Risk Classification for Claim Counts and Losses Using Regression Models for Location, Scale and Shape

George Tzougas[1], Spyridon Vrontos[2] and Nicholas Frangos[3]
Department of Statistics, School of Mathematical Sciences,
University College Cork[1]
Department of Mathematical Sciences, University of Essex[2]
Department of Statistics, Athens University of
Economics and Business[3]

April 25, 2016

1 Introduction

- The idea behind a priori risk classification is to split an insurance portfolio into classes that consist of risks with all policyholders belonging to the same class paying the same premium. In view of the economic importance of motor third party liability (MTPL) insurance in developed countries, actuaries have made many attempts to find a probabilistic model for the distribution of the number and costs of claims reported by policyholders.

- Recent actuarial literature research assumes that the risks can be rated a priori using generalized linear models, GLM (see Nelder and Wedderburn, 1972) and generalized additive models, GAM (see Hastie and Tibshirani, 1990). For motor insurance, typical response variables in these regression models are the number of claims (or claim frequency) and its corresponding severity. The models briefly described above assume that only the mean is modelled as a function of risk factors. However, any model for the mean in terms of a priori rating variables indirectly yields a model for scale and/or shape. Also, even if the mean is the most commonly used measure of the expected claim frequency and of the expected claim severity it does not provide a good description of a distribution's scale and shape. Specifically, the scale and shape parameters are not adequately described due to the unobserved heterogeneity changes with explanatory variables.

- In this study, we extend this setup by assuming that all the parameters of the claim frequency/severity distributions can be modelled as functions of explanatory variables with parametric linear functional forms. Joint modelling of all the parameters in terms of covariates improves rate making and estimation of the scale and shape of the claim frequency/ severity distributions. Specifically, we model the claim frequency using the Negative Binomial Type II, Sichel and Zero-Inflated Poisson models and the claim severity using the Gamma, Weibull, and Generalized Pareto models. Our contribution puts focus on the comparison of these models through their variance values and not only the mean values as usually considered in risk classification literature. To the best of our knowledge, it is the first time that the variance of the claim frequency and severity is modelled in the context of ratemaking.

- The differences between these models are analyzed through the mean and the variance of the annual number of claims and the costs of claims of the policyholders who belong to different risk classes, which are formed by dividing the portfolio into clusters defined by the relevant ratemaking factors. Finally, the resulting premium rates are calculated via the expected value and standard deviation principles with independence between the claim frequency and severity components assumed.

2 Regression Models for Location, Scale and Shape

2.1 Frequency Component

- Consider a policyholder i whose number of claims, denoted as K_i, are independent, for $i = 1, .., n$. The probability that the policyholder i has reported k claims to the insurer, $k = 0, 1, 2, ...$, is denoted by $P\left(K_i = k\right)$.

- The pdf of Negative Binomial Type II (NBII) distribution is given by

$$P\left(K_i = k\right) = \frac{\Gamma\left(k + \frac{\mu}{\sigma}\right)\sigma^k}{\Gamma\left(\frac{\mu}{\sigma}\right)\Gamma\left(k+1\right)\left[1+\sigma\right]^{k+\frac{\mu}{\sigma}}}, \tag{1}$$

for $\mu > 0$ and $\sigma > 0$. Following Rigby and Stasinopoulos (2005 and 2009), we assume that $\mu_i = \exp(c_{1i}\beta_1)$ and $\sigma_i = \exp(c_{2i}\beta_2)$, where $c_{ji}\left(c_{ji,1}, ..., c_{ji,J_j^{/}}\right)$ and $\beta_j^T\left(\beta_{j,1}, ..., \beta_{j,J_j^{/}}\right)$ are the $1 \times J_j'$ vectors of the a priori rating variables and the coefficients respectively, for $j = 1, 2$. The mean and the variance of K_i are given by

$$E(K_i) = \exp(c_{1i}\beta_1) \tag{2}$$

and

$$Var(K_i) = \exp(c_{1i}\beta_1)\left[1 + \exp(c_{2i}\beta_2)\right]. \tag{3}$$

- The pdf of the Sichel distribution is given by

$$P\left(K_i = k\right) = \frac{\left(\frac{\mu}{c}\right)^k K_{k+\nu}\left(a\right)}{k!\left(a\sigma\right)^{k+\nu} K_\nu\left(\frac{1}{\sigma}\right)}, \tag{4}$$

where $\sigma > 0$ and $-\infty < \nu < \infty$ and where $c = \frac{K_{\nu+1}\left(\frac{1}{\sigma}\right)}{K_{\nu}\left(\frac{1}{\sigma}\right)}$, where

$$K_{\nu}(z) = \frac{1}{2}\int_{0}^{\infty} x^{\nu-1} \exp\left[-\frac{1}{2}z\left(x + \frac{1}{x}\right)\right] dx, \tag{5}$$

is the modified Bessel function of the third kind of order ν with argument z and where $a^2 = \sigma^{-2} + 2\mu\,(c\sigma)^{-1}$. Following Rigby and Stasinopoulos (2008 and 2009), we assume that $\mu_i = e_i \exp(c_{1i}\beta_1)$, $\sigma_i = \exp(c_{2i}\beta_2)$ and $\nu_i = c_{3i}\beta_3$, where $c_{ji}\left(c_{ji,1}, \ldots, c_{ji,J_j^{/}}\right)$ and $\beta_j^T\left(\beta_{j,1}, \ldots, \beta_{j,J_j^{/}}\right)$ are the $1 \times J_j'$ vectors of the a priori rating variables and the coefficients respectively, for $j = 1, 2, 3$. The mean and variance of K_i are given by

$$E(K_i) = \exp(c_{1i}\beta_1) \tag{6}$$

and

$$\begin{aligned} Var(K_i) &= \exp\left(c_{1i}\beta_1\right) \qquad (7)\\ &\quad + \left[\exp\left(c_{1i}\beta_1\right)\right]^2 \left\{ \frac{2\exp\left(c_{2i}\beta_2\right)\left[c_{3i}\beta_3+1\right]}{c_i} + \frac{1}{c_i^2} - 1 \right\}, \end{aligned}$$

where $c_i = \frac{K_{c_{3i}\beta_3+1}\left(\frac{1}{\exp(c_{2i}\beta_2)}\right)}{K_{c_{3i}\beta_3}\left(\frac{1}{\exp(c_{2i}\beta_2)}\right)}$.

- The pdf of the Zero-Inflated Poisson (ZIP) distribution is given by

$$P\left(K_i = k\right) = \begin{cases} \pi + (1-\pi)\, e^{-\mu}, \text{ if } k = 0 \\ (1-\pi)\, \frac{e^{-\mu}\mu^k}{k!}, \text{ if } k = 1, 2, 3, \ldots \end{cases} \qquad (8)$$

Following Rigby and Stasinopoulos (2005 and 2009), we assume that

$\mu_i = e_i \exp\left(c_{1i}\beta_1\right)$ and $\pi_i = \frac{\exp(c_{2i}\beta_2)}{1+\exp(c_{2i}\beta_2)}$, where $c_{ji}\left(c_{ji,1}, ..., c_{ji,J_j^/}\right)$ and $\beta_j^T\left(\beta_{j,1}, ..., \beta_{j,J_j^/}\right)$ are the $1 \times J_j'$ vectors of the a priori rating variables and the coefficients respectively, for $j = 1, 2$. The mean and the variance of K_i are given by

$$E(K_i) = e_i \exp\left(c_{1i}\beta_1\right)\left[1 - \exp\left(c_{2i}\beta_2\right)\right] \tag{9}$$

and

$$\begin{aligned} Var(K_i) &= e_i \exp\left(c_{1i}\beta_1\right)\left[1 - \exp\left(c_{2i}\beta_2\right)\right] \\ &\quad \left[1 + e_i \exp\left(c_{1i}\beta_1\right)\exp\left(c_{2i}\beta_2\right)\right]. \end{aligned} \tag{10}$$

2.2 Severity Component

- Let $X_{i,k}$ be the cost of the kth claim reported by policyholder i, $i = 1, ..., n$ and assume that the individual claim costs $X_{i,1,}X_{i,2}, ...$ are independent and identically distributed (i.i.d).

- The pdf of the Gamma distribution is given by

$$f\left(x\right) = \frac{1}{\left(s^2 m\right)^{\frac{1}{s^2}}} \frac{x^{\frac{1}{s^2}-1} \exp\left(-\frac{x}{s^2 m}\right)}{\Gamma\left(\frac{1}{s^2}\right)}, \tag{11}$$

for $X_{i,k} > 0$, where $m > 0$ and $s > 0$. Following Rigby and Stasinopoulos (2009), we assume that $m_i = \exp\left(d_{1i}\gamma_1\right)$ and $s_i = \exp\left(d_{2i}\gamma_2\right)$, where $d_{ji}\left(d_{ji,1}, ..., d_{ji,J_j^/}\right)$ and $\gamma_j^T\left(\gamma_{j,1}, ..., \gamma_{j,J_j^/}\right)$ are the $1 \times J_j'$ vectors of the exogenous variables and the coefficients respectively, for $j = 1, 2$. The mean and the variance of $X_{i,k}$ are given by

$$E(X_{i,k}) = \exp\left(d_{1i}\gamma_1\right) \tag{12}$$

and

$$Var(X_{i,k}) = \left[\exp\left(d_{2i}\gamma_2\right)\right]^2 \left[\exp\left(d_{1i}\gamma_1\right)\right]^2 . \tag{13}$$

- The pdf of the Weibull distribution is given by

$$f\left(x\right) = \frac{sx^{s-1}}{m^s} \exp\left[-\left(\frac{x}{m}\right)^s\right], \tag{14}$$

where $m > 0$ and $s > 0$. Following Rigby and Stasinopoulos (2009), we assume that $m_i = \exp\left(d_{1i}\gamma_1\right)$ and $s_i = \exp\left(d_{2i}\gamma_2\right)$, where $d_{ji}\left(d_{ji,1}, ..., d_{ji,J_j^/}\right)$

and $\gamma_j^T \left(\gamma_{j,1}, ..., \gamma_{j,J_j^{/}}\right)$ are the $1 \times J_j'$ vectors of the exogenous variables and the coefficients respectively, for $j = 1, 2$. The mean and the variance of $X_{i,k}$ are given by

$$E(X_{i,k}) = \exp\left(d_{1i}\gamma_1\right)\Gamma\left(\frac{1}{\exp\left(d_{2i}\gamma_2\right)} + 1\right) \tag{15}$$

and

$$\begin{aligned} Var(X_{i,k}) &= \left[\exp\left(d_{1i}\gamma_1\right)\right]^2 \\ &\left\{\Gamma\left(\frac{2}{\exp\left(d_{2i}\gamma_2\right)} + 1\right) - \left[\Gamma\left(\frac{1}{\exp\left(d_{2i}\gamma_2\right)} + 1\right)\right]^2\right\}. \end{aligned} \tag{16}$$

- The pdf of the Generalized Pareto distribution is given by

$$f\left(x\right) = \frac{\Gamma\left(n+t\right)}{\Gamma\left(n\right)\Gamma\left(t\right)} \frac{m^t x^{n-1}}{\left(x+m\right)^{n+t}}, \tag{17}$$

where $m > 0$, $n > 0$ and $t > 0$. Following Rigby and Stasinopoulos (2008), we assume that $m_i = \exp(d_{1i}\gamma_1)$, $n_i = \exp(d_{2i}\gamma_2)$ and $t_i = \exp(d_{3i}\gamma_3)$, where $d_{ji}\left(d_{ji,1}, ..., d_{ji,J_j^/}\right)$ and $\gamma_j^T\left(\gamma_{j,1}, ..., \gamma_{j,J_j^/}\right)$ are the $1 \times J_j'$ vectors of the exogenous variables and the coefficients respectively, for $j = 1, 2, 3$. The mean and the variance of $X_{i,k}$ are given by

$$E(X_{i,k}) = \frac{\exp(d_{1i}\gamma_1)\exp(d_{2i}\gamma_2)}{\exp(d_{3i}\gamma_3) - 1} \tag{18}$$

and

$$\begin{aligned} Var(X_{i,k}) &= \frac{[\exp(d_{1i}\gamma_1)]^2\exp(d_{2i}\gamma_2)}{\exp(d_{3i}\gamma_3) - 1} \\ &\quad \left\{\frac{\exp(d_{2i}\gamma_2) + \exp(d_{3i}\gamma_3) - 1}{[\exp(d_{3i}\gamma_3) - 1][\exp(d_{3i}\gamma_3) - 2]}\right\}. \end{aligned} \tag{19}$$

3 Application

- The data were kindly provided by a Greek insurance company and concern a motor third party liability insurance portfolio observed during 3.5 years. The data set comprises 15641 policies. Both private cars and fleet vehicles have been considered in this sample. The available a priori rating variables we employ are the Bonus Malus (BM) class, the horsepower (HP) of the car and gender of the driver.

- The Bonus-Malus class consists of four categories: A, B, C and D, where: A = "drivers who belong to BM classes 1 and 2", B = "drivers who belong to BM classes 3-5", C ="drivers who belong to BM classes 6-9 & 11-20" and D = "drivers who belong to BM class 10". The horsepower of the car consists of three categories: A, B and C, where: A = "drivers who had a car with a HP between 0-33 & 100-132", B = "drivers who had a car with a HP between 34-66" and C = "drivers who had a car with a HP between 67-99". The gender consists of two categories: M= "male" and F = "female" drivers. Regarding the amount paid for each claim, there were 5590 observations that met our criteria. The Bonus-Malus class consists of three categories: A, B and C, where: A = "drivers who belong to BM classes 1 and 2", B = "drivers who belong to BM classes 3-5 & 6-9 & 11-20" and C = "drivers who belong to BM class 10". The horsepower of the car consists of four categories A, B, C and D, where: A = "drivers who had a car with a HP between 100-110 & 111-121 & 122-132", B = "drivers who had a car with a HP between 0-33 & 34-44 & 45-55 & 56-66", C = "drivers who had a car with a HP between 67-74" and D = "drivers who had a car with a HP between 75-82 & 83-90 & 91-99". Finally, the gender consists of three categories: M = "male", F = "female" and B = "both", since in this case, data for fleet vehicles used by either male or female drivers were also available, i.e. shared use.

- So far, we have several competing models for the claim frequency and severity components. The differences between models produce different premiums. Consequently, to distinguish between these models, this section compares them so as to select the best for each case.

- The resulting Global Deviance, AIC and SBC are given in Table 1 for the different claim frequency (Panel A) and claim severity (Panel B) fitted models.

Table 1: Models Comparison

Panel A: Claim Frequency Models				
Model	df	Global Deviance	AIC	SBC
NBII	11	28323.32	28345.32	28429.55
Sichel	11	28348.97	28370.97	28455.20
ZIP	12	28503.22	28527.22	28619.11
Panel B: Claim Severity Models				
Model	df	Global Deviance	AIC	SBC
Gamma	16	69665.05	69697.05	69803.11
WEI	16	70794.96	70826.96	70933.02
GP	22	69582.12	69526.12	69771.96

- Overall, with respect to the Global Deviance, AIC and SBC indices, from Panel A we observe the best fitted claim frequency model is the Negative

Binomial Type II model. From the claim severity models in Panel B we see that the best fitting performances are provided by the Generalized Pareto model.

- The final a priori ratemaking for the claim frequency models contains 24 classes. As expected, the variance of the NBII, Delaporte, Sichel and ZIP model exceeds the mean and these models allow for overdispersion. Furthermore, we observe that the biggest differences lie in the variance values of these models. For example, the in the following Table one can see the mean and the variance of the expected number of claims for a man who belongs to BM category A and has a car that belongs to HP category A, i.e. for the reference class, in the case of the NBII, Sichel and ZIP model respectively.

Table 2: A Priori Risk Classification Using Claim Frequency Models

Risk		NBII		Sichel		ZIP	
Class		Mean	Var	Mean	Var	Mean	Var
1	BMA, HPA, M	0.1267	0.2140	0.1258	0.1884	0.1261	0.1391
2	BMA, HPA, W	0.1357	0.1964	0.1377	0.2128	0.1414	0.1507
3	BMA, HPB, M	0.1001	0.1318	0.0984	0.1046	0.0983	0.1062
4	BMA, HPB, W	0.1072	0.1293	0.1078	0.1152	0.1102	0.1158
5	BMA, HPC, M	0.1178	0.1592	0.1166	0.1260	0.1148	0.1256
6	BMA, HPC, W	0.1261	0.1550	0.1277	0.1390	0.1288	0.1365
7	BMB, HPA, M	0.2385	0.4029	0.2383	0.4629	0.2742	0.2777
8	BMB, HPA, W	0.2555	0.3699	0.2610	0.5302	0.2527	0.2543
9	BMB, HPB, M	0.1885	0.2483	0.1863	0.2089	0.2136	0.2158
10	BMB, HPB, W	0.2020	0.2435	0.2040	0.2311	0.1969	0.1980
11	BMB, HPC, M	0.2217	0.2998	0.2208	0.2548	0.2496	0.2524
12	BMB, HPC, W	0.2375	0.2918	0.2418	0.2825	0.2300	0.2314

- The final a priori ratemaking for the claim severity models contains 36 classes. For example, the in the following Table one can see the mean and the variance of the expected cost of claims for a man who belongs to BM category A and has a car that belongs to HP category A in the case of the Gamma, WEI and Generalized Pareto model.

Table 3: A Priori Risk Classification Using Claim Severity Models

Risk		GA		WEI		GP	
Class		Mean	Var	Mean	Var	Mean	Var
1	BMA, HPA, B	584.00	135347.30	597.96	169637.36	583.03	142078.20
2	BMA, HPA, M	521.75	78621.46	526.73	110315.30	514.78	89891.64
3	BMA, HPA, W	543.92	82108.76	546.89	118812.19	536.75	95624.76
4	BMA, HPB, B	294.89	18453.33	295.51	19539.26	300.72	26138.91
5	BMA, HPB, M	263.46	10719.29	263.36	13061.64	265.51	16207.29
6	BMA, HPB, W	274.65	11194.75	273.45	14069.47	276.84	17199.88
7	BMA, HPC, B	326.75	19827.00	326.18	23575.68	333.03	29934.69
8	BMA, HPC, M	291.93	11517.24	290.72	15762.85	294.05	18551.62
9	BMA, HPC, W	304.32	12028.09	301.85	16979.11	306.59	19686.62
10	BMA, HPD, B	388.27	36033.34	390.33	43363.58	394.23	46566.35
11	BMA, HPD, M	346.88	20931.28	346.96	28820.08	348.08	29009.46
12	BMA, HPD, W	361.62	21859.70	360.26	31043.01	362.94	30803.37

- The claim frequency and severity models are better compared through their variance values, leading to a better classification of the policyholders and thus modelling jointly the location, scale and shape parameters in terms of a priori rating variables is justified because it enables us to use all the available information in the estimation of these values through the use of the important a priori rating variables for the number and the costs of claims respectively.

3.1 Calculation of the Premiums According to the Expected Value and Standard Deviation Principles

- The premium rates calculated according to the expected value principle are given by

$$P_1 = (1 + w_1)\, E(K_i)\, (1 + w_2)\, E(X_{i,k}), \tag{20}$$

where $w_1 > 0$ and $w_2 > 0$ are risk loads.

- The premium rates calculated according to the standard deviation principle are given by

$$P_2 = \left[E(K_i) + \omega_1\sqrt{Var(K_i)}\right]\left[E(X_{i,k}) + \omega_2\sqrt{Var(X_{i,k})}\right], \qquad (21)$$

where $\omega_1 > 0$ and $\omega_2 > 0$ are risk loads.

- In the following example (Table 4), two different groups of policyholders have been considered. In Table 4 a ‘YES’ indicates the presence of the characteristic corresponding to the column.

Table 4: The Six Different Groups of Policyholders to Be Compared

Group	BM Category A	HP 0-33	HP 34-66	HP 100-132	Male	Female
1	YES	YES	NO	NO	YES	NO
2	YES	YES	NO	NO	NO	YES

- We will calculate the premiums P_1 and P_2 that must be paid by a specific group of policyholders based on the alternative models for assessing claim frequency and the various claim severity models. We assume that $w_1 = w_2 = \omega_1 = \omega_2 = \frac{1}{10}$. The premiums P_1 and P_2 are obtained in Table 5 .

Table 5: Premium Rates Calculated Via the Expected Value and Standard Deviation Principles

Group	NBII-GA		NBII-WEI		NBII-GP	
	P_1	P_2	P_1	P_2	P_1	P_2
1	40.3903	47.3588	40.3750	47.5275	40.7045	48.1246
2	45.0967	51.3464	44.9000	51.3610	45.4563	52.1968
Group	SI-GA		SI-WEI		SI-GP	
	P_1	$P2$	P_1	P_2	P_1	P_2
1	40.1034	46.3306	40.0881	46.4957	40.4154	47.0800
2	45.7614	52.4340	45.5614	52.4489	46.1263	53.3025
Group	ZIP-GA		ZIP-WEI		ZIP-GP	
	P_1	P_2	P_1	P_2	P_1	P_2
1	40.1990	44.7401	40.1837	44.8994	40.5118	45.4635
2	46.9910	51.4043	46.7857	51.4189	47.3657	52.2557

4 Conclusions

- In this paper, we examined the use of regression models for location, scale and shape for pricing risks through ratemaking based on a priori risk classification.

- The resulting a priori premiums rates were calculated via the expected value and standard deviation principles with independence between the claim frequency and severity components assumed.

- Extensions to other frequency/severity regression models for location scale and shape can be obtained in a similar straightforward way. Moreover, these models are parametric and a possible line of further research is to explore the semiparametric approach and go through the ratemaking exercise when functional forms other than the linear are included, based on the generalized additive models for location scale and shape (GAMLSS) approach of Rigby and Stasinopoulos (2001, 2005 and 2009). Also see, for example, a recent paper by Klein et al. (2014) in which Bayesian GAMLSS models are employed for nonlife ratemaking and risk management.