; Overdispersion; Road accident

1. Introduction

Road traffic accidents claim the lives of approximately 1.3 million individuals annually. Between 20 and 50 million more people suffer nonfatal injuries, with many developing a disability as a result of their injuries. More than half of all road traffic fatalities involve vulnerable road users such as pedestrians, cyclists, and motorcycle riders. Despite the fact that low- and middle-income countries host approximately 60% of the world's vehicles, these countries account for 93% of all fatalities that occur on the world's roadways [1].

***Corresponding author:** janjira.pi@buu.ac.th
doi: xx.xxxxx/scitechasia.xxxx.xx

According to the data on the number of road accident injuries collected by the Injury Surveillance System of the Department of Disease Control, Ministry of Public Health (MOPH), Thailand [2], between 2018 and 2022, specifically for outpatient departments (OPD) found that males had a higher number of injuries from road accidents than females. Motorcycle accidents were to blame for a sizable portion of these injuries. The majority of the injuries occurred in the 15–19 age group. Therefore, the number of injuries caused by road accidents is an important indicator in assessing travel safety and this justifies the importance of analyzing the factors influencing the number of injuries occurring.

The study of accidents on Thailand's roads and the important indicators has been a popular topic and the subject of extensive researcharticle over the years [3–10]. These studies also examined the relevant indicators such as residential area by region, economic factors, energy consumption, population, age, gender, vehicle registration, road and environmental factors, festival and holiday seasons, etc. These variables affect various response variables, including the mortality rate, the number of fatalities, the number of accidents, and the number of accident injuries.

In the past, there have been studies of the number of people injured in road accidents with and without the use of models to examine the factors or indicators that affect this number [11–13]. A few studies have examined the relationship between the number of road accident injuries and factors such as road factors [6, 7], festival periods [6], and economic factors [10]; however, demographic factors (such as gender and age ranges) and accident types have not been investigated.

The number of road accident-related injuries is count data with non-negative integer values and a right-skewed or nonnormality distribution. Consequently, the occurrence of road accidents is not a continuous variable, and it is common to encounter data with a large number of zero values. To comprehend the relationship between independent (relevant factors) and count-dependent (road accident injuries) variables, several count regression models were studied to predict and explain the aforementioned relationship.

Poisson (POI) and negative binomial (NB) regression are two commonly used models for count regression. The POI regression model makes the assumption that the mean and variance of the count variable are equal. However, this may not be the case for many datasets since these datasets frequently exhibit overdispersion and underdispersion, which renders the Poisson regression model inappropriate [6, 14, 15]. For data with overdispersion, the NB regression model is the most frequent solution [14]. In addition, when there are numerous zeros for the dependent variable and a POI distribution is assumed, the accuracy of the POI regression model will be diminished. Hence, the zero-inflated Poisson (ZIP) regression model was developed to treat excess zeros [15–17]. Comparable to the ZIP regression model, the zero-inflated negative binomial (ZINB) regression model accounts for superfluous zeros in the count variable that assumes a NB distribution, and it can deal with the issue of overdispersion [18]. Furthermore, an alternative that is more flexible than the POI regression model is the Conway-Maxwell-Poisson (CMP) regression model. This model can be applied to both overdispersed and underdispersed data [6, 8].

As a result, the purpose of this study was to determine the relationship between the number of road traffic injuries and demographic factors (gender and age ranges) as well as the different accident types that occurred in Thailand between 2018 and 2022 by employing the POI, NB, ZIP, ZINB, and CMP regression models. In addition, we evaluated these models to find a suitable model for predicting the number of injuries in road accidents, which can be used for planning purposes to cope with the events and the number of road accident injuries that will occur in the future.

2. Materials and Methods

In this section, the data and the models employed will be summarized briefly.

2.1 Data source

The data used in this study are secondary data showing the number of road accident injuries (RAIs) from the Injury Surveillance System of the Department of Disease Control, MOPH, Thailand, between 2018 and 2022, specifically for the OPD [2]. The number of RAIs is the dependent variable (y), while the categorical independent variables (x) include gender (x_1), age ranges (x_2), and accident types (x_3).

2.2 Utilised models

For the sample size n, let y_i be the number of RAIs at the ith observation, and $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\top$ be the vector of p independent variables for the ith observation. Also, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ is a $p \times 1$ vector of the unknown regression coefficient. For $y_i = 0, 1, 2, \ldots$ and $i = 1, 2, \ldots, n$, the POI, NB, ZIP, ZINB, and CMP regression models can be summarized in the following subsection.

2.2.1 Poisson (POI) regression model

Let y_i be the independent POI random variable with mean parameter μ_i, the probability function for y_i given x_i is defined as

$$f(y_i|\boldsymbol{x}_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, \qquad (2.1)$$

where $\mu_i = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})$ and $\mu_i > 0$. The conditional mean and variance of the dependent variable are $E[y_i|\boldsymbol{x}_i] = Var[y_i|\boldsymbol{x}_i] = \mu_i$.

Under the assumption of independent observation, the log-likelihood function of the POI regression model is given by

$$\ln L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(-e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}} + y_i \boldsymbol{x}_i^\top \boldsymbol{\beta} - \ln y_i! \right). \qquad (2.2)$$

2.2.2 Negative Binomial (NB) regression model

The probability function of the NB regression model can be expressed as a function of μ_i and θ as follows:

$$f(y_i|\boldsymbol{x}_i) = \frac{\Gamma\left(y_i + \frac{1}{\theta}\right)}{\Gamma\left(\frac{1}{\theta}\right) y_i!} \left(\frac{1}{1+\theta\mu_i}\right)^{\frac{1}{\theta}} \times \left(\frac{\theta\mu_i}{1+\theta\mu_i}\right)^{y_i}, \qquad (2.3)$$

with $E[y_i|\boldsymbol{x}_i] = \mu_i$ and $Var[y_i|\boldsymbol{x}_i] = \mu_i(1+\theta\mu_i)$, where y_i is the negative binomial distribution count and θ is the overdispersion parameter. As $\theta > 0$, we can observe that $Var[y_i|\boldsymbol{x}_i] > E[y_i|\boldsymbol{x}_i]$. When $\theta = 0$, the POI distribution is a special case of the NB distribution under the POI assumption that $E[y_i|\boldsymbol{x}_i] = Var[y_i|\boldsymbol{x}_i]$.

Assuming that the independent variables are independent, the log-likelihood function of the NB regression model is given by

$$\ln L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ln\left[\Gamma\left(y_i + \frac{1}{\theta}\right)\right] - \sum_{i=1}^{n} \ln y_i!$$

$$-\sum_{i=1}^{n}\ln\left[\Gamma\left(\frac{1}{\theta}\right)\right]+\sum_{i=1}^{n}y_i\ln\theta+\sum_{i=1}^{n}y_i\boldsymbol{x}_i^\top\boldsymbol{\beta}$$
$$-\sum_{i=1}^{n}\left(y_i+\frac{1}{\theta}\right)\ln\left(1+\theta e^{\boldsymbol{x}_i^\top\boldsymbol{\beta}}\right). \qquad (2.4)$$

2.2.3 Zero-Inflated Poisson (ZIP) regression model

The ZIP model is a mixture model consisting of a Poisson distribution and a degenerate distribution at zero. When considering an independent random variable y_i with a ZIP distribution, it is assumed that the occurrence of zeros is linked to two different underlying states as follows:

$$y_i \sim \left\{\begin{matrix} k_{i0} & \text{, with probability } \pi_i \\ \text{POI}(\mu_i), & \text{with probability } 1-\pi_i, \end{matrix}\right\}, \qquad (2.5)$$

where k_{i0} is a degenerate distribution at zero. Then, the probability function describing the two components is as follows:

$$f(y_i|\boldsymbol{x}_i) = \begin{cases} \pi_i + (1-\pi_i)e^{-\mu_i}, & \text{if } y_i = 0 \\ (1-\pi_i)\frac{e^{-\mu_i}\mu_i{}^{y_i}}{y_i!}, & \text{if } y_i > 0, \end{cases} \qquad (2.6)$$

where $\mu_i > 0$ and $0 \leq \pi_i \leq 1$. The conditional mean and variance of the ZIP regression model are $E[y_i|\boldsymbol{x}_i] = \mu_i(1-\pi_i)$ and $Var[y_i|\boldsymbol{x}_i] = \mu_i(1-\pi_i)(1+\pi_i\mu_i)$, respectively.

In order to implement the ZIP regression model in practical modeling situations, the logit link is defined as follows:

$$\text{logit}(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{G}_i^\top\boldsymbol{\gamma}, \qquad (2.7)$$

where $\boldsymbol{G}_i^\top$ is the vector of independent variables, while $\boldsymbol{\gamma}$ is the vector of unknown parameters. Therefore,

$$\pi_i = \frac{e^{\boldsymbol{G}_i^\top\gamma}}{1+e^{\boldsymbol{G}_i^\top\gamma}}. \qquad (2.8)$$

In the event that π_i is not a function of μ_i, the log-likelihood function for the ZIP model is assumed below.

$$\ln L(\boldsymbol{\beta},\boldsymbol{\gamma}) = \sum_{\{i:y_i=0\}}\ln\left(e^{\boldsymbol{G}_i^\top\gamma}+e^{-e^{\boldsymbol{x}_i^\top\boldsymbol{\beta}}}\right)$$
$$-\sum_{\{i:y_i=0\}}\ln\left(1+e^{\boldsymbol{G}_i^\top\gamma}\right)-\sum_{\{i:y_i>0\}}\ln(y_i!)$$
$$+\sum_{\{i:y_i>0\}}\left(y_i\boldsymbol{x}_i^\top\boldsymbol{\beta}-e^{\boldsymbol{x}_i^\top\boldsymbol{\beta}}\right). \qquad (2.9)$$

2.2.4 Zero-Inflated Negative Binomial (ZINB) regression model

The ZINB model is proposed in order to illustrate variables with an excess of zeros and overdispersion. This model provides a better fit for the dependent variable with overdispersion than the ZIP model. Suppose that $y_i \sim \text{ZINB}(\mu_i,\theta,\pi_i)$, we can define the probability function of the ZINB regression model as follows.
If $y_i = 0$,
we get

$$f(y_i|\boldsymbol{x}_i) = \pi_i+(1-\pi_i)\left(\frac{1}{1+\theta\mu_i}\right)^{\frac{1}{\theta}}, \qquad (2.10)$$

while $y_i > 0$,

$$f(y_i|\boldsymbol{x}_i) = (1-\pi_i)\frac{\Gamma\left(y_i+\frac{1}{\theta}\right)}{\Gamma\left(\frac{1}{\theta}\right)y_i!}\left(\frac{1}{1+\theta\mu_i}\right)^{\frac{1}{\theta}}$$
$$\times\left(\frac{\theta\mu_i}{1+\theta\mu_i}\right)^{y_i}. \qquad (2.11)$$

The conditional mean and variation of the ZINB regression model are $E[y_i|\boldsymbol{x}_i] = \mu_i(1-\pi_i)$ and $Var[y_i|\boldsymbol{x}_i] = \mu_i(1-\pi_i)(1+\pi_i\mu_i+\theta\mu_i)$, respectively.

We obtain the log-likelihood function of the ZINB regression model in the same way as we did for the ZIP regression model. For $y_i = 0$, we obtain

$$\ln L(\boldsymbol{\beta},\boldsymbol{\gamma}) = \sum_{i=1}^{n}\ln\left(e^{\boldsymbol{G}_i^\top\gamma}+\left(\frac{1}{1+\theta e^{\boldsymbol{x}_i^\top\boldsymbol{\beta}}}\right)^{\frac{1}{\theta}}\right)$$

$$-\sum_{i=1}^{n}\ln\left(1+e^{\boldsymbol{G}_i^\top\gamma}\right). \quad (2.12)$$

The log-likelihood function is as follows for $y_i > 0$:

$$\begin{aligned}\ln L(\boldsymbol{\beta},\boldsymbol{\gamma}) &= -\sum_{i=1}^{n}\ln\left(1+e^{\boldsymbol{G}_i^\top\gamma}\right)\\ &+\sum_{i=1}^{n}\ln\left[\Gamma\left(y_i+\frac{1}{\theta}\right)\right]-\sum_{i=1}^{n}\ln\left[\Gamma\left(\frac{1}{\theta}\right)\right]\\ &-\sum_{i=1}^{n}\ln y_i!+\sum_{i=1}^{n}y_i\ln\theta+\sum_{i=1}^{n}y_i\boldsymbol{x}_i^\top\boldsymbol{\beta}\\ &-\sum_{i=1}^{n}\left(y_i+\frac{1}{\theta}\right)\ln\left(1+\theta e^{\boldsymbol{x}_i^\top\boldsymbol{\beta}}\right). \quad (2.13)\end{aligned}$$

2.2.5 *Conway-Maxwell-Poisson (CMP) regression model*

The CMP regression model, an extension of the POI regression with the parameters mean μ_i and dispersion parameter ν, can accommodate both overdispersion and underdispersion in the data. The probability function of the CMP model can be expressed as follows:

$$f(y_i|\boldsymbol{x}_i)=\frac{\mu_i^{y_i}}{(y_i!)^\nu}\frac{1}{Z(\mu_i,\nu)}, \quad (2.14)$$

where $Z(\mu_i,\nu)=\sum_{m=0}^{\infty}\mu_i^m/(m!)^\nu$, $\mu_i > 0$ and $\nu \geq 0$. If $\nu = 1$ and as $\nu \to \infty$, the CMP distribution becomes the standard POI distribution. On the other hand, $\nu < 1$ and $\nu > 1$ indicate overdispersion and underdispersion, respectively.

With the use of an asymptotic estimate for $Z(\mu_i,\nu)$, the conditional mean and the variance can be approximated as follows:

$$E[y_i|\boldsymbol{x}_i]\approx\mu_i^{\frac{1}{\nu}}-\frac{\nu-1}{2\nu}, \quad (2.15)$$

and $$Var[y_i|\boldsymbol{x}_i]\approx\frac{\mu_i^{\frac{1}{\nu}}}{\nu}. \quad (2.16)$$

The expression for the log-likelihood function for observation i of the CMP regression model is:

$$\begin{aligned}\ln L(\boldsymbol{\beta}) &= \sum_{i=1}^{n}\left(y_i\boldsymbol{x}_i^\top\boldsymbol{\beta}\right)-\nu\sum_{i=1}^{n}\ln y_i!\\ &-\sum_{i=1}^{n}\ln\left[Z\left(e^{\boldsymbol{x}_i^\top\boldsymbol{\beta}},\nu\right)\right]. \quad (2.17)\end{aligned}$$

2.3 Parameter estimation

The maximum likelihood technique is used to estimate the parameter $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ by calculating the first derivative of the log-likelihood function with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, and setting it to zero. The score equations are given as follows:

$$\frac{\partial\ln L(\cdot)}{\partial\boldsymbol{\beta}}=0, \quad (2.18)$$

and $$\frac{\partial\ln L(\cdot)}{\partial\boldsymbol{\gamma}}=0, \quad (2.19)$$

where $L(\cdot)$ is $L(\boldsymbol{\beta})$, $L(\boldsymbol{\gamma})$, or $L(\boldsymbol{\beta},\boldsymbol{\gamma})$. The resulting score equations are nonlinear with respect to the coefficients and do not have a closed form to estimate the parameters in the POI, NB, ZIP, ZINB, and CMP regression models. As a result, numerical methods must be taken into consideration to produce the maximum likelihood estimates (MLEs), such as the iteratively reweighted least squares [19], Newton-Raphson [20], or EM (expectation maximization) [21,22] algorithms, etc., which are available in standard programs.

2.4 Model comparison criteria

In this article, the likelihood-ratio (LR) test and the Akaike information criterion (AIC) were utilized to choose the most suitable model to fit the data. The following will provide an explanation of these criteria.

2.4.1 Likelihood-ratio test for nested model

The LR test is a statistical test that compares the goodness of fit of two statistical models to determine if adding complexity to a model significantly improves its accuracy. Begin by considering the situation in which there are two models, (1) simpler model (SM) or nested model and (2) complex model (CM), with the SM nested within the CM. The performance of the two models can be compared using the LR test statistic. In their most basic form, the hypotheses for the LR test are as follows:

H_0 : Data fit both the SM and CM.
The SM should be used.

H_1 : The data fit the CM better than the SM.
The CM should be used.

The likelihood-ratio test statistic is typically expressed as

$$\Lambda_{\mathrm{LR}} = -2\left[\ln L_0(\hat{\boldsymbol{\beta}}) - \ln L_1(\hat{\boldsymbol{\beta}})\right], \quad (2.20)$$

which is the difference between the log-likelihoods. Here, $\ln L_0(\hat{\boldsymbol{\beta}})$ and $\ln L_1(\hat{\boldsymbol{\beta}})$ represent the log-likelihood function values under the null and alternative hypotheses, respectively.

Assuming H_0 is true, the test statistic (Λ_{LR}) will be asymptotically chi-squared distributed (χ^2) with degrees of freedom equal to the dimensionality difference between CM and SM as n approaches ∞.

2.4.2 Vuong Test for Non-Nested Model

The Vuong non-nested test compares the predicted probabilities of two models that do not nest within one another. This test is typically applied to evaluate whether a count model with zero inflation is better than one with non-zero inflation. The hypotheses are as follows:

H_0 : Two models are comparable.

H_1 : Model 2 is superior to Model 1.

The relationship of the likelihood-ratio test is given by

$$m_i = \ln\left[\frac{f_1(y_i|\boldsymbol{x}_i)}{f_2(y_i|\boldsymbol{x}_i)}\right]. \quad (2.21)$$

Therefore, the Vuong test statistic is

$$V = \frac{\sum_{i=1}^{n} m_i}{sd_m\sqrt{n}}, \quad (2.22)$$

where sd_m is the sample standard deviation of m_i. For a large sample size and under the null hypothesis, the Vuong test statistic is asymptotically normally distributed by the central limit theorem.

2.4.3 Akaike Information Criterion

In statistics, the AIC is used to compare various possible models to find the best fit. The best-fit model is the one that explains the most variation with the fewest independent variables, and the better a model is, the lower the AIC. The formula for the AIC is as follows:

$$\mathrm{AIC} = -2\ln L(\hat{\boldsymbol{\beta}}) + 2q, \quad (2.23)$$

where q is the number of predictors in the model and $L(\hat{\boldsymbol{\beta}})$ is the value of the log-likelihood function at the model's estimated parameter vector ($\hat{\boldsymbol{\beta}}$).

2.5 Software utilized

The statistical software `R-4.2.3` and three major packages for count regression models (`MASS`, `pscl`, and `COMPoissonReg`) were used to analyse the data. In the `MASS` package, the `glm()` and `glm.nb()`

functions were used to construct the POI and NB regression models within a generalized linear model framework. The `zeroinfl()` function in the `pscl` package was used to build the ZIP and ZINB regression models. The `glm.cmp()` function in the `COMPoissonReg` package was used to construct the CMP regression model.

3. Empirical Results

This section presents the empirical findings of this study's analysis. Two major components comprise the analysis: descriptive analysis and model analysis.

3.1 Results of descriptive analysis

There were a total of 1,050 observations in the data set. When separated by gender, age ranges, and accident types, it was found that the number of RAIs ranged from 0 to 103,939. Moreover, no injuries were reported in some of the incidents accounting for 39 observations or roughly 3.71% of the total dataset.

Analyses of the data found that males accounted for 59.64% of RAIs, while females accounted for only 40.36%. RAIs were most common in those between the ages of 15 and 19, who accounted for 17.61%. Next came those aged 20-24 (12.25%), then those aged 10-14 (9.56%). Furthermore, the number of RAIs due to various modes of transportation can be broken down based on the accident types as follows: motorcycles tallied 82.69%, followed by bicycles (9.18%), cars (4.81%), pedestrians (3.01%), and buses (0.31%).

Table 1 shows that the number of RAIs had a larger variance (96,424,826) than their mean (3,699.61), indicating the count dependent variable's value has overdispersion. The histogram in Fig. 1 also deviates from the normal distribution because it is right-skewed (skewness > 0) and has a larger kurtosis than zero.

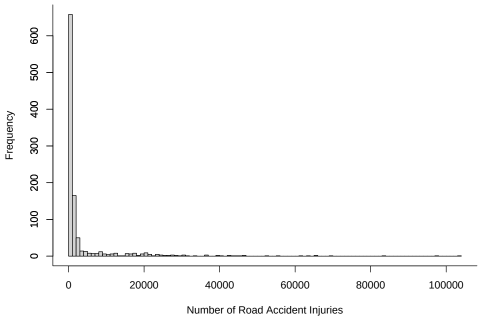

Fig. 1. Histogram for the number of road accident injuries.

Table 1. Summary statistics for the number of road accident injuries.

Mean :	3,699.61
Variance :	96,424,826.00
Skewness :	4.95
Kurtosis :	31.95

3.2 Results of model analysis and performance comparison

The relationship between the number of RAIs and independent variables such as gender, age ranges, and accident types was analyzed using count regression models. Cramer's V was used to analyze the intercorrelations of the independent variables before the establishment of the model. The investigation showed no significant correlations between variables. These results demonstrate that the independent variables are not multicollinear, supporting the regression models.

The estimated regression coefficients with standard errors (SE) are shown in Table 2, and can be used to create confidence intervals for the coefficients. This table also includes the AIC and log-likelihood values from the analysis. The count part modeling employed POI, NB, and CMP regression.

Table 2. Estimated parameters and standard errors from the Poisson, negative binomial, zero-inflated Poisson, zero-inflated negative binomial, and Conway-Maxwell-Poisson regression models.

Parameter	POI		NB		ZIP		ZINB		CMP	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Count Part										
Intercept	0.3245***	0.0420	0.5426***	0.1175	0.3441***	0.0424	0.5765***	0.1225	-0.7327***	0.0311
Gender										
Female (baseline)										
Male	0.3907***	0.0010	0.3432***	0.0342	0.3907***	0.0010	0.3440***	0.0344	1.2290***	0.0012
Age Ranges										
≥ 100 (baseline)										
95 - 99	-1.0934***	0.0836	-0.3499*	0.1609	-1.1129***	0.0838	-0.3843*	0.1722	-0.1730***	0.0489
90 - 94	0.8035***	0.0504	1.6254***	0.1413	0.7839***	0.0507	1.5911***	0.1462	-0.3640***	0.0564
85 - 89	2.4228***	0.0437	3.0485***	0.1365	2.4032***	0.0441	3.0143***	0.1409	-1.2030***	0.1222
80 - 84	3.5503***	0.0425	4.0282***	0.1349	3.5307***	0.0429	3.9941***	0.1391	-0.4395***	0.0600
75 - 79	4.2730***	0.0422	4.6332***	0.1343	4.2534***	0.0425	4.5992***	0.1384	2.1298***	0.0311
70 - 74	4.8799***	0.0420	5.1448***	0.1339	4.8603***	0.0424	5.1109***	0.1378	2.8579***	0.0310
65 - 69	5.3267***	0.0420	5.5915***	0.1338	5.3071***	0.0424	5.5578***	0.1377	0.6611***	0.0342
60 - 64	5.6198***	0.0420	5.8473***	0.1337	5.6002***	0.0423	5.8136***	0.1375	-0.5092***	0.0637
55 - 59	5.8258***	0.0419	6.0401***	0.1336	5.8062***	0.0423	6.0065***	0.1375	-0.2931***	0.0534
50 - 54	5.9228***	0.0419	6.0903***	0.1336	5.9032***	0.0423	6.0568***	0.1374	2.6009***	0.0310
45 - 49	5.9395***	0.0419	6.0564***	0.1336	5.9199***	0.0423	6.0229***	0.1373	-3.3497***	0.5371
40 - 44	5.9861***	0.0419	6.0671***	0.1336	5.9665***	0.0423	6.0336***	0.1373	1.6932***	0.0313
35 - 39	6.0420***	0.0419	6.1283***	0.1336	6.0224***	0.0423	6.0948***	0.1373	0.6902***	0.0340
30 - 34	6.1154***	0.0419	6.1686***	0.1336	6.0958***	0.0423	6.1351***	0.1373	-1.5492***	0.1649
25 - 29	6.4025***	0.0419	6.3619***	0.1336	6.3829***	0.0423	6.3285***	0.1371	3.9046***	0.0310
20 - 24	6.7277***	0.0419	6.4917***	0.1335	6.7081***	0.0423	6.4585***	0.1368	-1.7436***	0.1929
15 - 19	7.0900***	0.0419	6.7168***	0.1335	7.0704***	0.0423	6.6836***	0.1367	2.3196***	0.0311
10 - 14	6.4794***	0.0419	6.5359***	0.1335	6.4598***	0.0423	6.5020***	0.1374	1.7751***	0.0312
5 - 9	5.6120***	0.0420	6.3008***	0.1336	5.5924***	0.0423	6.2664***	0.1383	1.4730***	0.0315
0 - 4	5.0319***	0.0420	5.6685***	0.1337	5.0123***	0.0424	5.6344***	0.1380	0.5438***	0.0351
Accident Types										
Pedestrians (baseline)										
Bicycles	1.1160***	0.0034	0.9308***	0.0535	1.1160***	0.0034	0.9316***	0.0544	-0.6074***	0.0034
Motorcycles	3.3142***	0.0030	2.8610***	0.0531	3.3142***	0.0030	2.8585***	0.0554	1.1644***	0.0015
Buses	-2.2732***	0.0096	-2.5211***	0.0563	-2.2732***	0.0096	-2.5216***	0.0575	0.8956***	0.0016
Cars	0.4703***	0.0037	0.1758**	0.0539	0.4703***	0.0037	0.1755**	0.0556	0.1573***	0.0020
Zero Part										
Intercept					-2.4843**	0.7661	-3.1610*	1.4270		
Age Ranges										
≥ 100 (baseline)										
95 - 99					-12.4801	353.0238	-22.4150	91402.1440		
90 - 94					-12.9149	349.3678	-21.8190	44456.0350		
85 - 89					-11.8795	207.9169	-14.5440	1215.1000		
80 - 84					-20.7188	15452.7476	-20.7190	21678.6890		
75 - 79					-20.7188	15452.7476	-20.7190	21678.6890		
70 - 74					-20.7188	15452.7476	-20.7190	21678.6890		
65 - 69					-20.7188	15452.7476	-20.7190	21678.6890		
60 - 64					-20.7188	15452.7476	-20.7190	21678.6890		
55 - 59					-20.7188	15452.7476	-20.7190	21678.6890		
50 - 54					-20.7188	15452.7476	-20.7190	21678.6890		
45 - 49					-20.7188	15452.7476	-20.7190	21678.6890		
40 - 44					-20.7188	15452.7476	-20.7190	21678.6890		
35 - 39					-20.7188	15452.7476	-20.7190	21678.6890		
30 - 34					-20.7188	15452.7476	-20.7190	21678.6890		
25 - 29					-20.7188	15452.7476	-20.7190	21678.6890		
20 - 24					-20.7188	15452.7476	-20.7190	21678.6890		
15 - 19					-20.7188	15452.7476	-20.7190	21678.6890		
10 - 14					-20.7188	15452.7476	-20.7190	21678.6890		
5 - 9					-20.7188	15452.7476	-20.7190	21678.6890		
0 - 4					-20.7188	15452.7476	-20.7190	21678.6890		
AIC	735027.30		14184.61		735062.20		14225.84		19566081.00	
Log Lik	-367487.60		-7065.31		-367484.10		-7064.92		-9783013.00	

Note: *** p < 0.001, ** p < 0.01, * p < 0.05, and SE is standard error.

Zero-inflated models using ZIP and ZINB regression were implemented to improve model fitting in zero-part modeling, where the age ranges are the factor that affects the occurrence of actual zero. Comparing all models in Table 2, it was determined that the AIC for the NB regression model was 14,184.61, which was less than the other regression models. This indicates that the NB regression model may be the most appropriate.

In order to determine if the negative binomial regression model is appropriate for use, it was necessary to determine whether or not the dependent variable displayed overdispersion and zero-inflation. This was accomplished through the use of hypothesis tests that compared the proposed negative binomial model to the POI and ZINB regression models shown in Table 3.

If the mean and variance of the number of RAIs were equal or the dispersion ratio is near one, the POI regression model fit the data. For the overdispersion test with the hypotheses:

H_0 : POI fits the data nicely.

H_1 : NB would suit the data better.

Table 3 demonstrates that the NB regression model would suit the data more than the POI regression model ($p < 0.001$).

It indicates that the overdispersion parameter is greater than zero, corresponding to Table 1, and that the mean number of RAIs is less than its variance. This clearly shows that there is overdispersion.

Moreover, the AIC value of the NB regression model is close to that of the ZINB regression model (AIC = 14,225.84). As a result, a hypothesis test was carried out to validate these findings by applying the Vuong test. The hypotheses for this test are as follows:

H_0 : Both NB and ZINB are equivalent.

H_1 : ZINB is better than NB.

The results in Table 3 show that the Vuong z-statistic between the NB and ZINB regression models was -0.4008, indicating that both models performed similarly ($p >$ 0.05). Because of this, we go with the model that is the least complicated, which is the NB regression model.

Table 3. Overdispersion and zero-inflation testing.

Testing	Overdispersion	Zero-inflation
Method	LR Test	Vuong Test
Model 1	POI	NB
Model 2	NB	ZINB
Statistic	720845	-0.4008
p-value	< 0.001***	0.344

For the NB regression model, as can be shown in Table 2, the number of RAIs is significantly correlated with the independent variables of gender, age ranges, and accident types. When all other variables are constant, males had 1.41 ($e^{0.3432}$) times the number of RAIs compared to females. In addition, there is a chance that the number of RAIs between the ages of 15 and 19 is 826.17 ($e^{6.7168}$) times as high as compared to those ≥ 100 years old when other variables are held constant. Moreover, motorcycle accidents resulted in a 17.48 ($e^{2.8610}$) times higher number of RAIs compared to pedestrian accidents when other variables are held constant.

3.3 Results of data prediction

After comparing different models, it was determined that the NB regression model was the most appropriate model to be used to predict the number of RAIs.

The train-test split is a common method for evaluating a model's performance, specifically its accuracy in making predictions that estimate how well the model performs on new data (unseen data)—data not used in the training process. In this investigation, the data for 2018-2021 were contained in the training set, while the data for 2022 were included in the test set. The predicted values derived from the test set were compared to the actual values using the graphical representations and root mean square error (RMSE).

Table 4. Estimated parameters, standard errors, and the exponential of estimated parameters from negative binomial regression models using data from the training set (Year 2018-2021).

Parameter	Estimate	SE	exp(Est.)
Intercept	0.7115***	0.1164	2.0370
Gender			
Female (baseline)			
Male	0.3379***	0.0335	1.4020
Age Ranges			
≥ 100 (baseline)			
95 - 99	-0.4998**	0.1625	0.6066
90 - 94	1.5299***	0.1402	4.6175
85 - 89	2.9551***	0.1350	19.2029
80 - 84	3.9308***	0.1333	50.9472
75 - 79	4.5511***	0.1326	94.7321
70 - 74	5.0467***	0.1323	155.5059
65 - 69	5.5051***	0.1320	245.9535
60 - 64	5.7754***	0.1320	322.2653
55 - 59	5.9630***	0.1319	388.7711
50 - 54	6.0232***	0.1319	412.8958
45 - 49	5.9924***	0.1319	400.3574
40 - 44	5.9949***	0.1319	401.3578
35 - 39	6.0572***	0.1319	427.1835
30 - 34	6.0957***	0.1319	443.9552
25 - 29	6.2900***	0.1318	539.1458
20 - 24	6.4349***	0.1318	623.1923
15 - 19	6.6542***	0.1318	776.0681
10 - 14	6.4725***	0.1318	647.1245
5 - 9	6.2548***	0.1318	520.5059
0 - 4	5.6300***	0.1320	278.6474
Accident Types			
Pedestrians (baseline)			
Bicycles	0.9517***	0.0524	2.5901
Motorcycles	2.8748***	0.0520	17.7214
Buses	-2.5028***	0.0552	0.0819
Cars	0.1914***	0.0527	1.2109
AIC	11363.52		
Log Lik	-5654.76		

Note: *** $p < 0.001$, ** $p < 0.01$, and SE is standard error.

The results from the analysis are displayed in Table 4, and the NB regression model is obtained, as displayed in the following equation.

$$\ln(\mu_i) = 0.7115 + (0.3379)(\text{Male}) + \sum_{j=2}^{21} \beta_{2_j}\text{Age Ranges}_j + \sum_{j=2}^{5} \beta_{3_j}\text{Accident Types}_j. \tag{3.1}$$

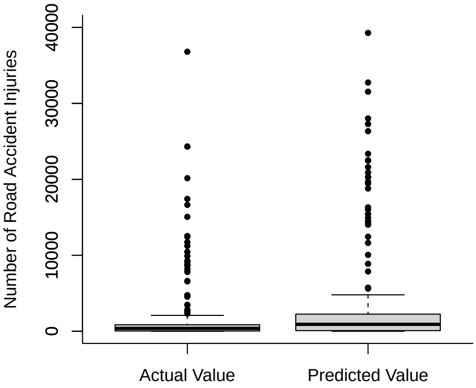

Fig. 2. Box plot of actual and predicted values for the number of road accident injuries in 2022.

To predict the number of RAIs in 2022, we used Eq. (3.1) derived from the NB regression model with the relevant independent variables, as shown in Table 4, for this prediction.

A comparison of the data in Fig. 2 reveals that the actual and predicted values for the number of RAIs differ and that the RMSE, which measures the average difference between the two values, equals 3,567.29 (as shown in Table 5). This difference can be attributed to the dispersion of the data as well as outliers in the actual values, which can range anywhere from 0 to 36,804 in 2022, leading to greater prediction errors.

When comparing the actual and predicted values, segmented by accident types,

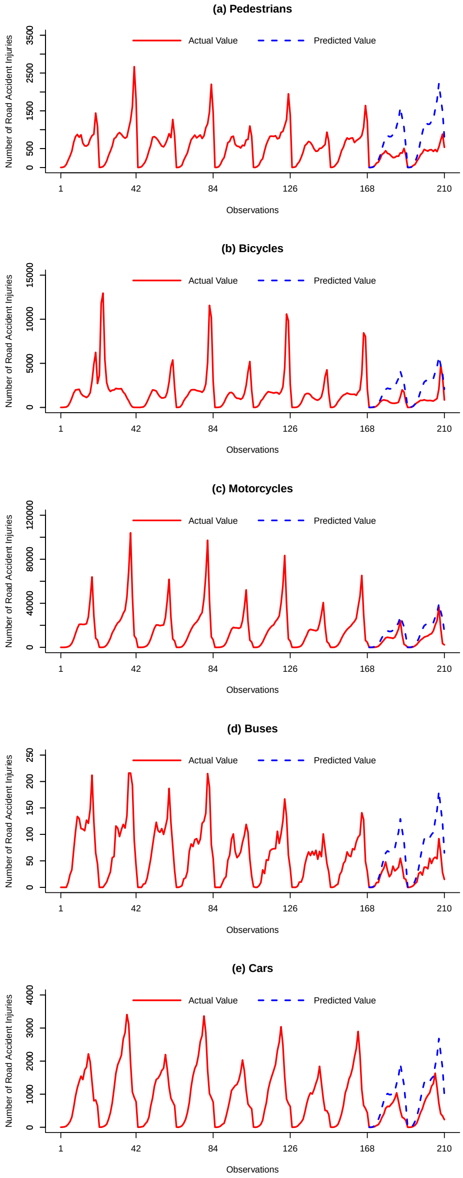

Fig. 3. Actual and predicted values for the number of road accident injuries by the accident types: (a) pedestrians, (b) bicycles, (c) motorcycles, (d) buses, and (e) cars.

Table 5. RMSEs based on accident types.

Accident Types	RMSE
Pedestrians	637.55
Bicycles	1,642.71
Motorcycles	7,750.32
Buses	47.08
Cars	673.14
Overall	3,567.29

as shown in Fig. 3 and indicated by the RMSEs in Table 5, it is obvious that this model is accurate in its prediction of the number of RAIs involving accident types such as buses (RMSE = 47.08), pedestrians (RMSE = 637.55), and cars (RMSE = 673.14). Nonetheless, there are large differences between the two values; the RMSE for motorcycle accidents is notably high at 7,750.32, followed by bicycle accidents (RMSE = 1,642.71).

4. Conclusions

In this article, the POI, NB, ZIP, ZINB, and CMP regression models were investigated for the number of RAIs that were treated in outpatient departments in Thailand between 2018 and 2022.

According to the results, the NB regression model is best suited for predicting the number of RAIs in this study because this count variable displays overdispersion but does not contain excessive zeros. In addition, we came to the conclusion that the number of RAIs is significantly influenced both by demographic factors (such as gender and age ranges) as well as the different accident types.

When using the NB regression model to predict the number of RAIs in 2022, it is clear that the model is more accurate at predicting accidents involving buses, pedestrians, and cars than accidents involving motorcycles and bicycles. In addition, the number of injuries sustained in all categories of road accidents in 2022 has decreased compared to the period from 2018 to 2021. Consequently, this is one of the reasons for the discrepancy when using the

NB regression model trained on data from those years to predict the test dataset in 2022. However, as it uses a few independent variables, this model can still be used as a preliminary tool for predicting the number of RAIs.

References

[1] World Health Organization (WHO). Road traffic injuries. Available from: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

[2] The Injury Surveillance System of the Department of Disease Control. The number of injuries from road accidents from 2018 to 2022. Available from: https://dip.ddc.moph.go.th/new/%E0%B8%9A%E0%B8%A3%E0%B8%B4%E0%B8%81%E0%B8%B2%E0%B8%A3/43-opd-dashboard.

[3] Chaiyapet C, Phakdeekul W, Kedthongma W. Risk factors of severity of road accident injury incidence at Kut Bak district Sakon Nakhon province, Thailand. Res Militaris, 2022;12(5):835-45.

[4] Jomnonkwao S, Uttra S, Ratanavaraha V. Forecasting road traffic deaths in Thailand: applications of time-series, curve estimation, multiple linear regression, and path analysis models. Sustainability, 2020;12:395.

[5] Khemthong P. Accident prediction model for two-lane highways in Surin province [M.Eng. thesis]. Bangkok: King Mongkut's Institute of Technology Thonburi; 2013.

[6] Lerdsuwansri R, Phonsrirat C, Prawalwanna P, Wongsai N, Wongsai S, Simmachan T. Road traffic injuries in Thailand and their associated factors using Conway-Maxwell-Poisson regression model. Thai Journal of Mathematics, 2022;Special Issue:240-9.

[7] Ngamchan S. Accident Prediction Models for Expressways: Case Studies of the First and the Second Stage Expressway Systems [M.Sc. thesis]. Nakhonratchasima: Suranaree University of Technology; 2010.

[8] Simmachan T, Wongsai N, Wongsai S, Lerdsuwansri R. Modeling road accident fatalities with underdispersion and zero-inflated counts. PLoS one, 2022;17(11):e0269022.

[9] Sriwattanapongse W, Prasitwattanaseree S, Khanabsakdi S, Wongtra-ngan S. Mortality rate model due to transportation accidents in Thailand. Silpakorn University Science and Technology Journal, 2013;7(1):9-18.

[10] Suphanchaimat R, Sornsrivichai V, Limwattananon S, Thammawijaya P. Economic development and road traffic injuries and fatalities in Thailand: an application of spatial panel data analysis, 2012–2016. BMC public health, 2019;19:1-15.

[11] Chadbunchachai W, Suphanchaimaj W, Settasatien A, Jinwong T. Road traffic injuries in Thailand: current situation. Journal of the Medical Association of Thailand, 2012;95(Suppl 7):S274-81.

[12] Suriyawongpaisal P, Kanchanasut S. Road traffic injuries in Thailand: trends, selected underlying determinants and status of intervention. Injury control and safety promotion, 2013;10(1-2):95-104.

[13] Tanaboriboon Y, Satiennam T. Traffic accidents in Thailand. IATSS research, 2005;29(1):88-100.

[14] Prasetijo J, Musa WZ. Modeling zero–inflated regression of road accidents at johor federal road F001. In MATEC web of conferences. EDP Sciences, 2016;47;03001.

[15] Worku G, Tesfaw D. The application of count regression models on traffic accidents in case of Addis Ababa, Ethiopia.

Abyssinia Journal of Science and Technology, 2020;5(1):26-33.

[16] Numna S. Analysis of extra zero counts using zero-inflated Poisson models [M.Sc. thesis]. Songkla: Prince of Songkla University; 2009.

[17] Srisuradetchai P, Junnumtuam S. Wald confidence intervals for the parameter in a bernoulli component of zero-inflated Poisson and zero-altered Poisson models with different link functions. Science & Technology Asia 2020;25(2):1-14.

[18] Saputro MIA, Qudratullah MF. Estimation of zero-inflated negative binomial regression parameters using the maximum likelihood method (case study: factors affecting infant mortality in Wonogiri in 2015). In Proceeding International Conference on Science and Engineering, 2021;4:240-54.

[19] Sellers KF. The Conway–Maxwell–Poisson distribution. Cambridge: Cambridge University Press; 2023.

[20] Draper NR, Smith H. Applied regression analysis. New York: John Wiley & Sons; 1998.

[21] Garay AM, Hashimoto EM, Ortega EMM, Lachos VH. On estimation and influence diagnostics for zero-inflated negative binomial regression models. Computational Statistics & Data Analysis, 2011;55(3):1304-18.

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