Insurance portfolios rate making: Quantile regression approach

Alicja Wolny–Dominiak[1], Agnieszka Ornat–Acedańska[2], Grażyna Trzpiot[3]

Abstract. Insurance portfolios rate–making frequently based on different multivariate regression models which is more sensitivity to the assumptions which significantly restrict the area of their applications. When error term is non-normal, asymmetric, fat–tailed or in presence of outliers it may have serious consequences for correct inference on the factors impact on endogenous variable. In this paper we analyze three models: generalized linear model (GLM), hierarchical generalized linear model (HGLM) and exponential quantile regression model (EQRM) in rate-making process. The approach using EQRM is robust to deviations from the classical assumptions (a distribution of error terms is left unspecified, which is the main virtue of the method as far as robustness to outliers is concerned). The aim of this paper is to applied GLM, HGLM and EQRM for rate–making and analyzed real insurance automobile data set. For this data set we adopted cross–validation procedure to compare results of these models according to cross–validation Root MSE criteria.

Keywords: rate–making, quantile regression, hierarchical generalized linear models, cross–validation.

JEL Classification: C21, G22
AMS Classification: 65C60

1 Introduction

The rate–making process is one of the most important problem in insurance portfolios issues. The techniques of rate–making are actually based on loss distribution or their moments, which are estimates using historical data. The key problem is to choose the correct model for estimation of loss value. Insurance portfolios rate–making is frequently based on different multivariate regression models which allow to investigate rating factors. Nevertheless, ordinary multivariate regression model has some crucial disadvantages – it is sensitive to the assumptions which significantly restrict the area of their applications. In insurance data case, when error term is non-normal, asymmetric, fat-tailed or in presence of outliers it may have serious consequences for correct inference on the factors impact on endogenous variable. Moreover, ordinary multivariate regression model often ignores the specific feature of the insurance data used. For example, for the real insurance portfolio, there are: possibility of catastrophic losses, the dependence of insured objects on each other (i.e. cumulating risk) or information shortfall to verify the statistical significance of model chosen [5]. Therefore, for modeling insurance data it is important to use models and estimators that are more robust to restrictive classical regression assumptions.

GLM is a good example of such model and therefore it is used by actuaries [7], [8], [1], [9]. However there are some problems connected with GLM. First problem is in choosing the predictors' distribution in GLM. It can be solve with simulation procedure based on the Monte Carlo method [20]. Second problem is in independency assumption for the value of claims. In such a situation HGLM model is recommended. The other approach proposed for modeling insured data (in particular expected net premium rates) is quantile regression – see Kudryavtsev [5]. This approach is consistent with the idea of using the distribution quantile for rate-making. Additional advantage of this method is fact that it allows to estimate the net premium rates including safety loadings and it may be estimated as a quantile of loss distribution.

The first section contains the description of Generalized Linear Model and Hierarchical Generalized Linear Model for the rate-making. Then we discuss the methodology of the quantile regression including it's special case – the exponential quantile regression model as the model which is used for the rate-making. The next section contains the description of the Cross-validation procedure. The empirical results are presented in section four.

[1] University of Economics in Katowice, Faculty of Economics, Department of Statistical and Mathematical Methods in Economics, Bogucicka 14, 40–226 Katowice, alicja.wolny-dominiak@ue.katowice.pl
[2] University of Economics in Katowice, Faculty of Informatics and Communication, Department of Demography and Economic Statistics, Bogucicka 14, 40–226 Katowice, agnieszka.orwat@ue.katowice.pl
[3] University of Economics in Katowice, Faculty of Informatics and Communication, Department of Demography and Economic Statistics, Bogucicka 14, 40–226 Katowice, trzpiot@ue.katowice.pl

2 Generalized linear model and hierarchical generalized linear model for rate–making

Currently in practice of insurance rate-making GLM models are applied. In GLM, a continuous dependent variable Y_i is treated as the value of claims in portfolio and categorical explanatory variables $X_{i1},....,X_{im}$, $i=1,2,\ldots,n$ are treated as rating factors. Observations $Y_1,...,Y_n$ are assumed to be independent. In insurance data the dependent variable Y_i is usually non-negative and skewed to the right. That is why the Gamma GLM for rate-making is applied in practice and is of the form $\mathbf{Y}=g^{-1}(\mathbf{X\beta})$, where $Y\sim G(\mu,\sigma)$ and the link function $g(x)=\ln x$ [1]. Parameters $\beta_1,...,\beta_m$ measure the impact of rating factors on the value of claims (constant for all categories). Tariff rates are calculated by the formula:

$$t=\exp(\mathbf{X\beta}) \tag{1}$$

and show how to change the base premium calculated as $\exp(Intercept)$ for every category of rating factors.

In practice independent assumption for the variable Y is unrealistic. In this case the linear mixed models are useful, where one categorical explanatory variable is assumed to be a random effect. The example of such a variable is area or vehicle model in third party claims. Then the rate-making process is carried out for every category of random effect separately. In this process the HGLM gamma-gamma model can be applied [19] and is of the form [6]:

- conditional on random effects u, the responses Y_i follow a GLM gamma family, satisfying:

$$\mathbf{Y}=g^{-1}(\mathbf{X\beta}+Zv),\ v(u)=\ln u \tag{2}$$

- the random effect u follows a distribution conjugate to a GLM gamma family

The fixed effects $\beta=[\beta_1,...,\beta_l]$ and random effects $v=[v(u_1),....,v(u_K)]$ of the model have the following interpretation:

- parameters β_i, $i=1,...,m$, measure the impact of i-th rating factor on the value of claims (constant for all categories)
- parameters $v(u_k)$, $k=1,...,K$, measure the level of risk within the category (inconstant for all categories)

Similarly as in GLM gamma, tariff rates are determined by the formula:

$$t=\exp(\mathbf{X\beta})\exp(Zv) \tag{3}$$

and show how to change the base premium calculated as $\exp(Intercept)$ for every category of fixed rating factors adjusted for the tariff cell for random effect.

3 Quantile regression

In quantile regression method [2] we analyze a problem of estimation of a vector of parameters $\boldsymbol{\beta}$ for a sample of independent observations y_i, $i=1,2,\ldots,n$ of a sequence of random variables $Y_1,Y_2,..,Y_n$ taken with distribution $P(Y_i<y)=\Im(y-\mathbf{x}_i'\boldsymbol{\beta})$, where $\mathbf{x}_i=(1,x_{i1},x_{i2},..,x_{im})'$ is a column $(n\times m+1)$-matrix of observations $\mathbf{X}$ and the distribution $\Im$ is unknown. The linear quantile regression model (LQRM) of order τ, $0<\tau<1$ is given by the formula [3]:

$$Q_\tau(Y_i|\mathbf{x}_i)=\mathbf{x}_i'\boldsymbol{\beta}^{(\tau)} \tag{4}$$

where $Q_\tau(Y_i|\mathbf{x}_i)$ indicates conditional quantile of random variable Y_i for probability τ provided vector $\mathbf{x}_i$, $\boldsymbol{\beta}^{(\tau)}=(\beta_0^{(\tau)},\beta_1^{(\tau)},\beta_2^{(\tau)},..,\beta_m^{(\tau)})'$ is vector of regression coefficient. The LQRM corresponding to the to linear regression model (LRM) can be expressed as $Y_i=\beta_0^{(\tau)}+\beta_1^{(\tau)}X_{i1}+\ldots+\beta_{im}^{(\tau)}X_{im}+\varepsilon_i^{(\tau)}$, where $\varepsilon_i^{(\tau)}$ is error term. Then $Q_\tau(\varepsilon_i^{(\tau)}|\mathbf{x}_i)=0$. A distribution of independent random variables $\varepsilon_i^{(\tau)}$ is left unspecified, which is the main virtue of the method as far as robustness to outliers is concerned. Koenker and Basset [2] defined a τ-th quantile regression estimator of $\boldsymbol{\beta}^{(\tau)}$, that its value $\mathbf{b}$ solves the problem:

$$\min_{\mathbf{b}\in\Re^{m+1}}\left[\sum_{i\in\{i:\,y_i\geq x_i'\mathbf{b}\}}\tau\left|y_i-\mathbf{x}_i'\mathbf{b}\right|+\sum_{i\in\{i:\,y_i<x_i'\mathbf{b}\}}(1-\tau)\left|y_i-\mathbf{x}_i'\mathbf{b}\right|\right] \tag{5}$$

The problem (5) has always a solution; for continuous distributions it is unique. Since the problem (5) can be transformed to a linear optimization problem its solution can be find using an internal point method [12].

Because the error distribution term is unspecified, statistical inference is based on nonparametric approach – bootstrap or Monte Carlo method. In bootstrap approach samples are drawn with replacement from analyzed data set. Based on the sample of n observations we form a bootstrap sample drawing *n* observations from the original sample. The procedure is repeated *N* times $(N \geq 1000)$. For every bootstrap sample *k* we calculate estimates $\beta_{kj}^{(\tau)}$. Then for hypothesis testing of parameter significance we calculate fraction of samples for which $\beta_{kj}^{(\tau)} = 0$ (null hypothesis $H_0 : \beta_j^{(\tau)} = 0$, alternative hypothesis $H_1 : \beta_j^{(\tau)} \neq 0$) and treat it as an empirical *p-value* of the test.

The quantile approach detects relationships missed by traditional data analysis methods. Robust estimates detect the influence of the bulk of the data, whereas quantile estimates detect the influence of co-variates on alternative parts of the conditional distribution. In practice, the distribution-free approach is often used for estimation – see for example Koenker, Halllock [4], Koenker [3]. Applications of the quantile regression for Polish capital market can be found in Trzpiot [13], Trzpiot [14], Trzpiot [15], Trzpiot [16], Trzpiot [17] and modifications of quantile regression in Orwat–Acedańska, Trzpiot [10], Orwat–Acedańska, Trzpiot [11] and Trzpiot [18].

3.1 Exponential quantile regression model for rate–making

Taking more general type of model ($\mathbf{Y} = g^{-1}(\mathbf{X}\boldsymbol{\beta})$) then linear models allows an actuary to take into account the influence of risk factors on the loss amount in the framework of linear form while the model is non-linear [5]. Combining this formulae with model (4) one can use the method of quantile regression in the following way $Q_\tau(Y_i|\mathbf{x}_i) = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta}^{(\tau)})$.)]. One of the possibilities is taking logarithm as a function *g*.[4] Then $\mathbf{Y} = \exp(\mathbf{X}\boldsymbol{\beta})$.

In this paper we taking as function $g(\cdot)$ logarithm and we use following the exponential quantile regression model (EQRM) of order τ^*:

$$Q_{\tau^*}(Y_i|\mathbf{x}_i) = \exp(\mathbf{x}_i'\boldsymbol{\beta}^{(\tau^*)}), \tag{6}$$

where $Q_{\tau^*}(Y_i|\mathbf{x}_i)$ indicates conditional quantile of random variable Y_i for probability τ^*, $0 < \tau^* < 1$ provided vector of rating factors $\mathbf{x}_i$, and $\boldsymbol{\beta}^{(\tau^*)} = (\beta_0^{(\tau^*)}, \beta_1^{(\tau^*)}, \beta_2^{(\tau^*)}, \ldots, \beta_m^{(\tau^*)})'$ is vector of regression coefficient of order τ^*.

The exponential quantile approach to making net premium rate is base on probability: [5].

$$\tau^* = \frac{\tau - p}{1 - p}, \text{ where } 0 < \tau^* < 1 \tag{7}$$

where *p* is fraction of policies with no claims incurred.

Model (6) of order τ^* in the form (7) gives the tariff rate estimators that are convenient for practical use: the estimators are conditional quantiles (given observable rating factors $\mathbf{x}_i$ known) of probability τ^* for *i*-th policy before loss occurs [5]. Finally tariff rates are determined by the formula (1).

4 Cross–validation procedure

In order to unify the process of comparing presented models, the choice of the model for rate-making is supported by statistical learning methods. 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 it has been taken from the population with a multidimensional distribution defined by an unknown density function:

$$p(x, y) = p(x)p(y \mid x) \tag{8}$$

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 to find the best element. Using the model $f(x,\varpi) \in H$, which is always a simplified equivalent of the analysed phenomenon, we accept some errors that are just the consequence of taking theoretical values instead of real values for response variable. These errors (for a given observation) are measured by 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*) . This paper uses 5–fold cross–validation algorithm for all models, i.e.:

[4] Function $g(\cdot)$ is usually differentiable and monotonic.

1. randomly divide the training set into $k = 5$ approximately equally sized parts;
   (n – the training set size, m_l – the size of the l – th subset, $l = 1,...,5$);
2. build 5 times every model using 4 of 5 parts ($n - m_l$ observations), treating excluded observations as validation set;
3. calculate 5 times the value of the mean squared error $RMSE_l = \sqrt{\frac{\sum (y - \hat{\mu}_l)^2}{m_l}}$ using the validation set;
4. estimate the cross–validation error: $cv = \sum_{i=1}^{5} \frac{m_l}{n} MSE_l$;

The model with the smallest cv value is selected.

5 Results of empirical analysis

In order to illustrate the process of rate–making with EQRM, GLM and HGLM models, the empirical example was calculated using the automobile insurance data set from literature [9]. The following variables from the data set have been considered in models:

1. `Driver.age` – age of the insured person (driver):
2. `Region`: classes from A to G
3. `MC.class`: classes from A to G

   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.

For the data set GLM, HGLM and EQRM models were applied with following assumptions:

- GLM – all rating factors are fixed effects with Gamma distribution (gamma model)
- HGLM – the risk factor `Region` is the random effect with Gamma distribution (Gamma–Gamma model)
- EQRM – the parameter $\tau = 0.99$, and on the base of empirical data we computed the fraction of policies with "no losses" incurred: p = 0.96. Thus, according to formula (6) we implemented EQRM of order $\tau^* = 0.75$. Moreover, p–values in EQRM were calculated using bootstrap method.

For GLM and HGLM models estimation we used a free software environment R CRAN, the package `{stats}` (`glm`) and the package `hglm` (function `hglm`). In the case EQRM we our own procedures, which were created in Matlab program in order to parameters estimate and to test of parameter significance. The *p-values* were calculated by means of bootstrap method. The estimated tariff rates in analyzed threes models are in table 1.

In GLM and HGLM model the base premium is equal to $P_{GLM} = 18354.07$ and $P_{GLM} = 18370.64$ while in EQRM model at a much higher level $P_{EQRM} = 25167.3$. The Similar situation is in the case of structural parameters of the most rating factors which are generally higher in EQRM model while lower or similar for GLM and HGLM models. Different situation is for the factor „`RegionG`", which treated as the random effect takes much higher value compared to the value in EQRM and GLM models, see Table 1.

	EQRM	**p-value**	**GLM**	**p-value**	**HGLM**	**p-value**
`Intercept`	25	0.00	18 354.07	0.00	18 370.64	0.00
`Driver.age`	1.00	-	1.00	-	1.00	-
`Driver.age`	1.62	0.03	1.62	0.04	1.63	0.05
`Driver.age`	2.37	0.00	2.47	0.00	2.37	0.00
`Driver.age`	2.23	0.01	2.18	0.00	2.11	0.01
`Driver.age`	1.52	0.08	1.81	0.01	1.74	0.03
`Driver.age`	1.51	0.17	1.56	0.08	1.51	0.12
`Driver.age`	0.75	0.17	0.65	0.25	0.68	0.32
`RegionA`	1.00	-	1.00	-	1.09	-
`RegionB`	1.12	0.21	1.14	0.36	1.25	-
`RegionC`	0.85	0.08	0.87	0.35	0.91	-
`RegionD`	0.86	0.14	0.90	0.44	0.99	-
`RegionE`	0.67	0.26	0.61	0.27	0.90	-
`RegionF`	0.54	0.12	0.65	0.19	0.90	-
`RegionG`	0.01	0.37	0.01	0.00	0.95	-

MC.classA	1.00	-	1.00	-	1.00	-
MC.classB	1.15	0.37	1.14	0.61	1.10	0.72
MC.classC	1.67	0.09	1.66	0.02	1.44	0.11
MC.classD	1.46	0.13	1.17	0.51	1.12	0.64
MC.classE	1.66	0.04	1.26	0.30	1.17	0.51
MC.classF	2.13	0.01	1.68	0.02	1.58	0.04
MC.classG	3.72	0.19	1.94	0.25	2.11	0.24
Veh.ageA	1.00	-	1.00	-	1.00	-
Veh.ageB	0.96	0.46	0.91	0.56	0.93	0.65
Veh.ageC	0.50	0.02	0.55	0.00	0.57	0.00
Veh.ageD	0.23	0.00	0.27	0.00	0.28	0.00

Table 1 Tariff rates for EQRM, GLM and HGLM model

In order to compare models by means of a unified measure, the 5–fold cross–validation procedure was applied. RMSE error in each validation set and Cross–validation RMSE (cv) are presented in Table 2 and Table 3.

Validation set	RMSE EQRM	RMSE GLM	RMSE HGLM
ValidPart1	46440.6	44242.1	44163.6
ValidPart2	35833.4	34029.2	34005.4
ValidPart3	42178.1	34799.3	34607.6
ValidPart4	41178.4	44823.1	44870.5
ValidPart5	45142.9	48603.8	48546.5

Table 2 RMSE for EQRM, GLM and HGLM model

Model	Cross–validation RMSE
EQRM	42 154.7
GLM	41 299.5
HGLM	41 238.7

Table 3 Cross–validation RMSE for EQRM, GLM and HGLM model

For the analyzed data set the lowest error cv obtained HGLM model. Therefore in his case, for further calculations of tariff rates this model should be used. Using cross–validation procedure gives a rather demonstrative result that may be a prelude to further analysis and verification of models. The problem lies in the selection of unified tests that would allow the final choice of the method for rate–making.

6 Conclusions

In this paper we presented briefly the quantile regression and the generalized regression models. After that we analyzed the capabilities of application both models in insurance rate–making process and we computed the numerical example. In order to estimate models and than realized the cross–validation procedure, the computer implementation of some algorithms was necessary.

There are few reasons for which we tested the capabilities of the quantile regression in rate-making, a specially several important statistical advantages. Firstly a distribution of error terms is left unspecified, which is the main virtue of the method as far as robustness to outliers is concerned. Secondly quantile estimates detect the influence of covariates on alternate parts of the conditional distribution, which we can choose arbitrarily (various orders of quantile). Thus it can be recommended in cases of non–normal asymmetric distributions – asymmetric or fat–tailed distributions. Thirdly there is the possibility to take into consideration the polices with no claims by τ parameter, which is impossible for HGLM Gamma–Gamma model. That is why in our empirical example the base premium and tariff rates are higher for EQRM than in GLM and HGLM.

Acknowledgements

This research is supported by the grant of Polish Ministry of Science and Higher Education (no. NN111461540).

References

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

[2] Koenker, R., and Basset, B.: Regression Quantiles, *Econometrica*, **46** (1978), 33–50.

[3] Koenker, R.: *Quantile regression*. Cambridge University Press, Cambridge, 2005.
[4] Koenker, R., and Hallock, K. F.: Quantile regression. *Journal of Economic Perspectives*, **15**(4) (2001), 143–156.
[5] Kudryavtsev, A. A.: Using quantile regression for rate–making. *Insurance: Mathematics and Economics* **45** (2009), 296–304.
[6] Lee, Y., and Nelder, A. J., and Pawitan, Y.: *Generalized Linear Models with Random Effects*, Monographs on Statistics and Applied Probability 106, Chapman & Hall\CRC, 2006.
[7] McCullagh, P., and Nelder, J. A.: *Generalized Linear Models*. Chapman & Hall/CRC, New York, 1999.
[8] McCulloch, Ch. E., and Searle, Sh. R.: *Generalized, Linear, and Mixed Models*, Wiley Series in Probability and Statistics, John Wiley & Sons, INC., 2001.
[9] Ohlsson, E., and Johansson, B.: *Non–Life Insurance Pricing with Generalized Linear Models*, Springer-Verlag, Berlin, 2010.
[10] Orwat–Acedańska, A., and Trzpiot, G.: Quantile regression in management Style Analysis of mutual balanced funds (in Polish). *Financial Investments and Insurances – World Trends and Polish Market*, University of Economics in Wrocław, **183** (2011a), 415–425.
[11] Orwat-Acedańska, A., and Trzpiot, G.: The classification of Polish mutual balanced funds based on the management style – quantile regression approach. *Theory and Applications of Quantitative Methods, Econometrics*, **31**(194) (2011b), 9–23.
[12] Portnoy, S., and Koenker, R.: The Gaussian Hare and the Laplacian Tortoise: Computability of Squared-Error Versus Absolute–Error Estimators, with Discussion, *Statistical Science*, **12** (1997).
[13] Trzpiot, G.: The Implementation of Quantile Regression Methodology in VaR Estimation (in Polish). *Studies and Researches of Faculty of Economics and Management University of Szczecin*, Szczecin, (2008), 316–323.
[14] Trzpiot, G.: Quantile Regression Model versus Factor Model Estimation. *Financial Investments and Insurances – World Trends and Polish Market* **60** (2009a), University of Economics in Wrocław, 469–479.
[15] Trzpiot, G.: Application weighted VaR in capital allocation. *Polish Journal of Environmental Studies*, Olsztyn, **18**, 5B, (2009b), 203–208.
[16] Trzpiot, G.: Estimation methods for quantile regression, *Economics Studies* **53** (2009c), Karol Adamiecki University of Economics in Katowice, 81–90.
[17] Trzpiot, G.: Quantile Regression Model of Return Rate Relation – Volatility for Some Warsaw Stock Exchange Indexes (in Polish). *Finances, Financial Markets and Insurances. Capital Market*, University of Szczecin, **28** (2010), 61–76.
[18] Trzpiot, G.: Bayesian Quantile Regression, *Economics Studies* **65** (2011), Karol Adamiecki University of Economics in Katowice, 33–44.
[19] Wolny–Dominiak, A.: Rate-making in Automobile Insurance with HGLM (in Polish), *Proceedings of I Scientific Spatial Econometrics and Regional Economic Analysis,* Lodz, 2010.
[20] Wolny-Dominiak, A., and Trzęsiok, M.: Monte Carlo Simulation Applied to A'Priori Rate Making, *Proceedings of 26th International Conference Mathematical Methods in Economics*, Liberec, 2008.