Modeling of claim counts using data mining procedures in R CRAN

Alicja Wolny-Dominiak[1]

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 tariffs group. For this purpose, in this paper we analyze models applicable to the modelling of counter 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 polices using estimated parameters of the model, which give the smallest cross-validation mean squared error and we classify using the regression tree. In the paper we use a real-world data set taken from literature. For all computations we used a free software environment R.

Keywords: claim counts, ZIP, HGLM, R CRAN

JEL Classification: C15, C88
AMS Classification: 65C60

1 Introduction

Every person, when applying for an insurance policy, is assigned to a class, that is homogeneous in terms of the rate-making process. One of the criteria used for assigning an individual to a certain class is the number of claims. Thus it is 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 a ranking of insurance policies and also for classifying them due to the number of claims. It allows a preliminary classification of a new policy to a group with an adequate premium level.

The very common choice of a method for modelling the number of claims is a regression model as in [1] with the use of Poisson distribution, which is a special case of a Generalized Linear Model (GLM Poisson), see in [8]. In regression claims modelling, dependent variables may be interpreted as risk factors. For the selection of these variables into the model one may use traditional methods from [9] or adopt genetic algorithms as in [3]. 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. It means that the data contains lots of zeros and, as a consequence, the Poisson regression may not give satisfactory results which is shown in [11]. Also these two models are with fixed effects so the assumption of independence among the responses is necessary which is sometimes unrealistic for insurance data set. To avoid this problems, except GLM Poisson model we considered ZIP Poisson model according to [6] and HGLM Poisson – Gamma model with a random effect as in [7].

The ranking creation procedure used a *k*–fold cross–validation and furthermore the ranking was discretized due to a parameter λ. We build many different models and then we use a 10–fold cross–validation in order to recognize which rating variables have an impact on the presence of zeros in the policies portfolios. Finally in order to simplify the ranking and classification we applied a regression tree. The data for the illustrative example has been taken from the literature [10]. 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 have been written in R language.

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

2 Modelling the number of claims

The generalized linear models (GLM) are used for creating a raking of insurance policies due to the number of claims. In GLM 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 as in [2]:

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

where Y_i is the number of claims for the *i*–th insured person, $y_i,..., y_n$ 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 predictors X_j, $j = 1,...,k$ that describe the insured individual or vehicle, e.g. sex, age, engine capacity. The logarithm is used as a link function as follows:

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

When creating the ranking we used $\min \lambda_i$ as a criterion.

The independence assumption in the above model may not be fulfilled. In that case the solution is to use a HGLM Poisson-Gamma and introducing a random effect v as in [5]. In case of automobile insurance data "Region" or "Vehicle model" can be treated as a random effect v. The HGLM Poisson-Gamma model has a form according to [7]:

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

where $\beta = [\beta_1,..., \beta_I]$, $u = [u_1,..., u_K]$ and **X** is the model matrix. The structural parameters of a model have a 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).

Another model used for modelling the number of claims is ZIP model, where counting response variable has a lot of zero values. This is exactly the case when modelling the number of counts. Analysing different risk portfolios it can be noticed 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 $1-\varpi_i$. It can be written in a form as in [6]:

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

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

$$\ln\left(\frac{\varpi_i}{1-\varpi_i}\right) = \sum_{j=1}^{t} \gamma_{ji} Z_{ji}$$
$$\ln \lambda_i = \sum_{j=1}^{k} \beta_{ji} X_{ji},$$

where $Z_1,..., Z_l$ are the dependent variables for the first equation and $X_1,..., X_k$ for the second one. Similarly to Poisson regression case, in the ZIP model we assume that the average number of claims equals the variance. The solution to a problem when overdispersion occurs is the use of negative-bimodal distribution (ZINB model), see in [6].

3 Procedure of creating ranking of property insurance policies and classification of these policies

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

1. Estimating λ parameter for every policy in the portfolio using 3 different models: GLM, HGLM and ZIP/ZINB model;

2. Applying 10–fold cross–validation procedure to every model from Step 1 [4]:
   a) randomly divide the training set into $k = 10$ approximately equally sized parts (n - the training set size, m_l - the size of the l -th subset, $l = 1,...,10$),
   b) build 10 times a model using 9 of 10 parts ($n - m_l$ observations), treating excluded observations as validation set,
   c) calculate 10 times the value of the mean squared error $MSE_i = \frac{\sum (y - \mu_l)^2}{m_l}$ using the validation set,
   d) estimate the cross–validation error: $cv = \sum_{l=1}^{10} \frac{m_l}{n} MSE_l$. The model with the smallest cv value is selected,
3. Choosing the model with the smallest cv error;
4. Creating the ranking of insurance policies for every combination of predictor variables X_i , using as a criterion $MIN\lambda$;
5. Discretizing the ranking due to the values of parameters λ and thus obtaining insurance risk classification which allow to classify a new policy to a group with an adequate premium level.

Based on the estimated parameter λ 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 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 object from different risk classes in a 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.

4 Case study for automobile insurance data set

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 has been used for computations [10]. The following variables form the data set and have been 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 which 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

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)
```

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

```
Library(hglm)
data(dataset)
```

```
hglm.model2=hglm(fixed= Num.claims~Driver.age+Region+MC.class+Veh.age, random=~1|Region, family=Poisson(link="log"), rand.family = family=Gamma(link="log"), data=dataset)
summary(hglm.model2)
```

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

```
Library(pscl)
data(dataset)
ZIP.model3=zeroinfl(formula=Num.claims~Driver.age+Region+
MC.class+Veh.age| Driver.age+Region+MC.class+Veh.age, data=dataset)
summary(ZIP.model3)
```

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

2. Ten fold cross–validation procedure was applied to every model from Step 1, obtaining corresponding *cv* errors which are shown in Table 1.

Model	cross-validation error
GLM Poisson	10.76
HGLM Poisson-Gamma	2.15
ZIP	0.89

Table 1 Cross-validation errors

3. The smallest value of *MSE cv* was obtained for the Model 3., i.e. for the zero–inflated generalized linear model. Thus this model was used further in the ranking creation steps. The results are presented in Table 2.

variables	parameters	standard error	tariffs
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

Table 2 Parameters for Model 3 - ZIP

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

4. After discretization every combination was assigned a label representing a risk class: from A – the lowest risk of claim to occur, to J – the highest risk of claim to occur. The distribution of risk classes is as follows in Table3.:

risk class	number of policies in classes	% of policies in classes
A	949	69.17%
B	241	17.57%
C	100	7.29%
D	45	3.28%
E	17	1.24%
F	10	0.73%
G	6	0.44%
H	2	0.15%
J	2	0.15%

Table 3 Distribution for risk classes - Model 3

So finally we received 9 risk classes. The number of combinations of different empirical values of predictor variables X_i equals 1372. In order to obtain a more synthetic description of each risk class, the classification tree model was used. The description of the this tree are presented in Table 4.

class description	class
If (Region=EFG)	A
If [(Region=ABCD) and (MC.class=AG)]	A
If [(Region=ABCDJ) and (MC.class=BCDEGF) and (Drive.age=G)]	A
If [(Region=ABCDJ) and (MC.class=BCDEF) and (Drive.age=GABCDF) and (Veh.age=ABC)]	A
If [(Region=ABCDJ) and (MC.class=AGCDEF) and (Drive.age=EF) and (Veh.age=ABC)]	A
If [(Region=ABCD) and (MC.class=BCDEFAG) and (Drive.age=ABCDEF) and (Veh.age=ABC)]	B
If [(Region=ABCDJ) and (MC.class=F) and (Drive.age=EF) and (Veh.age=ABC)]	B
If [(Region=ABCD) and (MC.class=CDEF) and (Drive.age=BCE) and (Veh.age=ABC)]	B
If [(Region=ABD) and (MC.class=CDEF) and (Drive.age=ABCDFE) and (Veh.age=ABC)]	B
If [(Region=ABD) and (MC.class=BCDEF) and (Drive.age=ABCDEF) and (Veh.age=D)]	B
If [(Region=ABD) and (MC.class=BCDEF) and (Drive.age=ABCDEF) and (Veh.age=D)]	B
If [(Region=ABD) and (MC.class=CDEF) and (Drive.age=BCE) and (Veh.age=ABC)]	C
If [(Region=ABD) and (MC.class=F) and (Drive.age=ABCDF) and (Veh.age=ABC)]	C
If [(Region=ABD) and (MC.class=CDEF) and (Drive.age=ABCDEF) and (Veh.age=D)]	C
If [(Region=ABD) and (MC.class=BCDEF) and (Drive.age=BCE) and (Veh.age=D)]	E

Table 4 Regression tree for risk classes

R Code

```
library(rpart)
library(e1071)
data(dataset)
data.tree=NULL
data.tree=data[,c(1:4, ncol(data))]
model.formula=paste(names(data)[ncol(data)] ," ~ .", sep="")
model.rpart=tune.rpart(eval(parse(text=model.formula)), data=data.tree, minsplit=3:10, cp=c(0.01, 0.03, 0.05))#,
method = "class")
summary(model.rpart)
table(dane[,ncol(data)])
print(model.rpart$best.model)
plot(model.rpart$best.model)
text(model.rpart$best.model)
```

5 Summary

The procedure for recognizing risk classes in the insurance policies portfolios proposed in the paper allows to differentiate policies with no claims observed in the insurance history. The minimum value of λ criterion used in classification causes that the risk classes and associated premiums are fairer for individuals applying for an insurance policy. Essentially the main disadvantage of ZIP model, that turned out to be the best in terms of *cv* error criterion, is that within every risk class the policies have equal expected number of claims, which is an unrealistic assumption. The solution to this issue may be using the mixed Poisson model and introducing a random effect that would differentiate policies (ZIP regression with random effect). However estimating that type of model is computationally very demanding what discourages from using in real world applications.

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.: *Generalized Linear Models for Insurance Data*. Cambride University Press, Cambrige, 2008.

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

[3] Gamrot W.: Representative Sample Selection via Random Search with Application to Surveying Communication Lines, *Proceedings of 26th, International Conference on Mathematical Methods in Economics 2008, (Rehorova P., Marsikova K, Hubinka Z., eds.)*, Technical University of Liberec, 127-132.

[4] Gatnar E.: *Ensemble Approach in Classification and Regression* (in Polish). Wydawnictwo Naukowe PWN, Warszawa 2008.

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

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

[7] Lee Y., Nelder A. J., Pawitan Y.: *Generalized Linear Models with Random Effects*. Monographs on Statistics and Applied Probability 106, Chapman&Hall\CRC, New York, 2006.

[8] McCullagh P., Nelder J. A.: *Generalized Linear Models*. Chapman & Hall/CRC, New York, 1999.

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

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

[11] Wolny-Dominiak A.: 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** (2011), 21-30.