International Journal of Data Science and Analysis
2020; 6(5): 137-144
http://www.sciencepublishinggroup.com/j/ijdsa
doi: 10.11648/j.ijdsa.20200605.13
ISSN: 2575-1883 (Print); ISSN: 2575-1891 (Online)

Modeling Zero Inflation and Over-Dispersion in Domestic Package Insurance Claims Portfolio: A Case of Madison Insurance Company-Kenya

Polycarp Nyabuto[1, *], Anthony Wanjoya[2], Antony Ngunyi[2]

[1]Department of Statistics and Actuarial Science, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya

[2]Department of Statistics and Actuarial Science, Dedan Kimathi University of Technology, Nyeri, Kenya

Email address:

polycarpn75@gmail.com (P. Nyabuto), awanjoya@gmail.com (A. Wanjoya), antonyngunyi@gmail.com (A. Ngunyi)

[*]Corresponding author

To cite this article:

Polycarp Nyabuto, Anthony Wanjoya, Antony Ngunyi. Modeling Zero Inflation and Over-Dispersion in Domestic Package Insurance Claims Portfolio: A Case of Madison Insurance Company-Kenya. *International Journal of Data Science and Analysis*. Vol. 6, No. 5, 2020, pp. 137-144. doi: 10.11648/j.ijdsa.20200605.13

Received: September 29, 2020; **Accepted**: October 14, 2020; **Published**: October 21, 2020

Abstract: The standard Poisson distribution is widely used as a mechanism for regression modeling of count data outcomes. However, the suitability of this modeling technique is only limited to equi-dispersed count data outcomes. This is due to the fact that this modeling technique does not take into account the problems associated with over dispersion and excess zeros in many data sets as with insurance claims data. The study objective is to model domestic package insurance claims frequency using zero inflated and hurdle models since insurance portfolios are characterized by the non-occurrence of claims over a given time interval. This non-occurrence of claims over a given time interval usually leads to the Zero-Inflation and Dispersion associated with insurance claims data. The study consequently evaluates the performance of the Poisson, Zero Inflated Poisson (ZIP) and Hurdle Poisson (HP) models in determining the model that best models the domestic package insurance claims data. This is then used to estimate, predict and determine the heterogeneity of occurrence of the aforementioned insurance claims. The statistical Hosmer-Lemeshow tests is used to define the suitability of the fitted model to estimate the zero-inflation and over-dispersion characteristic of the data. To determine the presence of outliers and the distribution of residuals, the Residual Pearson and Deviance statistics are used. Data on a number of claims for domestic package insurance policy from Madison Insurance ltd, Kenya spanning from 2014 to 2018 (261 weeks) is used in the study.

Keywords: Zero-Inflation, Dispersion, Insurance Claims, Poisson Distributions

1. Introduction

The insurance product is distinctive in nature in that its quality can only be judged when something goes wrong. For this reason, the way in which a claim is handled has important market repercussions for an insurer. A claim is a request by the insured to be indemnified by the insurer following a financial loss associated with the occurrence of the insured peril.

Non-life insurance companies (insurers) calculate probable income within a given period by offsetting premiums receivable against claims payable thus the need for them to strike a balance between the premiums receivable and claims payable [5]. The accurate estimation of premium expenses is thus a vital task for all insurance stakeholders and this has been traditionally accomplished through the degeneration of the overall claim expenses into claim frequency and claim amount which are thus the two key risk drivers in any insurance business [5].

This study concerns itself with modeling claim frequency (number of claims) modeling as it is an indispensable component for premium determination, a vital yet a difficult undertaking, by the insurers in insurance industry. Traditionally this was achieved by use of the classical statistical Poisson regression model by which different rating factors were justified by the use of a regression coefficient. However, this methodology proved not to provide accurate results since insurance claims count data possess a specific characteristic of having an excess number of zeros for a

particular time interval which is not catered for by the classical Poisson regression coefficients.

This excess number of zero claims can be attributed to policyholders willingly failing to report small claims to the insurance company for not getting deductibles and claim bonuses for reduction of the payable premiums for the forthcoming period i.e. year [3]. In this regard, claims count data modeling is viewed as a being a mixture of a degenerate distribution at 0 to model the excess zero claims and a positive continuous part to model the positive non-zero claims.

With the presence of zero claims in the claims data, this study compares the Poisson, Zero-Inflated and Hurdle models to determine an appropriate statistical distribution to model the excess zeros correctly. The over-dispersion in the data, which goes against the equi-dispersion property of the Poisson regression modeling, shall also be looked into. This is due to the randomness of occurrence of claims since an insurer is not able to determine precisely the number and amount of claims that will occur in the following period.

To aid the determination of the best fit statistical count data model for the domestic package insurance claims data, this research shall apply the following steps;

i. Selecting two subsets of variables, one for rating factors and the other for zero-inflation occurrence for each model.
ii. Hypothesis testing for the occurrence of zero inflation and over-dispersion effect for the domestic package insurance claims.
iii. Specifying the model selection criteria that shall be used in selecting the statistical distribution that best models the claims data.
iv. Estimating the parameters and calculating goodness-of-fit measures.

The study uses the term zero-inflation to put emphasis on the exceedance of the probability mass at count zero in comparison to that handled by the standard count distributions. If inadequately modeled, it can lead to the invalidation of the data analysis results thus endangering the reliability of the scientific inferences. Insurance portfolios are characterized by the non-occurrence of claims over a given time interval thus making zero-inflation and over-dispersion common phenomenon in Insurance Portfolios.

1.1. Domestic Package Insurance Claims

Domestic package insurance is an insurance package that covers accidental loss or damage to a residential home (private dwelling used for domestic purposes only) [3]. It covers loss attributed to any of the contents (household goods) of the residential home and personal effects which could belong to the owner of the home or a third party associated to the owner and residing in the residential home [3]. This policy can be extended to cover expenses incurred that may arise due to the death/sickness of the homeowner's domestic servant while on duty as defined in the Work Injury Benefits (WIBA) Act [3].

Domestic package insurance claims are formal requests by the insured to an insurer for indemnification upon the occurrence of the insured peril. Upon verification and approval of the claim by the insurer, payments are made to the insured so as to cover for the loss attributed to the occurrence of the insured peril. This policy has an excess which is the first amount of each loss that the insured must bearer for every claim made and when the loss is equal or less than the excess the insured bearers the whole loss.

The domestic package insurance policy covers six insurance perils namely; fire, burglary, theft, explosion, riot and strike and floods. Domestic package insurance claims occur randomly hence the insurer cannot predict the next occurrence and the magnitude of occurrence over a given period. This claims thus have special characteristics that shall be considered when choosing a distribution that will fit the data. This are;

i. The exclusion of some rating factors justifies the inclusion of the component of heterogeneity for the regression model and for some rating factors being the main cause of domestic package insurance claims compared to others.
ii. The high number of zero count data motivates zero-inflated and hurdle models to be fitted hence the worthiness of the fit can be explained by the insured's behavior that is modified once a claim has been reported in the year.
iii. The randomness of occurrence of claims and the associated assumption of independency of claim occurrence. This is due to the inverted business cycle where premiums are received before any costs are to be paid.

1.2. Statement of the Problem

The estimation of expected claim frequency and severity of any Insurance Policy enables insurers to make decisions on asset pricing, allocation and claim payment with high accuracy since insurance business is characterized by an inverted product cycle. This calls for optimal claim frequency modeling which is the purpose of this study to give a theoretical framework to explore and derive the same. It is in this regard that mixed random effect models are thought of providing the best fit for modeling insurance data as they are able to handle correlation, over-dispersion and zero-inflation in the data.

A recent study by Asmussen and Albrecher (2010) showcase this by using the additive and multiplicative random effect Poisson Model with the Gamma and Inverse Gaussian distributions to model risk premiums [14]. Wolny and Dominiak (2013) propose a mixed Poisson regression with spatial random effects which handle zero-inflation and over-dispersion effects [1]. The study thus seeks to explore the zero-Inflated and hurdle regression models in statistical literature that can sufficiently fit the claims data and then be used to estimate and predict the expected domestic package insurance claim liability. This shall be extended to help determine the incidence of heterogeneity in the occurrence of the claims.

1.3. Objectives of the Study

1.3.1. General Objective

The general objective of the study is to model zero inflation and over-dispersion in domestic package insurance claims using the Zero-Inflated and Hurdle regression models.

1.3.2. Specific Objective

The specific objectives are;

i. To model the zero-inflation and over-dispersion in domestic package insurance claims data using the zero inflated and hurdle models.
ii. To determine the zero-inflated count model that best fits the domestic package insurance claims data and use the same to estimate and predict the associated number of claims.
iii. To perform the diagnostics of the zero-inflated regression models on domestic package insurance claims data.

1.4. Significance of the Study

Insurance plays a major role in a country's economic development and this can only be possible if the insurance companies are making profits. This is only achieved through proper claim management since claims make up the main cash outflow for insurance companies. With proper claim modeling techniques, insurance companies will be able to estimate and predict future claims and in the long run predict their profitability since most insurance products are characterized by an inverted business cycle where premiums are paid before any claims are to be paid usually for a period of one year. Forecasting of profitability in the insurance industry can thus be used by policy makers to determine the contribution of insurance to economic growth and towards achieving the Kenya Vision 2030.

1.5. Scope of the Study

The study focuses on using the Zero-Inflated/Hurdle regression modeling techniques to model the domestic package insurance claims data for Madison Insurance Ltd, Kenya. The claims experience shall consist of detailed information on the type of domestic package insurance and the corresponding number of claims.

2. Literature Review

2.1. Introduction

This chapter is established with the intention of pre-viewing past studies on insurance claim modeling so as to get appropriate theories and the experiential proves to substantiate this research.

2.2. Empirical Literature Review

In the modeling of insurance claims, Shevchenko (2010) defined the insurance loss function as a multiplicative function of claim frequency and average claim severity [14]. Alicja & Dominiak (2013) extended this by applying the parametric regression to claims frequency modeling [1]. Yulia et al. (2013) estimated the insurance claim cost for insurance claims data using the Zero Adjusted Gamma and Inverse Gaussian Regression Models with an application to Malaysian Motor Insurance Claims [7].

Kwame & Agbodah (2014) studied probability modeling and simulation of Insurance Claims in Ghana [6]. Vytaras & Andreas (2015) modeled severity and tail risk of Norwegian Fire Insurance Claims using the Pareto, Log normal Pareto and the folded C distributions [8]. Evgenii & Elena (2017) also modeled insurance claims using the Generalized Hurdle and Gamma Distributions, an application to the Russian motor own damage insurance data [12].

Sakthivel & Rajitha (2017) studied the zero-inflated and hurdle models in comparison to the Artificial Neural Network in modeling insurance claims [13]. Joseph & Christophe (2011) reviewed the zero inflated count models in modeling annual trends in incidences of some occupational allergic diseases in France as an extension of using mixed random effects methodologies to model other count data sets other than insurance claims [9].

Marjan (2019) considered the problem of modeling violation of claims with excess zeros in a liability insurance portfolio [11]. Yogita & Kamalja (2017) summarized the several dispersed and zero inflated count data distributions used to handle dispersion in count data [2]. Cheng (2018) studied Hurdle models for general insurance claims data modeling [17]. Lu Yang & Edward (2016) extended the literature on multivariate frequency-severity regression modeling of claim counts by introducing a copula for modeling the dependence of claims [4].

3. Methodology

3.1. Introduction

This chapter discusses the Zero-Inflated and Hurdle Distribution count data models used in the modeling of zero-inflation and over-dispersion effect in claim frequency data.

3.2. Data

Data for the study included weekly number of claims filed under the domestic package insurance policy for the time period of 261 weeks from Jan 2014 to Dec 2018 for Madison Insurance Company. This was secondary data obtained from Madison Insurance Company, Ltd.

3.3. Statistical Modeling

The study uses the term modeling to refer to the process of identifying a theoretical distribution that fits the domestic package claims data reasonably well. This involved the fitting of Zero-Inflated and Dispersed statistical distributions to the domestic package insurance claims data to determine the distribution that fits the data reasonably well. The Poisson,

Zero-Inflated and Hurdle Poisson distributions were fit to the data.

The Poisson distribution is the classical distribution for modeling count data and its density function is given as;

$$p(K_i = k_i) = \frac{\exp(-\mu_i){\mu_i}^{k_i}}{k_i!} \quad (1)$$

with mean equal to variance given as ${\sigma_i}^2 = \mu_i = k_i$.

The Zero-Inflated Poisson distribution is given as;

$$p(K_i = k_i) = \begin{cases} \pi_i + (1-\pi_i)\exp(-\mu_i) & k_i = 0 \\ (1-\pi_i)\left(\frac{\exp(-\mu_i){\mu_i}^{k_i}}{k_i!}\right) & k_i \geq 1 \end{cases} \quad (2)$$

with mean and variance respectively given as $[(1-\pi_i)\mu_i]$ and $[(1-\pi_i)\mu_i(1+\ \mu_i\pi_i)]$.

The Hurdle Poisson distribution is given as;

$$p(K_i = k_i) = \begin{cases} (1-\pi_i)\,\pi_i\exp(-\mu_i) & k_i = 0 \\ \pi_i\left(\frac{\exp(-\mu_i){\mu_i}^{k_i}}{(1-\exp(-\pi_i))k_i!}\right) & k_i \geq 1 \end{cases} \quad (3)$$

with mean $\left[\frac{\pi_i}{1-exp(\mu_i)}exp(\mu_i)\right]$ and variance $[m\,exp(\mu_i) + m\,exp\,2(\mu_i)\,(1-\mathrm{m})]$ for $m = \frac{\pi_i}{1-exp(-\mu_i)}$.

3.4. Parameter Estimation

The parameter estimates were obtained through the maximum likelihood methodology in which letting r to be the number of zeros in the sample k_i, the likelihood function is given as;

$$L = (1-\pi_i + \pi_i exp(\mu_i))^r {\pi_i}^{n-r}[m1] \quad (4)$$

Where

$$m1 = \mu_i exp(-\mu_i(n-r)) \Big/ \prod_{i=1}^{n} k_i!$$

The values of π_i and μ_i which maximize the likelihood function are given by the partial derivatives of the log-likelihood for which;

$$\widehat{\pi_i} = n - r \Big/ n(1-exp(-\widehat{\mu_i})) \quad (5)$$

and

$$\widehat{\mu_i} = [(\textstyle\sum_{\forall i} k_i)(n-r)^{-1}][1-exp(-\widehat{\mu_i})] \quad (6)$$

3.5. Model Selection

The study used the Akaike Information Criterion (AIC) as a model selection measure to select the model that best fits the claims data. The model with a smaller value of the information criterion shall be deemed to be the one that gives a best fit to the domestic package insurance claims data. If we let p to be the model parameters and L the to be the likelihood function then the AIC information criterion is thus given as $AIC = -2\ln(L) + 2p$.

3.6. Model Diagnostics

To evaluate the goodness of fit of the fitted distributions, the study used the Pearson Chi-Square, Pearson & Residual Deviance, Hosmer-Lemeshow and the Cameron Trivedi test statistics.

4. Data Analysis Results & Discussions

4.1. Introduction

This chapter deliberates the zero-inflated and hurdle models for modeling claim frequency data, a case of Madison Insurance.

4.2. Descriptive Data Analysis

The exploratory data analysis gave a detailed account of preliminary analysis of the findings of the study and this is illustrated as below in Table 1.

Table 1. *Summary of Descriptive Statistics.*

Statistic	Min	Med	Mean	Max	$1^{st}Q$	$3^{rd}Q$	Var
Value	0	1	3	14	0	4	10.1

A total of 664 claims on domestic package insurance for the period Jan 2014 to Dec 2018 were used for the study. This was given a graphical visualization as in Figure 1.

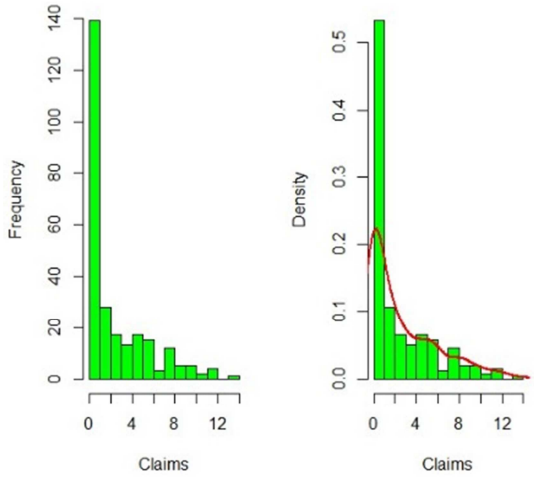

Figure 1. *Histogram and Density of Insurance Claims.*

The minimum, median, mean, maximum, 1^{st} quartile and 3^{rd} quartile number of claims made are 0, 1, 3, 14, 0 and 4 respectively. Skewness, kurtosis and variance are 1.2937, 0.8882 and 10.0644 respectively. The mean number of claims experienced is higher than the median number of claims thus the implication that most of the claims made are centered on the left of the mean value and that the extreme claim frequency values are on the right of the mean value. The skewness of the data is greater than zero and kurtosis is less than 3 thus giving direction that the claims frequency data follows a right skewed distribution that is leptokurtic. The standard deviation of 3.1724 (square root of variance) is close to the mean value of claims made thus implying that most of

the claim frequency values are close to the average number of claims i.e. there exists lower deviations in the data.

The variance of the claims data was found to be higher than the mean thus going against the equi-dispersion theory of the Poisson distribution thus the presence of over-dispersion in the data. The hanging rootogram was used to explain the dispersion in the claims data at different counts as given in Figure 2.

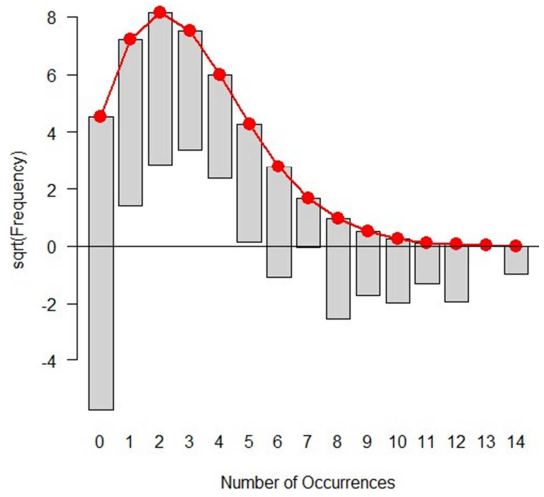

Figure 2. *Rootogram of Claims Data.*

At count zero there existed an almost equal over-dispersion and under-dispersion. For counts 1, 2, 3 and 4 there was a dramatic over-fitting and a pronounced under fitting at counts 6, 8-12 and 14. This under-fitting and over-fitting thus points to an over-dispersion and under-dispersion in the claims frequency data. The variability in the data is thus much higher that what the Poisson model can model efficiently hence the need to fit the hurdle models.

4.3. Fitted Model Coefficients

To model the insurance claims, the Poisson, Hurdle Poisson and Zero-Inflated Poisson model were fit to the data by regressing the number of domestic package insurance claims made on the regression coefficients. The regression covariates included Fire, Burglary, Theft, Explosion, Riots & Strikes and Floods.

The parameter estimates of the fitted models were estimated by the maximum likelihood estimation approach and given in Table 2 as;

Table 2. *Summary of Fitted Model Coefficients.*

	Poisson	**ZIP-T**	**ZIP-B**	**HP-T**	**HP-B**
Intercept	-0.3045	0.2510	-2.0508	0.1769	13.86
Fire	0.2354	0.1905	2.2951	0.1993	-33.58
Burglary	0.2403	0.1932	1.3395	0.1989	-33.51
Theft	0.2891	0.2276	0.8053	0.2340	-13.11
Explosion	0.3787	0.2740	1.2841	0.2784	-36.95
Riots	0.2232	0.1873	1.2850	0.1940	-17.72
Floods	0.3249	0.2097	2.0840	0.2214	-16.33

For the Poisson distribution; the estimated mean value of the claims data was found to have a 30.45% negative influence on the number of claims made. Fire, Burglary, Theft, Explosion, Riots and Floods had a 23.54%, 24.03%, 28.90%, 37.87%, 22.32% and 32.49% more chance of causing a financial loss to the insured which could result to a claim being made respectively. All the rating factors had a significant effect on the number of claims made.

For the Zero-Inflated Poisson model with log-link (ZIP-T), for an insured inclined to making a claim; Fire, Burglary, Theft, Explosion, Riots and Floods had a 19.93%, 19.89%, 23.40%, 27.84%, 19.40% and 22.14% chance for higher number of claims respectively as in Table 2. The estimated mean value of the claims data had a 17.69% positive influence on the number of claims made. All the rating factors had a significant effect on the number of claims made.

For the Zero-Inflated Poisson with Binomial logit-link model (ZIP-B), a unit increase in the number of claims recorded due to Fire, Burglary, Theft, Explosion, Riots and Floods had a respective -33.58, -33.51, -13.11, -36.95, -17.72 and -16.33 unit increase in the number of zero claims made. In this case, a unit increase in the estimated mean number of claims had a 13.86 unit increase on the number of zero claims made.

For the truncated Hurdle Model (HP-T) used to model the non-zero claims; Fire, Burglary, Theft, Explosion, Riots and Floods had a 19.05%, 19.32%, 22.76%, 27.40%, 18.73% and 20.97% respective chance for the insured to make additional non-zero claims to the already positive number of claims made. The estimated mean value of the claims data had a 25.10% positive influence on the number of claims made. All the rating factors had a significant effect on the number of claims made for this log-link hurdle coefficients.

For the Zero-Hurdle Poisson model (HP-B), a unit increase in the number of claims recorded due to Fire, Burglary, Theft, Explosion, Riots and Floods had a respective 2.2951, 1.3395, 0.8053, 1.2841, 1.2850 and 2.0840 unit increase in the number of zero claims made.

4.4. Results Discussion

This study's data exploration engrossed itself on modeling the number of insurance claims under domestic package insurance policy sold by Madison Insurance Company Ltd, Kenya with an aim of estimating the model parameters so as to aid insurance decision making. The Deviance & Pearson residuals, residual probability plots and the AIC & Log-likelihood model selection techniques were used to aid the data exploration.

4.4.1. Deviance and Pearson Residuals

Table 3. *Summary of Deviance and Pearson Residuals.*

	Min	**$1^{st}Q$**	**Med**	**$3^{rd}Q$**	**Max**
Poisson	-2.8546	-1.2145	-0.0205	0.6521	1.4679
HP	-1.8165	-0.3118	-0.0364	0.3495	0.9586
ZIP	-1.8131	-0.3016	-0.0007	0.4177	1.0390

Table 3 gave a summary of Deviance and Pearson Residuals of the fitted models in modeling insurance claims

frequency data. The median Deviance and Pearson residuals of the fitted models is close to zero and the residuals are to some extent symmetrical thus implying that the fitted models were not biased in modeling the claims data i.e. they did not overestimate or underestimate the model parameters.

4.4.2. Residual Probability Plots

As measure of goodness of fit of the fitted distributions to models claims frequency data, the quantile-quantile plots were used to give a visual representation of the residuals as given in Figure 3.

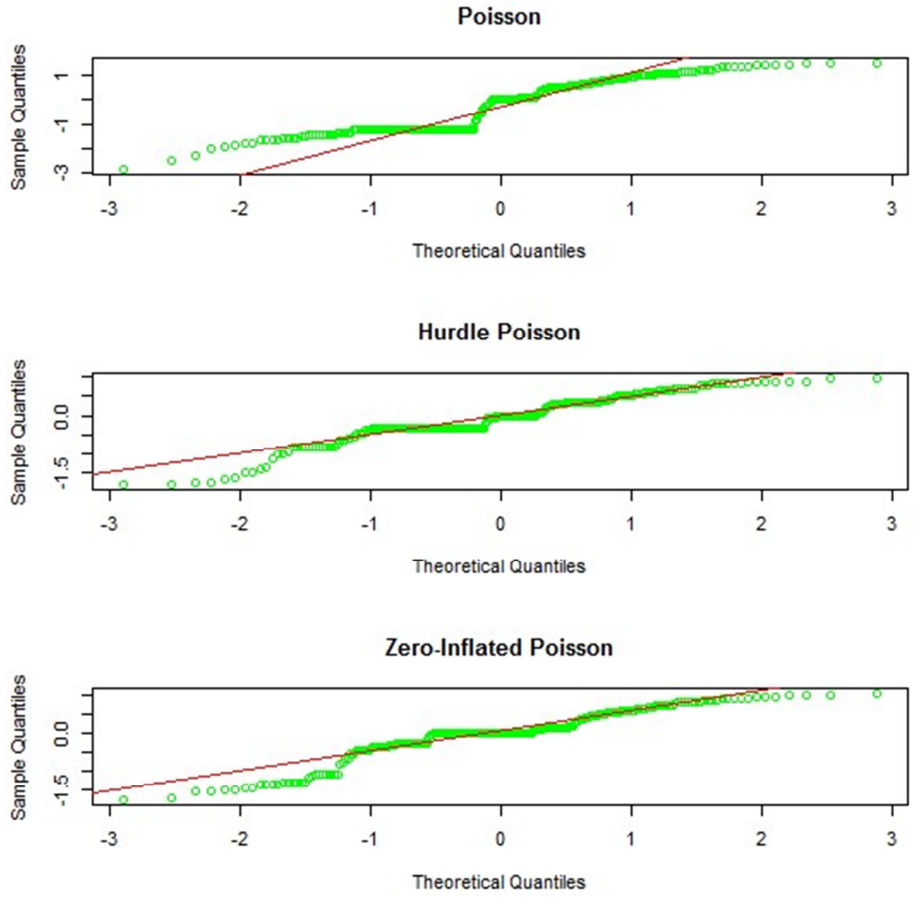

Figure 3. *Residual Probability Plots.*

A careful observation of the quantile-quantile plots of the data samples and the fitted models, revealed that the Hurdle Poisson and the Zero-Inflated Poisson models fitted the data well with a few outliers above and below the 45-degree reference line compared to the Poisson model. This gave an indication on the need to use zero-inflated and dispersed models to model claims frequency data-sets which are characterized by many zeroes.

4.4.3. Test for Dispersion

The Cameron & Trivedi (CT-1990) test was used to test for dispersion in the data. This test gave a dispersion value of 1.140364 which was greater than the reference value of 1 (equi-dispersion) for the Poisson distribution. This was a confirmatory test for the presence of dispersion in the data that was more than what the Poisson distribution can handle thus the need to fit higher forms of the Poisson distribution to the insurance claims data.

4.4.4. Model Selection

Model selection criteria was informed by the information criterion. Table 4 gives the values of AIC for the fitted models and the associated log likelihoods. The Zero-Inflated Poisson gave a better fit to the data as it had the lowest AIC and Log-likelihood.

Table 4. *Summary of AIC and Log-Likelihood.*

	Poisson	Z. I-Poisson	H-Poisson
AIC	782.255	653.2269	687.6441
Log-Likelihood	-	-312.6	-329.89

This was confirmed via the Hosmer-Lemeshow test which confirmed the appropriateness of the Zero-Inflated Poisson in the modeling of insurance claims frequency data as in Table 5.

Table 5. *Summary of Hosmer-Lemeshow test.*

	Poisson	H-Poisson	ZI-Poisson
X-Squared	1271.4	18.273	-39.662
Df	8	8	8
P-Value	2.2e-16	0.01927	1

The Hurdle Poisson and the Standard Poisson distributions had p-values of 2.2e-16 and 0.01927 respectively which were relatively small than that of the Zero-Inflated Poisson which was 1 at 8 degrees of freedom. Since small p-values for the

Hosmer-Lemeshow test indicate poor model fit, the Zero-Inflated Poisson is given as a model of good fit as it had a large p-value than the other models.

4.5. Claims Frequency Prediction

Since the Zero-Inflated Poisson gave a better fit to the data, it was used to predict the future number as given in Figure 4. Claims were predicted for a period of 24 weeks (8 months). The maximum and minimum predicted claims were six (7) and zero (0) respectively. Hence at any expected time the insurer would expect a maximum of seven (7) claims and a minimum of zero (0) claims. The expected mean number of claims was found to be two (2) claim counts with a standard deviation of 2.2161.

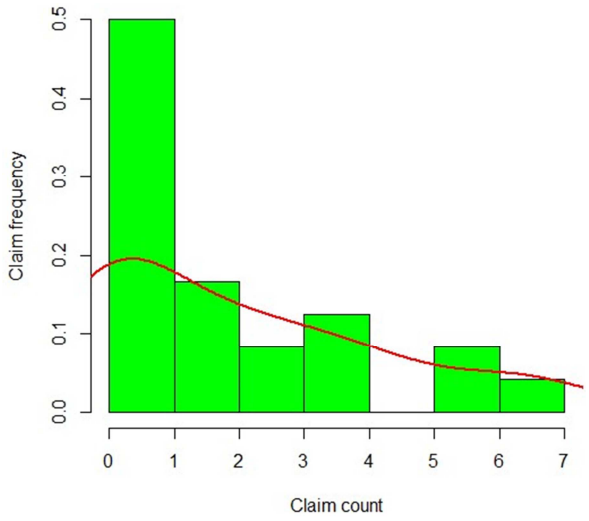

Figure 4. *Histogram of Predicted Number of Claims.*

This gave an implication that at any particular time interval, the insurer (Madison Insurance Company Ltd.) would expect at least two claims to be made with regard to the domestic package insurance policy thus enabling the insurer to set aside enough reserves to cater for the anticipated claims.

In order to determine the association between the observed and the predicted number of Domestic Package Insurance Claims, the Pearson Chi-Square test was used. Table 6 gave a summary of the Pearson Chi-Square test.

Table 6. *Chi-Square test Comparison for Observed and Expected Claims.*

	Statistic	Df	P-value
Value	498.59	13	2.2e-16

The observed and the expected number of Domestic Package Insurance Claims were found to be statistically significantly associated. This was evident as with the Pearson Chi-Square P-value that was close to zero (p-value ≈ 0) implying presence of association between the observed and expected number of claims.

This gave an implication that the observed and the expected claims came from the same distribution thus the appropriateness of the Zero-Inflated Poisson in the modeling and predication of insurance claims.

5. Conclusions and Commendations

5.1. Introduction

This chapter is the final stage of the study; it gives conclusions to the findings and recommendations for future research.

5.2. Conclusion

Modeling is a vital aspect for fair pricing of insurance products in the insurance industry. It helps determine the amount of premium to be charged and the appropriate policy exclusions (inclusions) for any given insurance policy. The study used the Poisson, Hurdle Poisson and the Zero-inflated Poisson in an application to modeling domestic package insurance claims frequency data in which a total of 664 claims from Madison Insurance for the period Jan 2014 to Dec 2018 were enrolled for the study.

The study concludes that in the analysis of count data with indistinguishable sources of excess zeros, the Zero inflated regression models should be used to analyze them. This is due to reason that this models give a better account of the heterogeneity of claim occurrence, zero-inflation and dispersion associated with such data-sets. This study postulates the use of the Zero-Inflated Poisson model as one of the zero-inflated models that can be used to model count data outcomes with excess zeros such as the number of insurance claims.

5.3. Recommendations

The claim count modeling is one of the important steps in the insurance rate making process. Great work needs to be done to help model zero-inflated data counts as with insurance data-sets with the expansion of this modeling techniques to capture other types of zero-inflated models. This would include the Generalized Quasi Poisson and the Discrete Weibull models in order to give a better fit of the model parameter estimates. This study gave an application of the Poisson, Zero-Inflated Poisson and Hurdle Poisson models to the modeling of the number of domestic package insurance claims. To improve on this application, other techniques like the bootstrapping can be considered in future researches with an application to longitudinal data with indistinguishable sources of excess zeros.

Acknowledgements

Many thanks to my noble, hardworking and kind hearted supervisors Dr. Antony Wanjoya and Dr. Anthony Ngunyi for their time, piece of advice, encouragement and constructive criticism throughout the research period. Profound gratitude also goes to my course mates, the members of the school of mathematical sciences and the entire JKUAT University staff for their support and encouragement during my graduate study. Much thanks also goes to Damackline Bundi of Madison, my mum, dad and all my siblings without forgetting my profound love (Fanis) and all love ones. I owe it to you all.

References

[1] Alicja, Wolny-Dominiak. (2013). Zero-inflated claim count modeling and testing–a case study. Ekonometria, 1 (39), 144-151. ISSN 1507-3866.

[2] Yogita, S. W., & Kamalja, K. K. (2017). Modeling Auto Insurance Claims in Singapore. Sri Lankan Journal of Applied statistics, 18 (2); 105–118. doi: 10.4038/sljastats.v18i2.7957.

[3] Insurance Regulatory Authority. (2017). Insurance Annual Report. Nairobi: Insurance Regulatory Authority. https://www.ira.go.ke.

[4] Edward, W. F., Gee, Lee. & Lu, Yang. (2016). Multivariate Frequency-Severity Regression models in Insurance. Department of Statistics, University of Wisconsin-Madison, 1300 University Avenue, Madison, WI 53706, USA.

[5] Feldman, J. A., & Robert, L. B. (2005). Risk and Insurance. Education and Examination Committee for the Society of Actuaries. https://www.soa.org/globalassets/assets/files/edu/P-21-05.pdf.

[6] Kwame, G. A., ELVIS, D., & Agbodah, K. (2014). Probability Modeling and Simulation of Insurance Claims in Ghana. Global Journal of Commerce and Management Perspective, 3 (5); 41-49. ISSN 2319-7285.

[7] Yulia, R., Noriszura, I., & Saiful, H. J. (2013). Estimation of Claim Cost Data Using Zero Adjusted Gamma and Inverse Gaussian Regression Models. Journal of Mathematics and Statistics, 9 (3); 186-192. doi: 10.3844/jmssp.2013.186.192.

[8] Vytaras, B., & Andreas, K. (2015). Measuring Severity and Tail Risk of Norwegian Fire Insurance Claims. North American Actuarial Journal, 20 (1); 1-16. doi: 10.1080/10920277.2015.1062784.

[9] Joseph, N. W., & Christophe, P. (2011). On the Zero-Inflated Count Models with Application to Modeling Annual trends in Incidences of some Occupational Allergic Diseases in France. Journal of Data Science 9 (2011); 639-659.

[10] Rotimi, F. A., Oyindamola B. Y., & Ayoola S. A. (2018). Modeling excess zeros in count data with application to Antenatal Care Utilization. International Journal of Statistics and Probability, 7 (3); 22-35. Doi: 105539/ijsp.v7n3p22.

[11] Marjan, Q. (2019). On the Violation of Claims with Excess Zeros in Liability Insurance: A comparative study. Open access journal, 7 (3); 1-17.

[12] Evgenii, V. G., & Elena, A. M. (2017). Modern Claim frequency and severity models: An application to the Russian motor and own damage insurance market. Cogent Economics & Finance, 5 (1). doi: 10.1080/23322039.2017.1311097.

[13] Sakthivel, K. M., & Rajitha, C. S. (2017). A Comparative Study of Zero-Inflated, Hurdle Models with Artificial Neural Network in Claim Count Modeling. International Journal of Statistics and Systems, 12 (2); 265-276. ISSN 0973-2675.

[14] Asmussen, S., & Albrecher, H. (2010). Ruin Probabilities (2Nd Edition). Advanced series on statistical science & applied probability, 14 (2); World Scientific Publishing Co. ISBN-13: 978-9813203617.

[15] Staub, k., & Winkelmann, R. (2013). Consistent estimation of zero inflated count models. Health Economics, (22); 673-686. doi: 10.1002/hec.2844.

[16] Klugman, S. A., Panjer, H. H., & Willmot, G. E. (2008). Loss Models: From Data to Decisions (Thirded.) John Wiley & Sons, Inc. ISBN: 978-1-119-52378-9.

[17] Cheng Tian. (2018). Hurdle Models in Non-Life Insurance. Foculty of Mathematics and Physics, Charles University. Thesis Id: 188550.