Warsaw School of Economics

Collegium of Economic Analysis

Self-report

to the doctoral thesis of M.Sc. Norbert Paska

under the title „Application of generalized linear mixed models in the pricing of motor insurance"

written under scientific guidance of
prof. dr hab. Bartosz Witkowski
and
dr hab. Michał Bernardelli, prof. SGH

Warsaw 2021

Table of contents

1. Justification for the topic selection

According to the information published by the Polish Financial Supervision Authority (2020), there were 33 insurance companies that operated on the Polish property insurance market in 2019, with gross written premium of PLN 31 billion. Therefore, as the insurance market can clearly be considered as competitive, the companies must compete for the target customer through, inter alia, commissions, bonus contracts for sellers, marketing activities and campaigns, wide protection scope of the offered products and the insurance price, that last one being one of the most important. When determining insurance premium, one must consider both: the client's willingness to pay and the insurance risk evaluation.

Therefore, it is crucial for an insurance company to segment the risk as accurately as possible. This goal can be achieved by using the pricing models, which allow the company to meet two postulates of the economics: justice and efficiency. The first one is fulfilled because people with risky driving profile (i.e., with high damage probability) must pay higher premiums as opposed to people driving carefully and with respect to traffic regulations (i.e., with lower damage probability). On the other hand, the company cares about efficiency by rewarding those who are less exposed to damage and punishing those who are driving in a reckless and dangerous manner. This way, the insurance company can maximize profitability – both through a higher sales volume and by reducing losses incurred as a consequence of claim payments.

As Antonio and Valdez (2012) pointed out, because of the risk portfolio heterogeneity, insurance companies cannot offer all insured persons the same insurance rates, as this could expose the company to the risk of adverse selection[1], leading to unfavorable financial consequences and ultimately also to insolvency, collapse of the insurance company and destabilization of the entire insurance market. Therefore, it is believed that the lack of appropriate pricing models may lead to inadequate ratio of bad and good risks[2] in the company's portfolio compared to the broader market. Antonio and Valdez (2012) also emphasized that in the longer term such situation could result in an average premium increase by the insurance company to maintain its solvency. This, in turn, will make the company even more

[1] Adverse selection is defined as a situation in which insurance is more likely to be bought by people with higher risk. This issue is closely related to the information asymmetry phenomenon between the policyholder and the insurance company (Sulewski and Meuwissen, 2014)

[2] Bad (good) risk is defined in a business practice as a policy characterized by higher (lower) than average probability of the damage.

uncompetitive, which will cause further loss of low-risk clients and persistent lack of profitability.

The effect of improving pricing models can be obtained in at least two ways. The first one consists in introducing new variables to the tariff which adequately determine the insurance risk. Husnjak, Peraković, Forenbacher and Mumdziev (2015) wrote that the analysis of driver behavior can give an insurance company a competitive advantage by proposing different insurance rates for drivers who belong to different insurance risk segment. The second way is to increase the predictive power of tariff models by improving the precision of the estimators. The second option is widely discussed in this doctoral thesis.

2. Subject and scope of the doctoral thesis

Widely used tools in actuarial science for predicting the number of claims and the loss amount are generalized linear models with Poisson and gamma distribution link function. Despite the considerable popularity of this class of models, there are several problems with their application described in the literature. Firstly, it may be questioned if modeling panel data containing the observations of the same driver in subsequent policy years does not violate one of the generalized linear models assumptions concerning the independence of variable distributions. In 1999 Frees, Yound and Luo wrote that insurance data, due to the presented problem, should be modeled by linear mixed models. A similar conclusion was presented in 2007 by Antonio and Beirlant who emphasized that it is not appropriate to use the generalized linear models (GLMs) in order to analyze the same observation in subsequent years – instead, the generalized linear mixed models (GLMMs) should be used. A similar conclusion, also in the insurance context, was presented in the paper of Antonio, Frees and Valdez (2010). The authors pointed out the possibility of using multidimensional GLMMs in the ratemaking. They proposed grouping the observations by random effects within the nested dimension of the vehicle owner (fleet), nested then within the insurance company.

Another questionable issue is the independence of variable distributions for observations that are represented by the same insurance agent, which was also presented by the author of this dissertation (Paska, 2018). This matter is explained primarily by the fact that the insurance agent, as an intermediary between the insurance company and the client, provides the information necessary to calculate the premium. The quality of the data depends significantly on insurance agent's inquisitiveness, thoroughness and truthfulness. Another important issue is also the customer-agent relationship, which may allow the agent to at least partially evaluate client's insurance risk. Depending on the character of the insurance agent relationship with particular insurance companies (e.g. form of sales contracts), agent can direct the specific type of client to different companies.

Moreover, another example is the publication of Wolny-Dominiak (2014) where the author pointed out that the possibility of an accident of two vehicles moving in the same region constitutes violation of the independence assumption because, inter alia, these vehicles are exposed to the same probability of bad weather conditions, which increases the probability of an accident. The risk portfolio then has a cluster structure, where the cluster is a set of policies from the same geographic region. In the author's opinion, the geographical dimension should therefore constitute be included in the random effects while performing the modeling process.

Another example of random effect, proposed innovative at doctoral dissertation, is the manager of insurance agent, nested with the agent. The rationale for choosing this dimension as a random effect is, firstly, the fact that a good manager can properly manage the sales network, minimizing the aforementioned risk of the agent's impact on loss ratio. Secondly, because managers of the agents operate locally, this dimension pursues similar goals to the geographical dimension proposed by Wolny-Dominiak (2014).

Another random effect proposed in the study is the car brand and car model of the insured vehicle. This is motivated by the fact, first, that car brands and car models are often linked to the personalities of the people who choose them (their driving style and, consequently, the number of accidents). Secondly, individual car brands and car models are stolen at different frequencies, which affects the risk of Autocasco (cf. Wolny-Dominiak, 2014). Thirdly, some car brands pay increased attention to the safety systems in vehicles that are supposed to minimize chance of causing an accident or reduce damage resulting due to an accident. At the same time, using this dimension as a fixed effect may be difficult due to the low size of each variable categories.

The last example of a random effect analyzed in the work is the customer dimension, described by Antonio and Valdez (2012). The authors noted that the unobservable characteristics of the driver or owner of the vehicle may affect the probability of incurring an accident in different years. Examples of such unobservable characteristics include guidelines on driving hours, mechanical check-ups and loading instructions.

Generalized linear mixed models that extend classical generalized linear models with non-zero variances random effects defined in a linear predictor, assuming directly unobservable heterogeneity within some regression coefficients are the solution for the abovementioned problems. Applications of GLMMs in the insurance context can be found in the papers of Garrido and Zhou (2009), Antonio and Valdez (2012) or Baumgartner, Gruber and Czado (2015). These models are also widely used in biological sciences (Bolker et al. 2009 as cited in Nakagawa and Schielzeth, 2013) and medical sciences (Gelman and Hill, 2007, Congdon, 2010, Snijders and Bosker, 2011 as cited in Nakagawa and Schielzeth, 2013).

Another problem that occurs in generalized linear models with the Poisson distribution, widely used in the actuarial science, is the overdispersion. This was noticed already in 1988 by Dionne and Vanasse. In 1989 McCullagh and Nelder even stated that due to its frequency the presence of overdispersion should be considered normal, while its absence would be therefore an exceptional situation. What is more, Frees (2010) stated that due to the concentration of the dependent variable observations in zero (i.e., no-claims observations), the variance must exceed the expected value. In 2015, David and Jemna demonstrated, using the example of the French insurance market, that the dispersion parameter in the claim frequency model is statistically significant, which indicates the correctness of using the negative binomial distribution. Similar conclusions were also presented in the papers of Antonio and Valdez (2012) and Shi and Valdez (2014).

Insurance data is also characterized by the excessive number of zeros in the dependent variable distribution that could be a problem in the claim frequency analysis. For this reason, widely described and used in the actuarial literature generalized linear models and their extensions (introducing only a random effect correction) may turn out to be insufficient. Yip and Yau (2005), by introducing zero-inflated models, emphasized that GLMs can be modified when there are too many zeros in the distribution of the dependent variable in order to avoid the effect of overdispersion. Antonio and Valdez (2012) proved that these models (proposed by Yip and Yau) are better than the "non-ZI" models.

Grize, Fischer and Luetzelschwab (2020) noted that attempts are being made to replace GLMs and GLMMs by another machine learning techniques, such as neural networks, random forests or extreme gradient boosting models. However, as Akinyemi and Leiser (2020) emphasized, these methods are still less often used due to lack of transparency and interpretability in relation to classic generalized linear models and their extensions in the form of generalized linear mixed models. Also Goldburd, Khare, Tevet and Guller (2020) pointed out that GLMs and their extensions not only deliver risk reflection results that are as satisfactory as showed by other advanced statistical models, but are also characterized by their high interpretability which in turn is often required by state insurers regulators. This conclusion is also confirmed in their paper of Henckaerts, Antonio, Clijsters and Verbelen (2018).

The purpose of the doctoral thesis is to compare different approaches used in the non-life insurance ratemaking to predict the claim frequency in Autocasco insurance. This issue is a key component in the insurance company activity – according to the information provided by the Polish Financial Supervision Authority (2021), the Autocasco contribute to approximately 26% of the technical result of non-life insurance companies. On the other hand, the technical result is created mainly on the applied premium tariff, and the adopted method of estimating the claim frequency may contribute to a more accurate assessment of the insurance premium and, consequently, increasing the company's profitability and protection against adverse selection. This thesis presents an overview of the actuarial literature on the methods used for this purpose.

3. Work structure

The work consists of five chapters: the first one presents a brief overview of the insurance history, with an emphasis on the pricing context, covering the times from ancient to modern and describing, among other things, the history of sea loans, co-insurance, reinsurance, marine insurance or insurance brokerage. Primarily, the history of motor insurance is described – the vehicles third party liability and legal obligatory issue of this insurance. Also, the history of motor accident insurance and Autocasco was presented. The focus of the next is on the history of international insurance agreements creating conditions for the free cross-border vehicle movement – from the Nordic Poole to the Green Card System. Further, the history of insurance pricing is widely described – from the work on calculating the probability of an event, through linear models, to the development of generalized linear models and their modifications. Then, the constraints of generalized linear models, especially in insurance context, were indicated (including the problem of overdispersion or too many zeros in the dependent variable). Attention was also paid to the panel nature of insurance data and to the dimensions grouping of the observations. Next, the literature on generalized linear mixed models was discussed to present methods for solving the above-mentioned problems. The consequences of not using generalized linear mixed models were summarized, as well as the advantages of these models over other machine learning techniques. This chapter also describes the motor claim frequency predictors found in the literature and presents a discussion on the legitimacy of using certain characteristics of the insurance subject owner in the pricing – such as, primarily, their gender. The current legal status concerning this issue was also indicated. Next, the issues of the key character of insurance ratemaking in the context of risk transfer, heterogeneity of the risk portfolio, adverse selection and financial consequences for the insurance company were discussed.

The second chapter provides general overview of the actuarial literature on the analysis of the claim frequency, with the focus on the problems arising in this analysis as well as the author's proposals for solving them. Additionally, the measures describing profitability, structure and size of the insurance company's portfolio, such as: written, earned and pure premium; the claim frequency, loss ratio, cost ratio and mixed ratio were described. The concept of technical result and the IBNR and IBNER claims components were also defined. These measures were used in the study to describe the financial consequences of introducing modifications to ratemaking models proposed by the author of this study. This chapter also describes the probability distributions used in frequency analysis – Poisson and negative binomial. Probability functions of distributions, as well as their expected values, variances, features and exemplary shaping depending on arbitrarily assumed parameter values were presented. In the further part, the generalized linear models were analyzed – their assumptions, constraints, problems, estimation methods, as well as their extensions, such as zero-inflated models, hurdle models and generalized linear mixed models (hierarchical, single and multi-dimensional). The concept of a fixed and random effect was introduced with actuarial examples of such effects. Moreover, the empirical study assessed significance of the random effects proposed by the author: the insurance agent, the agent's manager and the brand-model of the insured vehicle. The use of two nested random effects – both the insurance agent and its manager – should also be considered as added value in the context of actuarial applications. In particular, the author analyzed qualitatively and empirically the zero-inflated generalized linear mixed model with negative binomial distribution.

The third chapter presents the characteristics of the data used in the empirical study. This data describes 99.7 thousand annual motor policies – the information concerns both the subject of insurance (i.e., vehicle features) and the characteristics of the person who owns it as well as loss events (including the claim count and claim payments). Use of data from one of Polish insurance companies made it possible to demonstrate the advantage of the proposed methods on the example of the Polish insurance market. This chapter presents the histograms and descriptive statistics of the variables. In particular, the author emphasized the specificity of the dataset and the consequences resulting from the fact that the policies were purchased from the car dealers – firstly, high Casco saturation, secondly, lower car age compared to the population average. Then, the research methods used in the CASCO claim frequency analysis in chapter four were presented – the models analyzed in the study, determinants of the AC claim frequency, proposed random effects and measures used to select the optimal pricing model. The sub-portfolios tested in the empirical study were also identified and separated in order to indicate optimal areas for the use of new techniques and methods. The methods and the underlying assumptions of the financial simulation used in the work, aimed at illustrating the financial effects of implementing the proposed by the author solutions, were also described.

Chapter four presents and compares in the Autocasco context the results of parameters estimation and information criteria for:

1) generalized linear models (GLMs)
    a) with Poisson and negative binomial distribution,
    b) with zero-inflated Poisson and zero-inflated negative binomial distribution,
2) generalized linear mixed models (GLMMs) with insurance agent random effect
    c) with Poisson and negative binomial distribution,
    d) with zero-inflated Poisson and zero-inflated negative binomial distribution,

Based on the AIC and BIC information criteria, the zero-inflated generalized linear mixed model with the negative binomial distribution (ZINB GLMM AGENT) proposed by the author was recognized as the best. In the next step, it was compared with the models representing the same model class, but having different random effects (i.e., time and car brand-model).

The aim was to choose the optimal one, which turned out to be the insurance agent dimension (proposed by the author). The univariate random-effect model was then compared to the nested GLMM with both the insurance agent and agent's manager random effect. The results of the financial simulations were also presented, proving among others the significance of applying the model proposed by the author in the property pricing. Then, sensitivity analysis of financial simulations was carried out in order to determine how fixed and variable costs and the relation of competitor's premium changes affect the competitiveness of the insurance company's offer and its financial results. For the purpose of further model verification, the results of matching the curves of the optimal frequency model were compared with the empirical values as well as MSE and RMSE on the validation dataset were checked. In the next step, the application of the optimal loss frequency model on 12 sub-portfolios was analyzed. The results were compared using the R^2 determination coefficient given by Nakagawa and Schielzeth (2013) for mixed models:

- a sub-portfolio of new business risks,
- a sub-portfolio of renewal risks,
- a sub-portfolio of vehicles older than 4 years,
- a sub-portfolio of vehicles not older than 4 years,
- a sub-portfolio of vehicles registered in urban areas,
- a sub-portfolio of vehicles registered in rural and urban-rural areas,
- a sub-portfolio of vehicles registered in voivodeship capital cities,
- a sub-portfolio of vehicles registered in places other than voivodship cities,
- a sub-portfolio of customers who had, in the 4 last years, a claim in the Casco risk,
- a sub-portfolio of claims-free clients during the last 4 years,
- a sub-portfolio of company and leasing vehicles,
- a sub-portfolio of both non-company and non-leased vehicles.

It has been shown that thanks to the use of the ZINB GLMM AGENT model for the portfolio of vehicles registered in the urban areas, the insurance company can further increase profits from the implementation of the solutions proposed at this doctoral thesis.

Chapter five summarizes the work: research hypotheses are verified, the conclusions of the research and further development directions are presented.

4. Purpose

Finding the optimal class of frequency models is important for the insurance company because the claim frequency models are key components in premium estimation. The aim of the study was to propose a method of estimating the claim frequency models that solves the problems of insurance data analysis encountered in the literature, which in turn allows for a better adjustment of model parameters to empirical data and ultimately enables insurance company to maximize the profit from its basic activity.

The following research hypotheses were verified in the study:

1) The Casco claim frequency prediction errors are lower for generalized linear mixed models (GLMMs) than for currently used in the insurance ratemaking generalized linear models (GLMs).
2) Due to the characteristics of insurance data, i.e., large share of claims-free observations and overdispersion, the ZINB GLMM (zero-inflated negative binomial generalized linear mixed model) introduced by the author to the insurance tariff is better suited to empirical claims data than other GLMMs.
3) The implementation of models with insurance agent random effect and, separately, car brand-model random effect in the insurance pricing, allows better adjustment of the model parameters to the empirical claim data in comparison with the models with the widely described in the literature time random effect.
4) The GLMM with a nested random effect achieves lower prediction error compared to the single-level model

First hypothesis has been confirmed – it has been shown that the values of the AIC and BIC information criteria are lower for the generalized linear mixed models compared to the generalized linear models.

Second hypothesis was also confirmed by the values of the AIC and BIC information criteria – they were the lowest for the ZINB GLMM model. It was also verified that the ZINB GLMM AGENT model is not overfitted and that it fits well with the empirical data. This model achieves the highest value of the coefficient of determination R^2, describing the variance of the dependent variable to the greatest extent. It has also been proven that its use translates into higher financial results compared to the use of other pricing models.

Third hypothesis was also confirmed based on the AIC and BIC information criteria which are in accordance with the method adopted in the literature. The model with the insurance agent random effect reached the lowest values of the information criteria. The second was the model with the car brand-model random effect. The highest AIC and BIC values were achieved in the empirical study by the model with the time random effect (widely described in the actuarial literature).

Fourth hypothesis has not been confirmed – the single-level generalized linear mixed model fits the data better for the Autocasco claim frequency analysis than the multi-level model with the nested random effect of the insurance agent and the insurance agent's manager.

Summing up, the solutions proposed by the author of this dissertation constitute a vital component of the development of the actuarial literature. The aim of this doctoral thesis, which was to propose a method of estimating the claim frequency technique which solves the problems of analyzing insurance data, was achieved. Therefore, it is now possible to estimate a more adequate net premium, which allows the insurance company to gain market advantage by proposing a premium that maximizes profits and sales volumes.

Self-report bibliography

Akinyemi, K., Leiser, B. (2020). The Use of Advanced Predictive Analytics for Rate Making in Insurance, *Actuarial technology today*, p. 1-4.

Antonio, K., Beirlant, J. (2007). Actuarial statistics with generalized linear mixed models, *Insurance Mathematics and Economics*, vol. 40, p. 58-76.

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

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

Baumgartner, C., Gruber, L. F., Czado, C. (2015). Bayesian total loss estimation using shared random effects, *Insurance: Mathematics and Economics*, vol. 62, p. 194-201.

David, M., Jemna, D. (2015). Modeling the frequency of auto insurance claims by means of Poisson and negative binomial models, *Annals of the Alexandru Ioan Cuza University-Economics*, vol. 62(2), p. 151-168.

Dionne, G., Vanesse C. (1988). A generalization of automobile insurance rating models: the negative binomial distribution with a regression component, *Astin Bylletin*, vol. 19(2), p. 199-212.

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

Frees, E. W., Yound V. R., Luo Y. (1999). A longitudinal data analysis interpretation of credibility models, *Insurance: Mathematics and Economics*, vol. 24, p. 229-247.

Garrido, J., Zhou, J. (2009). Full credibility with generalized linear and mixed models, *Astin Bulletin*, vol. 39(1), p. 61-80.

Goldburd, M., Khare, A., Tevet, D., Guller, D. (2020). *Generalized linear models for insurance rating*. Virginia: Casualty Actuarial Society.

Grize, Y., Fischer, W., Lueztelschwab, C. (2020). Machine learning applications in nonlife insurance, *Appl Stochastic Models Bus Ind.*, vol. 36, p. 523-537.

Henckaerts, R., Antonio, K., Clijsters, M., Verbelen, R. (2018). A data driven binning strategy for the construction of insurance tariff classes, *Scandinavian Actuarial Journal*, vol. 2018(8), p. 681-705.

Husnjak, S., Peraković, D., Forenbacher, I., Mumdziev, M. (2015). Telematics System in Usage Based Motor Insurance, *Procedia Engineering*, vol. 100, p. 816-825.

Komisja Nadzoru Finansowego. (2020). *Raport o sektora ubezpieczeń po III kwartałach 2019 roku*. Warszawa: Urząd Komisji Nadzoru Finansowego.

Komisja Nadzoru Finansowego. (2021). *Raport o sektora ubezpieczeń po III kwartałach 2020 roku*. Warszawa: Urząd Komisji Nadzoru Finansowego.

McCullagh, P., Nelder, J. A. (1989). *Generalized Linear Models.* 2nd ed. Londyn: Chapman and Hall.

Nakagawa, S., Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models, *Methods in ecology and evolution*, vol. 4(2), p. 133-142.

Paska, N. (2018). Zastosowanie modeli ZINB GLMM z efektem losowym agenta w taryfikacji ubezpieczeń majątkowych, *Roczniki Kolegium Analiz Ekonomicznych SGH*, vol. 53, p. 63-76.

Shi, P., Valdez, E. A. (2014). Multivariate negative binomial models for insurance claim counts, *Insurance: Mathematics and Economics*, vol. 55, p. 18-29.

Sulewski, P., Meuwissen, M. (2014). Fundusze ubezpieczeń wzajemnych jako forma ograniczania ryzyka w rolnictwie, *Zagadnienia ekonomiki rolnej*, vol.339(2)., p. 127-143.

Wolny-Dominiak, A. (2014). *Taryfikacja w ubezpieczeniach majątkowych z wykorzystaniem modeli mieszanych*. Katowice: Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach.

Yip, K C. H., Yau, K. K. W. (2005). On modeling claim frequency data in general insurance with extra zeros, *Insurance Mathematics and Economics*, vol. 36, p. 153-163.

Norbert Paska

18.08.2021