Int. J. Anal. Appl. (2026), 24:274

International Journal of Analysis and Applications

Overdispersed and Zero-Truncated Count Models for Hospitality Insurance Claim Frequencies

Wikanda Phaphan[1,2], Samach Sathitvudh[3], Peang-or Yeesa[4], Sudarat Nidsunkid[5,*]

[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]*School of Science and Technology, The University of the Thai Chamber and Commerce, Bangkok 10400, Thailand*

[4]*Faculty of Science and Technology, Rajamangala University of Technology Srivijaya, Nakhon Si Thammarat 80110, Thailand*

[5]*Department of Statistics, Faculty of Science, Kasetsart University, Bangkok 10900, Thailand*

[*]*Corresponding author:* sudarat.n@ku.th

Abstract. Insurance claim frequencies are commonly modeled by Poisson regression, although the equidispersion assumption is often violated in real insurance portfolios. This paper studies count regression models for hospitality insurance claim counts under overdispersion and strictly positive sampling. We formulate the Poisson model as a benchmark and motivate the negative binomial model through its conditional mean-variance structure and Poisson–gamma mixture representation. Due to positive-count sampling, zero-truncated extensions are also contemplated. The empirical analysis is based on multi-year hospitality insurance claim data and statistically compares the benchmark, negative binomial along with Conway–Maxwell–Poisson, geometric, and zero-truncated specifications using dispersion diagnostics, likelihood-based criteria, and cross validation. The result shows that the negative binomial substantially improves the Poisson benchmark reducing the cross validated AIC by 1,003.46 on average. Among the extended models, the zero-truncated negative binomial model achieves the best likelihood-based fit, supported by the lowest AIC and BIC. These findings indicate that both unobserved heterogeneity and zero-truncation are crucial features of hospitality insurance claim frequencies.

Received: Jun. 22, 2026.

2020 *Mathematics Subject Classification.* 62E10, 62J05, 62P05.

Key words and phrases. claim frequency; Poisson regression; negative binomial regression; zero-truncated count models; likelihood-ratio test.

https://doi.org/10.28924/2291-8639-24-2026-274
ISSN: 2291-8639

1. Introduction

Hospitality industry such as hotel business regularly encounters various type of risks or unexpected accidents. Fire, inundation and related incidents occurred lead the business to be more cautious about risk management along with financial loss. However, having accurate and reliable claim frequency data plays a crucial role in the management and premium calculation policy. Insurance companies provide insurance to the policy holders, and in turn, the policy holders have to pay insurance premium periodically. In a competitive market, charging a fair premium according to the expected loss of the policyholder is profitable for the insurer. The premium paid by the policyholders is determined from estimates of their expected claim frequencies and the severity of the claims in non-life insurance business.

Many classical statistical methods have been applied in the industry for predicting the claim frequency of a policy, such as Poisson regression models, generalized linear models (GLMs) [1], and Bayesian models. Among these models, the Poisson model is considered as the most commonly used modeling representative of the frequency of claims [2]. An important development of Poisson regression models for claim counts is the contribution of Cameron and Trivedi [3], they managed to highlight the particularities of Poisson regression approach in modeling the claim frequency as a particular case of GLMs. However traditional methods for modeling the claim frequency is GLMs [4], for instance Renshaw [5] focused on the considerable potential of GLMs as a comprehensive modelling tool for the study of the claims process in the presence of covariates; Haberman and Renshaw [6] demonstrated the versatility of GLMs and the statistical package GLIM (generalized linear interactive modeling) for tackling a range of important practical problems arising in actuarial science.

In the case of non-life insurance, Dionne and Vanasse [7] developed a statistical model that adequately integrates risk classification and presented Poisson and negative binomial models with regression component in order to use all available information in the estimation of accident distribution for automobile insurance. Denuit and Lang [8] used Bayesian generalized additive models for non-life insurance ratemaking and proposed the flexible, semi-parametric approach to analyze claim frequencies. Their method allowed for simultaneous estimation of linear and non-linear effects, including spatial variations and risk factors. Gourieroux and Jasiak [9] introduced a dynamic model for modeling claim counts in car insurance that accounts for policyholder heterogeneity using a latent variable approach, specifically focusing on handling overdispersion, unobserved risk factors, and autocorrelation in claim data. They addressed the limitation of standard Poisson models by incorporating unobserved heterogeneity in risk profiles, often using dynamic mixtures.

The contribution of this paper is threefold as follows:

(1) Theoretical contribution: we formalize the role of overdispersion in hospitality insurance claim counts using the Poisson-negative binomial relationship including zero-truncated count models.

(2) Model selection contribution: we propose a likelihood-based decision procedure combining dispersion diagnostics, Akaike information criterion (AIC), log-likelihood, and likelihood-ratio testing.
(3) Applied contribution: we provide empirical evidence from regression-based methodologies and reveal the reliable benchmark among model candidates for further claim frequency predictions.

2. Statistical Framework

This section presents the statistical framework used to model claim count frequencies in hospitality insurance. Due to non-negative response variable, the analysis is framed and formulated with the class of generalized linear models (GLMs) for count data. We begin by introducing the recognized benchmark, Poisson (POI) regression model. Then, we extend to the negative binomial (NB) regression model to accommodate overdispersion. This model further motivates a Poisson-gamma mixture representation to encourage the presence of unobserved heterogeneity in insurance claim frequencies theoretically. Ultimately, likelihood and information based criteria are used to compare the competing model specifications.

2.1. **Poisson distribution.**

The Poisson distribution is commonly used for count data. The Poisson distribution is one of the earliest probability distributions used to model the number of events occurring within a fixed time interval or a specified region. There are simple assumptions that justify the application of the distribution. Events occur independently and the average rate is constant. Due to these assumptions, the structure of the distribution is simple and is readily understandable, which contributes to its widespread use in practical applications. Practically, it is commonly used as a baseline distribution against which other more complex distributions of count may be compared [10].

This distribution was introduced by Siméon Denis Poisson in 1837. A Poisson random variable represents the number of occurrences of a particular event within a fixed interval of time or space [10]. Let X be the Poisson random variable with parameter μ, i.e. $X \sim Poi(\mu)$. Then, the distribution has the probability mass function of the form

$$\mathbb{P}(X = k) = \frac{e^{-\mu}\mu^k}{k!}$$

where $k = 0, 1, 2, ...$ Further the expected value and variance are computed as

$$\mathbb{E}[X] = \mathbb{V}ar[X] = \mu$$

One well-known property of the Poisson distribution is that its mean and variance are equal. Although this feature makes the distribution analytically convenient, it can also be restrictive when working with real data. In many applications, count data tend to show more variability than the Poisson model permits, leading to a poor fit in practice. This limitation highlights the need for alternative count distributions that allow the variance to exceed the mean.

2.2. **Poisson regression as baseline model.** Let Y_i denote the claim count for observation $i = 1, \ldots, n$ and let $x_i = (1, x_{i1}, \ldots, x_{ip})^\top$ be a vector of explanatory variables. The Poisson regression model assumes that, conditioned on x_i,

$$Y_i \mid x_i \sim Poi(\mu_i)$$

where the conditional mean is related to the covariates through the log-link function

$$\log(\mu_i) = x_i^\top \beta$$

Equivalently,

$$\mu_i = \exp(x_i^\top \beta).$$

The conditional probability mass function is

$$\frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}, \qquad y_i = 0, 1, 2, \ldots$$

The log-likelihood function for the Poisson regression model is thus

$$\ln L(\beta) = \sum_{i=1}^{n} \left(-e^{x_i{}^\top \beta} + y_i x_i^\top \beta - \ln y_i! \right)$$

Because the response variable is assumed to follow a Poisson distribution,

$$\mathbb{E}[Y_i \mid x_i] = \mathbb{V}ar]Y_i \mid x_i] = \mu_i$$

demonstrating the equidispersion property.

This property makes Poisson model simple and interpretable, but it is often not practical for insurance claim data where unobserved heterogeneity such as risk characteristics, claim behavior, or business size can encourage the observed variance is much larger than the conditional mean. Therefore, although it is practical to start with Poisson model as a benchmark specification, more flexible count models should also be considered and evaluated throughout the analysis.

Despite its limitation, Poisson regression continues to be widely used in insurance and actuarial research to model claim frequencies under structured risk classifications [11]. In transportation safety studies, it has been applied to analyze traffic accident counts across different road segments and exposure levels [12]. The model is also frequently used in public health research to examine disease incidence rates over time and across regions [13]. More recently, it has been applied in operational risk and reliability modeling to analyze event frequencies within structured risk management frameworks [14].

2.3. **Negative binomial regression for overdispersed counts.** To allow for overdispersion, the negative binomial regression model extends the specification of Poisson by embedding an additional dispersion parameter. The conditional mean is specified as

$$\mathbb{E}[Y_i \mid x_i] = \mu_i = \exp(x_i^\top \beta),$$

while the conditional variance is allowed to exceed the mean:

$$\mathbb{V}ar[Y_i \mid x_i] = \mu_i + \alpha\mu_i^2, \qquad \alpha > 0.$$

Here, α represents the overdispersion parameter. When $\alpha = 0$, this model deduces to the regular Poisson regression model. In other words, the negative binomial model is a generalized version of the Poisson model.

Under the parameterization, the conditional probability mass function can be written as

$$f(Y_i \mid x_i) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(\alpha^{-1})y!}\left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1} + \mu_i}\right)^{y_i}, \qquad y_i = 0, 1, 2, \ldots.$$

Then the corresponding log-likelihood function is

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

The negative binomial model is especially suitable when claim counts show extra-Poisson variation. Overdispersion may arise regularly in the present insurance setting due to unobserved risks in many aspects as mentioned. These latent sources of variability can cause claim counts to vary more widely than the Poisson model allows. This particularly means that the negative binomial model provides a theoretically justified extension of the Poisson benchmark as a standard alternative to the Poisson model under overdispersion.

Researchers widely use the negative binomial model for count data, particularly in areas where the data shows high variation. In health research, it has been applied to analyze hospital visit counts and other medical event data that show overdispersion [15]. The model has also been employed in natural hazard studies to examine the number of power outages caused by hurricanes [16]. In insurance and risk modeling, the negative binomial regression model has been widely adopted to analyze claim frequencies that exhibit substantial variability across policyholders. It has been used to model automobile insurance claim counts, where heterogeneity among drivers leads to overdispersion in observed claims [17]. In non-life insurance studies, the model has been applied to assess claim frequency while accounting for differences in exposure and policy characteristics [18]. The negative binomial regression model has also been employed in actuarial pricing to improve premium estimation when claim counts deviate from Poisson assumptions [11]. In more recent years, methodological developments have strengthened its application in insurance frequency modeling by addressing issues such as multicollinearity and robust estimation [19, 20]. It has further been used in actuarial ratemaking and multi-dimensional risk classification problems [21]. In tourism and forecasting research, negative binomial regression has been applied to model event counts with heterogeneous demand patterns [22].

Overall, the negative binomial regression model provides the necessary flexibility to study complex count data when overdispersion causes it to deviate from standard patterns. By allowing the variance to exceed the mean, the model is better suited to real-world data that exhibit unobserved heterogeneity or excess variability. This added flexibility often leads to improved model fit

and more reliable inference compared with the Poisson regression model. For these reasons, the negative binomial regression model has become a standard choice in applied count data analysis.

2.4. **Poisson-gamma mixture model.** The negative binomial model can be theoretically motivated as a mixed Poisson model. It follows the classical mixed Poisson and negative binomial regression framework [3, 23] aligning with actuarial modeling practice. Suppose that, conditioned on an unobserved random effect U_i, the claim count follows a Poisson distribution:

$$Y_i \mid U_i, x_i \sim Poi(\mu_i U_i),$$

where

$$\mu_i = \exp(x_i^\top \beta).$$

The random effect U_i captures unobserved heterogeneity across insured units. Assume that U_i follows a gamma distribution with

$$\mathbb{E}[U_i] = 1, \qquad \mathbb{V}ar[U_i] = \alpha.$$

Then the conditional mean of Y_i is

$$\mathbb{E}[Y_i \mid x_i] = \mu_i,$$

while the conditional variance is

$$\mathbb{V}ar[Y_i \mid x_i] = \mu_i + \alpha\mu_i^2.$$

This mixture representation plays an important role in insurance claim modeling [24]. The mixture model provides a theoretical explanation for why the NB model is expected to outperform the POI model when the data contain unobserved risk heterogeneity.

2.5. **Overdispersion diagnostics and model selection.** The comparison between the Poisson and negative binomial model is guided by both theoretical and empirical considerations. The first step is to evaluate whether the equidispersion assumption holds. A simple diagnostic is the ratio of the sample variance to the sample mean:

$$D = \frac{s_Y^2}{\bar{Y}}.$$

Values of D larger than one indicate overdispersion. Also, in regression settings, overdispersion can be examined using the Pearson chi-square statistics [25],

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i},$$

and the dispersion estimate

$$\hat{\phi} = \frac{\chi^2}{n-p},$$

where p is number of estimated regression parameters. Under a correctly specified POI model, the estimate should be close to one.

Model comparison is then performed using likelihood-and-information-based criteria. First, the Akaike Information Criterion is defined as

$$AIC = -2\ell(\hat{\theta}) + 2p_k,$$

where $\ell(\hat{\theta})$ is the maximized log-likelihood and p_k is the number of estimated parameters. Smaller AIC refers to better balance between goodness of fit and model complexity which is then preferred.

In practice, a decision must be made based on both goodness of fit and model complexity. By adding a penalty for the number of estimated parameters, the criterion discourages the use of unnecessarily complex models. A more complex model is preferred only if it provides a clear improvement in goodness of fit. Otherwise, a model may fit the observed data well but perform poorly when used for prediction. It does not test a hypothesis in the traditional sense; instead, it provides a relative measure that allows models to be ranked according to their information loss [26]. For this reason, AIC has become a standard tool in regression model comparison and is frequently used in count data modeling to identify the most appropriate specification among alternatives [27].

Next, the likelihood-ratio statistics can also be used to compare the model:

$$\mu_{LR} = -2\left[\ln \sup_{\theta\in\Theta_0} L(\theta) - \ln \sup_{\theta\in\Theta} L(\theta)\right]$$

A large value μ_{LR} indicates that the negative binomial model provides a substantial improvement in likelihood over the Poisson model.

In practice, the LRT provides a structured way to decide whether additional parameters improve a model. By comparing the fit of a simpler model with that of a more complex one, the test helps ensure that added complexity is justified by the data rather than introduced unnecessarily. Under regularity conditions, the test statistic follows a chi-square distribution [28], which allows straightforward interpretation and decision making. For this reason, the LRT is widely used in regression modeling, particularly when evaluating logistic regression models, Poisson regression models, and negative binomial regression models.

2.6. **Extensions for dispersion and truncation.** In the present setting, overdispersion is expected because insurance claim counts may reflect unobserved heterogeneity in policyholder risk, business characteristics, location-specific hazards, and claim-reporting behavior. In addition to overdispersion, the support of the response variable must also be considered. Since the observed claim counts in this study are strictly positive, zero-truncated count models provide a relevant extension by conditioning the likelihood on $Y_i > 0$ [29]. This framework is particularly useful when zero counts are structurally unobserved or excluded by the data collection mechanism. Therefore, the empirical analysis considers both ordinary and zero-truncated count models.

2.7. **Conway–Maxwell–Poisson extension.** Beyond the negative binomial model, another flexible extension of the Poisson distribution is the Conway–Maxwell–Poisson (COM-Poisson) distribution

[30]. Let Y be a non-negative integer-valued random variable. The COM-Poisson distribution [31] is defined by

$$P(Y = y \mid \mu, \nu) = \frac{\mu^y}{(y!)^\nu Z(\mu, \nu)}, \qquad y = 0, 1, 2, \ldots,$$

where $\mu > 0$, $\nu \geq 0$, and

$$Z(\mu, \nu) = \sum_{j=0}^{\infty} \frac{\mu^j}{(j!)^\nu}$$

is the normalizing constant. The parameter μ controls the intensity of the count process, while ν controls the dispersion [32].

When $\nu = 1$, the COM-Poisson distribution reduces to the ordinary Poisson distribution. Thus, the Poisson model is obtained as a special case of the COM-Poisson family. The dispersion parameter ν determines the departure from the Poisson assumption. Values of $\nu < 1$ correspond to overdispersion, whereas $\nu > 1$ correspond to underdispersion relative to the Poisson distribution.

2.8. **Materials.** According to the objective, the regression models for predicting non-life insurance claim frequency are compared in terms of their effectiveness. We apply the actual data from the insurance company between 2017 and 2019. The study focuses on the Poisson and negative binomial regression models. For *count_lossdate* variable, is considered as the dependent variable. Independent variables include Cause of Loss, Risk Code, Sum Insured, and Rate, as shown in Table 1. Missing values are imputed using the mean, and the dataset is cleaned and organized prior to analysis.

Table 1. Independent variables

Variable name	Description
Cause of Loss	Type of event that caused damage to the insured property
Risk Code	Code used to categorize groups of insured properties or activities according to their level of risk
Sum Insured	Total sum insured under the policy (in millions of baht), representing the maximum amount the insurer will pay in compensation.
Rate	Premium rate used as an indicator of the level of risk assigned by the company to the policy

3. Empirical Results

This section empirically reports the evidence for claim count modeling in the hospitality insurance portfolio. The analysis is ranging from descriptive statistics of the claim count variable, exploratory diagnostics, overdispersion evidence to cross-validated model comparison. The goal is to connect the empirical results to the statistical structure of the data.

3.1. **Descriptive statistics.** Let Y_i denote the observed number of claims for policy observation i. The original dataset contains $n = 1,007$ observations. Table 2 reports the descriptive statistics of the claim frequency variable.

Table 2. Descriptive statistics of claim counts

Statistic	Value
Number of observations	1,007
Minimum	1
First quartile	1
Median	2
Mean	6.9553
Third quartile	4
Maximum	3,502
Standard deviation	110.3876
Variance	12,185.4165

As shown in Table 2, the distribution of the claim counts is substantially right-skewed. The median number of claims is only 2, while the mean is 6.9553, indicating that the average is strongly affected by extreme observations. The maximum observed claim count is 3,502, which is far above the upper quartile of 4. This large gap between the center and the upper tail indicates that a simple equidispersed count model (e.g., Poisson model) may not be sufficient.

3.2. **Distribution of claim frequencies.** Table 3 reports the empirical frequency distribution of the claim count variable for the most common claim frequencies. The distribution is also illustrated in Figure 1.

Based on the results, almost half of the observations have only one claim. Specifically, 47.77% of the sample have $Y_i = 1$ while 64.75% of the observations have either one or two claims. This concentration, along with the presence of extreme values at the tail, indicates that the claim count distribution is strongly asymmetric and heavy-tailed which is common in insurance data where a small number of policies generate uncommon high number of claims.

3.3. **Multicollinearity.** Before fitting the regression models, multicollinearity among the predictors is evaluated using the Variance Inflation Factor (VIF). The results are reported in Table 4.

All VIF values are close to one, indicating no evidence of multicollinearity in the explanatory variables. This supports the stability of the regression-based comparisons. The explanatory variables used in the model comparison are Cause of Loss, Risk Code, Sum Insured, and Rate.

3.4. **Benchmark Poisson and negative binomial models.** Since the Poisson model is first fitted as the baseline count model, the descriptive analysis suggests that overdispersion should be accounted for, making the negative binomial model more appropriate than the Poisson model

Table 3. Frequency distribution of claim counts

Number of claims	Frequency	Percentage
1	481	47.77
2	171	16.98
3	87	8.64
4	58	5.76
5	38	3.77
6	28	2.78
7	30	2.98
8	25	2.48
9	11	1.09
10	9	0.89
More than 10	69	6.85
Total	1,007	100.00

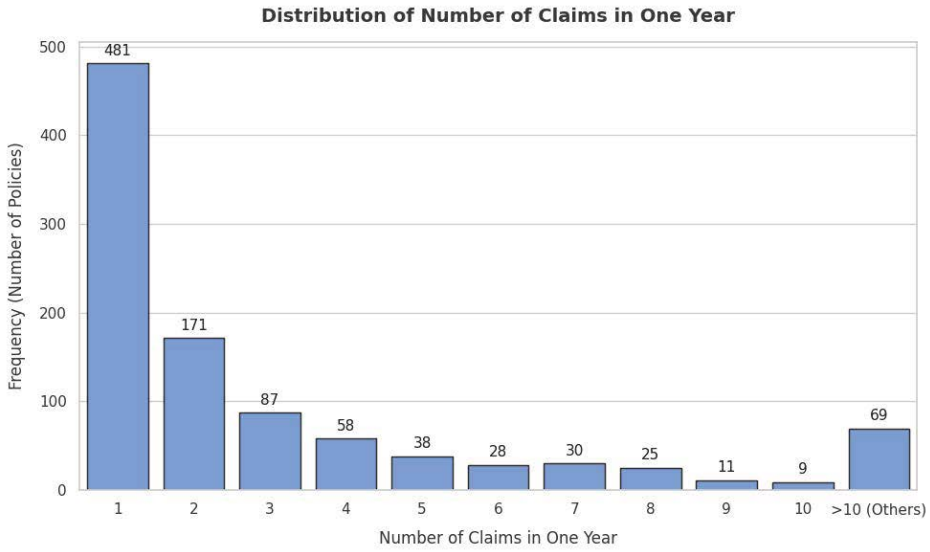

Figure 1. Distribution of claim frequencies

under the equidispersion assumption. The five-fold cross-validation results for these two models are summarized in Table 5.

The results clearly indicate that the negative binomial model provides a substantially better likelihood-based fit than the Poisson model, as reflected in lower AIC and BIC values and a higher log-likelihood.

This implies that:

$$\Delta AIC = AIC_{\text{POI}} - AIC_{\text{NB}} = 4,456.2978 - 3,452.8378 = 1,003.46$$

Table 4. Variance Inflation Factors for explanatory variables

Variable	VIF
Cause of Loss	1.30
Risk Code	1.02
Sum Insured	1.21
Rate	1.09

Table 5. Cross-validated performance of Poisson and negative binomial models

Model	RMSE	MAPE	AIC	BIC	Log-likelihood
Poisson	4.9288	117.3400	4,456.2978	4,479.4958	-2,223.1489
Negative binomial	4.9730	117.9023	3,452.8378	3,476.0359	-1,721.4189

while

$$LR = 2\{\ell_{\text{NB}} - \ell_{\text{POI}}\} = 2\{-1,721.4189 - (-2,223.1489)\} = 1,003.4599.$$

The corresponding p-value is less than 0.0001, leading to rejection of the Poisson model in favor of the negative binomial model. This result is consistent with the strong empirical overdispersion observed in the descriptive statistics.

3.5. **Extended count models and zero-truncation.** In addition to the POI and NB models, we also study zero-truncated count models since the observed claim counts are strictly positive in the dataset. The models include Conway-Maxwell-Poisson model, geometric model, zero-truncated Poisson model, zero-truncated negative binomial model, zero-truncated geometric model, and zero-truncated Conway-Maxwell-Poisson model.

Table 6 presents the average cross-validated performance of all proposed models.

Table 6. Average performance of competing count models under cross-validation

Model	RMSE	MAPE	AIC	BIC	Log-likelihood
Poisson	4.9288	117.3400	4,456.2978	4,479.4958	-2,223.1489
Negative binomial	4.9730	117.9023	3,452.8378	3,476.0359	-1,721.4189
COM-Poisson	4.9232	122.1679	3,401.8294	3,423.0275	-1,694.9147
Geometric	4.9197	121.6707	3,468.4191	3,489.6172	-1,728.2096
Zero-truncated Poisson	4.9592	116.4666	4,301.6487	4,324.8468	-2,145.8244
Zero-truncated negative binomial	6.9720	133.0791	2,743.2401	2,764.4362	-1,365.6200
Zero-truncated geometric	5.3111	82.7337	2,850.5573	2,873.7554	-1,420.2787
Zero-truncated COM-Poisson	5.5880	120.1881	2,917.2732	2,945.1109	-1,452.6366

The results reveal that the zero-truncated negative binomial model achieves the best likelihood-based performance, with highest average log-likelihood -1,365.62, lowest AIC 2743.2401, and

lowest BIC 2764.4362. Specifically, compared to the ordinary NB model, the AIC is reduced by 3,452.8378 - 2,743.2401 = 709.5977. While the AIC of zero-truncated POI model increased significantly by 4,456.2978 - 2,743.2401 = 1,713.0577.

These results confirm that both overdispersion and zero truncation are important features of the data. The strong variance-to-mean ratio supports overdispersion, while the strictly positive observations justify the use of zero-truncated models. Therefore, models conditioning on $Y_i > 0$ better capture the data-generating process than non-truncated alternatives.

On the other hand, the geometric model and the zero-truncated geometric model achieve the lowest RMSE (4.9197) and MAPE (82.7337), respectively. This reflects the fact that likelihood-based measures and point-prediction errors evaluate different aspects of model performance. Consequently, cross-validated RMSE and MAPE are reported as complementary metrics rather than substitutes for likelihood-based model selection. Therefore, the choice of model depends on whether the focus is on statistical fit or predictive accuracy. If prediction is prioritized, geometric and COM–Poisson models may be preferred.

3.6. **Equations of the fitted models.** The fitted equations of the candidate count models including zero-truncated versions are summarized in coefficient-table form. For each model, the fitted conditional mean is written as

$$\widehat{\mu}_i = \exp\left(\widehat{\beta}_0 + \widehat{\beta}_1 X_{1i} + \widehat{\beta}_2 X_{2i} + \widehat{\beta}_3 X_{3i} + \widehat{\beta}_4 X_{4i}\right),$$

where X_{1i}, X_{2i}, X_{3i}, and X_{4i} denote, respectively, cause of loss, risk code, insured sum in million baht, and rate. The estimated coefficients are reported in Table 7. For models marked as standardized, the predictors are standardized before estimation. That is,

$$Z_{ji} = \frac{X_{ji} - \bar{X}_j}{s_{X_j}}, \qquad j = 1, \ldots, 4.$$

Table 7. Estimated Coefficients for All Count Regression Models (Full Dataset, $N = 956$)

	Standard Models				Zero-Truncated Models			
Variable	Poisson	Neg. Bin.	COM-Poisson	Geometric	ZT-Poisson	ZT-Neg. Bin.	ZT-CMP	ZT-Geometric
Intercept	-5.2967^{***}	-5.9273^{***}	-1.3093	1.1566^{***}	0.9795^{***}	-7.2285^{**}	-0.4521^{***}	0.5770^{***}
	(0.5210)	(1.3030)	(0.9380)	(0.0310)	(0.0240)	(2.6390)	(0.0166)	(0.0429)
Cause of Loss	0.0079^{***}	0.0086^{***}	0.0030^{*}	0.0774^{***}	0.2004^{***}	0.3874^{***}	0.0409^{***}	0.3033^{***}
	(0.0010)	(0.0020)	(0.0010)	(0.0250)	(0.0150)	(0.0810)	(0.0055)	(0.0429)
Risk Code	-0.0002^{***}	-0.0002^{***}	-0.0001^{**}	-0.0993^{***}	-0.2229^{***}	-0.3358^{***}	-0.0775^{***}	-0.2976^{***}
	(0.0000)	(0.0000)	(0.0000)	(0.0310)	(0.0290)	(0.0700)	(0.0211)	(0.0488)
Sum Insured[a]	0.0008^{***}	0.0009^{***}	0.0006^{***}	0.2494^{***}	0.3673^{***}	0.8492^{***}	0.0900^{***}	0.5967^{***}
	(0.0000)	(0.0000)	(0.0001)	(0.0250)	(0.0170)	(0.0910)	(0.0074)	(0.0453)
Rate	-0.3888^{***}	-0.2945	-0.2063	-0.0479	-0.2597^{***}	-0.1835^{*}	-0.1853^{***}	-0.2255^{**}
	(0.1350)	(0.1960)	(0.1540)	(0.0350)	(0.0700)	(0.0790)	(0.0535)	(0.0785)

Standard errors in parentheses.

* $p < 0.05$; [**] $p < 0.01$; [***] $p < 0.001$.

[a] Policy Sum Insured (in million baht).

Geometric, ZT-Poisson, ZT-Neg. Binomial, ZT-CMP, and ZT-Geometric were fitted on standardized covariates; Poisson, Neg. Binomial, and COM-Poisson were fitted on the original scale. Zero-truncated models use only observations with $Y \geq 1$.

3.7. **Likelihood-ratio tests for nested model comparisons.** We applied the likelihood-ratio tests to compare nested model pairs. Table 8 reports the results.

Table 8. Likelihood-ratio tests for nested model comparisons

Restricted model	Unrestricted model	df	LR statistic	Result
Poisson	Negative binomial	1	1,003.4599	Reject H_0
Poisson	COM-Poisson	1	1,056.4684	Reject H_0
Zero-truncated Poisson	Zero-truncated negative binomial	1	1,560.4087	Reject H_0
Zero-truncated Poisson	Zero-truncated COM-Poisson	1	1,386.3755	Reject H_0

This report clearly reveals that the restricted models are rejected at conventional significance levels. The comparison between the POI and NB models confirms that adding an overdispersion parameter significantly improves model fit. This is also the case for restricted and unrestricted zero-truncated models. Because some dispersion parameters lie on the boundary of the parameter space under the restricted model, the likelihood-ratios should be interpreted along with the AIC, BIC, log-likelihood, and cross validation metrics.

3.8. **Predictive performance and model interpretation.** The cross validation results indicate that no single model dominates under every evaluation criterion. The zero-truncated NB model provides the strongest likelihood-based fit backed up by AIC, BIC, and log-likelihood while the geometric produces the lowest RMSE, and the zero-truncated geometric model produces the lowest MAPE. The implication depends on the objective in the industry. Likelihood-based criteria are important when the full probability model is considered. That is, the zero-truncated NB model is the strongest candidate in this case. On the other hand, in forecasting setting, the point prediction error is the main concern. In this case, the geometric and COM-Poisson models should be the strongest candidate instead.

4. Discussion and conclusion

This study develops an overdispersion-based count modeling framework for hospitality insurance claim frequencies. The descriptive results show that the claim-count distribution is highly right-skewed and strongly overdispersed, with a mean of 6.9553, variance of 12,185.4165, and variance-to-mean ratio of 1,751.9580. These findings provide clear evidence against the Poisson equidispersion assumption.

The model comparison results show that the negative binomial model provides a substantially better benchmark than the Poisson model. The ordinary negative binomial model reduces AIC by 1,003.46 relative to the Poisson model, while the zero-truncated negative binomial model achieves the strongest likelihood-based performance among the candidate models. These results support the theoretical argument that unobserved heterogeneity and positive-count sampling structure are important features of hospitality insurance claim data.

Although the zero-truncated negative binomial model provides the best likelihood-based fit, its numerical stability should be examined further because convergence issues appear in some cross validation folds. The ordinary negative binomial model therefore remains a stable and interpretable benchmark. Future research may extend the present analysis by incorporating additional exposure variables, richer policyholder characteristics, and alternative count models such as hurdle, zero-inflated, or random-effects specifications.

Acknowledgments: This research is supported by the Department of Statistics, Faculty of Science, Kasetsart University. Also, the research budget is allocated by the National Science, Research and Innovation Fund (NSRF), and King Mongkut's University of Technology North Bangkok with (Project no. KMUTNB-FF-69-B-19).

Conflicts of Interest: The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] E. Ohlsson, B. Johansson, Non-Life Insurance Pricing with Generalized Linear Models, Springer Berlin Heidelberg, 2010. https://doi.org/10.1007/978-3-642-10791-7.

[2] K. Antonio, E.A. Valdez, Statistical Concepts of a Priori and a Posteriori Risk Classification in Insurance, AStA Adv. Stat. Anal. 96 (2012), 187–224. https://doi.org/10.1007/s10182-011-0152-7.

[3] A.C. Cameron, P.K. Trivedi, Regression Analysis of Count Data, 2nd ed., Cambridge University Press, 2013. https://doi.org/10.1017/cbo9781139013567.

[4] J.A. Nelder, R.W.M. Wedderburn, Generalized Linear Models, J. R. Stat. Soc. Ser. A Gen. 135 (1972), 370–384. https://doi.org/10.2307/2344614.

[5] A.E. Renshaw, Modelling the Claims Process in the Presence of Covariates, ASTIN Bull. 24 (1994), 265–285. https://doi.org/10.2143/AST.24.2.2005070.

[6] S. Haberman, A.E. Renshaw, Generalized Linear Models and Actuarial Science, Statistician 45 (1996), 407–436. https://doi.org/10.2307/2988543.

[7] G. Dionne, C. Vanasse, Automobile Insurance Ratemaking in the Presence of Asymmetrical Information, J. Appl. Econom. 7 (1992), 149–165. https://doi.org/10.1002/jae.3950070204.

[8] M. Denuit, S. Lang, Non-Life Rate-Making with Bayesian GAMs, Insur. Math. Econ. 35 (2004), 627–647. https://doi.org/10.1016/j.insmatheco.2004.08.001.

[9] C. Gourieroux, J. Jasiak, Heterogeneous INAR(1) Model with Application to Car Insurance, Insur. Math. Econ. 34 (2004), 177–192. https://doi.org/10.1016/j.insmatheco.2003.11.005.

[10] F.A. Haight, Handbook of the Poisson Distribution, John Wiley and Sons, 1967.

[11] E.W. Frees, Regression Modeling with Actuarial and Financial Applications, Cambridge University Press, 2009. https://doi.org/10.1017/cbo9780511814372.

[12] D. Lord, F. Mannering, The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives, Transp. Res. Part A Policy Pract. 44 (2010), 291–305. https://doi.org/10.1016/j.tra.2010.02.001.

[13] J.M. Hilbe, Modeling Count Data, Cambridge University Press, 2014. https://doi.org/10.1017/cbo9781139236065.

[14] P. Shi, E.A. Valdez, Multivariate Negative Binomial Models for Insurance Claim Counts, Insur. Math. Econ. 55 (2014), 18–29. https://doi.org/10.1016/j.insmatheco.2013.11.011.

[15] A.L. Byers, H. Allore, T.M. Gill, P.N. Peduzzi, Application of Negative Binomial Modeling for Discrete Outcomes, J. Clin. Epidemiol. 56 (2003), 559–564. https://doi.org/10.1016/s0895-4356(03)00028-3.

[16] H. Liu, R.A. Davidson, D.V. Rosowsky, J.R. Stedinger, Negative Binomial Regression of Electric Power Outages in Hurricanes, J. Infrastruct. Syst. 11 (2005), 258–267. https://doi.org/10.1061/(ASCE)1076-0342(2005)11:4(258).

[17] J.-P. Boucher, M. Denuit, M. Guillén, Risk Classification for Claim Counts: A Comparative Analysis of Various Zero-Inflated Mixed Poisson and Hurdle Models, N. Am. Actuar. J. 11 (2007), 110–131. https://doi.org/10.1080/10920277.2007.10597487.

[18] M. Denuit, X. Maréchal, S. Pitrebois, J.-F. Walhin, Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems, Wiley, 2007. https://doi.org/10.1002/9780470517420.

[19] S.C.K. Lee, Delta Boosting Implementation of Negative Binomial Regression in Actuarial Pricing, Risks 8 (2020), 19. https://doi.org/10.3390/risks8010019.

[20] A.F. Lukman, O. Albalawi, M. Arashi, J. Allohibi, A.A. Alharbi, R.A. Farghali, Robust Negative Binomial Regression via the Kibria–Lukman Strategy: Methodology and Application, Mathematics 12 (2024), 2929. https://doi.org/10.3390/math12182929.

[21] G. Tzougas, A. Pignatelli di Cerchiara, The Multivariate Mixed Negative Binomial Regression Model with an Application to Insurance a Posteriori Ratemaking, Insur. Math. Econ. 101 (2021), 602–625. https://doi.org/10.1016/j.insmatheco.2021.10.001.

[22] S. Sathitvudh, P. Wongwiwat, W. Phaphan, Enhancing Tourist Forecasting in Thailand's National Parks with Zero-Inflated Models, WSEAS Trans. Environ. Dev. 21 (2025), 284–292. https://doi.org/10.37394/232015.2025.21.25.

[23] J.F. Lawless, Negative Binomial and Mixed Poisson Regression, Can. J. Stat. 15 (1987), 209–225. https://doi.org/10.2307/3314912.

[24] S. Hu, T.B. Murphy, A. O'Hagan, Bivariate Gamma Mixture of Experts Models for Joint Insurance Claims Modeling, arXiv:1904.04699, 2019. https://doi.org/10.48550/arXiv.1904.04699.

[25] Z. Hossain, Maria, Analyzing Overdispersed Antenatal Care Count Data in Bangladesh: Mixed Poisson Regression with Individual-Level Random Effects, Austrian J. Stat. 50 (2021), 78–90. https://doi.org/10.17713/ajs.v50i4.1163.

[26] H. Akaike, A New Look at the Statistical Model Identification, IEEE Trans. Autom. Control 19 (1974), 716–723. https://doi.org/10.1109/TAC.1974.1100705.

[27] P. Stoica, Y. Selén, Model-Order Selection: A Review of Information Criterion Rules, IEEE Signal Process. Mag. 21 (2004), 36–47. https://doi.org/10.1109/MSP.2004.1311138.

[28] S.S. Wilks, The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses, Ann. Math. Stat. 9 (1938), 60–62. https://doi.org/10.1214/aoms/1177732360.

[29] J.T. Grogger, R.T. Carson, Models for Truncated Counts, J. Appl. Econom. 6 (1991), 225–238. https://doi.org/10.1002/jae.3950060302.

[30] K.F. Sellers, S. Borle, G. Shmueli, The COM-Poisson Model for Count Data: A Survey of Methods and Applications, Appl. Stoch. Model. Bus. Ind. 28 (2012), 104–116. https://doi.org/10.1002/asmb.918.

[31] G. Shmueli, T.P. Minka, J.B. Kadane, S. Borle, P. Boatwright, A Useful Distribution for Fitting Discrete Data: Revival of the Conway–Maxwell–Poisson Distribution, J. R. Stat. Soc. Ser. C Appl. Stat. 54 (2005), 127–142. https://doi.org/10.1111/j.1467-9876.2005.00474.x.

[32] K.F. Sellers, B. Premeaux, Conway–Maxwell–Poisson Regression Models for Dispersed Count Data, WIREs Comput. Stat. 13 (2021), e1533. https://doi.org/10.1002/wics.1533.