MATHEMATICAL ECONOMICS

No. 8(15) 2012

RANKING AND CLASSIFICATION OF AUTOMOBILE INSURANCE POLICIES ACCORDING TO THE NUMBER OF CLAIMS

Alicja Wolny-Dominiak

Abstract. In the ratemaking process, the ranking which takes into account the number of claims generated by a policy in a given period of insurance, may be helpful. For example, such a ranking allows to classify the newly concluded insurance policy to the appropriate tariff groups and to differentiate policies with no claims observed in the insurance history. For this purpose, in this paper we analyze models applicable to the modeling of count variables. In the first part of the paper, we present the classical Poisson regression and a modified regression model for data, where there is a large number of zeros in the values of the counter variable, which is a common situation in the insurance data. In the second part, we expand the classical Poisson regression by adding the random effect. The goal is to avoid an unrealistic assumption that in every class all insurance policies are characterized by the same expected number of claims. In the last part of the paper, we propose to use *k*-fold cross-validation to identify the factors which influence the number of insurance claims the most. Then, setting the parameters of the Poisson distribution, we create the ranking of policies using the estimated parameters of the model, which give the smallest cross-validation mean squared error. In the paper we use a real-world data set taken from literature. For all computations we used the free software environment R.

Keywords: claims counts, Poisson regression, zero-inflation effect.

JEL Classification: C25, C13.

1. Introduction

Every person, when applying for an insurance policy, is assigned to a class that is homogeneous in terms of the system of tariffs. One of the criteria used for assigning an individual to a certain class is the number of claims observed in a certain period of time. Thus, it is the insurance companies' very important task to model the number of claims in a given insurance portfolio. In the paper we propose a simple procedure for creating

Alicja Wolny-Dominiak
Department of Statistical and Mathematical Methods in Economics, University of Economics in Katowice, 1 Maja Street 50, 40-287 Katowice, Poland.
E-mail: woali@ae.katowice.pl

a ranking of insurance policies and also for classifying them according to the number of claims. This allows a preliminary classification of a new policy to a group with an adequate premium level.

A very common choice of method for modeling the number of claims is a regression model using the Poisson distribution (Poisson regression), which is a special case of the Generalized Linear Model (GLM). 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. This means that the data contains lots of zeros and, as a consequence, the Poisson regression may not give satisfactory results. Therefore, when creating the ranking, the GLM model and ZIP model (zero-inflated Poisson) and the model with a random effect were considered. The ranking creation procedure used a *k*-fold cross-validation and furthermore the ranking was discretized due to a parameter λ. We built many different models and then we useda 10-fold cross-validation in order to recognize which rating variables have an impact on the presence of zeros in the policies' portfolios. The data for the illustrative example was taken from the literature (Ohlsson, Johansson, 2010). All the computations were conducted in R – the free software environment. The procedure for building a model with random effect and a cross-validation technique was written in R language.

2. Modeling the number of claims

The linear regression models are used for creating a ranking of insurance policies due to the number of claims.

Generalized Linear Models (GLM): In these models we assume that the number of claims is a dependent variable Y that follows a Poisson distribution and it depends on a certain system of predictors (Denuit et al., 2007):

$$P(Y_i = y_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!},\ i = 1,...,n,$$

where Y_i is the number of claims for the *i*-th insured person, $Y_1, ..., Y_n$ are independent and have equal variances, and the average number of claims is equal to the variance. The λ_i parameter is the expected number of claims and it depends on k predictors X_j (risk factors), $j = 1, ..., k$ which describe the insured individual or vehicle, e.g. sex, age, engine capacity. These predictors are categorical variables. The logarithm is used as a link function:

$$\ln \lambda_i = \mathbf{X\beta},$$

where $\mathbf{X}$ is the design matrix and $\boldsymbol{\beta}$ is the vector of regression coefficients. We can see that for every linear combination of predictors the expected value of the number of claims is always positive. The λ_i parameter is adjusted with the use of d_i – exposition to risk factor for the *i*-th policy. This factor expresses what part of the analyzed period of time was covered by a given policy:

$$\tilde{\lambda}_i = d_i e^{\beta_0 + \sum_{j=1}^{k} \beta_{ij}} .$$

When creating the ranking, the goal is to minimize the number of claims, so we used $\min \tilde{\lambda}_i$ as a criterion.

The independence assumption in the above model may not be fulfilled (Hall, 2000). In that case the solution is to use a mixed model and introduce a random effect *v*.

Hierarchical Generalized Linear Models (HGLM): In the case of automobile insurance data, the region or the vehicle model can be treated as a random effect v. Hierarchical generalized linear model (HGLM) with variable $y\,|\,u$ following the Poisson distribution, has a form (Lee, Nelder, Pawitan, 2006):

$$\mu = E(y\,|\,u) = e^{X\beta + v},$$

$$Var(y\,|\,u) = \phi V(\mu),$$

$$v = \log u,$$

where $\beta = [\beta_1, ..., \beta_I]$, $u = [u_1, ..., u_K]$ and $\boldsymbol{X}$ is the model matrix. The distribution of the random effect may belong to the exponential dispersion family of distributions, e.g. the gamma distribution with parameter α:

$$E(u) = \psi$$

$$Var(u) = \alpha V(\psi).$$

The structural parameters of a model have the following interpretation:

– parameter β_i, $i = 1, ..., I$, measures the influence of the *i*-th predictor on the number of claims;

– parameter u_k, $k = 1, ..., K$, measures the risk level for every category (which is different for every category).

Zero-inflated Poisson model: Another model used for estimating the number of claims is the ZIP model, where the counting response variable has a lot of zero values (Lambert, 1992). This is exactly the case when modeling the number of counts. On analyzing different risk portfolios, one can notice that for many policies there is no claim observed and if the claims occur their number is one, two or three and very rarely more. In the ZIP model the independent variables Y_i take zero values $Y_i \sim 0$ with the probability ϖ_i or values from Poisson distribution $Y_i \sim Pois(\lambda_i)$ with probability $1-\varpi_i$. This can be written in the following form:

$$P(Y_i = y_i) = \begin{cases} \varpi_i + (1-\varpi_i)e^{-\lambda_i}, y_i = 0 \\ (1-\varpi_i)\dfrac{e^{(-\lambda_i)}\lambda_i^{y_i}}{y_i!}, y_i > 0 \end{cases}, \ i = 1,...,n .$$

Thus, in the ZIP model we have two parameters: λ_i and ϖ_i. Both parameters, as in the case of the Poisson regression, are linked with predictor variables with the following link functions:

$$\ln(\frac{\varpi_i}{1-\varpi_i}) = \sum_{j=1}^{t}\gamma_{ji}Z_{ji},$$

$$\ln\lambda_i = \sum_{j=1}^{k}\beta_{ji}X_{ji} ,$$

where $Z_1, ..., Z_l$ are the dependent variables for the first equation and $X_1, ..., X_k$ for the second one. The expected value and variance of the number of claims for the *i*-th policy in the ZIP model are, respectively:

$$E(Y_i) = \lambda_i(1-\varpi_i) ,$$

$$D^2(Y_i) = (1-\varpi_i)(\lambda_i - \varpi_i\lambda_i^2) .$$

Similarly to the Poisson regression case, in the ZIP model we assume that the average number of claims equals the variance. The solution to a problem when over-dispersion occurs, is the use of negative binomial distribution (Lambert, 1992).

3. Cross-validation procedure

In order to meaningfully compare the presented models, the choice of the model for the number of claims and the choice of the combination of predictor variables that generate zero counts in the claims for policies, were supported by statistical learning methods (Picard, Cook, 1984). In general, in these methods we assume we are given a training data set $D = \{(x^i, y^i), i = 1, ..., N\}$, where $x^i, y^i \in R$. Moreover, we assume that data is i.i.d. (independent and identically distributed) and has been taken from the population with a multidimensional distribution defined by an unknown density function:

$$p(x,y) = p(x)p(y\,|\,x)\,.$$

The task is to search a given set of functions $H = \{f(x,\varpi) : \varpi \in \Omega\}$, where ϖ is a model parameters vector, and find the best element. Using the model $f(x, \varpi) \in H$, which is always a simplified equivalent of the analyzed phenomenon, we accept some errors that are just the consequence of taking theoretical values instead of real values for the response variable. These errors (for a given observation) are measured by the so-called loss functions $L(y, f(y,\varpi))$. In the concept of statistical learning, the risk functional is considered which measures the overall loss, i.e. the sum of errors for all possible observations. One of the methods of estimating the value of the risk functional is the cross-validation method (CV) (Gatnar, 2008; Picard, Cook, 1984). This paper uses 10-fold cross-validation algorithm, i.e.:

a) randomly dividing the portfolio of policies (training set) into $k = 10$ approximately equally sized parts, where n is the training set size, m_l – the size of the l-th subset, $l = 1,...,10$;

b) building 10 times a model using 9 of 10 parts ($n - m_l$ observations), treating excluded observations as validation set;

c) calculating 10 times the value of the mean squared error MSE_l using the validation set;

d) estimating the cross–validation error: $cv = \sum_{l=1}^{10} \frac{m_l}{n} MSE_l$.

The model with the smallest cv value is selected.

4. Procedure of creating ranking of property insurance policies and classification of these policies

The procedure of building a ranking of policies using the linear models presented in the previous part of the paper may be formulated in a few steps:

Step 1. Estimating $\tilde{\lambda}_i$, $i = 1,...,n$ parameter for every policy in the portfolio using three different models: generalized linear model, hierarchical generalized linear model and zero-inflated generalized linear model.

Step 2. Applying cross-validation procedure to every model from Step 1.

Step 3. Choosing the model with the smallest cv error.

Step 4. Creating the ranking of insurance policies using as a criterion:

$$\min \tilde{\lambda}_i$$

Step 5. Discretizing the ranking according to the values of parameters $\tilde{\lambda}_i$ and thus obtaining insurance risk classification, which allows to classify a new policy to a group with an adequate premium level.

Based on the estimated parameter $\tilde{\lambda}_i$ for a chosen model, we have created ranking and conducted discretization in order to obtain different classes of insurance risk. Discretization means dividing the ordered set of values of a given continuous variable onto a finite number of disjoint intervals. Labels can be assigned to these intervals, e.g. high insurance risk level, neutral to risk, etc. The problem is how to determine the cut points. These cut points should separate the objects from different risk classes in the best possible way. There are two main approaches in discretization: agglomerative and divisive. The first one starts with every single empirical value of the continuous variable belonging to a different interval, and then neighbouring intervals are merged iteratively until the maximum value of a homogeneity of subsets measure is reached. The second approach starts with one big interval covering all empirical values of the continuous variable, and then it is iteratively divided using previously determined cut points.

5. Empirical example

In order to illustrate the process of creating the ranking and discretizing it, the necessary procedures were implemented in R environment. The automobile insurance data set, including information about the number of claims, was used for computations (Ohlsson, Johansson, 2010). The following variables from the data set were considered in the model:

1. *Driver_age* – age of the insured person (driver);
2. *Region*: classes from 1 to 7;
3. *MC_class*: classes from 1 to 7.

These classes were created based on the EV coefficient defined as:

$$EV = \frac{\text{engine capacity in kW x 100}}{\text{vehicle weight in kg} + 75},$$

where 75 kg is the average weight of a driver:

4. *Veh_age* – age of the vehicle;
5. *Num_claims* – number of claims – the sum within the class.

Procedure for creating the ranking

Step 1. We model the number of claims with the use of three types of models presented above.

Model 1. GLM for the variable *Num_claims* assuming Poisson distribution

R Code

```
data(dataset)
glm.formula=Num_claims~Driver_age+Region+MC_class+Veh_age
glm.model1=glm(glm.formula, family=Poisson(link="log"),
data=dataset)
summary(glm.model1)
```

Table 1. Parameters for Model 1

	β_i	Standard error	e^{β_i}
Intercept	– 2.362	0.235	0.0942
Driver_ageA	0	–	1
Driver_ageB	– 0.256	0.182	0.7741
Driver_ageC	– 0.5	0.182	0.6065
Driver_ageD	– 1.273	0.196	0.28
Driver_ageE	– 1.298	0.181	0.2731
Driver_ageF	– 1.435	0.192	0.2381
Driver_ageG	– 1.961	0.291	0.1407
RegionA	0	–	1
RegionB	– 0.396	0.108	0.673
RegionC	– 0.817	0.118	0.4418
RegionD	– 0.909	0.104	0.4029
RegionE	– 1.843	0.342	0.1583
RegionF	– 1.455	0.248	0.2334
RegionG	– 2.065	1.002	0.1268
MC_classA	0	–	1
MC_classB	0.307	0.2	1.3593
MC_classC	0.081	0.168	1.0844
MC_classD	– 0.011	0.181	0.9891
MC_classE	0.554	0.171	1.7402
MC_classF	1.035	0.168	2.8151
MC_classG	– 0.499	0.437	0.6071
Veh_ageA	0	–	1
Veh_ageB	– 0.456	0.122	0.6338
Veh_ageC	– 0.769	0.125	0.4635
Veh_ageD	– 1.239	0.108	0.2897

Source: own elaboration.

The following combination has been chosen as reference categories: *Driver_ageA*, *RegionA*, *MC_classA*, *Veh_ageA*.

Model 2. HGLM of a type POISSON-GAMMA for the variable *Num_claims* assuming Poisson distribution and treating variable *Region* as a random effect with gamma distribution

R Code

```
Model.Poisson.gamma=function(X=X, Z=Z, Y=Y, data-
set.letters= dataset.letters,
glm.formula=Num_claims~Driver_age+Region+MC_class+Veh_age)
```

Table 2. Parameters for Model 2 – fixed effects

	β_i	Standard error	e^{β_i}
Intercept	– 2.520	0.565	0.080
Driver_ageA	0.000	–	1.000
Driver_ageB	1.215	0.164	3.371
Driver_ageC	1.221	0.164	3.390
Driver_ageD	0.590	0.178	1.804
Driver_ageE	1.195	0.164	3.303
Driver_ageF	0.693	0.175	2.000
Driver_ageG	– 0.988	0.268	0.372
MC_classA	0.000	–	1.000
MC_classB	0.207	0.184	1.229
MC_classC	1.276	0.154	3.582
MC_classD	0.752	0.165	2.122
MC_classE	1.190	0.156	3.289
MC_classF	1.339	0.153	3.816
MC_classG	– 1.929	0.401	0.145
Veh_ageA	0.000	–	1.000
Veh_ageB	0.124	0.113	1.132
Veh_ageC	0.062	0.114	1.064
Veh_ageD	0.787	0.099	2.197

Source: own elaboration.

Table 3. Parameters for Model 2 – random effect *Region*

	β_i	Standard error	e^{β_i}
RegionA	0.618	0.527	1.855
RegionB	0.508	0.528	1.663
RegionC	0.197	0.530	1.218
RegionD	0.670	0.527	1.954
RegionE	– 2.304	0.603	0.100
RegionF	– 1.670	0.565	0.188
RegionG	– 3.795	0.908	0.022

Source: own elaboration.

Table 4. Parameters for Model 3

	β_i	Standard error	e^{β_i}
[1] Intercept	–1.179	0.303	0.308
Driver_ageA	0.000	–	1.000
Driver_ageB	– 0.269	0.189	0.764
Driver_ageC	– 0.514	0.189	0.598
Driver_ageD	– 1.281	0.202	0.278
Driver_ageE	– 1.305	0.187	0.271
Driver_ageF	– 1.447	0.198	0.235
Driver_ageG	– 1.976	0.296	0.139
RegionA	0.000	–	1.000
RegionB	– 0.385	0.112	0.681
RegionC	– 0.807	0.121	0.446
RegionD	– 0.898	0.108	0.407
RegionE	– 1.831	0.345	0.160
RegionF	– 1.446	0.251	0.235
RegionG	– 2.048	1.011	0.129
MC_classA	0.000	–	1.000
MC_classB	0.320	0.204	1.377
MC_classC	0.081	0.171	1.084
MC_classD	– 0.007	0.183	0.993
MC_classE	0.560	0.174	1.751
MC_classF	1.046	0.172	2.846
MC_classG	– 0.479	0.444	0.619
Veh_ageA	0.000	–	1.000
Veh_ageB	– 0.459	0.127	0.632
Veh_ageC	– 0.771	0.129	0.463
Veh_ageD	– 1.241	0.112	0.289

Source: own elaboration.

The probability that variable *Num_claims* takes zero value equals 82%.

Model 3. Model ZIP takes into account a large number of zero values for variable *Num_claims*.

R Code

```
data(dataset)
ZIP.model3=zeroinfl(formula=Num_claims~Driver_age+Region+
MC_class+Veh_age|1, data=dataset)
summary(ZIP.model3)
```

Function `zeroinfl` is from the library `{pscl}`

Step 2. Ten fold cross-validation procedure was applied to every model from Step 1, obtaining corresponding *cv* errors.

Model1

```
[1] MSE on one of 10 validation parts in CV method:  9,86491
[1] MSE on one of 10 validation parts in CV method:  6,05339
[1] MSE on one of 10 validation parts in CV method:  5,91436
[1] MSE on one of 10 validation parts in CV method:  8,99081
[1] MSE on one of 10 validation parts in CV method:  40,3884
[1] MSE on one of 10 validation parts in CV method:  6,60433
[1] MSE on one of 10 validation parts in CV method:  6,80834
[1] MSE on one of 10 validation parts in CV method:  6,98802
[1] MSE on one of 10 validation parts in CV method:  9,73839
[1] MSE on one of 10 validation parts in CV method:  6,33045
[1] MSE_CV for the model of the form:
Num_claims ~ Driver_age + Region + MC_class + Veh_age
[1]
**************************************************************
[1] MSE_CV for the model equals: 10,7681
```

Model2

```
[1] MSE on one of 10 validation parts in CV method:  2,13794
[1] MSE on one of 10 validation parts in CV method:  1,67301
[1] MSE on one of 10 validation parts in CV method:  2,08547
[1] MSE on one of 10 validation parts in CV method:  3,00046
[1] MSE on one of 10 validation parts in CV method:  1,54829
[1] MSE on one of 10 validation parts in CV method:  2,08823
[1] MSE on one of 10 validation parts in CV method:  2,03896
[1] MSE on one of 10 validation parts in CV method:  1,977
[1] MSE on one of 10 validation parts in CV method:  3,28651
[1] MSE on one of 10 validation parts in CV method:  1,679
[1]
**************************************************************
[1] MSE_CV for the model equals: 2,15149
```

```
Model3
[1] MSE on one of 10 validation parts in CV method:  0,680905
[1] MSE on one of 10 validation parts in CV method:  0,740123
[1] MSE on one of 10 validation parts in CV method:  0,863056
[1] MSE on one of 10 validation parts in CV method:  1,27631
[1] MSE on one of 10 validation parts in CV method:  0,596856
[1] MSE on one of 10 validation parts in CV method:  1,0285
[1] MSE on one of 10 validation parts in CV method:  0,974237
[1] MSE on one of 10 validation parts in CV method:  0,667925
[1] MSE on one of 10 validation parts in CV method:  1,49571
[1] MSE on one of 10 validation parts in CV method:  0,674469
[1] MSE_CV for the model of the form:
Num_claims ~ Driver_age + Region + MC_class + Veh_age | 1
[1]
*************************************************************
[1] MSE_CV for the model equals: 0,89981
```

Step 3. The smallest value of *MSE cv* was obtained for Model 3, i.e. for the zero-inflated generalized linear model. Thus, this model was used further in the ranking creation steps.

Step 4/Step 5. After discretization, every combination was assigned a label representing a risk class: from 10 – the lowest risk of claim to occur, to 1 – the highest risk of claim to occur. The first five combinations of categories in the ranking were presented in Table 5 for illustration.

Table 5. Part of the ranking and classification based on Model 3

Driver_age	***Region***	***MC_class***	***Veh_age***	***Lambda***	***Risk class***
Driver_ageG	*RegionG*	*MC_classG*	*Veh_ageD*	0.000296	10
Driver_ageG	*RegionE*	*MC_classG*	*Veh_ageD*	0.000369	10
Driver_ageG	*RegionG*	*MC_classG*	*Veh_ageC*	0.000473	10
Driver_ageG	*RegionG*	*MC_classD*	*Veh_ageD*	0.000482	10
Driver_ageG	*RegionG*	*MC_classA*	*Veh_ageD*	0.000487	10

Source: own elaboration.

The number of combinations of different empirical values of predictor variables X_i equals **1372**.

6. Summary

The procedure for recognizing risk classes in insurance policies portfolios proposed in the paper enables to differentiate policies even in the event of observing no claims in the insurance history. The minimum value of λ criterion used in classification means that the risk classes and associated premiums are more equitable for individuals applying for an insurance policy. Essentially, the main disadvantage of the ZIP model – which turned out to be the best in terms of *cv* error criterion – is that within every risk class the policies have an equal expected number of claims, which is an unrealistic assumption. The solution to this issue may be in using the mixed Poisson model and introducing a random effect that would differentiate policies (ZIP regression with random effect). Further work can be done in implementing the R procedure to estimate the latter model. Even if it is computationally very demanding, its value in real word applications could be investigated and compared to the presented approach.

Literature

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

Gatnar E. (2008). *Ensemble Approach in Classification and Regression* (in Polish). Wydawnictwo Naukowe PWN. Warszawa.

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

Lambert D. (1992). *Zero-Inflated Poisson regression, with an application to defects in manufacturing*. Technometrics. Vol. 34. No. 1.

Lee Y., Nelder A.J., Pawitan Y. (2006). *Generalized Linear Models with Random Effects*. Monographs on Statistics and Applied Probability 106. Chapman & Hall\CRC.

Ohlsson E., Johansson B. (2010). *Non-Life Insurance Pricing with Generalized Linear Models*. Springer-Verlag. Berlin.

Picard R., Cook D. (1984). *Cross-validation of regression models*. Journal of the American Statistical Association. Vol. 79(387).