EKONOMETRIA ECONOMETRICS 1(39) • 2013

ISSN 1507-3866

Alicja Wolny-Dominiak
University of Economics in Katowice
alicja.wolny-dominiak@ue.katowice.pl

ZERO-INFLATED CLAIM COUNT MODELING AND TESTING – A CASE STUDY

Abstract: In this paper the application of parametric count data models in claim counts modeling is investigated. 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. As the zero-inflation and over-dispersion effects are a common situation in insurance portfolios, three models: zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB) and zero-inflated generalized Poisson regression (ZIGP) are tested against the classic Poisson model. The 4-step procedure for modeling zero-inflation effect is proposed. This procedure is applied in the case study. For all calculations the R CRAN software was used.

Keywords: claim counts, ZIP, ZINB, ZIGP.

1. Introduction

In insurance practice, the important part of ratemaking is to model the claims count distribution, where a regression component is included to take the individual characteristics into account. A very common method chosen for modeling a claim count is a regression model, as in [Denuit et al. 2007] with the use of Poisson distribution, which is a special case of a Generalized Linear Model (GLM Poisson). In claim count regression, independent variables (rating variables) may be interpreted as risk factors. For the selection of these variables into the model, one may use traditional methods from [Miller 1990] or adopt genetic algorithms as in [Gamrot 2008]. Literature review reveals that, in particular, attempts are undertaken to find a probabilistic model for the claims count distribution in motor third-party liability, where usually the claims count 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. This means that the data contains lots of zeros and, as a consequence, the Poisson regression may not give satisfactory results. In order to allow the presence of excess zeros in an insurance portfolio, the zero-inflated models are applied [Wolny--Dominiak 2011]. The classic model is the zero – inflated Poisson model (ZIP)

[Lambert 1992], which is a mixture of a Poisson distribution and a zero point mass. The generalization of this model is possible and then the zero-inflated generalized Poisson model (ZIGP) is obtained. 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, Yau 2005].

The other problem often existing in insurance data is the incidence of over-dispersion, which means that data exhibit greater variability than allowed for the Poisson model (the mean is not equal to variance). The reason for that may be the disregarding of some latent factors affecting the claims occurrence. Usually in the case of over-dispersion in the ZIP model, the zero-inflated negative binomial (ZINB) model is used [Hall 2000]. In literature there are some simulation studies with the score test for over-dispersion based on the ZIGP model, which illustrate that the ZIGP model has higher empirical power than the ZINB model [Yang et al. 2009]. This paper aims to propose and apply a procedure to select between the ZIP, ZINB and ZIGP models in the case of the zero – inflation and the over-dispersion occurrence in insurance portfolio.

This paper is organized into five sections, and the introduction. In Section 2 the automobile insurance data used in the case study is presented. Section 3 includes a brief description of the ZIP, ZINB and ZIGP models for insurance data; Section 3 contains three hypothesis tests, which can be made to accept or reject models. In section 4 the estimation MLE method is discussed. For all the calculations the R CRAN software is used.

2. Motor Third-Party liability insurance dataset

To present the modeling process with the zero-inflated and over-dispersion occurrence, a case study based on the real-world dataset is analyzed. The data contains information about Motor Third-Party liability insurance from a Swedish insurance company [De Jong, Heller 2008]. There are five exogenous variables for every policy, as well as the total number of claims at fault that were reported within a yearly period. The following list of variables were considered in claims count models:

1. *age* – age band of policy holder A (youngest), B, C, D, E, F,
2. *gender* – gender male, female,
3. *area* – area of residence A, B, C, D, E, F,
4. *veh.age* – vehicle age A (new), B, C, D,
5. *veh.body* – vehicle body type bus, convertible, coupe, hatchback, hardtop, motorized caravan/combi, minibus, panel van, roadster, sedan, station wagon, truck, utility,
6. *num.clams* – The number of claims: 0, 1, 2, 3, 4.

Every variable can have an influence on the number of claims (called rating variable) as well as on the occurrence of the zero-inflation effect. The maximum

number of reported claims recorded is four and the rate of the total zero claims is 93.19%. The empirical distribution is shown in Table 1.

Table 1. The number of claims in the dataset

Number of claims	Observed	% observed
0	63 232	93.19
1	4333	6.39
2	271	0.40
3	18	0.03
4	2	0.00
Total	67 856	

Source: own elaboration.

In modeling, three zero-inflated models are investigated against the Poisson model: ZIP, ZINB and ZIGP. To select the optimal model, the following four-step procedure is applied: 1. selection of two subsets of variables, the subset of rating variables and the subset of variables affecting the zero–inflation occurrence, for each ZIP, ZINB and ZIGP model separately; 2. performance of hypothesis testing for the zero-inflation effect and over-dispersion occurrence; 3. comparison between every pair of ZIG, ZINB and ZIGP models using the Vuong test; 4. maximum likelihood (ML) estimation of structural parameters and calculating goodness-of-fit measures.

For all calculations, the software R CRAN was applied. In order to execute the ML estimation, a few packages were used: {pscl} package for ZIP/ZINB models, the Vuong test and {ZIGP} package for the ZIGP model. The score test is achieved via the implementation of the score statistics $S(\hat{\beta})$ (see Section 3).

3. Claims count models

Claims data are discrete count data for which, in most cases, the Poisson distribution is applied for modeling. Let the response variable Y_i, $i = 1,...,n$ be the number of claims in the portfolio of n risk classes. If Y_i follows the Poisson distribution, the mean claims count $E(Y_i) = \lambda_i$ is assumed to be constant in every risk class, which implies that $E(Y_i) = \text{var}(Y_i) = \lambda_i$. As was mentioned above, the assumption about equality of mean and variance usually is not satisfied for insurance data, which often display over-dispersion. This means that the heterogeneity within risk classes is allowed by defining a distribution for the parameter λ_i. Typically, assuming λ_i to be the Gamma distribution with two first moments $E(\lambda_i) = \mu_i$, $\text{var}(\lambda_i) = \frac{\mu_i^2}{\alpha}$ and $Y_i \mid \lambda_i$ to be the Poisson distribution with $E(Y_i \mid \lambda_i) = \lambda_i$, the number of claims Y_i follows a negative binomial distribution with probability density function given by [Lawless 1987]:

$$f(\mu_i,\alpha,y_i)=\frac{\Gamma(y_i+\frac{1}{\alpha})}{y_i!\Gamma(\frac{1}{\alpha})}(\frac{1}{1+\alpha\mu_i})^{\frac{1}{\alpha}}(\frac{\alpha\mu_i}{1+\alpha\mu_i})^{y_i},\ i=1,...,n.$$

This distribution is also known as gamma-Poisson mixture distribution, where the mean claims count is $E(Y_i)=\mu_i$ and the variance is $\text{var}(Y_i)=\mu_i(1+\alpha\mu_i)$.

The other model, which can be fitted in the case of the over-dispersion effect, is the generalized Poisson distribution given by [Famoye, Singh 2006]:

$$f(\mu_i,\alpha,y_i)=(\frac{\mu_i}{1+\alpha\mu_i})^{y_i}\frac{(1+\alpha y_i)^{y_i-1}}{y_i!}[\frac{-\mu_i(1+\alpha y_i)}{1+\alpha\mu_i}],\ i=1,...,n$$

with the mean and the variance respectively: $E(Y_i)=\mu_i$, $\text{var}(Y_i)=\mu_i(1+\alpha\mu_i)^2$. If $\alpha=0$, the above model reduces to the Poisson distribution with no over-dispersion effect. The general form of zero-inflated (ZI model) distribution can be expressed as follows:

$$P(Y=y_i)=\begin{cases}\varpi+(1-\varpi)f(\mu_i,\alpha,y_i), & y_i=0\\ (1-\varpi)f(\mu_i,\alpha,y_i), & y_i>0\end{cases},\ i=1,...,n,$$

where ϖ is the probability of zero claims count and $f(\mu_i,\alpha,y_i)$ is the probability density function, most often generalized Poisson or negative binomial. The zeros from the first equation of the generalized Poisson distribution are called "structural" zeros and from the second equation – "sampling" zeros. The first two moments in zero-inflated models are: $E(y_i)=(1-\varpi)\mu_i$ and $\text{var}(y_i)=E(y_i)[(1+\alpha\mu_i)^2+\varpi\mu_i]$.

The variable Y_i depends on rating variables X_{ij}, $j=1,...,k$, which influence the claims count (e.g. in automobile insurance: the age of driver or the engine capacity) by the log link function $\mu_i=\exp(\sum_{j=1}^{k}X_{ij}\beta_j)$. In turn, the parameter ϖ_i are linked with explanatory variables via the logit function $\ln(\frac{\varpi}{1-\varpi})=\sum_{j=1}^{t}\gamma_j Z_{ij}$, where Z_{ij}, $j=1,...,t$ are variables affecting the occurrence of "sampling" zeros. When the distribution $f(\mu_i,\alpha,y_i)$ is assumed to be Poisson, the ZIP model is received, for a negative binomial – the ZINB model and for a generalized Poisson – the ZIGP model.

For the analyzed insurance dataset, all ZIP, ZINB and ZIGP models are investigated against the Poisson model. Firstly, the selection of the subset of variables is made by the backward elimination technique. This technique starts by estimating ZIP, ZINB and ZIGP models with all combinations (1024 possibilities) of rating variables and variables affecting the occurrence of "sampling" zeros. For further analysis only statistically significant variables in each model are taken into account, which are: *age* and *veh.body* as rating variables and *Intercept* for modeling "sampling" zeros.

4. Score tests in claim count modeling

The goal of this section is to give a brief description of the course of conduct in the claims count model testing. As was mentioned above, the insurance portfolios usually have more zero claims than are expected in the Poisson model, which means the zero–inflation occurrence. For testing whether there are many observed zeros, the score test in which the probability ϖ_i, $i = 1,...,n$ is assumed to be constant across observations (in our case – all policies) can be applied [Van den Broek 1995]. The null hypothesis takes a form $H_0 : \varpi = 0$. The score statistic is defined as:

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

where $\bar{y}$ is the average of the claims count. The statistic $S(\hat{\beta})$ follows an asymptotic χ^2 distribution with 1 degree of freedom. In cases of dataset with the zero-inflation occurrence, the over-dispersion effect occurred as the consequences of high variability for zero claims. So in fact if the null hypothesis is rejected in the above score test that means the over-dispersion occurrence. For such data, the Poisson model is not appropriated and leads to a serious underestimation of standard errors and regression parameters. The solution is to apply a ZIP, ZINB or ZIGP model. To select one of those three models, the Vuong test can be applied [Vuong 1989].This test compares two models based on Kullback–Leibler information criteria as a measure of the distance between these models. Assuming that $f_1(x,\beta^1)$ and $f_2(x,\beta^2)$ are two competing claims count models the null hypothesis is as follows: $H_0 : LR \equiv \max_{\beta^1} E[\log f_1(x,\beta^1)] - \max_{\beta^2} E[\log f_2(x,\beta^2)] = 0$. If the consi-dered models are non-nested, under the null hypothesis H_0 the statistic $\sqrt{n}\sum_{i=1}^{n}[\log f_1(x_i,\hat{\beta}_n^1) - \log f_2(x,\hat{\beta}_n^2)]$ follows a normal distribution $N(0,w^2)$, where $w = E[\log f_1(x,\beta^1) - \log f_2(x,\beta^2)]$.

In the case study, firstly the test for the zero-inflation occurrence is applied in the Poisson model. According to the score statistic $S(\hat{\beta}) = -0.0046$, the null hypothesis should be rejected (significant at level 5%), which means that the zero-inflated effect occurs in the data. A comparison of zero-inflated models using the Vuong test yields the following results:

Table 2. Vuong test value for ZIP, ZINB and ZIGP

Model1	Model2	Vuong Statistics *LR*
ZIP	ZINB	1.36 (ZINB > ZIP)
ZIP	ZIGP	1.39 (ZIGP > ZIP)
ZIGP	ZINB	1.34 (ZINB > ZIGP)

Source: own elaboration.

Summarizing, the two tests show that in the analyzed case study, the zero-inflation effect occurs and the ZINB model provides the best results. But the Kullback–Leibler distances for all pairs of models vary slightly, so in the estimation all models are still taken into consideration.

5. Parameters estimation

Parameters in the zero-inflated models are estimated by the maximum likelihood estimation method (MLE). The log-likelihood functions of the ZIP and ZINB models are given by [Yip, Yau 2005]:

$$\ln L^{ZIP} = -n\ln(1+\frac{\varpi}{1-\varpi}) + \sum_{y_i=0}\ln(\frac{\varpi}{1-\varpi} + e^{-\lambda_i}) + \sum_{y_i>0}[-\lambda_i + \ln(\frac{\lambda_i^{y_i}}{y_i!})]$$

$$\ln L^{ZINB} = \sum_{y_i=0}\ln[\varpi + (1-\varpi)(\frac{1}{1+\alpha\mu_i})^{\frac{1}{\alpha}}] +$$

$$+\sum_{y_i>0}[\ln(1-\varpi) + \ln\frac{\Gamma(y_i+\frac{1}{\alpha})}{\Gamma(y_i+1)\Gamma(\frac{1}{\alpha})} + \ln(\frac{1}{1+\alpha\mu_i})^{\frac{1}{\alpha}} + \ln(\frac{\alpha\mu_i}{1+\alpha\mu_i})^{y_i}].$$

The parameters estimation of the ZIP and ZINB models is achieved via R CRAN using the *zeroinfl* function from the {pscl} package (see Table 1). To obtain the maximum likelihood estimates for the ZIGP model, the log-likelihood function is as follows [Famoye, Singh 2006]:

$$\ln L^{ZIGP} = -n\ln(1+\frac{\varpi}{1-\varpi}) + \sum_{y_i=0}\ln(\frac{\varpi}{1-\varpi} + e^{-(\frac{\mu_i}{1+\alpha\mu_i})}) +$$

$$+\sum_{y_i>0}[y_i\ln(\frac{\mu_i}{1+\alpha\mu_i}) + (y_i-1)\ln(1+\alpha y_i) - \ln(y_i!) - \frac{\mu_i}{1+\alpha\mu_i}(1+\alpha y_i)].$$

In this case, the *est.zigp* function from the {ZIGP} package is applied.

Table 3. Results of fitting ZIP, ZINB and ZIGP models

	ZIP	s.e. ZIP	ZINB	s.e. ZINB	ZIGP	s.e. ZIGP
1	2	3	4	5	6	7
Intercept	–0.8043	0.3449	–1.3497	0.3548	–1.3505	0.5086
veh.bodyBUS	0.0000	–	0.0000	–	0.0000	–
veh.bodyCONVT	–1.7059	0.6770	–1.7200	0.6790	–1.6962	0.6652
veh.bodyCOUPE	–0.7462	0.3600	–0.7623	0.3632	–0.7707	0.3449
veh.bodyHBACK	–1.0471	0.3406	–1.0611	0.3439	–1.0576	0.3251
veh.bodyHDTOP	–0.8560	0.3508	–0.8687	0.3541	–0.8579	0.3354
veh.bodyMCARA	–0.4631	0.4334	–0.4719	0.4368	–0.4616	0.4177

Table 3, cont.

1	2	3	4	5	6	7
veh.bodyMIBUS	–1.1667	0.3723	–1.1807	0.3754	–1.1749	0.3579
veh.bodyPANVN	–0.8093	0.3618	–0.8251	0.3650	–0.8366	0.3470
veh.bodyRDSTR	–0.5734	0.6935	–0.6138	0.6935	–0.7253	0.7019
veh.bodySEDAN	–0.9955	0.3405	–1.0110	0.3437	–1.0204	0.3249
veh.bodySTNWG	–0.9598	0.3407	–0.9740	0.3440	–0.9756	0.3252
veh.bodyTRUCK	–1.0109	0.3512	–1.0267	0.3544	–1.0382	0.3361
veh.bodyUTE	–1.2228	0.3450	–1.2379	0.3482	–1.2421	0.3297
agecatA	0.0000	–	0.0000	–	0.0000	–
agecatB	–0.1693	0.0559	–0.1712	0.0560	–0.1795	0.0553
agecatC	–0.2040	0.0545	–0.2058	0.0547	–0.2129	0.0539
agecatD	–0.2314	0.0544	–0.2335	0.0545	–0.2428	0.0538
agecatE	–0.4268	0.0608	–0.4285	0.0609	–0.4330	0.0603
agecatF	–0.4365	0.0692	–0.4388	0.0694	–0.4491	0.0689
AIC	36 041		36 042		36 030	

Source: own elaboration.

The AIC=36 030.5 is the lowest value between ZIP, ZINB and ZIGP, which suggests that the ZIGP model gives the best results for this dataset and should be applied in the ratemaking process. Probably the lower AIC can be achieved in the case of a repeal of the assumption of a constant value ϖ. Unfortunately, at this time, {pscl} and {ZIGP} packages do not give such technical possibilities.

6. Conclusions

The claim count modeling is one of the important steps in the ratemaking process, especially in bonus-malus system. So the researchers are still continuing and developing the technique of this type of modeling. Because the zero-inflation and the over-dispersion effects in insurance datasets are a common situation, the enlargement of the ZIP and ZINB models to a ZIGP model could give better estimations. This paper shows a short procedure of testing such effects and estimating models which can be applied in insurance practice. To make this procedure more sophisticated, in place of the van der Broek and Vuong tests, other techniques like bootstrapping [Yang et al. 2009] or ordered statistics can be used.

Literature

De Jong P., Heller G.Z. (2008), *Generalized Linear Models for Insurance Data*, Cambridge University Press, Cambridge.

Denuit M., Marechal X., Pitrebois S., Walhin J. (2007), *Actuarial Modelling of Claims Count*, John Wiley&Sons.

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.

Gamrot W. (2008), Representative sample selection via random search with application to surveying communication lines, [in:] P. Rehorova, K. Marsikova, Z. Hubinka (eds.), *Proceedings of 26th International Conference on Mathematical Methods in Economics 2008*, Technical University of Liberec, pp. 127–132.

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

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

Lawless J.F. (1987), Negative binomial and mixed Poisson regression, *The Canadian Journal of Statistics* 15 (3): 209–225.

Miller A. (1990), *Subset Selection in Regression*, Chapman and Hall, London.

Van den Broek J. (1995), A score test for zero inflation in a Poisson distribution, *Biometrics* 51: 738––743.

Vuong Q. (1989), Likelihood ratio tests for model selection and non-nested hypotheses, *Econometrica* 57: 307–33.

Wolny-Dominiak A. (2011), Zmodyfikowana regresja Poissona dla danych ubezpieczeniowych z dużą liczbą zer, [in:] *Prognozowanie w zarządzaniu firmą*, Prace Naukowe Uniwersytetu Ekonomicznego nr 185, pp. 21–30.

Yang Z., Hardin J.W., Addy Ch.L. (2009), Testing over-dispersion in the zero-inflated Poisson model, *Journal of Statistical Planning and Inference* 139: 3340–3353.

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.

MODELOWANIE LICZBY SZKÓD Z UWZGLĘDNIENIEM EFEKTU NADMIERNEJ LICZBY ZER ORAZ NADMIERNEJ DYSPERSJI – STUDIUM PRZYPADKU

Streszczenie: W niniejszej pracy rozważamy zastosowanie parametrycznych modeli, służących do estymacji zmiennych licznikowych, w procesie modelowania liczby szkód w zakładzie ubezpieczeń. Portfele ubezpieczeniowe mają specyficzny charakter, a mianowicie dla bardzo dużej liczby polis nie następuje żadna szkoda, co oznacza, iż w danych występuje duża liczba zer. Zatem modelując liczbę szkód, należy brać pod uwagę ten efekt. Dlatego też w pracy testujemy trzy modele uwzględniające efekt nadmiernej liczby zer: ZIP, ZINB oraz ZIGP w porównaniu z klasyczną regresją Poissona w proponowanej 4-etapowej procedurze modelowania liczby szkód. Procedurę tę stosujemy w studium przypadku. Do wszelkich obliczeń wykorzystujemy program R CRAN.

Słowa kluczowe: liczba szkód, ZIP, ZINB, ZIGP.