The Poisson regression with fixed and random effects in non-life insurance ratemaking

Alicja Wolny-Dominiak[1], Daniel Sobiecki[2]

Abstract

Important part of data analysis in insurance business is the construction of a fair tariff structure called classification ratemaking. The goal of this classification is partition all policies in particular portfolio into homogeneous classes. Within every class, all policyholders pay the same pure premium. To design classification rating plans, actuaries use the generalized linear models (GLM) technique. In GLM model, the dependent variable is usually the claim severity or the claim frequency for ith policy. In the paper we focus on the claim frequency. The rating variables are the categorical variables with few categories like e.g. gender or a large number of categories like e.g. spatial variables. The GLM model assumes observed responses are independent. However many portfolios of policies yield correlated observations. Correlation results from the sampling design or the way data are collected. The aim of this paper is to propose mixed models based on Poisson regression which are useful in claim frequency modelling. In this models we take into account specificities of insurance data: correlation, overdispersion and zero-inflation effects in data. The case study demonstrate the validity of the application of these models.

Keywords: *claim frequency, mixed model, Poisson regression, ZIP, ZIGP*

JEL Classification: C21, C51

1. Introduction

Modelling claim frequency in the portfolio of policies is an essential part of non-life insurance ratemaking in the portfolios of policies. The ratemaking is defined as risk classification, which involves the grouping policies into various classes that share a homogenous set of characteristics influences claim frequency. In every class the same net premium, calculated as the expected claim frequency in this case, is than reasonable. The ratemaking is usually done in two steps (Denuit, Maréchal et al. 2007; Boucher and Guillén, 2009). In the first step, called a priori ratemaking, policies in the portfolio are classified according measurable information about the policyholder and insured object (Antonio and Beirlant, 2006; de Jong and Heller, 2008; Wolny-Dominiak and Studnik, 2013). After a priori classification, the portfolio is divided into homogenous groups, but only of the observable factors. Some important hidden characteristics still generate heterogeneity in every group of policies (e.g. in

[1] University of Economic in Katowice, Department of Statistical and Mathematical Methods in Economics, 1 Maja 50, 40-287 Katowice, Poland; alicja.wolny-dominiak@ue.katowice.pl

[2] Warsaw School of Economics, Collegium of Economic Analysis, Madalińskiego 6/8, 02-513 Warsaw; sobieckid@gmail.com

automobile insurance – the behavioral characteristics of driver). That is why the history of claims as it emerge for individual policy is taken into account in second step of ratemaking, called a posteriori ratemaking (Antonio and Valdez, 2012). In a priori ratemaking cross-sectional data are used while in a posteriori ratemaking rather longitudinal structure.

As the claim frequency is an example of count data there are few problems in modelling of such a data. Literature review reveals that, in particular, attempts are undertaken to find a probabilistic model for the claim frequency distribution, where usually this distribution is assumed to be Poisson. However the insurance portfolios have a very specific characteristic, i.e. for many policies there are no claims observed in the insurance history for a given period of time. It means that the data contains lots of zeros and, as a consequence, the Poisson regression may not give satisfactory results (zero-inflation effect). In order to allow the presence of excess zeros in insurance portfolio, the zero-inflated models are applied (Lambert, 1992; Yip and Yau, 2005; Wolny-Dominiak, 2013). The classic model is the zero – inflated Poisson model (ZIP), which is a mixture of a Poisson distribution and a zero point mass. The other problem often existing in insurance data is the incidence of overdispersion, which means that data exhibit greater variability than allowed to the Poisson model and the mean is not equal to variance (overdispersion effect). The reason of that may be the disregarding some latent factors affecting the claims occurrence. The generalization of the Poisson model is possible and than the generalized Poisson model (GP) is received (Consul and Famoye, 1992). The generalized Poisson distribution usually is used when the occurrence of claims is probably dependent, which is a common situation in non-life insurance (Yip and Yau, 2005). Usually in case of overdispersion in ZIP model, zero-inflated negative binomial (ZINB) model is used (Hall, 2000), but the zero-inflated generalized Poisson (ZIGP) is also possible. In literature there are some simulation studies with the score test for overdispersion based on ZIGP model, which illustrate that ZIGP model has higher empirical power than ZINB model (Yang et al., 2009).

Within the context of a priori ratemaking, it is becoming a standard norm in practice to use Generalized Linear Models (GLMs) where cross-sectional data is modelled within the class of exponential dispersion distributions (Haberman and Renshaw, 1996; Ohlsson, 2008). The GLM model assumes observed responses are independent. However many portfolios of policies yield correlated observations. Correlation results from the sampling design or the way data are collected. The correlated data typically has a clustered structure, e.g. the automobile portfolio with the spatial variable like geographical regions. In this case the source of correlation could be explain in following way: each region is likely to experience roughly the same weather conditions and hence different policies in the same region are likely to have a similar claims experience. Similarly for a posteriori ratemaking, as in longitudinal studies, where the observations represent repeated outcomes from individual subjects, the correlation response at one time is correlated with the response at another time. The longitudinal data is a special case of clustered data in which the cluster is the policyholder. Ignoring the correlation can lead to erroneous conclusions (de Jong and Heller, 2008). That is why we propose appropriate mixed models based on Poisson mixed regression models in ratemaking.

In the paper we propose two models based on the mixed Poisson regression model, which can be used in a priori ratemaking as well as a posteriori. In the case study we consider the real portfolio of automobile policies taken from polish insurance company. As the portfolio contains the cross-sectional data structure we analyze zero-inflation and overdispersion effects in the portfolio given the spatial variable as random effect. The estimation is done with marginal likelihood method (MLM) using NLMIXED procedure from SAS (the integration method is Adaptive Gaussian Quadrature) (see Littell, 2006).

2. The mixed models based on Poisson regression

Assume the portfolio of n policies. Let N denotes the claim frequency for individual policy. Suppose *j*th response variable from *i*th cluster follows a Poisson distribution:

$$\Pr_{POIS}[N_{ij} = n_{ij}] = \frac{e^{-\lambda_{ij}} \lambda_{ij}^{n_{ij}}}{n_{ij}!}, \tag{1}$$

$i = 1,...,p$ and $j = 1,...,n_i$, where p is the number of clusters and n_i is the number of observations within cluster *i*. In case of the overdispersion effect, the probability density function (1) should be extended to the generalized Poisson distribution given by:

$$\Pr_{GP}[N_{ij} = n_{ij}] = (\frac{\lambda_{ij}}{1+\alpha_{ij}\lambda_{ij}})^{n_{ij}} \frac{(1+\alpha_{ij}n_{ij})^{n_{ij}-1}}{n_{ij}!} \exp[\frac{-\lambda_{ij}(1+\alpha_{ij}n_{ij})}{1+\alpha_{ij}\lambda_{ij}}] \tag{2}$$

with the mean and the variance respectively: $E(N_{ij}) = \lambda_{ij}$, $Var(N_{ij}) = \lambda_{ij}(1+\alpha_{ij}\lambda_{ij})^2$. If $\alpha_{ij} = 0$, the above model reduces to the Poisson distribution with no overdispersion effect. In the situation of excess zeros in the portfolio, the ZI-probability density function is recommended to use. The general form of $\Pr_{ZI}[N_{ij} = n_{ij}]$ can be express as follows:

$$\Pr_{ZI}[N_{ij} = n_{ij}] = \begin{cases} \varpi_{ij} + (1-\varpi)\Pr[N_{ij} = n_{ij}], & n_{ij} = 0 \\ (1-\varpi_{ij})\Pr[N_{ij} = n_{ij}], & n_{ij} > 0 \end{cases}, \tag{3}$$

where ϖ_{ij} is the probability of zero claim frequency for i th policy in j th cluster and $\Pr[N_{ij} = n_{ij}]$ is the probability density function (1) or (2). The zeros from first equation are called "structural" zeros and from second equation – "sampling" zeros. First two moments in zero - inflated models are:

$$E[N_{ij}] = (1-\varpi_{ij})\lambda_{ij},\ Var[N_{ij}] = (1-\varpi_{ij})(\lambda_{ij}{}^{2} + \lambda_{ij}) \tag{4}$$

for the ZIP distribution and

$$E[N_{ij}] = (1-\varpi_{ij})\lambda_{ij},\ Var[N_{ij}] = E[N_{ij}][(1+\alpha_{ij}\lambda_{ij})^{2} + \varpi_{ij}\lambda_{ij}] \tag{5}$$

for the ZIGP distribution.

In ratemaking we are interested in extended probability density functions (1) – (3) to models with explanatory variables. In the regression setting, the mean λ_{ij}, zero proportion ϖ_i and α_{ij} are related to the covariates vectors $\mathbf{x}_{ij}$, $\mathbf{z}_{ij}$ and $\mathbf{w}_{ij}$ respectively. Responses within the same cluster are likely to be correlated. To accommodate the inherent correlation, random effects u_i are incorporated in the linear predictors $\eta_{ij} = \mathbf{x}_{\mathbf{ij}}^{'}\boldsymbol{\beta}$. Given the vector $u = (u_1,...,u_p)^{'}$ of random effects, we propose following models:

1. The mixed Poisson regression with random intercept (denoting by *POIS-M*):

$$\begin{cases} N_{ij} \sim Pois(\lambda_{ij}) \\ \lambda_{ij}(\boldsymbol{\beta}, X_1,...,X_k) = e^{\mathbf{x}_{\mathbf{ij}}^{'}\boldsymbol{\beta}+u_i}, \\ u_i \sim N(0,\sigma_u^2) \end{cases} \tag{6}$$

2. The mixed Poisson regression when zip-inflation occurs (denoting by GP-M):

$$\begin{cases} N_{ij} \sim ZIP(\lambda_{ij},\varpi_{ij}) \\ \lambda_{ij}(\boldsymbol{\beta}, X_1,...,X_p) = e^{\mathbf{x}_{\mathbf{ij}}^{'}\boldsymbol{\beta}+u_i} \\ \varpi_{ij}(\boldsymbol{\gamma}, Z_1,...,Z_q) = \dfrac{e^{\mathbf{z}_{\mathbf{ij}}^{\mathbf{T}}\boldsymbol{\gamma}}}{1+e^{\mathbf{z}_{\mathbf{ij}}^{\mathbf{T}}\boldsymbol{\gamma}}} = \dfrac{1}{1+e^{-\mathbf{z}_{\mathbf{ij}}^{\mathbf{T}}\boldsymbol{\gamma}}}, \\ u_i \sim N(0,\sigma_u^2) \end{cases} \tag{7}$$

3. The mixed Poisson regression when overdispersion and zero-inflation occur (denoting by ZIGP-M):

$$\begin{cases} N_{ij} \sim ZIGP(\lambda_{ij}, \alpha_{ij}, \varpi_{ij}) \\ \lambda_{ij}(\boldsymbol{\beta}, X_1,\dots,X_k) = e^{\mathbf{x}'_{ij}\boldsymbol{\beta}+u_i} \\ \alpha_{ij}(\varphi, W_1,\dots,W_t) = 1+e^{\mathbf{w}'_{ij}\varphi} \\ \varpi_{ij}(\boldsymbol{\gamma}, Z_1,\dots,Z_q) = \dfrac{e^{\mathbf{z}'_i\boldsymbol{\gamma}}}{1+e^{\mathbf{z}'_i\boldsymbol{\gamma}}} = \dfrac{1}{1+e^{-\mathbf{z}'_i\boldsymbol{\gamma}}} \\ u_i \sim N(0,\sigma_u^2) \end{cases}, \tag{8}$$

The link function in ZIGP-M model for α_{ij} parameter are taken from (Czado et al, 2007). The random effects u_i are assumed to be independent and normally distributed, $u_i \sim N(0,\sigma_u^2)$.

Based on the generalized linear mixed model formulation (Breslow and Clayton, 1993; McCulloch, 2006), the marginal likelihood method (MLM) estimates of (6)-(8) mixed regression models parameters can be obtain. The general form of the marginal likelihood can be than express as follows:

$$l(\psi) = \sum_{i=1}^{m} \log L_i(\psi) = \sum_{i=1}^{m} \log \int \prod_{j=1}^{n_i} \Pr[N_{ij} = n_{ij} \mid u_i] f(u_i) du_i \,, \tag{9}$$

where ψ is the vector of parameters in the model and f denotes the normal density function for the random effects u_i. The function $L_i(\psi)$ is of the form (up to the model):

$$L_{(POIS-M)i}(\boldsymbol{\beta}, \sigma_u^2) = \int \prod_{j=1}^{n_i} \Pr_{POIS}[N_{ij} = n_{ij}] f(u_i) du_i \,, \tag{10}$$

$$\begin{aligned} L_{(ZIP-M)i}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma_u^2) = \int \prod_{j=1}^{n_i} [(\varpi_{ij} + (1-\varpi_{ij}) e^{-\lambda_{ij}}) I_{(n_{ij}=0)}] \times \\ \times [(1-\varpi_{ij}) \Pr_{ZIP}[N_{ij} = n_{ij}] I_{(n_{ij}>0)}] f(u_i) du_i \end{aligned}, \tag{11}$$

$$\begin{aligned} L_{(ZIGP)i}(\boldsymbol{\beta}, \varphi, \boldsymbol{\gamma}, \sigma_u^2) = \int \prod_{j=1}^{n_i} [(\varpi_{ij} + (1-\varpi_{ij}) \exp[\frac{-\lambda_{ij}}{1+\alpha_{ij}\lambda_{ij}}]) I_{(n_{ij}=0)}][(1-\varpi_{ij}) \times \\ \times \Pr_{GP}[N_{ij} = n_{ij}] I_{(n_{ij}>0)}] f(u_i) du_i \end{aligned}, \tag{12}$$

where $I_{Y>0} = \begin{cases} 0 & dla \;\; y \le 0 \\ 1 & dla \;\; y > 0 \end{cases}$.

There are few methods to estimate above mixed models (Wolfinger and O'connell, 1993; McCulloch, 1997; Littell, 2006). Up to the method there is a need to obtain the marginal likelihoods (10)-(12) by integrating out random effects u_i. As the analytical solution of integrals is intractable, one first apply numerical approximation: Laplace transformation or Gauss-Hermite quadrature. In claim frequency modeling we use marginal likelihood method (MLM) with NLMIXED procedure from SAS (the integration method is Adaptive Gaussian Quadrature).

3. The mixed Poisson regression – a priori ratemaking

We analyze cross-sectional sample of the automobile insurance portfolio of a company operating in Poland. Only private-use cars are considered in this sample. There are 4 categorical exogeneous variables as well as the claim frequency for every policy in the portfolio at fault that were reported within the yearly period: *CLIENT_AGE* (the policyholder's age, 6 categories), *CAR_AGE* (the car's age, 3 categories), *POWER* (the engine power, 3 categories), *VOIVODESHIP* (the region in Poland, 16 categories). We model the variable *CLAIM_COUNT* (claim frequency). The exogenous information is coded by means of binary variables. We consider the portfolio with 51 557 policies. As the fraction of zeros is 95.81 % in the portfolio (policies with no claims) we suspect that zero-inflation and overdispersion effects appear in data. Consequently we investigate ZIP-M and ZIGP-M models with random effect assumed to be the spatial variable *VOIVODESHIP*.

In order to analyze the validity of the application of POIS-M, ZIP-M and ZIGP-M we estimate parameters of models for the whole portfolio. This allows us to test statistical significance of parameters ϖ, α, σ_u^2. The results are shown in Tab. 1 and Tab. 2.

We observe that the variance component σ_u^2 is statistically significant in all three models as well as other parameters: ϖ in ZIP-M model and ϖ, α in ZIGP-M model which that using mixed models take into consideration zero-inflation and overdispersion effects is reasonable. In our portfolio the model ZIP-M against ZIGP-M is preferable according AIC.

Parametr	**Estimate (s.e.)**	**p-value**
	POIS-M	
β_0	-3.0134 (0.05046)	<.0001
σ_u^2	0.1804 (0.0394)	0.0004
-2log-likehood	20 187	-
AIC	20 191	-

Table 1 Parameter estimates for POIS-M.

Parametr	Estimate (s.e.) ZIGP-M	p-value	Estimate (s.e.) ZIP-M	p-value
ϖ	0.9747 (0.0005)	<.0001	0.6681 (0.0237)	<.0001
α	-0.1624 (0.0019)	<.0001	-	-
β_0	-0.3545 (0.0249)	<.0001	-1.9095 (0.0872)	<.0001
σ_u^2	0.0548 (0.0255)	0.0487	0.1795 (0.0399)	0.0004
-2log-likehood	26 170	-	20 024	-
AIC	26 178	-	20 030	-

Table 2 Parameter estimates for ZIGP-M and ZIP-M models with no covariates.

Conclusions

Ratemaking is an extremely important part of establishing reasonable classification for a portfolio of insurance policies. In the literature there is a lot of regression models proposed to be used in this problem. We focused on the modified mixed Poisson models with spatial random effect which handle with zero-inflation and overdispersion effects. In the case study we investigated the validity of this approach by testing the statistical significance of parameters σ_u^2, ϖ and α. This is preliminary analysis which is useful in the final selection of the model in ratemaking of particular portfolio of policies.

Acknowledgements

The work of Alicja Wolny-Dominiak was supported by grant NN 111461540 from Polish National Science Center.

References

Antonio, K., & Valdez, E. A. (2012). Statistical concepts of a priori and a posteriori risk classification in insurance. *AStA Advances in Statistical Analysis*, *96*(2), 187-224.

Antonio, K., & Beirlant, J. (2006). *Risk Classification in Nonlife Insurance.* Encyclopedia of Quantitative Risk Analysis and Assessment.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*(421), 9-25.

Boucher, J. P., & Guillén, M. (2009). A survey on models for panel count data with applications to insurance. *RACSAM-Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas, 103*(2), 277-294.

Consul, P. C., & Famoye, F. (1992). Generalized Poisson regression model. *Communications in Statistics-Theory and Methods, 21(1),* 89-109.

Czado, C., Erhardt, V., Min, A., & Wagner, S. (2007). Zero-inflated generalized Poisson models with regression effects on the mean, dispersion and zero-inflation level applied to patent outsourcing rates. *Statistical Modelling*, *7*(2), 125-153.

De Jong, P., & Heller, G. Z. (2008). *Generalized linear models for insurance data*. Cambridge University Press.

Denuit, M., Maréchal, X., Pitrebois, S., & Walhin, J. F. (2007). *Actuarial modelling of claim counts: Risk classification, credibility and bonus-malus systems*. John Wiley & Sons.

Haberman, S., & Renshaw, A. E. (1996). Generalized linear models and actuarial science. *Statistician*, *45*(4), 407-436.

Hall, D. B. (2000). Zero ‐ inflated Poisson and binomial regression with random effects: a case study. *Biometrics*, *56*(4), 1030-1039.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, *34*(1), 1-14.

Littell, R. C. (2006). *SAS for mixed models*. SAS institute.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American statistical Association*, *92*(437), 162-170.

McCulloch, C. E. (2006). *Generalized linear mixed models*. John Wiley & Sons, Ltd.

Ohlsson, E. (2008). Combining generalized linear models and credibility models in practice. *Scandinavian Actuarial Journal*, *2008*(4), 301-314.

Wolfinger, R., & O'connell, M. (1993). Generalized linear mixed models a pseudo-likelihood approach. *Journal of statistical Computation and Simulation*, *48*(3-4), 233-243.

Wolny-Dominiak, A. (2013). Zero-inflated claim count modeling and testing–a case study. *Ekonometria*, *1*(39), 144-151.

Wolny-Dominiak, A., & Studnik, J. (2013). *Estimation of claim counts quantiles*. In Papież, M. & Śmiech, S. (Eds.), *Proceedings of 7th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena*. Cracow University of Economics, Poland, 198-204.

Yang, Z., Hardin, J. W., & Addy, C. L. (2009). Testing overdispersion in the zero-inflated Poisson model. *Journal of Statistical Planning and Inference*, *139*(9), 3340-3353.

Yip, K. C., & Yau, K. K. (2005). On modeling claim frequency data in general insurance with extra zeros. *Insurance: Mathematics and Economics*, *36*(2), 153-163.