INVENTOPIA 2025

FBM-SEREMBAN INTERNATIONAL INNOVATION COMPETITION (FBM-SIIC)

INNOVATION IN ACTION: TURNING IDEAS INTO REALITY

Chapter 49

Detariffed Premium Estimation for Automobile Insurance: Modelling Frequency and Severity Using Poisson, Lognormal, Pareto and Generalized Linear Model

Syazreen Niza Shair, Nur Alya Safiyya binti Shamsul Rijal, Nurin Qistina binti Nor Azlin, Nazhatul Ulya binti Mohd Farid & Nur Aliya Syuhada binti Saleh

Centre for Actuarial Studies
Faculty of Computer and Mathemtical Sciences
Universiti Teknologi MARA (UiTM) Shah Alam

syazreen@uitm.edu.my

ABSTRACT

The transition from a tariff-based to a de-tariffed automobile insurance market in Malaysia has significantly reshaped premium estimation thus requiring for more accurate risk assessment among insurance companies. This project provides significant milestones by investigating the effects of de-tariffication by modelling claim frequency and severity using real-world insurance data. Advanced statistical models have been applied for robust premium pricing in which Poisson and Zero-Modified Poisson (ZMP) models were employed to model claim frequency. Moreover, Lognormal and Pareto distributions models were adopted to model the respective claim severity. The Lognormal distribution best captured small to moderate claims, whereas the Pareto model better represented extreme losses. Model evaluations favored the Lognormal distribution due to its superior goodness-of-fit metrics. We propose Generalized Linear Models (GLMs) to estimate pure premiums. The tariff model included factors like Cylinder Capacity and vehicle value, while the de-tariff model incorporated additional risk variables such as driver age, experience, vehicle age and power, second driver availability, area classification, and vehicle value. Findings confirm that risk-based pricing allows for fairer, more personalized premiums but also introduces new challenges like price competition and adverse selection. This research offers important insights for insurers, actuaries, and regulators striving for pricing sustainability.

Key Words: De-tariffication, Premium estimation, Claim frequency, Claim severity, Generalized Linear Model (GLM).

1. INTRODUCTION

The automobile insurance industry plays a vital role in managing risks, particularly in Malaysia, where road accidents are on the rise. In 2017, the industry experienced a significant transformation with the implementation of a de-tariffed model, which introduced risk-based pricing. This shift increased the need for more accurate premium estimation methods. Accurate claim prediction has become essential for ensuring financial stability and fairness in pricing. This study sets out to address these challenges with four key objectives. Firstly, this study aims to model the frequency of automobile insurance claims using both the Poisson and Zero-modified Poisson models. Secondly, it seeks to model claim severity using Lognormal and Pareto distributions. Thirdly, this study focuses on modelling the pure premium of automobile insurance in the context of de-tariffication, employing a Generalized Linear Model (GLM). The findings are expected to benefit a range of stakeholders, including regulators, insurers, and academic researchers.

2. LITERATURE REVIEW

This study examines the impact of de-tariffication on Malaysia's motor insurance, focusing on claim frequency, severity, and premium modelling. It evaluates statistical models such as Poisson, Zero-Modified Poisson, Lognormal, Pareto, and Generalized Linear Models (GLM) to enhance risk-based pricing and insurer sustainability. Malaysia's transition to a risk-based pricing model in 2017 replaced fixed tariffs, promoting fairer premiums and market competition. However, challenges such as price instability and imperfect competition persist. Effective risk assessment methods are crucial to ensuring balanced premium structures and market stability. Several factors influence claim frequency and severity. The zero-claim phenomenon arises when policyholders avoid small claims to retain No-Claim Discounts (NCD), leading to excess zeros in datasets. The Zero-Inflated Poisson model effectively accounts for this issue. Claim frequency and severity are also interdependent, requiring two-part models for accurate risk assessment. Additionally, outliers and skewed data impact risk modelling, necessitating proper statistical methods to ensure fair premium calculations. Risk assessment and pricing rely on statistical models that incorporate factors such as demographics and vehicle characteristics. Advanced modelling approaches, including Conditional Probability Decomposition and Copula Models, improve claim predictions. Since de-tariffication, insurers have adjusted their rating factors, with studies identifying key determinants like driver age, vehicle type, mileage, and NCD levels. This research extends previous works by applying Poisson and Zero-Modified Poisson (ZMP) models to analyze claim frequency and Lognormal and Pareto distributions are used, Lognormal for moderate claims and Pareto for extreme losses, ensuring accurate premium estimation. In addition, premium calculation is performed using Generalized Linear Models (GLM), which account for the relationship between claim frequency and severity, improving pricing accuracy. This study provides a statistical foundation for risk-based pricing in Malaysia's de-tariffed motor insurance market, offering insights for regulators, insurers, and policymakers.

3. METHODOLOGY

This study uses a 2018 Spanish motor insurance dataset comprising 28,037 policy records (risk type 3 – passenger cars) with 19 attributes including policyholder details, vehicle information, and claims history as in the Table 1.

Table 1 List of attributes and explanations from data

Attribute	Explanation	Variable
Policy Year	Year of the policy.	
Year of Birth	Policyholder's years of birth.	
Age	The age of policyholder in the year of 2018.	x_1
Driving Experience	How many years the policyholders had their driver license since 2018.	x_2
Annual Claim Cost	Total cost of claims incurred for the insurance policy during the current year.	
Number of Claims	Total number of claims incurred for the insurance policy during the current year.	
Area Type	Dichotomous variable indicates the area. 0 for rural and 1 for urban (more than 30,000 inhabitants) in terms of traffic conditions.	x_3
Second Driver	1 if there are multiple regular drivers declared, or 0 if only one driver is declared.	x_4
Vehicle Registration Year	Year of registration of the vehicle (YYYY).	
Vehicle Age	The age of the car in year 2018 (in years)	x_5
Power	Vehicle power measured in horsepower.	x_6
Cylincer Capacity (CC)	Cylinder capacity of the vehicle.	x_7
Vehicle Market Value	Market value of the vehicle on 31/12/2019.	x_8

Claim frequency is modelled using Poisson distribution, appropriate for count data. To handle excess zeros, a Zero-Modified Poisson (ZMP) model is also applied. Parameters were estimated via Maximum Likelihood Estimation (MLE). Claim severity is modelled using Lognormal and Pareto distributions. Lognormal captures right-skewed, multiplicative losses, while Pareto handles heavy tails and extreme claims. The Kolmogorov-Smirnov (K-S) test is used to compare empirical and theoretical cumulative distribution functions, verifying whether the chosen distributions align with the observed data. The AIC and BIC are employed to measure the fitness of the models. Pure premium is estimated by multiplying claim frequency and severity, as expressed by $PP = E(N) \times E(X)$. In the tariff-based approach, premium calculation is primarily influenced by vehicle characteristics such as cylinder capacity and the insured value of the vehicle. In contrast, using Generalised Liner Model, the de-tariffed approach adopts a risk-based framework, incorporating a broader set of factors.

4. RESULTS AND DISCUSSION

Evaluation of models resulted in the Zero-Modified Poisson (ZMP) model outperforming the standard Poisson model for claim frequency withl ower AIC and BIC values compared to Poisson. This result highlights ZMP's effectiveness in handling excess zeros in the data, making it the more suitable model for insurance claim frequency analysis. The comparison of

severity models shows that while both capture the heavy-tailed nature of the data, the Lognormal model shows a better fit with lower AIC and BIC, and a smaller K-S statistic values than Pareto. This suggests the Lognormal distribution is more accurate than Pareto particularly for moderate-to-large claims in the datasets. A Poisson Generalized Linear Model (GLM) was applied to estimate claim frequency under both tariff and de-tariff frameworks. In the tariff model, cylinder capacity and vehicle value were used as predictors, with results indicating both variables significantly influenced claim frequency. For the de-tariff model, additional factors such as driver age, experience, car age, second driver availability, and area were included. All were statistically significant except car power. Fitted equations from both models were derived as follows.

Tariff model frequency fitted equation:

$$E(N) = \exp(-2.608881 + 0.0003962x_7 + 0.0000002x_8)$$

De-Tariff model frequency fitted equation:

$$E(N) = \exp(-2.673473 + 0.0060594x_1 - 0.0207724x_2 + 0.1296949x_3 + 0.4234535x_4 + 0.0183851x_5 + 0.0003042x_7 + 0.0000086x_8)$$

A Lognormal Generalized Linear Model (GLM) was used to estimate claim severity, excluding zero claims for consistency. In the tariff model, cylinder capacity (CC) and vehicle value were significant predictors, aligning with traditional premium rules. The de-tariff model included broader variables, but only car age, second driver, and vehicle value were statistically significant. Fitted equations were derived as below.

Tariff model severity fitted equation:

$$E(X) = \exp(5.6994233 - 0.0002686x_7 + 0.0000183x_8)$$

De-Tariff model severity fitted equation:

$$E(X) = \exp(5.955632 - 0.1970641x_4 - 0.0220186x_5 + 0.0000102x_8)$$

Pure premiums were calculated by combining fitted frequency and severity models using GLM. Under the tariff model, premiums were based on cylinder capacity and vehicle value. In the de-tariff model, broader risk factors were included, with the second driver and area having the largest impact, while driving experience reduced premiums. These results highlight how risk-based pricing under de-tariffication enables more accurate and equitable premium estimation. The results illustrate the shift from fixed-rate to risk-based pricing, offering more accurate premium estimation aligned with individual risk profiles.

Table 2 Premium estimated using Tariff model

Value of vehicle (€) \ CC	**0 – 1000**	**1001 – 1500**	**1501 – 2000**	**> 2000**
270 – 10000	30.06	30.06 - 32.04	32.04 - 34.15	>38.8
10001 – 20000	30.06 - 36.16	30.06 - 38.55	32.04 - 41.09	>34.15
20001 – 100000	36.16 - 158.77	36.17 - 169.23	38.55 - 180.39	> 41.09

Table 3 Premium estimated using De-Tariff model

Value of vehicle (€) \ CC	0 – 1000	1001 – 1500	1501 – 2000	> 2000
270 – 10000	35.34	35.35 - 41.15	41.16 - 47.91	> 64.94
10001 – 20000	35.34 - 42.66	35.35 - 49.67	41.16 - 57.82	> 47.91
20001 – 100000	42.66 - 192.20	42.67 - 223.77	49.68 - 260.53	> 57.83

5. CONCLUSION AND RECOMMENDATION

This study proposed insurance premium modelling in a de-tariffed market, with a focus on comparing traditional tariff-based pricing to risk-based pricing. The Zero-Modified Poisson (ZMP) model was found to be effective in capturing claim frequency, highlighting that younger, inexperienced drivers and older vehicles tend to generate higher claim counts. For claim severity, the Lognormal distribution provided the best fit, effectively reflecting the variation in claim sizes across different driver age groups and vehicle conditions. The estimated pure premiums from the Generalized Linear Model (GLM) model demonstrated that the de-tariffed model produced more accurate and individualized premium estimates. Although risk-based pricing offers a more equitable and precise premium structure, it is crucial to maintain transparency and ensure consumer protection in this evolving market landscape. Future research should consider additional risk factors such as driving behaviour, accident history and road conditions. Incorporating telematics data could further enhance risk assessment and improve the accuracy of premium predictions in a risk-based pricing environment.

REFERENCES

Boucher, J., Denuit, M., & Guillen, M. (2009). Number of Accidents or Number of Claims? An Approach with Zero-Inflated Poisson Models for Panel Data. Journal of Risk & Insurance, 76(4), 821–846. https://doi.org/10.1111/j.1539-6975.2009.01321.x

Garrido, J., Genest, C., & Schulz, J. (2016). Generalized linear models for dependent frequency and severity of insurance claims. Insurance Mathematics and Economics, 70, 205–215. https://doi.org/10.1016/j.insmatheco.2016.06.006

Ismail, N., & Zamani, H. (2013). Estimation of claim count data using negative binomial, generalized Poisson, zero-inflated negative binomial, and zero-inflated generalized Poisson regression models. Casualty Actuarial Society E-Forum, Spring, 1- XX.

U. S. Pasaribu, H. Husniah, R. Abubakar, A. Sonhaji, & N. F. Sa'idah. (2024). Determining fire insurance premium in Indonesia based on severity and frequency claim distributions. MATEMATIKA, MJIM, Volume 40, Number 2, 49–60, Volume 40. https://matematika.utm.my/index.php/matematika/article/view/1490

Yogita, S. W., & Kirtee, K. K. (2017). Zero-inflated models and estimation in zero- inflated Poisson distribution. Communications in Statistics - Simulation and Computation, 46(7), 5460-5475.

Zhang, P., Pitt, D., & Wu, X. (2024). A comparative analysis of several multivariate zero-inflated and zero-modified models with applications in insurance. Communications in Statistics - Theory and Methods, 1–28. https://doi.org/10.1080/03610926.2024.2360079