Estimation of claim counts quantiles

Alicja Wolny-Dominiak[1], Joanna Studnik[2]

Abstract

In this paper we analyze claim counts models applicable in many areas of non-life insurance practice, such as ratemaking, claims reserving or bonus-malus systems. The typical situation in portfolios of policies is the zero-inflated and the overdispersion effects occurrence. That is why, in claim counts modelling the zero-inflated Poisson regression (ZIP) is usually used. ZIP model gives the possibility to analyze the influence regressors on the location of the conditional distribution of claim counts. The additional information how the risk factors affects the entire shape of distribution gives the estimation of quantiles of the claim counts distribution. To estimate quantiles, we apply the quantile regression technique and the asymmetric maximum likelihood. The goal of this paper is to present the possibility of implementation those models in insurance practice.

Keywords*: claim counts, ZIP, quantile regression, asymmetric maximum likelihood*

JEL Classification: C51, C14, C21,

AMS Classification: 62M20, 62G05

1. Introduction

In many areas of non-life insurance practice, such as ratemaking, claims reserving or bonus-malus systems, the important problem is to model the claim frequency distribution, where a regression component is included to take into account the individual characteristics of policyholders. A very common method chosen for modelling a claim frequency data is a GLM Poisson regression model, see [2]. However, in non-life insurance portfolios, the typical situation is the zero-inflated and the overdispersion effects occurrence. The reason of that, from one hand may be the disregarding some latent factors affecting the claims occurrence and from the other hand no report of small claims, what is not cost-effective because of the bonus-malus system. In this case GLM Poisson regression gives unsatisfied results and typically ZI-models are used. In actuarial literature there are few comparative studies of such models [14], where zero-inflated Poisson (ZIP), zero-inflated negative-binomial (ZINB), zero-inflated generalized Poisson (ZIGP), zero-inflated double Poisson (ZIDP), hurdle and heterogeneous models are analyzed. Well known ZIP model proposed by Lambert is a mixture of a Poisson distribution and a zero point mass [6]. In case of the

[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] Katowice Institute of Information Technologies, Mickiewicza 29, 40-085 Katowice, Poland, joanna.studnik@wsti.pl

overdispersion effect, the problem can by solved by assuming the negative-binomial distribution for count claims and the ZINB model is received. More efficient is to model count claims with generalized Poisson distribution, specially when the occurrence of claims is probably dependent, see in e.g. [4], [9].

All models mentioned above gives the possibility to analyze the influence regressors on the location of the conditional distribution of claim counts. The investigation how the risk factor affects the entire shape of distribution gives the additional information. These are all useful tools that take into account the asymmetry of claim counts like quantile regression technique (QR) or asymmetric maximum likelihood (AML). In this paper we analyze those two distribution-free approaches QR and AML, which can be treated as additional tool in claim counts modelling except ZI-models. The case study based on the real-world insurance portfolio examines the four-steps process modelling: (i) testing zero-inflation and overdispersion effects (ii) risk factors selection (iii) ZIP claim counts estimation (iv) quantiles of claim counts estimation.

The reminder of this paper is organized into three sections. In Section 2 a brief description of ZIP, AML and QR models for insurance data are presented as well as the method of model's parameters estimation. Section 3 contains the case study based on the motorcycle insurance dataset taken from Ohlsson and Johansson [10]. For all calculation, the software R CRAN is applied. In order to execute the ML estimation, few packages are used: {pscl} package for ZIP/ZINB models, {quantreg} package for the QR model and {VGAM} package for AML.

2. Claim counts modelling

This section discusses two approaches in claim counts modelling. One is the generalized linear regression with the assumption of zero-inflated Poisson distribution (ZIP) and the other one is the distribution-free method of fitting regressions for the conditional percentiles of the response variable as the function of risk factors. Let consider the random variable Y_i, $i = 1, \ldots, n$ denoting the number of claims with independent realizations in the portfolio of policies and $X_1, ..., X_m$ denoting categorical variables interpreted as risk factors influenced claim counts. In ratemaking analysis usually the multiplicative relationship is used [10], so we assumed the link function between claim counts and risk factors as logarithm:

$$\log(\mu_i) = \mathbf{x}_i{}'\boldsymbol{\beta} \tag{1}$$

where $\mathbf{X}$ is a $(m+1)$-dimentional design matrix for risk factors and $\boldsymbol{\beta} = (\beta_1,...,\beta_m)$ is a vector of regression coefficients influence claim counts.

The first approach in claim counts modelling is well known ZIP model discussed e.g. in [12], [14]. In ZIP model the independent variables Y_i take zero values with the probability ϖ_i or values from Poisson distribution $Y_i \sim Pois(\lambda_i)$ with probability $1-\varpi_i$. Risk factors $X_1,...,X_m$ simultaneously affect the number of claims and fractions of no claim policies ϖ_i. The link between claim counts and parameters ϖ_i is assumed as logit:

$$\ln(\frac{\varpi_i}{1-\varpi_i}) = \mathbf{x}_i{}'\,\boldsymbol{\gamma}, \tag{2}$$

where $\boldsymbol{\gamma} = (\beta_1,...,\beta_m)$ is a vector of regression coefficients influence ϖ_i.

The second approach we consider is claim counts modelling in which there is no assumption about the distribution of Y and the asymmetry of data is taken into account. One method of such modelling was proposed by Efron [3], who introduced the variant of the maximum likelihood estimation called asymmetric maximum likelihood (AML), specially useful in generalized linear models with overdispersion effect. The other one, proposed by Machado and Silva [7] is to impose some data smoothness to transform discrete variables in continuous variables and apply the quantile regression.

Asymmetric maximum likelihood AML is the extension of asymmetric least squares regression. The idea of AML is based on minimize the asymmetric version of the deviance between any two members of the exponential family, depending on a positive constant $w > 0$, see in [3]:

$$D_w(\mu_1,\mu_2) = \begin{cases} D(\mu_1,\mu_2), & \mu_1 \le \mu_2 \\ wD(\mu_1,\mu_2), & \mu_1 > \mu_2 \end{cases}. \tag{3}$$

To find the AML estimator of vector $\boldsymbol{\beta}$, the iterative method is applied to minimizing the expression:

$$\hat{\boldsymbol{\beta}}_w = \min_{\boldsymbol{\beta}} \sum_{i=1}^{n} [y_i \mathbf{x}_i{}'\boldsymbol{\beta} - e^{\mathbf{x}_i{}'\boldsymbol{\beta}} - \ln(y_i)] w^{\mathrm{I}_{(y_i > e^{\mathbf{x}_i{}'\boldsymbol{\beta}})}}\ , \tag{4}$$

where the function $\mathrm{I}_{(condition)}$ takes the value 1 if the condition is true and 0 otherwise. For $w = 1$ we receive the usual maximum likelihood estimate of $\boldsymbol{\beta}$.

In Machado and Silva concept [7], the quantil regression (QR) is used to estimate claim counts. They applied the jittering technique introduced by Stevens (1950), in which the discrete data are modified by adding a noise U generated from a continuous distribution with support on [0,1]. According to Theorem 2 in Machado and Silva paper [7], there is one-to-one relationship between the conditional quantiles of variables Y and $Z = Y + U$. After smoothing, the quantile regression QR is able to apply. QR of order α, $0 < \alpha < 1$, for claim counts is given by the formula:

$$Q_{T(Z;\alpha)}(\alpha|\mathbf{x}_i) = \mathbf{x}_i'\boldsymbol{\beta}, \tag{5}$$

where $Q_{T(Z;\alpha)}(\alpha|\mathbf{x}_i)$ indicates conditional quantile of Y_i for probability α and $T(Z;\alpha)$ indicate some monotonic transformation of Z. The assumption about the differentiable sample objective function is now fulfill and the estimator $\hat{\boldsymbol{\beta}}_\alpha$ is efficient – see Theorem 4. [7]. In QR for claim counts we assume the T-transformation $T(\mathbf{X}\boldsymbol{\beta}) = \ln(\mathrm{Z})$, where $\mathbf{Z}$ is a vector of claim counts after the jittering process and the noise U is drawn from a uniform distribution. In order to obtain more efficient estimator than $\hat{\boldsymbol{\beta}}_\alpha$, we perform the k- iterative procedure for $j = 1,...,k$ in following steps:

(i) drawing the random sample u_{ij}, $i = 1,...,n$ of the noise $U \sim Uniform$

(ii) QR-estimation of $\hat{\boldsymbol{\beta}}_{\alpha j}$ for the variable $Z = Y + U_j$

(iii) repeating (i)-(ii) m - times

(iv) calculation the estimator $\hat{\boldsymbol{\beta}}_\alpha = \frac{\sum_{j=1}^{k} \hat{\boldsymbol{\beta}}_{\alpha j}}{k}$

The 100αth quantile depends on parameter w as follows:

$$\alpha = \frac{1}{n}\sum_{i=1}^{n} I_{Y_i \le \exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}}_w)} .$$

The drawbacks of AML regression is that 100αth quantiles cannot be computed for α smaller than proportion of zero’s in the sample, in our application: proportion of no claims policies in portfolio.

3. The case study - motorcycle insurance dataset

To present the claim counts modelling process with the zero – inflated and overdispersion occurrence, the case study based on the real-world dataset was analyzed. The dataset contains aggregate information about policies and claims from former Swedish insurance company WASA [10]. We analyzed 4 risk factors for every policy (the policyholder gender is omitted): driver’s age (from A (youngest) to G), the geographic zone (from A to G), MC class (from A to G) and vehicle’s age (from A (youngest) to C). The risk factor MC class is classified by the EV ratio, where $EV = \frac{\text{engine capacity in kW x 100}}{\text{vehicle weight in kg} + 75}$ (75 kg is the average weight of a driver). Every risk factor can have an influence on the number of claims as well as on the occurrence of the zero – inflation and the overdispersion effect. There are 97,67% of no claim policies.

In the first step of process modelling the zero-inflation effect was tested using van der Broek score test as in [11]. Under the null hypothesis $H_0 : \varpi = 0$ and the assumption of Poisson distribution of claim counts, the score statistics is define as follows:

$$S(\hat{\beta}) = \frac{(\sum_{\substack{i=1 \\ y_i=0}}^{n} \frac{1}{e^{-\hat{\lambda}_i}} - 1)^2}{(\sum_{i=1}^{n} \frac{1}{e^{-\hat{\lambda}_i}} - 1) - n\bar{y}}, \tag{6}$$

where $\bar{y}$ is the average of the number of claims. The statistics $S(\hat{\beta})$ follows an asymptotic χ^2 distribution with 1 degree of freedom. In analyzed dataset, the score statistics takes the value $S(\hat{\beta}) = 31{,}79$ (p-value less than 0.0001), which means that null hypothesis should be reject and the zero-inflation effect occurs. It can therefore be concluded that in the portfolios there are mostly insurance policies without an accident. In ratemaking process this fact should be taking into consideration.

In second step of claim counts modelling, risk factors were selected under the assumption of zero-inflated Poisson distribution of claim counts. We applied following procedure in the selection process, see [8]:

(i) estimating ZIP-model for every combination of risk factors,

(ii) selection of this subset of risk factors which gives the minimum value of the Akaike information criterion (AIC).

In analyzing insurance portfolio the lower AIC.min = 7369.06 reaches the subset {*veh.age*, *area*}, where *veh.age* is a variable in count model and *area* is the variable in zero-inflated model.

In next step of modelling we applied free-distribution regressions: AML and QR for claim counts with {*veh.age*, *area*} regressors. We restricted the estimation to 3rd quartile in QR and the parameter $w = 0.875$ in AML regression. Table 1 presents estimates of the partial effects of the risk factors in three investigated models. All factors are statistically significant (p-value less than 0.0001), except *area E* and *area G* in QR model.

	$\hat{\boldsymbol{\beta}}$	**Count Effect**	**Std. Error**	$\hat{\boldsymbol{\beta}}_w$	**Count Effect**	**Std. Error**	$\hat{\boldsymbol{\beta}}_{\alpha=0.75}$	**Count Effect**	**Std. Error**
veh.age A	0.00	1.00	-	0.00	1.00	-	0.00	1.00	-
veh.age B	-0.44	0.64	0.13	-0.44	0.64	0.13	-0.27	0.76	0.03
veh.age C	-1.05	0.35	0.11	-1.05	0.35	0.11	-0.33	0.72	0.02
area A	0.00	1.00	-	0.00	1.00	-	0.00	1.00	-
area B	-0.39	0.68	0.11	-0.39	0.67	0.11	-0.15	0.86	0.02
area C	-0.76	0.47	0.12	-0.77	0.46	0.13	-0.20	0.82	0.02
area D	-0.94	0.39	0.11	-0.94	0.39	0.11	-0.38	0.68	0.02
area E	-1.68	0.19	0.35	-1.69	0.18	0.37	-0.04	0.96	0.03
area F	-1.44	0.24	0.25	-1.45	0.23	0.26	-0.17	0.84	0.03
area G	-2.05	0.13	1.01	-2.05	0.13	1.07	-0.04	0.96	0.09

Table 1 Estimation results – ZIP, AML, QR.

The base variables are (*veh.age.*A, *area A*) with partial effects equal to one. The other partial effects show how each parameter impacts the number of claims in comparing to base variables. In our results, the estimated effects in ZIP and AML models are quite similar, so in this case for $w = 0.875$ this two models are equivalent. You can see, that all effects reduce the expected number of claims compared to base effects. Therefore, the least risky policies (generating the least expected number of claims) are: *veh.age B* and *area B*. The fraction of no claims policies is $\hat{\varpi} = 0.79$ (p-value less than 0.0001). Generally the regressors around mean have a little impact on the shape of the conditional distribution than in comparing to 3rd quartile.

4. Conclusions

The claim counts modelling process is the important part in the ratemaking process. That is why the researchers are still continuing and developing the technique of this type of modelling. In the paper, we presented that there is a possibility to analyze the regressors effects in different parts of claim counts distribution. The problem is how to compare the presented techniques of estimations. The good solution it seems to be taken as a goodness-of-fit measure the cross-validation error, what will be considered in future researches.

Acknowledgements

This research is supported by the grant of Polish Ministry of Science and Higher Education
(nr NN 111461540).

References

[1] De Jong, P., Heller, G.Z., 2008. Generalized Linear Models for Insurance Data. Cambrige: Cambridge University Press.

[2] Denuit, M., Marechal, X., Pitrebois, S., Walhin, J., 2007. Actuarial Modelling of Claims Count. John Wiley&Sons Ltd.

[3] Efron, B., 1992. Poisson Overdispersion Estimates Based on the Method of Asymmetric Maximum Likelihood. Journal of the American Statistical Association 87, (417), 98-107.

[4] Famoye, F., Singh, K.P., 2006. Zero-Inflated Generalized Poisson Regression Model with an Application to Domestic Violence Data. Journal of Data Science 4, 117-130.

[5] Hall, D.B., 2000. Zero-Inflated Poisson and Binomial Regression with Random Effects: A Case Study. Biometrics 56, 1030-1039.

[6] Lambert, D., 1992. Zero-Inflated Poisson Regression with an Application to Defects in Manufacturing. Technometrics 34, 1-14.

[7] Machado, J.A.F., Santos Silva, J.M.C., 2005. Quantiles for Counts. Journal of the American Statistical Association 100 (472), 1226-1237.

[8] Miller, A., 1990. Subset selection in Regression. London: Chapman and Hall.

[9] Min, A., Czado C., 2010. Testing for zero-modification in count regression models, Statistica Sinica 20, 323-341.

[10] Ohlsson, E., Johansson, B., 2010. Non-Life Insurance Pricing with Generalized Linear Models. Berlin: Springer-Verlag.

[11] van den Broek, J., 1995. A score test for zero inflation in a Poisson distribution. Biometrics 51, 738-743.

[12] Wolny-Dominiak, A. 2011. Zero-Inflated Poisson Model for Insurance Data with a Large Number of Zeros (in Polish). In: Forecasting in Management, Research Papers of Wroclaw University 185, 21-30.

[13] Yang, Z., Hardin, J.W., Addy, Ch.L., 2009. Testing overdispersion in the zero – inflated Poisson model. Journal of Statistical Planning and Inference 139, 3340-3353.

[14] Yip, K.C.H., Yau, K.K.W., 2005. On modeling claim frequency data in general insurance with extra zeros. Insurance: Mathematics and Economics 36, 153-163.