MODELLING MOTOR INSURANCE CLAIMS FREQUENCY USING POISSON AND NEGATIVE BINOMIAL MODELS

MESIKE, G.C[a], ADELEKE, I.A[b] and IBEKWE, U. A[c]
[a,b,c]Department of Actuarial Science and Insurance, Faculty of Management Sciences, University of Lagos, Lagos, Nigeria
Corresponding author: [a]mesikegodson@yahoo.co.uk

Abstract

In non-life insurance, the distinctive challenge of estimating the count variable of interest at inception, coupled with the variability of claim costs generally gives insurance companies considerable concern about the chances and sizes of large claims, particularly for automobile insurance where it is required to manage large number of scenarios with a wide variety of risks. These count variables of losses represent individual risks, and need to be predicted, predominantly when the risk premium is to be computed for new policyholders, or when future premiums are adjusted based on past experience. Statistical modelling of count data therefore denotes a fundamental step in pricing of non-life insurance as it allows the classification of the risk factors and the estimation of the expected frequency of claims given the risk characteristics. This study presents the actuarial modelling of motor insurance claim occurrence using Nigerian motor insurance portfolio, to verify and estimate empirically an econometric model and the risk factors influencing the frequency of claims. The log-likelihood ratio and the information criteria was used in choosing the best model and the profile of policyholders with the highest degree of risk is determined. The modelling results are suggested for insurance companies in establishing fair and equitable risk pricing as this will help in appropriate premium determination, alleviate the effect of possible adverse selection and ensure premiums stability in the individual and aggregate portfolio.

Keywords: Poisson model, negative binomial model, claim frequency, motor insurance.

1. INTRODUCTION

The distinctive challenge of estimating the cost of insurance at inception, coupled with the variability of claim costs generally gives insurance companies considerable concern about the chances and sizes of large claims, particularly for automobile insurance where it is required to manage large number of scenarios with a wide variety of risks. A major task of an actuary is the design of a tariff that fairly distributes the burden of claims among policyholders as the insurers aim to sell coverage at prices that are sufficient enough to compensate for the cost of capital necessary to support the sale of such coverage (Mesike, Adeleke & Ojikutu, 2019). In non-life insurance, for instance, the count variable of interest may possibly be the number of a claim made on a motor vehicle insurance policies or the number of losses due to the insurer or the insured in a year. These count variables of losses represent individual risks, and need to be predicted, predominantly when the risk premium is to be computed for new policyholders, or when future premiums are adjusted based on past experiences.

A foremost technique in determining the basic elements of the risk premium is multiplying the conditional expectation of the claim frequency with that of the expected cost of claims. Thus, according to David and Jemna (2015), statistical modelling of count data therefore denotes a fundamental stride in pricing of non-life insurance. Boucher and Guillen (2009) posited that count regression analysis allows the classification of the risk factors and the estimation of the expected frequency of claims given the risk characteristics. Given the economic importance of motor liability insurance in industrialized countries, many attempts have been made over the years within the actuarial literature to find a probabilistic model for the distribution of the number of claims reported by insured drivers (see for example, Nelder & Wedderburn, 1972; Gourieroux, Monfort & Trognon, 1984a, 1984b; Hausman, Hall & Griliches, 1984; McCullagh

& Nelder, 1989; Dionne & Vanasse, 1989, 1992; Gourieroux & Jasiak, 2004; Jong & Heller, 2013; Antonio & Valdez,2012; David, 2015; Mesike, Adeleke & Ojikutu 2020).

A major improvement in the development of models for count data according to Cameron and Trivedi (1988) is the emergence of Generalized Linear Models (GLMs). The theory and application advantages of the Poisson regression, which is a special instance of GLMs were developed by Nelder and Wedderburn (1972). A comprehensive analysis of the Poisson model was further explored in the works of Gourieroux *et al.* (1984a) and Hausman *et al*. (1984). Cameron and Trivedi (1998) demonstrated the particularities of Poisson regression approach in modelling claim frequency as a particular case of GLMs. Many studies within non-life insurance literature have underscored the theoretical and practical features of the GLMs technique (Poisson models) in estimating the frequency of insurance claims (see for example, Jong & Heller, 2013; Frees, 2010, Antonio & Valdez, 2012).

In spite of its prevalence as a foundation in the analysis of count data, due to its descriptive adequacy in the presence of randomness and the underlying homogeneity assumption, the Poisson regression model imposes a strong constraint of equidispersion. This makes it often inappropriate because of unobserved heterogeneity and failure of the independence assumption if the data consist of repeated observations on the same policyholders. (see, Hausman *et al*. 1984; Cameron & Trivedi, 1999; Vasechko, Grun-Rehomme & Benlagha, 2009; Gourieroux & Jasiak, 2001; Charpentier & Denuit, 2005; Jong & Heller, 2013; Hilbe, 2014; David & Jemna 2015; Mesike *et al*. 2020). One of the well-known consequences of unobserved heterogeneity in count data analysis is that the variance of the count variable always exceeds the mean (overdispersion). Other justification is presented by Jong and Heller (2013) who called the

overdispersion as extra-Poisson variation because this type of data displays far greater variance than that predicted by the Poisson model.

Failing to account for this overdispersion according to Denuit, Xavier, Pitrebois, and Walhin, (2007), yields underestimated standard errors, thereby conveying erroneously high levels of significance which might produce too many risk classes in a portfolio. A convenient way to take this overdispersion into account is by introducing a random effect and the alternative distributions used mostly to correct the phenomenon are known as compound or mixed Poisson distribution (see, Gourieroux *et al.* 1984a; 1984b; Dionne & Vanasse, 1989; Cameron & Trivedi, 1986; 1990, 1998; Winkelmann, 2004; Denuit, *et al.* 2007; Greene, 2008; Boucher, Denuit & Guillen, 2008; Hilbe, 2014). Poisson mixtures are well-known counterparts to the simple Poisson distribution for the description of inhomogeneous populations. In this case the probability distribution of the population can be regarded as a finite mixture of Poisson distributions. It is traditional to allow for unobserved heterogeneity by superposing a random variable (called a random effect) on the mean parameter of the Poisson distribution, yielding a mixed Poisson model. In a mixed Poisson process, the annual expected claim frequency itself becomes random.

A particular instance of this mixture model is the negative binomial distribution which is used widely, possibly because of its appealing properties and efficient techniques to relax the limitation of the Poisson distribution. There exists extant literature on many ways to construct the negative binomial distribution, of which the approach (NB1 and NB2) introduced by Cameron and Trivedi (1998) is widely used. A comprehensive image regarding the mixed Poisson model is given by Denuit *et al.* (2007), where they presented the negative binomial distribution as a satisfactory alternative to Poisson distribution in modeling the claim frequency for a motor insurance portfolio. Boucher, Denuit & Guillen (2007) uphold, using cross-

sectional data that the extra parameter of the negative binomial distribution enhances the fit of data when compared with the Poisson distribution. For more excellent description of claim frequency distributions regarding longitudinal data see Boucher *et al* (2008), Boucher and Guillen (2009) and Antonio and Valdez (2010), where they underscored the theoretical and practical application of negative binomial models for motor insurance data.

David and Jemna (2015) applied the negative binomial models (NB) to French motor insurance data and concluded that the NB models correct the overdispersion, and provide a better fit to the data in comparison to the Poisson model while Mesike et al (2020) using Nigerian motor insurance data, employed the negative binomial model to relax the equidispersion restriction of the Poisson model to estimate the average expected loss for determining equitable premium. This study presents the actuarial modelling of motor insurance claim frequency using Nigerian motor insurance portfolio to estimate empirically an econometric model for claim frequency.

2. DATA AND METHODS

Frequency modelling

The Poisson regression model is often suggested for count data but found to be inadequate because the data displays far greater variance than that predicted by the Poisson. Thus a Poisson model for the number of claims is inappropriate since the observed variance is much larger than the mean (Denuit *et al*. 2007). One alternative to Poisson regression is negative binomial regression. Within the actuarial literature, it has been shown that the negative binomial distribution may be viewed as a statistical model for counts, in the situation where overdispersion is explained by heterogeneity of the mean over the population (see, Denuit *et al*. 2007; Jong & Heller, 2013; David & Jemna, 2015; Mesike *et al*. 2020). Another alternative choice is the quasi-likelihood (Poisson variance). The negative binomial is intuitively more

appealing than quasi-likelihood, because it explains the mechanism underlying the overdispersion. In recent years the negative binomial has gained popularity as the distribution of choice when modelling overdispersed count data in many fields, possibly because of its simpler computational requirements and its availability in standard software.

Extant literature present various ways of constructing the negative binomial distribution, nevertheless Boucher, Denuit and Guillen (2008) argued that an intuitive way is the introduction of a random heterogeneity term θ with mean 1 and variance α in the mean parameter of the Poisson distribution. For an intensive discussion of this approach see Gourieroux *et al*. (1984a), Cameron and Trivedi (1990, 1998).

Poisson Model

Cameron and Trivedi (1998) demonstrated the particularities of Poisson regression approach in modelling claim frequency as a particular case of GLMs. With Poisson regression, the mean μ is explained in terms of explanatory variables *x* via an appropriate link,

If $y \sim P(\mu)$

$$f(y) = \mu^y \frac{e^{-\mu}}{y!} \quad , \qquad y = 0,1,2, \ldots, \tag{1}$$

Within the framework of GLMs, the mean of the response variable is related to the linear predictor through the log link function:

$$\mathrm{g}(\mu) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} = x_i' \beta \tag{2}$$

The estimation of the parameters is done by maximum likelihood and the likelihood function is defined as follows:

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} = \prod_{i=1}^{n} \frac{e^{-e^{x_i\beta}} (e^{x_i\beta})^{y_i}}{y_i!} \tag{3}$$

Using logarithm in both sides, the log-likelihood function is obtained as follows:

$$LL(\beta) = \sum_{i=1}^{n} (y_i ln\mu_i - \mu_i - lny_i!)$$

$$= \sum_{i=1}^{n} (y_i x_i \beta - e^{x_i'\beta} - y_i!) \quad (4)$$

The first two partial derivatives of the log-likelihood function which exists can be expressed as follows:

$$\frac{\partial LL(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} (y_i - \mu_i) x_{ij} = \sum_{i=1}^{n} (y_i - e^{x_i'\beta}) x_{ij} \quad (5)$$

$$\frac{\partial^2 LL(\beta)}{\partial \beta_j \beta_k} = -\sum_{i=1}^{n} \mu_i x_{ij} x_{ik} = -\sum_{i=1}^{n} \left(e^{x_i'\beta} x_{ij} x_{ik}\right) \quad (6)$$

The maximum likelihood estimators $\hat{\beta}_j$ are the solutions of the likelihood equations obtained by differentiating the log-likelihood with respect to the regression coefficients and solving them to zero. The resulting equation forming the system is solved numerically by using iterative algorithm such as Newton-Raphson or Fisher information (see, Charpentier & Denuit, 2005). Though Poisson distribution is often considered as a benchmark model in modelling claim count but in practice there are some idiosyncratic risks related to individual insurance contract that make the underlying assumption of the model seem quite unrealistic (see, Gourrieroux & Jasiak, 2007; Jong & Heller, 2013; David &Jemna, 2015).

Negative Binomial

Within the actuarial literature, the negative binomial distribution is employed as a functional form that relaxes the equidispersion restriction of the Poisson model. The negative binomial is derived from a Poisson-gamma mixture distribution. Given λ, if the count y is Poisson distributed; $y|\lambda \sim P(\lambda) \Rightarrow f(y|\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}$

Suppose λ is a continuous random variable with probability density function (pdf) $g(\lambda)$ where $g(\lambda) = 0$ for $\lambda < 0$, then the unconditional pdf of y is

$$f(y) = \int_0^{\infty} f(y|\lambda)\, g(\lambda) d\lambda \tag{7}$$

If $\lambda \sim G(\mu, \upsilon)$,

$$f(y) = \int_0^{\infty} \frac{e^{-\lambda}\lambda^{y}}{y!} \frac{\lambda^{-1}}{\Gamma(\nu)} \left(\frac{\lambda\nu}{\mu}\right)^{\upsilon} e^{-\lambda\nu/\mu} d\lambda$$

$$= \frac{1}{y!\,\Gamma(\nu)} \left(\nu/\mu\right)^{\nu} \int_0^{\infty} \lambda^{y+\nu-1}\, e^{-\lambda\left(1+\frac{\nu}{\mu}\right)} d\lambda$$

$$= \frac{\Gamma(\nu + y)}{y!\,\Gamma(\nu)} \left(\frac{\nu}{\nu+\mu}\right)^{\nu} \left(\frac{\mu}{\nu+\mu}\right)^{y} \quad y = 0,1,2,\ldots. \tag{8}$$

Substituting $\kappa = 1/\nu$ results in the $NB(\mu, \kappa)$ (see, Jong & Heller, 2013). The first two moments of the negative binomial are $E(y) = \mu,\ Var(y) = \mu(1 + \kappa\mu)$. The maximum likelihood estimator is the standard estimator for this model and the log-likelihood function is given as:

$$LL(\nu, \beta) = \sum_{k=1}^{n} \left\{ -\log(y_i) + \sum_{k}^{y_i} \log(\nu y_i - k\nu + 1) - (y_i + \nu^{-1}) \log(1 + \nu\mu_i) + y_i \log(y_i) \right\} \tag{9}$$

Criteria for Assessing the Models' Goodness of Fit

There exists many statistics in the literature that can be used to select and measure the performance of count regression models, however Denuit and Lang (2004) described the likelihood ratio (LR) as the standard measure of goodness of fit for assessing the adequacy of various models. The test statistics follows a $\chi^2_{\alpha,p}$ distribution for a significance level α of 0.05 and p degrees of freedom, where p represents the number of explicative variables included in the regression model. This statistics test is obtained from the difference between the deviance of the regression model without covariates (D_0) and that of the deviance of the model including the independent variables (D_p):

$$LR = D_0 - D_p \tag{10}$$

The deviance was defined by Charpentier and Denuit (2005) as twice the difference between the maximum log-likelihood $possible(y_i - \lambda_i)$ and the log-likelihood of the fitted model:

$$D = 2(LL(y_i|y_i) - LL(y_i|y_i) \quad (11)$$

A likelihood ratio value that is higher than the statistics theoretical value $(LR > \chi^2_{\alpha,p})$ indicates that the regression model explains well the fitted data. For comparison of the models, the log-likelihood function based test is used as a standard method of comparison between the Poisson and NB model. The test statistic follows the χ^2 distribution with one degree of freedom and a calculated test value higher than the theoretical value $(LR > \chi^2_{2\alpha,1})$ indicates that the NB model is chosen over the Poisson regression.

Data

The data used were extracted from the registered policies of motor insurance portfolio obtained from a Nigerian insurance service provider during the year 2018. The data set comprises 15,979 policies and the covariates considered in the policies are the explanatory variables used for this study which reflects the insured characteristics: policyholder's age (<24 years, 24-30 years, 31-60 years and > 60 years), gender (male, female, entity, couple), occupation (self-employed, publicly-employed, privately-employed, unemployed), the geo-political zone where the policyholder lives (federal capital territory, south-west, south-east, south-south, north-west, north-east, north-central), product type (commercial vehicle, comprehensive, third party, motorcycle), customer type (individual, companies, government, others account), nature of loss (theft, collision, accident, vandalisation, others). The preliminary descriptive analysis of the data showing the frequency distribution of policyholder in the portfolio is presented in Table 1. The observed mean claim frequency and mean claim cost for the portfolio are 14.09% and 284117.71 naira respectively. The age structure of the portfolio as described in Table 1 shows that most policyholders were middle-aged as 7730 insured drivers representing 48.4% of the portfolio were in the age bracket of 31 and 60 years. Only 1458 insured drivers signifying 9.1%

of the portfolio were over 60 years. The young drivers represent 28% of the portfolio (4472), and the remaining 2318 insured drivers (14.5% of the portfolio) were in the age range of 24 to 30 years. There were 9672 male policyholders (representing 60.5 % of the portfolio) and 4958 female policyholders (representing 31.0 % of the portfolio) while it is 1248 for an entity and 100 for couples (representing 7.8% and 0.6 % of the portfolio respectively).

Table 1: Frequency distribution of policyholder in the portfolio

Variables		***Frequency***	***Percentage***
Age group			
	Less than 24 years	4472	28.0
	24 - 30 years	2318	14.5
	31 - 60 years	7730	48.4
	61 years and Above	1458	9.1
Gender			
	Male	9672	60.5
	Female	4958	31.0
	Entity	1248	7.8
	Couple	100	.6
Geo-political zone			
	FCT	976	6.1
	South-west	13144	82.3
	South-east	327	2.0
	South-south	981	6.1
	North-east	57	.4
	North-west	296	1.9
	North-central	197	1.2
Occupation			
	Self-employed	1340	8.4
	Publicly employed	6078	38.0
	Privately employed	8210	51.4
	Unemployed	350	2.2
Product type			
	Commercial Vehicle	2783	17.4
	Comprehensive	12520	78.4
	Third party	641	4.0
	Motorcycle	34	.2
Nature of loss			
	Theft	306	1.9
	Collision	14261	89.3
	Accident	391	2.4
	Vandalisation	767	4.8
	Others	253	1.6
Customer type			
	Individual	13283	83.1
	Companies	2611	16.3
	Government	77	.5
	All account	7	.0

Source: Author's computation

3. RESULTS AND DISCUSSION

Descriptive Statistics for the Insured Portfolio

The preliminary exploratory data analysis shows that motor liability clim data are heavily tailed and highly peaked. The distribution of the claim frequency suggests that the portfolio is heterogeneous. From Table 2, it can easily be seen that on the average the frequency of claim reported by the insured decreases initially with age and then increases along the age group. This may be attributed to the fact that younger drivers on average have less driving experience and take more risks, while older individuals on the other hand are riskier drivers due to a deterioration of their cognitive and sensory skills (Kelly & Nielson, 2006). From the exploratory data analysis result displayed in Tables 2 to 4, very positive skewness and heavy tailed kurtosis were observed for all the rating factors. Surprisingly, the mean claim frequency for female was higher than for male and the female policyholders tends to report more claim than their male counterpart as presented in Table 3. The mean number of claims per product type was 24.91 for commercial vehicle, 10.72 in an auto comprehensive, 33.32 in third party liability and 6.26 for a motorcycle. On average, policyholders paid annual premiums of 9453698 naira in commercial vehicle, 1248764 naira in auto comprehensive, 20762294 naira in auto third party liability and 658981 in motorcycle.

Table 2: Descriptive analysis of claim cost, claim frequency and premiums by age group

AGE GROUP		*Mean*	*N*	*Std. Deviation*	*Kurtosis*	*Skewness*
< 24 years						
	CLAIM FREQUENCY	19.46	4472	39.524	11.676	3.320
	PREMIUM	7229804.8883	4472	13834803.33631	9.494	2.963
24 - 30 years						
	CLAIM FREQUENCY	9.08	2318	20.209	32.306	5.003
	PREMIUM	76074.0459	2318	114142.64100	114.794	9.288
31 - 60 years						
	CLAIM FREQUENCY	10.65	7730	25.706	30.778	5.119
	PREMIUM	115915.6951	7730	228360.25275	333.667	14.811
≥61 years						
	CLAIM FREQUENCY	23.78	1458	48.159	7.671	2.870
	PREMIUM	15000794.3528	1458	25093219.11026	.539	1.544

Source: Researcher's computation

Table 3: Descriptive analysis of claim cost, claim frequency and premiums by gender

GENDER		Mean	N	Std. Deviation	Kurtosis	Skewness
Male						
	CLAIM FREQUENCY	13.04	9672	30.293	22.010	4.404
	PREMIUM	2839125.8332	9672	10913413.8530	27.765	5.256
Female						
	CLAIM FREQUENCY	15.36	4958	34.884	16.641	3.923
	PREMIUM	5258711.5213	4958	13912482.8924	11.478	3.373
Entity						
	CLAIM FREQUENCY	17.76	1248	38.753	14.859	3.732
	PREMIUM	1388533.5632	1248	2333449.62718	6.463	2.584
couple						
	CLAIM FREQUENCY	6.26	100	11.105	17.809	3.912
	PREMIUM	96069.1945	100	82334.89117	8.006	2.182

Source: Researcher's computation

Table 4: Descriptive analysis of claim cost, claim frequency and premiums by product type

PRODUCT TYPE		Mean	N	Std. Deviation	Kurtosis	Skewness
Commercial Vehicle						
	CLAIM FREQUENCY	24.91	2783	47.319	7.647	2.828
	PREMIUM	9453698.6903	2783	19496649.0307	4.494	2.451
Comprehensive						
	CLAIM FREQUENCY	10.72	12520	25.639	29.731	5.007
	PREMIUM	1248764.7010	12520	6138406.65700	111.524	9.683
Third party						
	CLAIM FREQUENCY	33.32	641	50.857	4.327	2.183
	PREMIUM	20762294.4163	641	20224342.8616	-1.630	.441
Motor Cycle						
	CLAIM FREQUENCY	6.26	34	19.111	25.127	4.877
	PREMIUM	658981.0382	34	539777.63776	-1.905	.134

Source: Researcher's computation

The SPSS GENLIN procedure which enables the use of type 3 analyses that allows the impact assessment of each risk factor, considering all other explanatory variables is used to fit the Poisson and NB regression models in the framework of GLMs.

Poisson model

The type 3 analysis provides the values of Chi-square statistics for each variable by calculating two times the difference between the log-likelihood of the model which includes all the independent variables and the log-likelihood of the model obtained by deleting one of the specified variables. This test statistic value the impact of each risk factor on the studied interest and follow the asymptotic χ^2 distribution with p degrees of freedom, representing the number of parameters related to the analysed variable. The results of the type 3 analysis are presented

in Table 5. The p-value column indicates the probability associated to the likelihood ratio test which appreciates the impact of each risk factor on the studied event. It can be observed that all the rating variables are statistically significant with a p-value (<.0.05), which clearly underlines their influence on the claims frequency.

Table 5: Likelihood Ratio Statistics for Type 3 Analysis

Source	Likelihood Ratio Chi-Square	df	P-value
(Intercept)	787.572	1	.000
Age	2520.802	3	.000
Gender	419.494	3	.000
District	385.777	6	.000
Occupation	1008.901	3	.000
Product type	18012.051	3	.000
Loss type	37553.284	4	.000
Customer type	648.441	3	.000

Source: Researcher's computation

The goodness-of-fit statistics displayed in Table 6 provides measures that are useful for comparing competing models. Additionally, the Values for the Deviance and Pearson Chi-Square statistics divided by its degree of freedom gives corresponding estimates for the scale parameter. To verify if the data are overdispersed, the most common way is the interpretation of the deviance and Pearson statistics values. These values should be near 1.0 for a Poisson regression; the fact that they are greater than 1.0 (28.877 and 57.799 respectively) indicates an inequality between the mean and variance of the claim frequency, and thus the overdispersion hypothesis is confirmed.

Table 6: Goodness of fit test

Criterion	Value	df	Value/df
Deviance	460638.564	15952	28.877
Pearson Chi-Square	922017.507	15952	57.799
Log Likelihood	-256858.784		
Akaike's Information Criterion (AIC)	513769.568		
Finite Sample Corrected AIC (AICC)	513769.656		
Bayesian Information Criterion (BIC)	513969.222		
Consistent AIC (CAIC)	513995.222		

A formal test to determine whether there is overdispersion is to perform a likelihood ratio test between a standard Poisson regression and a negative binomial regression with all other settings equal. With a negative binomial fit, an estimated κ close to zero suggests a Poisson response. A formal test of $\kappa = 0$ is based on the likelihood ratio test. Since $\kappa = 0$ is at the boundary of the possible range $\kappa \geq 0$, the distribution of the test statistic is non-standard and requires care. The likelihood ratio test statistic is $-2(P - NB)$, where P and NB are the values of the log-likelihood under the negative binomial and Poisson models, respectively. The distribution of the statistic has a mass of 0.5 at zero, and a half Chi-square one degree of freedom distribution above zero. A test at the $100\alpha\%$ significance level, requires a rejection region corresponding to the upper 2α point of the Chi-square one degree of freedom distribution (Cameron and Trivedi 1998).The Poisson and negative binomial regressions yield $P = -256858.784$ and $NB = -56092.990$. Hence the likelihood ratio statistic is 401531.588. The hypothesis $\kappa = 0$ is rejected, at all significance levels. The conclusion is that overdispersion is indeed present. For a significance level $\alpha = 0.05$, the hypothesis $\kappa = 0$ is rejected if the likelihood ratio statistic is greater than the upper 10% point of the Chi-square one degree of freedom distribution, which is 2.71. The test statistics suggest very strong evidence against the fit of the Poisson model to the data, hence the need for the alternative mixed model.

Negative binomial

Table 7, 8, and 9 present the result of the claim frequency modelling based on the negative binomial model. The results show that the different age groups, gender, occupation, district, product type, loss type and customer type are significant in determining the frequency of claims reported. Considering the goodness of fit tests, the results presented indicate that the fitted model is significant at the value/df column for the Pearson chi-square test. The result of the type 3 analysis presented in Table 8 shows that all the predictor variables is statistically significant. The table includes the six degree of freedom test which indicates that as a whole, the rating variable district is a significant predictor of the number of claims occurrence. There

is a significant improvement of the fitted model over the model without any predictors as indicated by the p-value when the overall model is tested against the null model using the likelihood ratio chi-square test.

Table 7: Goodness of fit test

Criterion	Value	df	Value/df
Deviance	29164.489	15952	1.828
Pearson Chi-Square	68622.029	15952	4.302
Log Likelihood	-56092.990		
Akaike's Information Criterion (AIC)	112237.979		
Finite Sample Corrected AIC (AICC)	112238.067		
Bayesian Information Criterion (BIC)	112437.633		
Consistent AIC (CAIC)	112463.633		

Source: Researcher's computation

Table 8: Wald Statistics for Type 3 Analysis

Source	Wald Chi-Square	df	P-value
(Intercept)	387.920	1	.000
Age	183.715	3	.000
Gender	23.978	3	.000
District	36.240	6	.000
Occupation	116.094	3	.000
Product type	836.374	3	.000
Loss type	1469.071	4	.000
Customer type	86.584	3	.000

LR Chi-Square: (5406.714, p-value<0.000)

Analysing the results presented in Table 7, it is noted that the value of deviance and Pearson divided by the number of degrees of freedom are now closer to 1.0 (1.828 and 4.302 respectively). This is a significant improvement over the Poisson model. The analysis of parameter estimates table contains the Poisson and NB regression coefficients for each of the predictor variables along with their standard errors, Wald chi-square values and p-values for the coefficients. Analyzing the result from Table 9, a decrease of the claims frequency can be observed along with an increase in the age of the insured. On the contrary, when the gender coefficient increases, the frequency of claims increases as well. Furthermore, there is an estimate of the dispersion coefficient, (Negative binomial). The parameter 95% confidence interval does not include zero, suggesting that the model fitted is more appropriate than the

Poisson.

Table 9 : Analysis of Parameter Estimates

	Poisson				Negative Binomial			
Parameter	**Estimate**	**Std. Error**	**Wald Chi-Square**	**P-value**	**Estimate**	**Std. Error**	**Wald Chi-Square**	**P-value**
(Intercept)	0.659	0.172	14.661	0.00	0.672	0.4782	1.977	0.16
<24 years	-0.114	0.007	262.616	0.00	-0.01	0.0341	0.092	0.76
24 - 30 years	-0.498	0.0111	2018.28	0.00	-0.437	0.0448	95.416	0.00
31 - 60 years	-0.389	0.009	1872.95	0.00	-0.341	0.0399	72.874	0.00
≥61 years	0a				0[a]			
Male	0.374	0.0402	86.774	0.00	0.408	0.1093	13.944	0.00
Female	0.444	0.0403	121.714	0.00	0.452	0.11	16.868	0.00
Entity	0.322	0.0409	62.136	0.00	0.357	0.1142	9.749	0.00
Couple	0a				0[a]			
FCT	0.325	0.0249	170.439	0.00	0.196	0.0831	5.553	0.02
South-west	0.259	0.0234	121.855	0.00	0.045	0.0766	0.352	0.55
South-east	0.408	0.0272	224.298	0.00	0.144	0.0958	2.273	0.13
South-south	0.329	0.0251	172.113	0.00	0.181	0.0832	4.743	0.03
North-east	0.191	0.0456	17.525	0.00	0.134	0.1593	0.703	0.40
North-west	0.15	0.0291	26.731	0.00	-0.039	0.0979	0.158	0.69
North-central	0a				0[a]			
Self- employed	0.086	0.017	25.543	0.00	0.192	0.0637	9.042	0.00
Publicly- employed	0.087	0.016	29.827	0.00	0.183	0.0607	9.069	0.00
Privately-employed	-0.094	0.0156	36.444	0.00	-0.041	0.0579	0.491	0.48
Unemployed	0a				0[a]			
Commercial vehicle	1.523	0.0687	491.621	0.00	1.561	0.1879	69.045	0.00
Comprehensive	0.873	0.0687	161.364	0.00	0.922	0.1875	24.196	0.00
Third party	1.801	0.069	680.118	0.00	1.8	0.192	87.95	0.00
Motor cycle	0a				0[a]			
Theft	2.205	0.0249	7864.76	0.00	2.231	0.0891	626.905	0.00
Collision	0.408	0.0238	293.155	0.00	0.344	0.0683	25.267	0.00
Accident	1.147	0.0262	1914.12	0.00	1.121	0.0856	171.493	0.00
Vandalisation	-0.183	0.0286	40.726	0.00	-0.187	0.0785	5.704	0.02
Others	0a				0[a]			
Individual	0.011	0.1476	0.006	0.94	0.04	0.4067	0.01	0.92
Companies	0.035	0.1477	0.055	0.81	0.103	0.4078	0.063	0.80
Government	-1.081	0.1567	47.565	0.00	-1.115	0.427	6.815	0.01
All account	0a				0[a]			
(Scale)	1b				1[b]			
(Negative binomial)					1.71	0.0175		

Dependent Variable: CLAIMS FREQUENCY

a. Set to zero because this parameter is redundant.

b. Fixed at the displayed value.

Source: Researcher's computation

4. CONCLUSIONS

The basic idea of the entire process of non-life insurance pricing comprises of establishing an equitable premium payable by the policyholder for transferring contingent risk to the insurer. A major step in motor insurance pricing is the modelling of claim frequency which is very germane in determining a reasonable and equitable price. Hence, this study considered the analysis of classical and mixed count data models used to estimate the frequency of claim reported on motor insurance policies, using individual socio-demographic characteristics and motor risk factors.

A distinct analysis procedure that allows the impact assessment of each risk factor while considering all other explanatory variables was used to fit the Poisson and NB regression models in the framework of GLMs. The equidispersion assumption of Poisson distribution was tested and the test statistic indicates an inequality between the mean and variance of the claim frequency, and thus existence of overdispersion within the studied motor insurance portfolio. The NB model was used to correct the overdispersion and the test result showed that NB model provides a better fit to the motor insurance data compared to the Poisson model.

Using past claim data, the regression results revealed that the risk factors: the age of the insured, the gender, the geographical region where the insured resides, the occupation of policyholders, the product type, the nature of the loss and the customer type significantly explain the frequency of motor claims occurrence. The descriptive statistics shows that motor insurance claims data is highly peaked and heavily-tailed and also differ considerably across age groups, gender, occupation, nature of loss, geographical region, product type and customer type. The obtained results revealed that the frequency of claim decreases on the average with age initially but then increases along the age group, which support the fact noted in studies such as McKnight and McKnight (1999, 2003), Kelly and Nielson (2006) that younger motorists on the average have

more claims due to less driving skill and ability to take more risks, while older individual on the other hand are more dangerous drivers because of worsening cognitive and sensory skills. The modelling results are suggested for insurance companies in establishing fair and equitable risk pricing. It will help in appropriate premium determination, alleviate the effect of possible adverse selection and ensure premiums stability in the individual and aggregate portfolio. Feasible and sustainable motor liability insurance needs to be risk-based driven; hence pricing it involves thoughtful research and careful analysis of the complex function, and a large number of more detailed variables that need to be properly established and actuarially determined. This will enable motor insurance risks to be determined on a sustainable basis, and helps emerging economies improve their response to the challenge presented by motoring.

References

Antonio, K., Frees, E. W. & Valdez, E. A. (2010). A multilevel analysis of intercompany claim counts. *ASTIN Bulletin*, 40(1), 151-177.

Antonio, K. & Valdez, E. A. (2012). Statistical concepts of a priori and a posteriori risk classification in insurance.*AstA Advances in Statistical Analysis*, 96, 187 -224.

Boucher, J. P., Denuit, M., & Guillen, M., (2007). Risk classification for claims count- a comparative analysis of various zero-inflated mixed Poisson and hurdle models. North *American Actuarial Journal*, 11(4), 110-131.

Boucher, J. P., Denuit, M., & Guillen, M., (2008). Models of insurance claim counts with time dependence based on generalization of Poisson and negative binomial distributions. *Advancing the Science of Risk Variance*, 2(1), 135-162.

Boucher, J. P. & Guillen, M., (2009). A survey on models for panel count data with applications to insurance. *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales*, 103(2), 277-295

Cameron, A. C. & Trivedi, P. K. (1986). Econometric models based on count data, comparisons and applications of some estimators and tests. *Journal of Applied Econometrics*, 1(1), 29-53.

Cameron, A. C. & Trivedi, P. K. (1990). Regression-based tests for overdispersion in the Poisson model. *Journal of Econometrics*, 46(3), 347-364.

Cameron, A. C. & Trivedi, P. K. (1998).*Regression Analysis of Count Data*. New York: Cambridge University Press.

Cameron, A. C. & Trivedi, P. K. (1999). Essentials of count data regression. In B. B.H (Ed.), a companion to theoretical econometrics. Malden, MA: Blackwell Publishing Ltd.

Charpentier, A. & Denuit, D. (2004).Mathématiques de l‟Assurance Non-Vie, Tome I: Principe fondamentaux de théorie du risqué. Economica, Paris.

David, M. & Jemna, D. (2015).Modelling the frequency of auto insurance claims by means of poisson and negative binomial models. Scientific Annals of the "AlexandruIoanCuza" University of Iasi Economic Sciences, 62(2), 151-168

David, M. (2015). Automobile insurance pricing using generalized linear models. *Procedia Economics and Finance*, 20, 147-156

Denuit, M., Xavier, M. Pitrebois, S. & Walhin, J.F. (2007).*Actuarial modelling of claim counts*. Chichester: John Wiley & Sons.

Denuit, M., & Lang, S., (2004). Nonlife ratemaking with bayesian GAM's. *Insurance: Mathematics and Economics*, 35(3), 627-647.

Dionne, G., & Vanasse, C., (1989). A generalization of automobile insurance rating models: the negative binomial distribution with a regression component. *ASTIN Bulletin*, 19(2), 199-212.

Dionne, G., & Vanasse, C., (1992). Automobile insurance ratemaking in the presence of asymmetrical information. *Journal of Applied Econometrics*, 7(2), 149-165.

Frees, E. W. (2010). *Regression modeling with actuarial and financial applications*. Cambridge University Press, Cambridge.

Gourieroux, C., Montfort, A. & Trognon, A. (1984a). Pseudo maximum likelihood methods: theory, *Econometrica*, 52, 681-700.

Gourieroux, C., Montfort, A. & Trognon, A. (1984b), Pseudo maximum likelihood methods: application to Poisson models, *Econometrica*, **52**, 701-720.

Gourieroux, C. &Jasiak, J. (2004).Heterogeneous model with application to car insurance.*Insurance: Mathematics and Economics*, 34(2), 177- 192.

Gourieroux, C. & Jasiak, J. (2007).*The econometrics of individual risk: credit, insurance and marketing.* New Jersey: Princeton University Press.

Greene, W. H. (2008). Functional forms for the negative binomial model for count data. *Economics Letters,* 99(3), 585-590.

Hausman, J., Hall, B., & Griliches, Z. (1984). Economics model for count data with an application to the patents- R&D relationship. Ecometrica, 52(4), 909-938.

Hilbe, J. M. (2014). *Modelling count data.* New York: Cambridge University Press.

Jong, P. & Heller, G. (2013).*Generalized linear models for insurance data*. (5th ed.) New York: Cambridge University Press.

Kelly, M. & Nielson, N. (2006). Age as a variable in insurance pricing and risk classification. *The Geneva Papers: Issues and Practice*, 31(2), 212–232.

McCullagh, P. & Nelder, J.A., (1989). *Generalized linear models. 2nd ed*. London: Chapman and Hall.

McKnight, J. A. & McKnight, S. A. (1999).Multivariate analysis of age-related driver ability and performance deficits. Accident Analysis and Prevention, 31(5), 445–454.

McKnight, J. A. & McKnight, S. A. (2003). Young novice drivers: careless or clueless? Accident Analysis and Prevention, 35(6), 921–925.

Mesike, G. C., Adeleke, I. A. & Ojikutu, R. K. (2020). A risk-based model approach for motor insurance pricing in Nigeria. *Unilag Journal of Business,* 6(1), 125-144

Nelder, J.A. & Wedderburn, R.W. M. (1972).Generalized linear interactive models. *Journal of the Royal Statistical Society*, A 135(3), 370-384.

Vasechko, O. A., Grun-Rehomme, M., & Benlagha, N., (2009). Modelisation de la frequence des sinistres en assurance auto. *Bulletin Francais d'Actuariat,* 9(18), 41-63

Winkelmann, R. (2004). Co-payments for prescription drugs and the demand for doctor visits-evidence from a natural experiment. *Health Economics,* 13(11), 1081-1089.