Unilag Journal of Mathematics and Applications,
Volume 2, Issue 1 (2022), Pages 9–22.

ISSN: 2805 3966. URL: http://lagjma.edu.ng

CAPTURING EXCESS ZEROS IN MODELING AUTO-INSURANCE CLAIMS IN AN INDIGENOUS INSURANCE FIRM USING ZERO INFLATED MODELS AND HURDLE MODELS

MARY AKINYEMI*, ABISOLA RUFAI, AND NOFIU IDOWU BADMUS

Abstract. Count data occur naturally in a number of disciplines ranging from economics and social sciences to finance as well as medical sciences. Count data could be plagued with over-dispersion and excess zeros making it difficult to employ the use of classic linear models. Different models have been proposed to capture this peculiarity in count data, A number of classical regression models such as the generalized Poisson and negative binomial have been used to model dispersed count data. Hurdle and zero-inflated models are also said to be able to capture over-dispersion and excess zeros in count data. In this paper, we compare the performance of Poisson and Negative Binomial hurdle models, zero-inflated Poisson and Negative Binomial models, classical Poisson and Negative Binomial regression models as well as the zero-inflated compound Poisson generalized linear models to modeling frequency of auto insurance claims in a typical emerging market. The model parameters are estimated using the method of maximum likelihood. The models performances are compared based on some model selection criteria, including: Akaike and Bayesian information Criteria (AIC and BIC), and the model lift which was obtained by the Gini index score. The zero-inflated compound Poisson generalized linear models performed better than the other models considered.

1. Introduction

Count data occur naturally in a number of disciplines ranging from economics and social sciences to finance as well as medical sciences. Naturally, a typical data set containing the number of insurance claims made over a period is considered as count data (see [11], [6] and [7]). Modeling the number of claims is a crucial part of insurance pricing. Count regression analysis allows identification of risk

2010 *Mathematics Subject Classification.* Primary: 22E30. Secondary: 58J05.

Key words and phrases. At least 5 keywords (separated by semicolon) related to the article using the 2010 MSC must be included in your manuscript.

Submitted: January 21, 2022. Revised: June 30, 2022; August 31, 2022. Accepted: December 16, 2022.

* Correspondence.

factors and prediction of the expected frequency claims based on the type of policy taken out and the characteristics of the policy holders. Most insurers would calculate the premium by combining the expected claim amount with the conditional expectation of the number of claims given the risk characteristics. Some insurers may also consider experience rating when setting the premiums, so that the number of claims reported in the past can be used to improve the estimation of the conditional expectation of the number of claims for the following year [5].

Over the years, Insurers gradually amassed sizable longitudinal information on their policy holders, this somewhat availability of data has allowed research in this area to expand so that the literature on the application of count regression analysis in insurance line of discipline has grown considerably in the past years. [5] in their paper addressed panel count data models in the context of insurance, to showcase the advantages of using the information on each policy holder over time for modeling the number of claims. They argue that new panel data models presented in their work allow for time dependence between observations and are closer to the data generating process that one can find in practice.

1.1. **Literature Review.** Different models have been proposed to capture this peculiarity of excess zeros in count data. [20] apply the Poisson, Negative Binomial (NB), Generalized Poisson (GP), Zero Inflated Poisson (ZIP) and Zero Inflated Generalized Poisson (ZIGP). [12] fitted negative binomial and generalized Poisson regression models to Malaysian own damaged (OD) claim count data and zero-inflated negative binomial and zero-inflated generalized Poisson regression models were fitted to the German healthcare count data.[10] applies generalized hurdle models suitable for the analysis of over-dispersed or under dispersed count data allowing for asymmetric departures from the binary logit model to Medicaid utilization data. [22] investigate alternative approaches to constructing multivariate count models based on the negative binomial distribution. They considered two different methods of modeling multivariate claim counts using copulas. The first one works with the discrete count data directly with the mixture of max-id copulas that allows for flexible pairwise association as well as tail dependence. The second one employs elliptical copulas to join continuitized data while preserving the dependency among original counts. The empirical analysis looks into an insurance portfolio from a Singapore auto insurer where claim frequency of three types of claims (third party property damage, own damage, and third party bodily injury) are considered. The results demonstrate the superiority of the copula based approaches over the common shock model.

[17], applied zero-inflated Poisson models to a micro level data set and made comparisons to existing panel data models for count data. They showed that separately controlling for whether outcomes are zero or positive in one of the two years does make a difference for counts with a layer number of zeros. [14] studied recreation demands and visitor characteristics of urban green spaces in the Sapporo city area in Northern Japan. Recreation demands for 21 large urban green spaces were estimated using a zero-inflated negative binomial model. [23] applied the zero inflated poisson model to fishing data. [26] modelled accident

risk at road level of multiple road networks in multiple cities of the Valencian Community (Spain) using the ZINB model. [24], applied the zero inflated poisson factor model to microbiome read counts using data on oral infections, glucose intolerance and insulin resistance. Their model assumed that the microbiome counts follow a ZIP process with library size as offset and Poisson rates negatively related to the inflated zero occurrences. ([2]) provided and extensive study on the robustness of ZIP regression with varying zero inflated sub-models, they proposed a more flexible alternative link function for the ZIP model. [16] applied a multivariate zero inflated endemic epidemic model to measles count data from 16 Germany states. They found that by extending the HHH, endemic-epidemic model using the Zero inflated model, they were able to capture seasonality and serial correlation which in turn improves probability forecasts.In the application of a Poisson regression model to examine spatial patterns in antenatal care (ANC) utilization in Nigeria by 8, results revealed a significant difference in ANC between never-married and married-women respondents. Results also showed substantial spatial variation with a distinct north-south divide in ANC utilization. [18], explored disease mapping and regression with count data in the presence of over dispersion and spatial autocorrelation. The outcome suggested that modelling strategies based on the use of generalized Poisson and negative binomial with spatial autocorrelation worked well and provided a robust basis for inference. In 2016, [9] investigated the spatial distribution of antenatal care utilization in West Africa using a geo-additive zero-inflated count model and the results revealed a tie, transcending boundaries especially among regions of Mali, Niger and northern Nigeria where utilization remains persistently lower.

Although, the Nigerian economy seems to rock immensely our staggering population projected at over 180 million still makes us an attractive destination for consumer goods and services especially new and used automobiles. The Nigerian road use laws (http://www.highwaycode.com.ng/iv-vehicle-insurance.html) stipulate that an automobile user shall take out either third party or comprehensive insurance policies, so that most people typically subscribe to some insurance scheme majorly for statutory reasons. However, what we observed is that most automobile users do not make claims even if they can legitimately make one.

This study compares the performance of Poisson and Negative Binomial hurdle models, zero-inflated Poisson and Negative Binomial models, classical Poisson and Negative Binomial regression models as well as the zero-inflated generalized compound Poisson models to modeling number of auto insurance claims in Nigeria. The model parameters are estimated using the method of maximum likelihood. The models performances are compared based on the model selection criteria (AIC and BIC) and the Gini index which compares the lift of a model against another model.

The rest of this paper is structured as follows: In Section 2 , we discuss the models considered, we give useful details regarding the Zero Inflated models in Section 2.1 and that of the Hurdle models in Sections 2.2 . Section 3.1 describes the data used. The results are presented in Section 3. Finally, Section 5 concludes.

2. Materials and Methods

Count response variables are non-normal responses hence the need for the Generalized linear models(GLM) which extend standard linear regression models to incorporate non-normal response distributions. GLM has three components viz: the random component, the linear predictor and the link function given as:

$$f(\lambda) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n \tag{2.1}$$

Where $X_1, X_2, \cdots, X_n$ are explanatory variables, $\beta_i \quad , i = 0, 1, \cdots, n$ are the intercept and regression coefficients. λ is the link function. The random component of a GLM consists of a response variable y with independent observations $(y_1, \ldots, y_n)$. The conditional distribution of each y_i on a vector of regressors is a linear exponential family with probability density function

given by:

$$f(y; \lambda, \phi) = \exp\left\{\frac{y \ . \ \lambda - c(\lambda)}{\alpha(\phi)} + n(y, \phi)\right\}, \tag{2.2}$$

where λ is the canonical parameter or link function; $c(\lambda)$, the cummulant and $\alpha(\phi)$ is the scale parameter, set to one in discrete and count models and $n(y, \phi)$ is the normalization term. The exponential family of distributions include the Normal, poisson, Gamma, Binomial Negative Binomial etc. Common choices for the link function include, identity $(f(\lambda) = \lambda)$, log $(f(\lambda) = ln\lambda)$ and logit $(f(\lambda) = ln\frac{\lambda}{1-\lambda})$.

This paper considers the Poisson and Negative Binomial hurdle models, zero-inflated Poisson and Negative Binomial models, classical Poisson and Negative Binomial regression models as well as the zero-inflated generalized compound Poisson model with link functions given as:

- Poisson = $\log(\lambda)$: The link function here results in a log-linear relationship between mean and linear predictor. Recall that the variance in the Poisson model is identical to the mean, hence, the dispersion is fixed at $\phi = 1$.
- Negative binomial = $\log(\lambda)$: Similar to the Poisson model, the dispersion is fixed at $\phi = 1$.

2.1. **Zero Inflated models.**

Zero-inflated models have been proposed as a class of models more capable of dealing with excess zeros in count data than the classical GLMs ([19];[15]). They are two-component mixture models combining a point mass at zero with a count distribution such as Poisson, geometric or negative binomial. Thus, there are two sources of zeros: zeros may come from both the point mass and the count component. For modeling the unobserved state (zero vs. count), a binary model is used: in the simplest case only with an intercept but potentially containing regressors ([25]). Formally, Zero-inflated models mix a point mass at zero $I_0(y)$ and a count distribution $f_{count}(y; x, \beta)$. The probability of observing a zero count is inflated with probability $\pi = f_{zero}(0; x, \gamma)$:

$$\begin{aligned} f_{zeroinfl}(y; x, z, \beta, \gamma) = {} & f_{zero}(0; x, \gamma) \cdot I_0(y) \\ & + (1 - f_{zero}(0; x, \gamma)) \cdot f_{count}(y; x, \beta) \end{aligned} \tag{2.3}$$

Where $I(\cdot)$ is an indicator variable. The unobserved probability π of belonging to the point mass component is modeled by a binomial GLM $\pi = g^{-1}(z^{\intercal}\gamma)$. The corresponding regression equation for the mean is given as;

$$\mu_i = \pi_i \cdot 0 + (1 - \pi_i) \cdot \exp(\beta_i^{\top} x) \tag{2.4}$$

using the canonical log link. The vector of regressors in the zero-inflation model z_i and the regressors in the count component x_i need not to be distinct in the simplest case, $z_i = 1$ is just an intercept. The default link function $g(\pi)$ in binomial GLMs is the logit link, but other links such as the probit are also available. The full set of parameters of β, γ, and potentially the dispersion parameter ϕ (if a negative binomial count model is used) can be estimated by ML. Inference is typically performed for β and γ, while ϕ is treated as a nuisance parameter even if a negative binomial model is used.

2.1.1. *The likellihood and log likelihood models for the ZIP models.*

$$f_{zeroinfl}(y; x, z, \beta, \gamma) = f_{zero}(0; x, \gamma) \cdot I_0(y) + (1 - f_{zero}(0; x, \gamma)) \cdot f_{count}(y; x, \beta) \tag{2.5}$$

Where, β is the coefficient of the count model and γ is the coefficient of the Zero inflation model and

$$f_{count}(y; x, \beta) = \frac{exp(-exp(\beta^T x)(exp(\beta^T x)^y))}{y!} \tag{2.6}$$

Let $\theta = (\gamma^T, \beta^T)^T$ be the parameters to be estimated. The likelihood function for the ZIP model can be described as:

$$L(\theta) = \prod_{i=1}^{n} \left[\frac{f_{zero}(0; x, \gamma)}{f_{zero}(0; x, \gamma) + exp(\beta^T x_i)} \right]^{y_i=0} \tag{2.7}$$
$$\times \prod_{i=1}^{n} \left\{ [1 - f_{zero}(0; x, \gamma)] \frac{exp(-exp(\beta^T x_i)(exp(\beta^T x_i)^{y_i}))}{y_i!} \right\}^{y_i>0}.$$

The log-likelihood function of the ZIP model is given as:

$$l(\theta) = logL(\theta) = \sum_{i=1}^{n} l_i(\theta) \tag{2.8}$$
$$= \sum_{i=1}^{n} logL \left\{ [I_{y_i=0}) \left[log(f_{zero}(0; x, \gamma) - log(f_{zero}(0; x, \gamma) + exp(\beta^T x_i)) \right] \right\}$$
$$+ \sum_{i=1}^{n} logL \left\{ [I_{(y_i>0)}) \left[log(f_{zero}(0; x, \gamma) + (y_i \beta^T x_i + exp(\beta^T x_i) - log(y_i!)) \right] \right\}$$

2.2. **Hurdle models.** The hurdle model was originally proposed by [19]. They consist of two-component models viz:

(1) A truncated count component, such as Poisson, geometric or negative binomial, is employed for positive counts, and
(2) A hurdle component which models zero vs. larger counts.

For the latter, either a binomial model or a censored count distribution can be employed [25].

Hurdles models combine a count data model $f_{count}(y; x, \beta)$ and a zero hurdle model $f_{zero}(y; x, \gamma)$. The models are such that $f_{count}(y; x, \beta)$ is left truncated at $y = 1$ and $f_{zero}(y; x, \gamma)$ is right truncated at $y = 1$:

$$f_{hurdle}(y; x, z, \beta, \gamma) = \begin{cases} f_{zero}(0; x, \gamma) & \text{if } y = 0, \\ (1 - f_{zero}(0; x, \gamma)) \cdot \frac{f_{count}(y;x,\beta)}{1-f_{count}(0;x,\beta)} & \text{if } y > 0. \end{cases} \tag{2.9}$$

The model parameters β, γ, and potentially one or two additional dispersion parameters ϕ (if f_{count} or f_{zero} or both are negative binomial densities) are estimated by ML, where the specification of the likelihood has the advantage that the count and the hurdle component can be maximized separately. The corresponding mean regression relationship is given by

$$\log(\mu_i) = x_i^\top \beta + \log(1 - f_{zero}(0; z_i, \gamma)) - \log(1 - f_{count}(0; x_i, \beta)) \tag{2.10}$$

using the canonical log link. For interpreting the zero model as a hurdle, a binomial GLM is probably the most intuitive specification. Another useful interpretation arises if the same regressors $x_i = z_i$ are used in the same count model in both components $f_{count} = f_{zero}$: A test of the hypothesis $\beta = \gamma$ then tests whether the hurdle is needed or not.

2.3. **Model evaluation.** We compare model performances by employing the following penalised measures:

(1) Akaike Information Criterion (AIC): It penalizes the log-likelihood for additional model parameters. AIC provides an asymptotically unbiased estimator of the expected Kullback discrepancy between the generating model and the fitted approximating model. it is computed as follows:

$$AIC = -2 \ln L(\hat{\theta}_k \mid y) + 2k. \tag{2.11}$$

([21]). Where, k is a the number of estimated parameters in the model, $\hat{\theta}_k$ is the maximum likelihood estimator of θ, the vector of k parameters and $L(\hat{\theta}_k \mid y)$ is the log-likelihood function. Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value.

(2) Bayesian Information Criterion (BIC): It also penalizes the log-likelihood for additional model parameters, however this penalty increases as the number of records in the dataset increases. BIC provides a large-sample estimator of a transformation of the Bayesian posterior probability associated with the approximating model. it is computed as

$$BIC = -2 \ln L(\hat{\theta}_k \mid y) + k \ln n \tag{2.12}$$

([21]). Where, k is a the number of estimated parameters in the model, $\hat{\theta}_k$ is the maximum likelihood estimator of θ, the vector of k parameters and $L(\hat{\theta}_k \mid y)$ is the log-likelihood function. In a set of candidate models for the data, the one with the lowest BIC is preferred.

Note 2.1. It is noteworthy that AIC and BIC feature the same goodness-of-fit term, however, the penalty term of BIC is more stringent than the penalty term of AIC. (For $n \geq 8$, $k \ln n$ exceeds $2k$.) Consequently, BIC can be too restrictive and tends to favor smaller models than AIC.

2.3.1. *Gini Index.* We evaluate the model lift using the *Gini Index*, (also called the Gini coefficient or the Gini ratio). The Gini index is defined as

$$Gini = \frac{\int_0^1 (P - L(P))dP}{1/2}. \tag{2.13}$$

The index is usually is defined on the basis of the Lorenz curve and is a measure of the degree of income inequality in society. The Lorenz curve is best explained as follows: For a given population, let y be personal income, x a pre-specified level of income, $F(x)$ a fraction of the population with $y \leq x$ with density function $f(x) = F^{'}(x)$. Furthermore, denote the average income (assuming all income is negative) by $\bar{y} = \int_0^\infty yf(y)dy$. The *The lorenz function* is a function $L : [0, 1] \to \mathcal{R}$, satisfying,

$$P = F(x) \implies L(P) = \frac{\int_0^x yf(y)dy}{\bar{y}}. \tag{2.14}$$

Where P is a proportion of the said population ([3] and [1]). The *Lorenz curve* is simply the graph of $(P, L(P))$. The Lorenz curve is often accompanied by a straight diagonal line with a slope of 1, which represents perfect equality in income or wealth distribution; the Lorenz curve lies beneath it, showing the observed or estimated distribution. The area between the straight line and the curved line, expressed as a ratio of the area under the straight line, is the *Gini coefficient*, a scalar measurement of inequality. Although the Lorenz curve is mostly used to represent economic inequality, it can also demonstrate unequal distribution in any system as in the case of this paper where we compare equality in the performance of pairs of models. The farther the curve is from the baseline, represented by the straight diagonal line, the higher the level of inequality.

A high value of Gini index means high degree of inequality in the distribution of income. If everybody had the same income, then the Lorenz curve would coincide with the 45^o line and the Gini index would be zero. In the context of this paper, a high value of Gini means a model has a higher lift than the other and if the models were the same, the Lorenz curve would coincide with the 45^o line on the axis of the plot and the Gini index would be zero ([3]).

3. Result

3.1. **Data.** The data consists of 616 policies issued between 2011-2015 by an indigenous insurance company. The attributes available for consideration from the data source include: year policy was taken, gender of the policy holder, class of

the car (private or commercial), premium, insurance type (third party or comprehensive) and the number of claims. Table 1 below shows the summary of the data considered. From Table 1 we see that the highest number of policies were taken out between 2013 and 2014, more than 85% of these policies were taken out by men. We also observed that most of the customers who took out policies took them out on their private cars and there was a preference for comprehensive insurance policies (>97%). It was further observed that all 17 third party insurance policies recorded were private cars. Furthermore, we observed that about 91.72% of the policy holders made no claims so that the data does have many zero (i.e. it is zero inflated). We also observed that 89% of the comprehensive insurance policy holders had made no claims in the time period considered. In addition, of the 91.97 zero claims, about 72.2% of them were private car owners and the remaining 19.5% were commercial vehicles.

Table 1. Descriptive statistics of the data

Attribute	Factors	Frequency	Percent
Year	2011	72	11.69
	2012	101	16.40
	2013	158	25.65
	2014	200	32.47
	2015	85	13.80
Gender	Male	526	85.39
	Female	90	14.61
Motor class	Private	478	77.6
	Commercial	138	22.4
Insurance type	Third party	17	2.76
	Comprehensive	599	97.24
Claims	0	565	91.72
	1	48	7.79
	2	2	0.32
	3	1	0.16

Figure1 represents the distribution of claims by premium. It can be observed from Figure 1 that the bulk of the customers with no claims fall within the lower average premium bracket. Furthermore, since the data consists of 97% comprehensive and 3% third party insurance policy holders and none of the third party insurance policy holders made any claims in the time period considered, we thus base the analysis on comprehensive insurance policy holders only.

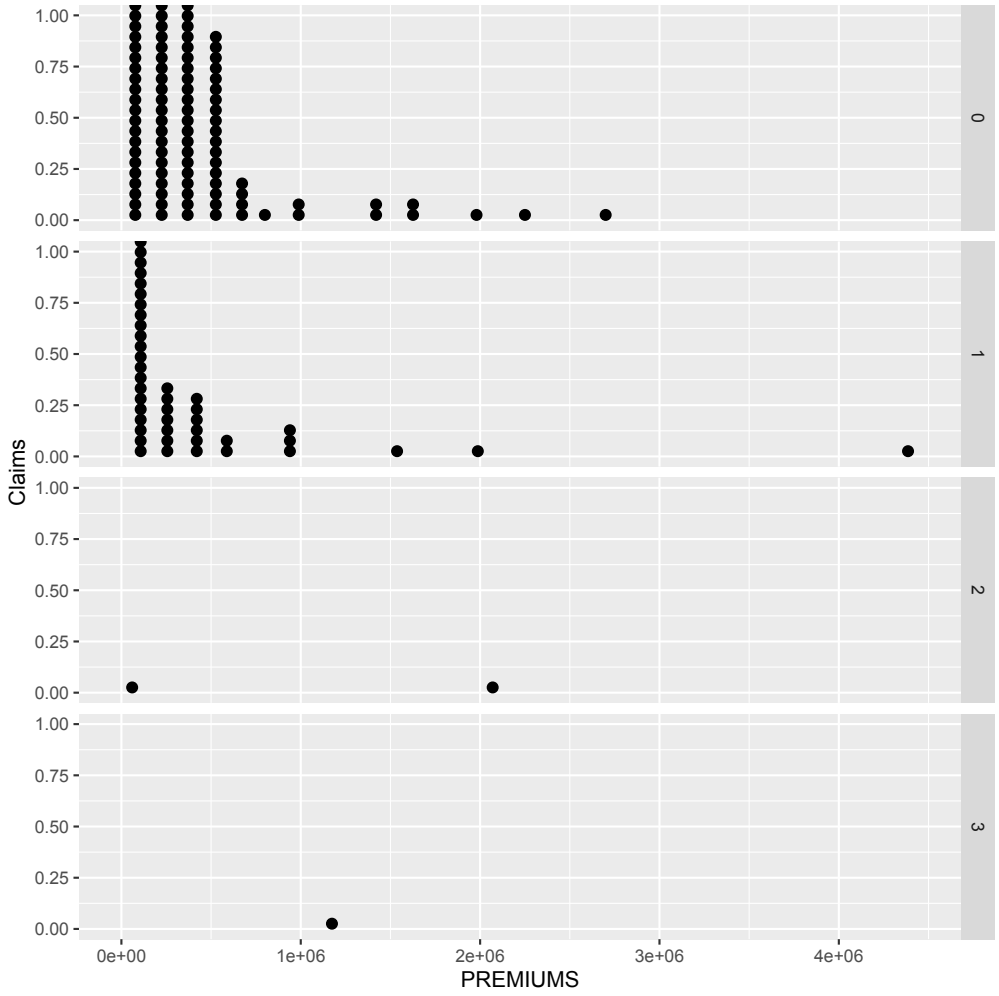

FIGURE 1. Premium amount by number of claims

3.2. **Model fit.** The data was fitted to Poisson Hurdle (HurdlePois) and Negative Binomial Hurdle (HurdleNB), Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB), classical Poisson (Poisreg) and Negative Binomial (NBreg) regression models as well as the Zero-Inflated Generalized Compound Poisson (ZIGCP) models. Table 2 shows the parameter estimates and the corresponding standard errors for the fitted models.

TABLE 2. Poisreg, NBreg, ZIP, ZINB, HurdlePois, HurdleNB, ZIGCP Models for Car Insurance Claims

	Poisreg	NBreg	ZIP	ZINB	HurdlePois	HurdleNB	ZIGCP
Intercept	130.7(226.9)	102.1(231.7)	-8.62(124.14)	54.22(254.34)	-2.23(0.53)	-2.23(0.53)	-2.15 (0.48)
Gender	-0.20(0.42)	-0.21(0.43)	9.87(124.13)	22.36(95.72)	-0.25(0.44)	-0.25(0.44)	-0.15(0.42)
Motor class	-0.48(0.31)	-0.46(0.32)	-0.53(1.41)	40.03(265.11)	-0.38(0.35)	-0.38(0.35)	-0.48(0.31)
premiums	0.0007(0.0002)	0.0008(0.0002)	-0.004(0.002)	-2.96(10.30)	0.001(0.0004)	0.001 (0.0004)	0.001(0.0002)

The Gini index (Table 3) and corresponding asymptotic standard errors (in parenthesis) were computed based on the ordered Lorenz curve (Figure 2) for

each of the 7 models considered. A pairwise comparison of the lift of the models was computed by computing the Gini index scores using Equation 2.13. The Gini index scores are reported in Table 3.

Greyed areas in Table 3 represents models with comparative better lift. We observed from Table 3 (Greyed areas represent better lift) that the zero inflated models as well as the hurdle models have better lift than the classical Poisson and Negative Binomial models. Furthermore, the classical Negative binomial model also has better lift than the classical Poisson model. The zero inflated models also have better lift than the hurdle models. The zero inflated generalised compound poisson model outperforms all the other models. In addition, according to the "min-max" argument, the selected best model is the Zero inflated generalised compound poisson model (ZIGCPM). It was observed that the GLM-type models had the least performance.

TABLE 3. Gini Index scores with Corresponding standard errors

	Poisreg	NBreg	ZIP	ZINB	HurdlePois	HurdleNB	ZIGCP
Poisson	0.00(0.0)	14.21(8.3)	16.99(8.4)	16.95(8.5)	20.65(7.7)	19.51(8.1)	19.80(7.9)
Negative binomial	-11.12(8.7)	0.00 (0.0)	16.30(8.5)	16.30 (8.5)	21.19(7.8)	20.19(7.9)	18.03(8.2)
Zero inflated poisson	-6.04(8.5)	-5.05(8.6)	0.00 (0.0)	-5.58(7.9)	0.59(8.8)	0.16(8.5)	5.84(7.7)
Zero inflated negative binomial	-6.01(8.5)	-5.06(8.6)	5.60(7.9)	0.00(0.0)	0.60(8.8)	0.13(8.5)	5.8(7.7)1
Hurdle poisson	-12.28 (7.2)	-13.49(7.8)	7.79(8.7)	7.81(8.7)	0.00(0.0)	7.06(7.8)	10.92(8.0)
Hurdle negative binomial	-11.37(8.2)	-12.73(7.9)	9.30(8.4)	9.33(8.4)	-4.70(7.9)	0.00(0.0)	11.90(7.9)
Zero inflated generalised compound poisson	-4.50(8.1)	-2.21(8.3)	2.11(7.9)	2.16(7.9)	1.58(8.2)	2.08(8.1)	0.00(0.0)

Figure 2 is the plot of the ordered Lorenz curves for the data.

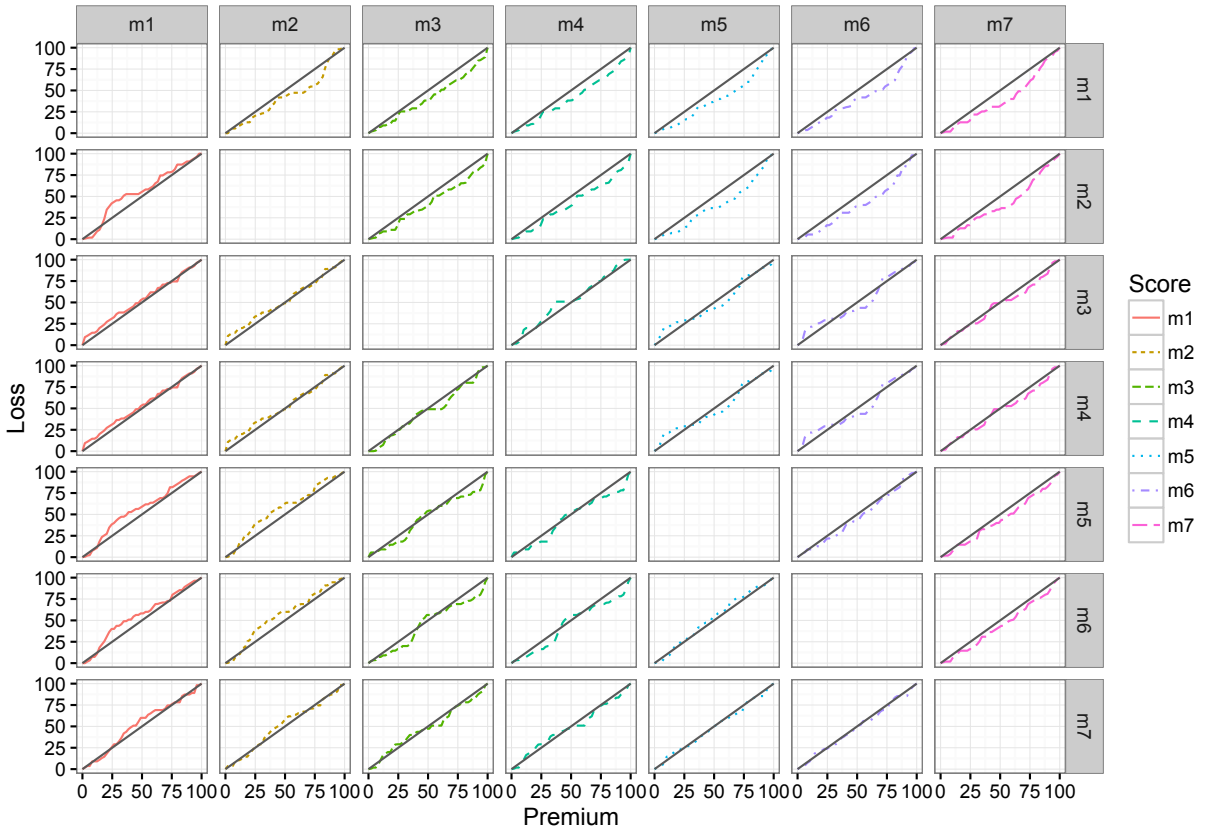

FIGURE 2. Lorenz curve

The results of the model selection criteria (Akaike information criteria (AIC) and Bayesian information criteria (BIC)) are presented in Table 4. It can be observed here that between the classical models, the Poisson model was selected as the better model (AIC=370.10 and BIC=392.08), between the Zero inflated models the Zero Inflated Poisson model was selected as the better model (AIC=374.11, BIC=413.67) and with the hurdle models, the Hurdle Poisson model was also selected as the better model. Overall, the results of the AIC and BIC reveal that has the Zero inflated negative binomial has the highest AIC and BIC (AIC =376.11 and BIC=413.67) while the the Zero inflated generalised compound poisson had the smallest AIC and BIC (AIC=234.93 and BIC=265.70) indicating that the Zero inflated generalised compound poisson is the best model for capturing excess zeros in the claims data considered. This result agrees with that of the Gini index as the ZIGCP model still shows up as the best model since it has the smallest AIC and BIC.

Table 4. AIC and BIC results for each model

	AIC	BIC
Poisson	370.10	392.08
Negative binomial	371.66	398.03
Zero inflated poisson	374.11	413.67
Zero inflated negative binomial	376.11	413.67
Hurdle poisson	372.91	408.07
Hurdle negative binomial	374.07	407.23
Zero inflated generalised compound poisson	**234.93**	**265.70**

4. Discussion

This paper applied poisson and Negative Binomial regression models, Zero Inflated Poisson and Zero Inflated Negative Binomial mdoels, Hurdle Poisson and Hurdle Negative Binomial models and zero Inflated Generalized Compound Poisson model to car insurance claims data to capture excess zeros in the dataset. It has been observed that count data frequently exhibit over-dispersion in addition to possible zero inflation. The selected dataset contains 91.72Hence the choice of Zero inflated and hurdle models. We apply the Gini index to score models that have a better lift (improvement on another model). As expected, the zero inflated models as well as the hurdle models have better lift than the classical Poisson and Negative Binomial models, reiterating the fact that classical regression models are not adequate for capturing the dynamics of zero inflated and over dispersed data. In addition, the results of the AIC and BIC reveal that has the Zero inflated negative binomial has the highest AIC and BIC (AIC =376.11 and BIC=413.67) while the the Zero inflated generalised compound poisson had the smallest AIC and BIC (AIC=234.93 and BIC=265.70) indicating that the Zero inflated generalised compound poisson is the best model for capturing excess zeros in the claims data considered.

5. Conclusion

This study applied seven models to claims data from an indigenous car insurance firm. 6 of the models are in 3 different classes and the generalised model. The models include the classical models class: poisson and Negative Binomial, zero Inflated models class: Zero Inflated Poisson and Zero Inflated Negative Binomial, Hurdle models class: Hurdle Poisson and and Hurdle Negative Binomial and zero Inflated Generalized Compound Poisson. A comparison between the Poisson and Negative Binomial model within each of the classes reveal the Poisson models to be consistently better than the Negative Binomial models. However, the best model overall model was was selected based on the lift which was determined from the Gini index score and the information criteria viz (AIC and BIC) The Gini index score as well as the AIC and BIC results selected the Zero Inflated Generalised Compound Poisson model as the optimal model.

Acknowledgment. The authors acknowledge with thanks the input of all the reviews of the manuscripts at all levels of evaluation

Authors Contributions.

- Dr. Mary Akinyemi: Research idea conception, Data Collection, part analysis and Document assembly.
- Abisola Rufai: some of the analysis and literature review.
- Dr. Badmus: Overall document review and publication readiness

Authors' Conflicts of interest. Authors have no conflicts of interest to declare.

References

[1] Aghion, P. and Durlauf, S. *Handbook of Economic Growth S.* 1(2005). Elsevier, 1 edition.

[2] Ali, E. *A simulation-based study of zip regression with various zero-inflated submodels.* Communications in Statistics - Simulation and Computation. 0(0) (2022), 1–16.

[3] Atkinson, A. and Bourguignon, F. *Handbook of Income Distribution.* 1 (2000). Elsevier, 1 edition.

[4] BBC. *The mint countries: Next economic giants?.* International Journal of Environmental Research and Public Health.(2014).

[5] Boucher, Jean-Philippe, M. D. and Guillen, M. *Models of insurance claim counts with time dependence based on generalization of poisson and negative binomial distributions.* Variance. 2(1)(2008), 132–162.

[6] Cameron, A. C. and Trivedi, P. K. *Count data models for financial data.* Handbook of Statistics. (1996), 363–391. Elsevier, North-Holland.

[7] Famoye, F. and Singh, K. P. *Zero-inflated generalized poisson regression model with an application to domestic violence data.* Journal of Data Science. 4(1) (2006), 117–130.

[8] Gayawan, E. *A poisson regression model to examine spatial patterns in antenatal care utilization in nigeria.* Population, Space and Place. 20 (2014), 485–497.

[9] Gayawan, E. and Omolofe, O. T. *Analyzing spatial distribution of antenatal care utilization in west Africa using a geo-additive zero-inflated count model.* Spatial Demography. 4 (2016), 245–262.

[10] Gurmu, S. *Generalized hurdle count data regression models.* Economics Letters. 58(3) (1998), 263 – 268.

[11] Hidayat, B. and Pokhrel, S. *The selection of an appropriate count data model for modeling health insurance and health care demand: Case of Indonesia.* International Journal of Environmental Research and Public Health. 7(1) (2010), 9–27.

[12] Ismail, N. and Zamani, H. *Estimation of claim count data using negative binomial, generalized poisson, zero-inflated negative binomial and zero-inflated generalized poisson regression models.*(2013)

[13] Kass, R. E. and Raftery, A. E. *Bayes factors.* Journal of the American Statistical Association. 90(430) (1995), 773–795.

[14] Kim, H., Shoji, Y., Tsuge, T., Aikoh, T., and Kuriyama, K. *Understanding recreation demands and visitor characteristics of urban green spaces: A use of the zero-inflated negative binomial model.* Urban Forestry & Urban Greening. 65 (2021), 127332.

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

[16] Lu, J. and Meyer, S. *A zero-inflated endemic-epidemic model with an application to measles time series in germany.*(2022).

[17] Majo, M. and Soest, A. *A fixed-effects zero-inflated poisson model with an application to health care utilization.* CentER Discussion Paper. (2011), 2011-083.

[18] Mohebbi, M., Wolfe, R., and Forbes, A. *Disease mapping and regression with count data in the presence of overdispersion and spatial autocorrelation: A bayesian model averaging approach.* International Journal of Environmental Research and Public Health. (2014).

[19] Mullahy, J. *Specification and testing of some modified count data models.* Journal of Econometrics. 33(3) (1986), 341–365.

[20] Ozmen, I. and Famoye, F. *Count regression models with an application to zoological data containing structural zeros.* Journal of Data Science. 5(4) (2007), 491–502.
[21] Schwarz, G. *Estimating the dimension of a model.* The Annals of Statistics. 6(2) (1978),461–464.
[22] Shi, P. and Valdez, E. A. *Multivariate negative binomial models for insurance claim counts.* Insurance: Mathematics and Economics. 55 2014), 18 – 29.
[23] Truong, B.-C., Pho, K.-H., Dinh, C.-C., and Mcaleer, M. *Zero-inflated poisson regression models: Applications in the sciences and social sciences.* Annals of Financial Economics. 16(2) (2021), 2150006.
[24] Xu, T., Demmer, R. T., and Li, G. *Zero-inflated poisson factor model with application to microbiome read counts.* Biometrics. 77(1) (2021), 91–101.
[25] Zeileis, A., Kleiber, C., and Jackman, S. *Regression models for count data in R.* Journal of Statistical Software, 27(1) (2008),1–25.
[26] lvaro Briz-Redn, Mateu, J., and Montes, F. *Modeling accident risk at the road level through zero-inflated negative binomial models: A case study of multiple road networks.* Spatial Statistics, 43 (2021), 100503.

Mary Akinyemi*
Department of Statistics, University of Lagos, Akoka, Lagos, Nigeria.
E-mail address: makinyemi@unilag.edu.ng

Abisola Rufai
Bank of America, Atlanta, Georgia, USA
E-mail address: bisolarufai@gmail.com

Nofiu Idowu Badmus
Department of Statistics, University of Lagos, Akoka, Lagos, Nigeria.
E-mail address: nibadmus@unilag.edu.ng