International Conference on Actuarial Sciences (ICActS)
Available on-line at: **www.sicash.sliit.lk**

A Poisson Mixture Model of Claim Counts to Improve Insurance Claim Predictions Using Incomplete Data/ Asymmetric Data: A Case Study with Telematics Insurance

K. G. H. S. Peiris[1]*, J. K. H. Sampath[1], L. P. Nadeeka D Premarathna[2]

[1] *Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, Canada*
[2]*Department of Mathematics, University of Kelaniya, Kelaniya, Sri Lanka*

Corresponding author*: hashan_peiris@sfu.ca

Abstract

In the evolving landscape of insurance analytics, integrating traditional and telematics data is pivotal for enhancing the accuracy of claim predictions. This study introduces a two-fold approach utilizing a Poisson mixture model to merge these distinct data streams effectively. Initially, we apply the Poisson mixture model to traditional insurance features common to both datasets, employing Hamiltonian Monte Carlo (HMC) and Metropolis-Hastings algorithms separately for model fitting. Subsequently, the predicted claim counts derived from the Poisson mixture model are used as an offset to fit a Poisson generalized linear model (GLM) exclusively with telematics-based features. Our focus is on assessing the suitability of HMC and Metropolis-Hastings for addressing data integration challenges within Poisson mixture frameworks. Comparative analysis reveals that while HMC demands more computational time to achieve convergence, it exhibits superior performance in parameter estimation in scenarios with increased model complexity. This study underscores the potential of advanced Monte Carlo methods in refining predictive models by leveraging the synergy between traditional and telematics data sources.

Keywords: Automobile insurance, Data integration, Driver telematics, Hamiltonian Monte Carlo, Metropolis-Hastings, Poisson mixture model.

Introduction

In the rapidly evolving landscape of the insurance industry, the diversity of automobile insurance products has increased, creating competition among companies to attract customers. These companies must also manage the cost of insurance to maintain competitive pricing. In this context, accurately predicting claim counts has become a focal point of innovation. Typically, this process utilizes the policyholder data corresponding to each product type.

Usage-Based Insurance (UBI) is a product that calculates the premium based on the driver's behaviour, as data streams generated by vehicles provide valuable insights for researchers and experts. They have developed robust methods to model and interpret this complex information. The literature on telematics data modelling encompasses diverse approaches aimed at leveraging real-time or behavioural data to improve risk assessment, personalize insurance policies, and ultimately transform the industry.

https://doi.org/10.54389/CMRP1573

Insurance companies generally have access to large datasets related to policyholders that contain traditional characteristics, such as driver demographics and vehicle features. Traditional features (X_1) are the data collected when offering an insurance contract. In contrast, telematics features (X_2) are collected during the contract period through a telematics device if the insured purchases a UBI product by giving consent to collect driving behaviour data.

However, a telematics dataset may have fewer data points than a traditional dataset, as the number of telematics-related policyholders is low. According to Peiris et al. (2024), the scarcity of telematics observations compared to traditional observations could result from privacy and security concerns or customers' reluctance to adopt new technology. Additionally, risky drivers might not purchase UBI as it results in a higher premium if their risk level increases. If the population of insureds can be observed as two-fold, low-risk and high-risk, one can expect most UBI policyholders to be low-risk drivers compared to traditional policyholders as their premiums depend on some feature of their driving behaviour. Thus, one can recognize two different subpopulations among drivers concerning the type of insurance product they have.

Though analysing the two datasets separately is common practice, there are different approaches available in literature that fit one model considering a finite population with both policies (Ayuso et al., 2019; Meng et al.,2022; Chan et al.,2022; Peiris et al.,2024). In this project we investigate the available datasets jointly to better understand the population characteristics, compared to a separate analysis of the traditional and telematics datasets. Thus, we model the number of claims (N) by incorporating both traditional and telematics insurance features, making this method a data integration approach to fit a Generalized Linear Model.

On the other hand, Poisson mixture models are statistical tools used to analyse count data where the population is assumed to be a mixture of two or more distinct subpopulations. Each subpopulation is modelled with its own Poisson distribution, characterized by different mean rates of occurrence. This model is particularly useful in situations where the data exhibits overdispersion, often indicative of an underlying heterogeneous population. Thus, we can use a Poisson mixture model in this case as the population of an automobile insurance company contains at least two subpopulations with possibly different risk profiles when both traditional and telematics policies are present.

For example, in insurance claim predictions, one subpopulation might represent low-risk clients with fewer claims, while the other represents high-risk clients with more frequent claims. The Poisson mixture model with two subpopulations allows for a more nuanced analysis, providing insights that can inform better decision-making. It helps in identifying and quantifying the different risk levels within the overall population, leading to more targeted and effective risk management strategies. Similar approaches can be seen in the works of Bermúdez et al. (2020) and Brown and Buckley (2015), who present a non-parametric Bayesian approach to model claim counts in different types of insurance using a Poisson mixture model, helping to determine group heterogeneity.

The integration of traditional data sources with telematics data has opened new avenues for enhancing prediction models. This study delves into the development and application of a Poisson mixture model and a Poisson Generalized Linear Model (GLM) for insurance claim counts, highlighting the potential to refine predictions by leveraging the rich, detailed insights provided by telematics data.

Materials and Methods

The study proposes a two-step Bayesian approach using a Poisson mixture model to analyse insurance claim counts from heterogeneous populations. A two-component Poisson mixture model is specified,

where each insured's claim count is generated from one of two Poisson distributions with distinct rate parameters that depend on covariates and an exposure term through a log-linear form. This setup allows the model to capture heterogeneity between traditional and telematics policyholders, with the expected claims from the mixture model later used as an offset in a Poisson GLM for correction.

Priors are defined using normal distributions for regression coefficients and a beta distribution for the mixture weight, and inference is carried out within a Bayesian framework using Markov Chain Monte Carlo (MCMC) sampling, particularly the No-U-Turn Sampler (NUTS) in Stan.

Posterior means, variances, and credible intervals are employed for parameter estimation and model evaluation. To fit the model, two sampling algorithms are considered: the classical Metropolis-Hastings algorithm and the more efficient Hamiltonian Monte Carlo method. A simulation study is then conducted using a finite population of 100,000 observations divided into two subpopulations, incorporating both traditional and telematics features. The dataset is split into a validation set, a telematics training set, and a traditional training set. The proposed mixture model is fitted using both MH and HMC approaches, and its predictive performance is compared against several benchmark models, namely a naive Poisson GLM with telematics data only, a traditional GLM with only traditional features, a full model using all available features, and a boosting model that combines traditional estimates with telematics adjustments similar to Peiris et al. (2024). Model performance is assessed using Root Mean Squared Error (RMSE) and Poisson Deviance (DEV) on the validation set, with bootstrap resampling repeated 1000 times to ensure robust average performance measures. This methodology provides a structured and flexible way to account for heterogeneity in insurance claim data while leveraging both traditional and telematics information.

Results and Discussion

The results of the study highlight differences in parameter estimation, processing time, and model performance between the Hamiltonian Monte Carlo (HMC) and Metropolis-Hastings (MH) algorithms. Parameter estimates obtained using both methods were compared against true parameter values, showing that HMC generally provided closer approximations for the S1 subpopulation, while both algorithms performed similarly for the S2 subpopulation. However, MH estimates exhibited lower variability compared to HMC, indicating greater stability in certain cases. In terms of computational efficiency, MH proved to be faster, requiring 2.81 hours compared to HMC's 5.45 hours, which reflects the higher computational cost of HMC due to its more detailed parameter space exploration. When benchmarked against standard models using Root Mean Squared Error (RMSE) and Poisson Deviance (DEV), both HMC and MH performed competitively, producing values like the naive model and outperforming the traditional model. The full model, while delivering the best predictive accuracy, is impractical in real applications. Overall, the findings suggest that the proposed mixture-based methods, particularly HMC, offer reliable and accurate estimation capabilities, though at the expense of longer processing times, whereas MH provides a faster but slightly less precise alternative.

Conclusions

The results suggest that while HMC offers more precise estimates, it does so at the cost of increased computational resources and time. In contrast, MH provides a more efficient, albeit slightly less accurate, alternative. The choice between these methods should consider the trade-offs between accuracy and computational efficiency, especially in scenarios where time or resources are limited. The choice of prior and proposal distributions significantly affects the performance of both algorithms.

The normal distributions used for parameters and the beta distributions for in this study appear to be well-suited, as indicated by the overall good performance in parameter estimates. The trace plots obtained for the MH algorithm indicate that the algorithm may not have converged. It is recommended to employ various priors and proposal distributions to test for convergence.

For practical applications, these findings suggest that HMC might be preferable in scenarios where the highest accuracy is required and sufficient computational resources are available. Also, it shows the ability to handle the high dimensional parameter space than MH. On the other hand, MH could be employed effectively in situations where quicker, albeit slightly less precise, results are needed. However, a significant problem that arises in practical situations is the lack of knowledge regarding parameter estimates other than the mixing proportion. Hence, it is suggested to determine the estimates with an acceptable level of convergence and then choose the best set of estimates with the mixing proportion estimate closest to the true value which is available to the company. This study highlights the effectiveness of sampling algorithms with the MCMC approach techniques and their applicability in real-world statistical problems. The choice of algorithm should be guided by specific project needs, balancing accuracy and computational efficiency. Further studies could explore alternative distributions, or hybrid approaches to optimize both performance metrics. Moreover, it is suggested to use the adaptive Metropolis-Hastings algorithm as an extension to the existing MH algorithm. Finally, given that the out-of-sample validation set consists of customers from both risk subpopulations, lower Avg_RMSE and Avg_DEV values suggest that the proposed method might be an effective way to incorporate telematics information and increase the prediction accuracy when UBI products are used by high-risk customers.

References

Ayuso, M., Guillen, M., & Nielsen, J. P. (2019). Improving automobile insurance ratemaking using telematics: Incorporating mileage and driver behaviour data. *Transportation, 46,* 735–752.

Bermúdez, L., Karlis, D., & Morillo, I. (2020). Modelling unobserved heterogeneity in claim counts using finite mixture models. *Risks, 8(*1*),* 10.

Brown, G. O., & Buckley, W. S. (2015). Experience rating with Poisson mixtures. *Annals of Actuarial Science, 9(*2*),* 304–321.

Burda, M., Harding, M., & Hausman, J. (2012). A Poisson mixture model of discrete choice. *Journal of Econometrics, 166(*2*),* 184–203.

Chan, J. S., Choy, S. B., Makov, U., Shamir, A., & Shapovalov, V. (2022). Variable selection algorithm for a mixture of Poisson regression for handling overdispersion in claims frequency modeling using telematics car driving data. *Risks, 10(*4*),* 83.

Meng, S., Wang, H., Shi, Y., & Gao, G. (2022). Improving automobile insurance claims frequency prediction with telematics car driving data. *ASTIN Bulletin: The Journal of the IAA, 52(*2*),* 363–391.

Owen, A. B. (2013). *Monte Carlo theory, methods and examples*. Retrieved from https://artowen.su.domains/mc/

Peiris, H., Jeong, H., Kim, J.-K., & Lee, H. (2024). Integration of traditional and telematics data for efficient insurance claims prediction. *ASTIN Bulletin*, 1–17. https://doi.org/10.1017/asb.2024.6

Yamada, T., Ohno, K., & Ohta, Y. (2022). Comparison between the Hamiltonian Monte Carlo method and the Metropolis-Hastings method for coseismic fault model estimation. *Earth, Planets and Space, 74,* 86.