ALICJA WOLNY-DOMINIAK
GRAŻYNA TRZPIOT

GLM and quantile regression models in *a priori* ratemaking[1]

Classification a priori *ratemaking in non-life insurance applies a different type of multivariate regression models, which is more sensitive to the assumptions that significantly restrict the area of their applications. When an error term is non-Gaussian, asymmetric, fat-tailed or in the presence of outliers, it may have serious consequences for the correct inference of the factor's impact on an endogenous variable. In this paper we analyze two types of regression, which take into account the mentioned problems. The first regression is based on the GLM technique while the second used the modified quantile regression technique. Since the quantile regression is a non-parametric method, there is no measure of the relative quality of the model. For this reason, we propose the cross-validation procedure to compare two models and choose the optimal in terms of minimum cross-validation error.*

Key words: Casualty/Property insurance, *a priori* ratemaking, GLMs, quantile regression, cross-validation.

Introduction

The ratemaking process is one of the most important factors in issues regarding insurance portfolios. The techniques of ratemaking are actually based on loss distribution or their moments, which are estimated using historical data. The key challenge is to choose the correct model for the estimation of loss value. Ratemaking of insurance portfolios is frequently based on different multivariate regression models which allow the investigation of rating factors[2]. Nevertheless, the ordinary multivariate regression model has some crucial disadvantages – it is sensitive to assumptions that significantly restrict the area of their applications. In an insurance data case, when an error

1. This research is supported by a grant from the Polish Ministry of Science and Higher Education (no. NN 111461540).

2. K. Antonio, and E. Valdez, "Statistical concepts of *a priori* and a posteriori risk classification in insurance," Volume 96 of *AStA Advances in Statistical Analysis* 2 (2012).

term is non-normal, asymmetric, fat-tailed or in the presence of outliers, it may have serious consequences for the correct inference of the factor's impact on an endogenous variable. Moreover, the ordinary multivariate regression model often ignores a specific feature of the insurance data such as: the possibility of catastrophic losses, the dependence of insured objects on each other (i.e., cumulative risk) or information shortfall to verify the statistical significance of the model chosen[3]. Therefore, it is important to use models and estimators that are more robust to restrictive classical regression assumptions for modeling insurance data. GLM is a good example of such a model and is therefore used by actuaries[4]. However, there are some problems connected with GLM. The main problem lies in choosing the predictors' distribution in GLM. This can be solved with the simulation procedure based on the Monte Carlo method[5]. The other approach proposed for modeling insured portfolios of policies is the quantile regression approach[6] . This is consistent with the idea of using the distribution quantile for ratemaking. The additional advantage of this method is the fact that it allows the estimation of the net premium rates including safety loadings and it may be estimated as a quantile of loss distribution.

In this paper we present two regression models in *a priori* ratemaking: model GLM and model EQRM. As there is no proper measure to compare GLM and EQRM models, we propose the cross-validation procedure based on RMSE error. In the case study we analyze an example of motor insurance portfolio taken from literature[7]. All computations are performed with the statistical software R.

1. Classical *a priori* ratemaking – model GLM

Nowadays, classical statistical techniques used in *a priori* ratemaking are GLM models. Let $Y_{1,}$ $E,$ Y_n be independent random variables with $y_{1,}$ $E,$ y_n realizations, where Y indicates the claim severity of i-th policy in an insurance portfolio. Further let us denote k risk factors as $X_1 EX_A$ and assume that Y follows three-parameter Tweedie distribution $Y \sim T(\mu, \varphi, p)$[8]. The first parameter φ is the dispersion parameter and the second parameter φ is the power in the variance of Y_i:

$$Var(Y_i) = \varphi \mu_i^p, i = 1, \ldots, n \qquad (1)$$

3. R. Koenker, and K. F. Hallock, "Quantile regression," *Journal of Economic Perspectives* 15(4) (2001).
4. P. McCullagh, and J. A. Nelder, Generalized Linear Models, New York: Chapman & Hall/CRC, 1999. Ohlsson, E., and B. Johansson, "Non-Life Insurance Pricing with Generalized Linear Models," Berlin: Springer–Verlag, 2010. De Jong P., and G. Z. Heller. "Generalized Linear Models for Insurance Data," Cambridge: Cambridge University Press, 2008.
5. A. Wolny-Dominiak, and M. Trzęsiok, "Monte Carlo Simulation Applied To A Priori Ratemaking," In Proceedings of 26th International Conference on Mathematical Methods in Economics, Liberec, 2008.
6. R. Koenker, and B. Basset, "Regression Quantiles," Econometrica 46 (1978). Koenker, R. Quantile regression. Cambridge: Cambridge University Press, 2005. R. Koenker, and K. F. Hallock, "Quantile regression," Journal of Economic Perspectives 15(4) (2001).
7. E. Ohlsson, and B. Johansson, "Non-Life Insurance Pricing with Generalized Linear Models," Berlin: Springer–Verlag, 2010.
8. B. Jørgensen, and M. De Souza, "Fitting Tweedie's compound Poisson model to insurance claims data," *Scandinavian Actuarial Journal* 1 (1994).

The proper GLM model for *a priori* ratemaking has the following components:

$$\begin{cases} Y_i \sim T(\mu_i, \varphi, p) \\ \eta_i = \mathbf{x}_i^T \boldsymbol{\beta} \\ \mu_i = \exp(\eta_i) \end{cases} \tag{2}$$

where $\mathbf{x}_i$ denotes the *i*-th row of the model matrix X for the *i*-th policyholder and $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)^T$ denotes the vector for fixed effects estimated directly from the data. The link function is taken as $\log(.)$ since the multiplicative model is usually applied.

In order to estimate the model (2), we propose a two-step procedure. In the first step, the parameters φ and p are estimated using maximum likelihood estimation and Fourier inversion[9]. In the second step, classical IWSL algorithm to estimate the vector $\boldsymbol{\beta}$ is used with $\hat{\varphi}$ and $\hat{p}$ plug-ins.

Based on the results of *a priori* ratemaking, the base premium for the whole portfolio and the tariff rate for *i*-th policyholder can be calculated:

$$\begin{cases} B = \exp(\hat{\beta}_0) \\ t_i = \exp(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}) \end{cases} \tag{3}$$

The indicator t_i shows the increase or decrease of the base premium B for *i*-th policyholder.

2. Alternative *a priori* ratemaking – model EQRM

The alternative approach for *a priori* ratemaking is the distribution-free quantile regression model[10]. Similarly as in the GLM model, the goal of the quantile regression model is to estimate vector $\boldsymbol{\beta}$ for a sample of realizations $y_1 \ldots y_n$ of a sequence of independent random variables $Y_{1,E,}Y_n$. The basic assumption is that random variables $Y_{1,E,}Y_n$ are taken with distribution $P(Y_i < y) = \Im(y - \mathbf{x}_i^T \boldsymbol{\beta})$ and the distribution $\Im$ is unknown. The linear quantile regression model of order τ, $0 < \tau < 1$ is given by the formula:

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

where $Q_\tau(Y_i | \mathbf{x}_i)$ indicates the conditional quantile of random variable Y_i for probability τ[11].

To apply the linear model (4) in *a priori* ratemaking and take into account the multiplicative model, some modification is necessary. Assuming logarithmic transformation of the conditional quantile $Q_\tau(Y_i | \mathbf{x}_i)$, we receive the exponential quantile regression model (EQRM) of order τ^* of the form:

$$Q_{\tau^*}(Y_i | \mathbf{x}_i) = \exp(\mathbf{x}_i^T \boldsymbol{\beta}^{(\tau^*)}) \tag{5}$$

where $Q_{\tau^*}(Y_i | \mathbf{x}_i)$ indicates the conditional quantile of random variable Y_i for probability τ^*, $0 < \tau^* < 1$, and $\boldsymbol{\beta}^{(\tau^*)} = (\beta_0^{(\tau^*)}, \beta_1^{(\tau^*)}, \beta_2^{(\tau^*)}, \ldots, \beta_m^{(\tau^*)})^T$ is the vector of parameters of order τ^* (see para. 6).

9. P. K. Dunn, and G. K. Smyth, "Evaluation of Tweedie exponential dispersion model densities by Fourier inversion." *Statistics and Computing* 18,1 (2008).
10. A. A. Kudryavtsev, "Using quantile regression for ratemaking," *Insurance: Mathematics and Economics* 45 (2009).
11. R. Koenker, and B. Basset, "Regression Quantiles," *Econometrica* 46 (1978).

According to para. 8, we define the τ^*-th quantile regression estimator of $\boldsymbol{\beta}^{(\tau^*)}$ as the vector b being the solution of the minimization problem:

$$\min_{\mathbf{b}\in\mathfrak{R}^{m+1}}\left[\sum_{i\in\{i:\,y_i\geq\mathbf{x}_i^T\mathbf{b}\}}\tau^*\left|y_i-\exp(\mathbf{x}_i^T\mathbf{b})\right|+\sum_{i\in\{i:\,y_i<\mathbf{x}_i^T\mathbf{b}\}}(1-\tau^*)\left|y_i-\exp(\mathbf{x}_i^T\mathbf{b})\right|\right] \tag{6}$$

Because the error distribution term is unspecified, statistical inference is based on nonparametric approach – the bootstrap or Monte Carlo method.

The results of *a priori* ratemaking are as with a classical approach: the base premium for the whole portfolio and the tariff rate for i-th policyholder:

$$\begin{cases} B^* = \exp(\hat{\beta}_0^{(\tau^*)}) \\ t_i^* = \exp(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}^{(\tau^*)}) \end{cases} \tag{7}$$

3. Cross-validation procedure

In order to unify the process of comparing the classical and alternative approach in *a priori* ratemaking, we propose applying the cross-validation procedure based on RMSE error[12]. In our case study we use a 5-fold cross-validation algorithm for models (2) and (5). The procedure is as follows:

(s1) randomly divide the training set into $k = 5$ approximately equally sized parts, (n – the training set size, m- the size of the l-th subset, $l = 1, \ldots, 5$)

(s2) build every model 5 times using 4 of 5 parts ($n - m_l$ observations), treating excluded observations as the validation set,

(s3) 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,

(s4) estimate the cross-validation error: $cv = \sum_{i=1}^{5}\frac{m_l}{n}RMSE_l$

The model with the smallest *cv* value is taken as a better estimation of the base premium in the portfolio and tariff rates.

4. Case study – automobile insurance portfolio

In order to illustrate the process of *a priori* ratemaking with GLM and EQRM models, we considered a motor insurance portfolio from the former Swedish insurance company Wasa, which concerns partial motor hull insurance, for motorcycles[13]. In the model we complied the following rating variables:

12. S. Portnoy, and R. Koenker, "The Gaussian Hare and the Laplacian Tortoise: Computability of Squared– Error Versus Absolute-Error Estimators, with Discusssion," Statistical Science 12 (1997).
13. E. Ohlsson, and B. Johansson, "Non-Life Insurance Pricing with Generalized Linear Models," Berlin: Springer–Verlag, 2010.

Rating variable	Rating variable (origin name)	Rating variable (description)
X_1	Agarald	owner's age
X_2	Kon	gender
X_3	Mcklass	MC class, a classification by the so called EV ratio, defined as (Engine power in kW × 100) / (Vehicle weight in kg + 75), rounded to the nearest lower integer. The 75 kg represents the average driver's weight.
X_4	Fordald	vehicle age
E	Duration	number of policy years (exposure)
Ω	Antskad	number of claims
Y	Skadkost	claim cost

Source: the author's own research.

Firstly, we considered the GLM model (2), and the parameters p and ϕ were estimated with R package {tweedie}. Following that, the GLM model was constructed with the following components:

$$\begin{cases} Y_i \sim T(\mu_i, \hat{\varphi} = 0.008, \hat{p} = 1.63) \\ \eta_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} \\ \mu_i = \exp(\eta_i) \end{cases} \tag{8}$$

Finally, we estimated the model (8) using an IWSL algorithm as well as the model (5) by minimization of the problem (6). The computations were performed with two R Packages: {stats} for GLM and {quantreg} for EQRM models.

At the initial stage of the estimation of all rating, variables $X_1, \ldots, X_4$ were included in the model, but only the "owner's age" proved to be statistically significant. Ultimately, in GLM and EQRM models only one rating variable was introduced. The results of the estimation are shown in Tab. 1 and Tab 2.

Table 1. Model parameters for GLM and EQRM models

	Estimate	s.e.	p-value	Estimate	s.e.	p-value
Intercept	9,44	0,28	0,00	10,47	0,45	0,00
The owner's age B	0,67	0,31	0,03	0,88	0,46	0,06
The owner's age C	1,10	0,31	0,00	1,26	0,51	0,01
The owner's age D	0,95	0,33	0,00	1,51	0,51	0,00
The owner's age E	0,66	0,31	0,03	1,29	0,53	0,02
The owner's age F	0,78	0,33	0,02	1,38	0,64	0,03
The owner's age G	–0,12	0,53	0,82	0,37	0,58	0,53

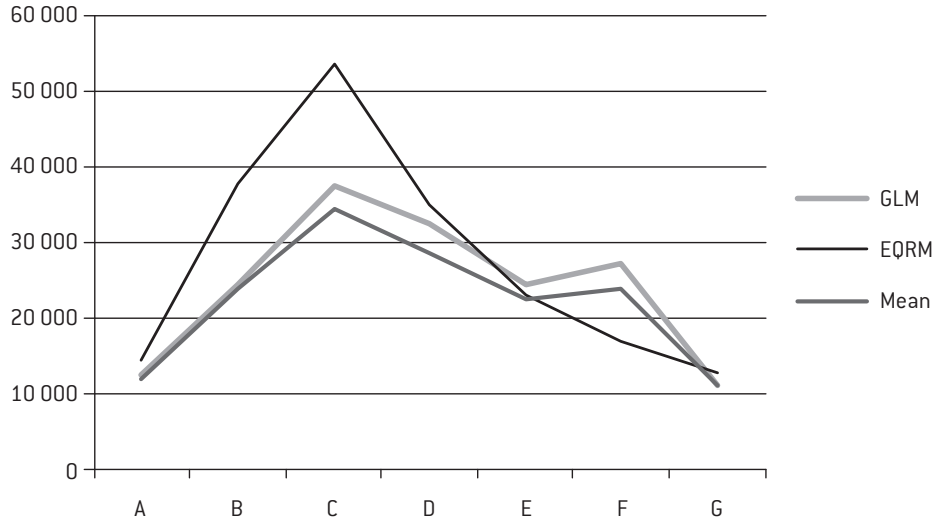
Source: the author's own research.

Table 2. The summary of the estimation of claim costs for the owner's age

The owner's age	The numbers of policies in the portfolio	Classes of the owner's age	GLM	EQRM	Mean Claim cost
17–20	1658 (2.66%)	*Base claim cost* A	12 536	14 472	11 956
21–25	5831 (9.36%)	B	24 408	37 798	23 874
26–30	7311 (11.73%)	C	37 578	53 637	34 421
31–40	9997 (16.04%)	D	32 500	34 892	28 681
41–50	19258 (30.9%)	E	24 292	23 156	22 516
51–50	13521 (21.7%)	F	27 221	16 815	23 827
61–	4746 (7.62%)	G	11 077	12 708	11 077

Source: the author's own research.

Figure 1. Estimated claim costs for the owner's age – the comparison



Source: the author's own research.

Since only one rating variable is statistically significant, we compare the estimated claim cost in the portfolio with the mean value in groups. In both analyzed models, the base claim cost is relatively higher than the mean claim cost. The similar situation is in the case of the fitted claim costs obtained in the GLM model. We also observe a large discrepancy in the results obtained in EQRM model when compared to the mean. In the preliminary comparison, both GLM and EQRM models (Fig. 1) clearly show that claim costs estimated by the GLM model are closer to mean costs than in the EQRM. 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 as follows:

Table 3. RMSE for GLM and EQRM models

Validation set	RMSE GLM	RMSE EQRM
ValidPart1	44 242,10	46 440,60
ValidPart2	34 029,20	35 833,40
ValidPart3	34 799,30	42 178,10
ValidPart4	44 823,10	41 178,40
ValidPart5	48 603,80	45 142,90

Source: the author's own research.

Table 4 Cross-validation GLM and EQRM model

Model	*Cross-validation RMS…*
GLM	41 299,5
EQRM	42 154,7

Source: the author's own research.

For the analyzed portfolio, the lowest cv error was obtained for the GLM model. Therefore, in this case, for further calculations of tariff rates and net premiums for the *i*-th policy, the GLM model should be used. Using the cross-validation procedure gives fairly demonstrative results that may be a prelude to further analysis and verification of the models. The problem lies in the selection of unified tests that would allow the final choice of the method for *a priori* ratemaking.

5. Conclusions

Nowadays, GLMs are standard industry practice for *a priori* ratemaking. These models extend the ordinary linear models to the class of the exponential dispersion family of distributions. However, problems with wrong-fitted distribution can still occur. That is why we tested the capabilities of the quantile regression in ratemaking. Firstly, the distribution of error terms is left unspecified – this is the main virtue of the method as far as robustness to outliers is concerned. Secondly, quantile estimates detect the influence of co-variates on alternate parts of the conditional distribution, which we can choose arbitrarily (by using various orders of quantile). Thus, quantile regression can be recommended in cases of non-normal asymmetric distributions – asymmetric or fat-tailed distributions. Despite these advantages, the GLM model can still be the better solution. A useful technique for a model selection is the cross-validation procedure.

Quantile regression is becoming more and more popular in practice, especially in finance theory. We suspect that it could also be a very useful tool in the insurance business[14]. We note that

14. A.A. Kudryavtsev, "Using quantile regression for ratemaking," *Insurance: Mathematics and Economics* 45 (2009).

the distribution-free approach is often used for estimation[15]. Applications of quantile regression for the Polish capital market can be found in w papers[16].

References

Antonio K., and Valdez E., "Statistical concepts of *a priori* and a posteriori risk classification in insurance," Volume 96 of *AStA Advances in Statistical Analysis* 2 (2012).

De Jong P., and Heller G. Z., "Generalized Linear Models for Insurance Data," Cambridge: Cambridge University Press, 2008.

Dunn P., K., and Smyth G. K., "Evaluation of Tweedie exponential dispersion model densities by Fourier inversion," *Statistics and Computing* 18,1 (2008).

Jørgensen, B., and De Souza M., "Fitting Tweedie's compound Poisson model to insurance claims data," *Scandinavian Actuarial Journal* 1 (1994).

Koenker, R., and Basset B., "Regression Quantiles," *Econometrica* 46 (1978).

Koenker, R., "Quantile regression", Cambridge: Cambridge University Press, 2005.

Koenker, R., and Hallock K. F., "Quantile regression," *Journal of Economic Perspectives* 15(4) (2001).

Kudryavtsev, A. A., "Using quantile regression for ratemaking," *Insurance: Mathematics and Economics* 45 (2009).

McCullagh, P., and Nelder J. A., "Generalized Linear Models," New York: Chapman & Hall/CRC, 1999.

Ohlsson, E., and Johansson B., "Non-Life Insurance Pricing with Generalized Linear Models", Berlin: Springer–Verlag, 2010.

Orwat-Acedańska, A., and Trzpiot G., "Quantile Regression in management style analysis of mutual balanced funds," *Financial Investments and Insurance – World Trends and the Polish Market* 183 (2011a).

Orwat-Acedańska, A., and Trzpiot G., "The classification of Polish mutual balanced funds based on management style – quantile regression approach," *Theory and Applications of Quantitative Methods, Econometrics* 31, 194 (2011b).

Portnoy, S., and Koenker R., "The Gaussian Hare and the Laplacian Tortoise: Computability of Squared–Error Versus Absolute-Error Estimators, with Discusssion," *Statistical Science* 12 (1997).

Trzpiot, G., "Quantile Regression Model of Return Rate Relation – Volatility for Some Warsaw Stock Exchange Indexes," *Finances, Financial Markets and Insurance. Capital Market* 28 (2010).

Trzpiot, G., "Bayesian Quantile Regression," Studia Ekonomiczne, *Zeszyty Naukowe* 65 (2011): 33–44.

Wolny-Dominiak, A., and Trzęsiok M., "Monte Carlo Simulation Applied To *A Priori* Ratemaking," in *Proceedings of 26th International Conference on Mathematical Methods in Economics*, Liberec, 2008.

15. R. Koenker, "Quantile regression," Cambridge: Cambridge University Press, 2005. R. Koenker and K. F. Hallock. "Quantile regression." *Journal of Economic Perspectives* 15(4) (2001).

16. A. Orwat-Acedańska and G. Trzpiot, "The classification of Polish mutual balanced funds based on management style – quantile regression approach," Theory and Applications of Quantitative Methods, *Econometrics* 31, 194 (2011b). G. Trzpiot, "Quantile Regression Model of Return Rate Relation – Volatility for Some Warsaw Stock Exchange Indexes," *Finances, Financial Markets and Insurance. Capital Market* 28 (2010). Trzpiot, G. "Bayesian Quantile Regression," Studia Ekonomiczne, *Zeszyty Naukowe* 65 (2011): 33–44.

Modele GLM i regresji kwantylowej w taryfikacji *a priori*

W procesie taryfikacji a priori *w ubezpieczeniach majątkowych wykorzystywane są głównie modele regresyjne klasy GLM, w których przyjmowane jest założenie odnośnie zmiennej objaśnianej umożliwiające przyjęcie w modelu innego rozkładu prawdopodobieństwa niż jedynie rozkład normalny. Zatem rodzi się problem wyboru rozkładu zakładanego w modelu. W niniejszym artykule rozpatrujemy możliwość zastosowania regresji kwantylowej, w której nie zakłada się żadnej postaci rozkładu, co eliminuje wspomniany wyżej problem. Rozważamy zarówno model GLM jak również model zmodyfikowanej regresji kwantylowej dla portfela polis ubezpieczeniowych. Jako że regresja kwantylowa jest modelem nieparametrycznym, nie zdefiniowano miary będącej odpowiednikiem kryterium AIC w modelu GLM. Powoduje to trudności w porównywaniu modeli, a dalej w ostatecznym wyborze modelu do taryfikacji. Dlatego w pracy proponujemy zastosowanie procedury kroswalidacji w celu porównania modeli GLM oraz regresji kwantylowej i dalej wyboru modelu lepszego tzn. takiego, który daje mniejszy błąd cv.*

Słowa kluczowe: ubezpieczenia majątkowe, taryfikacja *a priori*, model GLM, regresja kwantylowa, kroswalidacka.

ALICJA WOLNY-DOMINIAK, Ph.D. – University of Economics in Katowice, Department of Statistical and Matematical Methods in Economics.

PROF. GRAŻYNA TRZPIOT – University of Economics in Katowice, Department of Demography and Economic Staticstics.