RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS
No. 66 2009
Towards Information-based Welfare Society

Alicja Wolny-Dominiak
Katowice University of Economics, Katowice, Poland

MINIMUM BIAS METHODS IN *A PRIORI* RATE MAKING

Abstract

Insurance companies specializing in casualty insurance create their own rating systems for setting fair premiums for every risk for different kinds of insurance portfolios. The rating system is mostly based on the data analysis concerning the number and the value of claims for individuals or groups (classes) of insured people within a given portfolio. Based on a given rating system, the premium for a particular risk is calculated in two stages: *a priori* rating and *a posteriori* rating. In this paper, the process of a priori rating is analyzed with the emphasis on minimum bias methods used for modelling the rating variables.

1. Introduction

Insurance companies specializing in casualty insurance create their own rating systems for setting fair premiums for every risk for different kinds of insurance portfolios. The calculation of the pure premium for the overall portfolio of risks should meet two basic conditions: it should ensure that the insurance company will receive premiums at a level adequate to cover the claims and should fairly reflect the probability of an insured event for different groups of customers (i.e. a higher premium for the group of customers where the probability of an event or the sum of claims is higher). Thus, the development of the rating system involves classifying risks from the inhomogeneous overall portfolio in order to obtain homogeneous sub-portfolios. The homogeneous sub-portfolio is defined as a subset of insurance policies where claims are generated independently and the random variables – the number of claims – are identically distributed within the subset and the random variables – the value of compensation claimed – are identically distributed, too. The same pure premium is assigned to every policy in the homogeneous sub-portfolio [7].

The rating system is mostly based on the data analysis concerning the number and the value of claims for individuals or groups (classes) of insured people within a given portfolio. Based on a given rating system, the premium for a particular risk is calculated in two stages: *a priori* rating and *a posteriori* rating. *A priori* rating relies

on the base rate calculation taking into consideration factors which describe the insured person, the specification of the insured object (automobile, property, etc.) and the general insurance experience. *A posteriori* rating is the calculation of an additional rate (mostly calculated as a percentage of the base rate) where a number of claims made by the insured person in the past is taken into consideration. Therefore the "bonus-malus" system is mostly used, i.e. the system of discounts and increases in the premium according to an individual's claim experience [8].

As mentioned above, the value of claims in the portfolio depends on different predictors. Treating these predictors as factors (i.e. qualitative random variables) allows us to conduct a statistical data analysis in order to measure the influence of every predictor on the level of claims. Having this done the only problem in the *a priori* rating that needs to be solved is the estimation of the levels of predictors. Due to their nature, the predictors are called "rating variables". In the automobile insurances, for example, the standard rating variables are: the driver's region, gender, age and the engine capacity. In this paper, the process of *a priori* rating is analyzed with the emphasis on minimum bias methods used for modelling the rating variables.

2. Minimum bias methods used for estimating the rating variables

A basic method in *a priori* rate making is one-way analysis of loss data, for every rating variable independently. A one-way analysis summarizes insurance statistics, such as frequency or loss ratio, for each value of each rating variable, but without taking into account the effect of other variables. This kind of an analysis can be distorted by correlations between rating variables. For example, relativities based on one-way analyses of two different rating variables would double-count the effect of one of them. Traditional actuarial techniques for addressing this problem usually attempt to standardize the data in such a way as to remove the distorting effect of uneven business mix, for example by focusing on loss ratios on a one-way basis, or by standardizing for the effect of one or more rating variables. One-way analyses also do not consider interdependencies between variables in the way they affect claims experience [1].

More developed ratemaking technique is called minimum bias procedures. Actually this is rather a set of procedures linking the observed data, the rating variables, and relativities. An iterative procedure solves the system of equations by attempting to converge to the optimal solution.

There are many minimum bias procedures presented in the literature. Usually they are modifications of one of four basic approaches to the problem of analyzing the influence of rating variables on the value of claims. In all these procedures the influence is measured with the use of the so called *relativities*. These indexes show how to adjust the level of the premium to every sub-portfolio, i.e. how to change the level of the base rate [2].

Let us denote by:

1) X, Y – rating variables,
2) $x_1, \ldots, x_n$ – the relativities for variable X; $y_1, \ldots y_n$ – the relativities for variable Y,
3) r_{ij} – the average value of claims for the i-th value of X and the j-th value of Y,
4) n_{ij} – the average value of claims for the i-th value of X and the j-th value of Y,
5) B – the base average value of claims.

The first approach applies the *balance principle* for every value of the rating variable. The balance principle can be written in the following form [2]:

for the variable X: $\sum_j n_{ij} r_{ij} = \sum_j n_{ij} B x_i y_j$,

and for the variable Y $\sum_i n_{ij} r_{ij} = \sum_i n_{ij} B x_i y_j$,

The estimators for the relativities derived from the balance equations have the form:

$$\hat{x}_i = \frac{\sum_j n_{ij} r_{ij}}{\sum_j n_{ij} B y_j} \quad \text{and} \quad \hat{y}_j = \frac{\sum_i n_{ij} r_{ij}}{\sum_i n_{ij} B x_i} .$$

The second approach uses the least squares method, where the relativities are estimated by minimizing the squared differences between the average values of claims and their estimated values given by the model. Formally we search for the solution of the optimization task: $\min_{x,y} \sum_{i,j} n_{ij} (r_{ij} - B x_i y_j)^2$. Applying necessary condition for the local minimum of the functions of more than one variable we obtain [4]:

for the variable X: $\sum_j n_{ij} 2(r_{ij} - B x_i y_j)(-y_j) = 0$

and for the variable Y: $\sum_i n_{ij} 2(r_{ij} - B x_i y_j)(-x_i) = 0$.

The estimators for the relativities derived from these equations have the form:

$$\hat{x}_i = \frac{\sum_j n_{ij} r_{ij} B y_j}{\sum_j n_{ij} (B y_j)^2} \quad \text{and} \quad \hat{y}_j = \frac{\sum_i n_{ij} r_{ij} B x_i}{\sum_i n_{ij} (B x_i)^2} .$$

In the third approach estimation of the relativities is based on the minimization of the χ^2 statistic: $\min_{x,y} \sum_{i,j} n_{ij} \frac{(r_{ij} - B x_i y_j)^2}{B x_i y_j}$. Again using the theorem on local extrema we obtain the following estimators [4]:

$$\hat{x}_i = \sqrt{\frac{\sum_j n_{ij} r_{ij}^2 (By_j)^{-1}}{\sum_j n_{ij} By_j}} \text{ and } \hat{y}_j = \sqrt{\frac{\sum_i n_{ij} r_{ij}^2 (Bx_i)^{-1}}{\sum_i n_{ij} x_i}}.$$

The fourth approach uses the maximum likelihood method. It implies the need of making assumption about the claims' distribution. For insurance data the most often used distribution is Gamma distribution. In this case the maximum likelihood estimators for variables X and Y have the form [4]:

$$\hat{x}_i = \frac{\sum_j n_{ij} r_{ij} (By_j)^{-1}}{\sum_j n_{ij}} \text{ and } \hat{y}_j = \frac{\sum_i n_{ij} r_{ij} (Bx_i)^{-1}}{\sum_i n_{ij}}.$$

In order to calculate the values of the relativities we have to repeat the iterative algorithm computing x_i^k and y_j^k in every iteration $k \to \infty$. The starting point is set to x_0=1, y_0=1. In every iteration we use constant base B which is the weighted mean

$$B = \frac{\sum_{i,j} n_{ij} r_{ij}}{\sum_{i,j} n_{ij}}.$$

The convergence conditions for the iterative algorithm in all above models are: $x_i = \lim_{k->\infty} \hat{x}_i^k$, $y_j = \lim_{k->\infty} \hat{y}_j^k$ [6].

The weighted absolute percentage bias, which is the weighted average of absolute difference between the observations and fitted values [2]

$$d = \frac{\sum_{i,j} n_{ij} \frac{|r_{ij} - Bx_i y_j|}{Bx_i y_j}}{\sum_{i,j} n_{ij}},$$

is suggested as the criterion for the rational choice of the model for the relativities estimation.

3. Example for the minimum bias procedures

In this section we present numerical example illustrating how the minimum bias procedures work. The following assumptions were made for the example: claims are described by two variables – the value of the individual claim and the number of claims in the sub-portfolio. Moreover, two rating variables are considered in the multiplicative model.

Table 1. Average value of claim (r_{ij}) and the number of claims (n_{ij}) in the sub-portfolio

r_{ij}	17-20	21-24	25-29	30-34	35-39	40-49	50-59	60+
Not driving to work	250.48	213.71	250.57	229.09	153.62	208.59	207.57	192.00
Driving to work < 10 miles	274.78	298.6	248.56	228.48	201.67	202.8	202.67	196.33
Driving to work > 10 miles	244.52	298.13	297.90	293.87	238.21	236.06	253.63	259.79
Driving mostly on business	797.8	362.23	342.31	367.46	256.21	352.49	340.56	342.58
n_{ij}	17-20	21-24	25-29	30-34	35-39	40-49	50-59	60+
Not driving to work	21	63	140	123	151	245	266	260
Driving to work < 10 miles	40	171	343	448	479	970	859	578
Driving to work > 10 miles	23	92	318	361	381	719	504	312
Driving mostly on business	5	44	129	169	166	304	162	96

Source: [6].

Using the data from Table 1, four iterative algorithms were performed on the basis of four minimum biased models implemented in **R** – the language for statistical computing. The stopping criterion for the convergence of the algorithm was set to <0.0000001, which caused the algorithm to terminate after $k = 4$ iterations. The base was calculated to be $B = 241.46$.

Table 2. Relativities for the variable X

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Model 1	1.203546	1.207631	1.15438	1.123666	0.89052	0.970975	0.953408	0.921826
	1.259185	1.222628	1.136752	1.09986	0.878172	0.959818	0.97269	0.954536
	1.260293	1.222945	1.136482	1.099436	0.877956	0.959584	0.972997	0.955178
	1.260314	1.222951	1.136477	1.099428	0.877952	0.959579	0.973003	0.95519
Model 2	1.203546	1.207631	1.154380	1.123666	0.890520	0.970975	0.953408	0.921826
	1.289876	1.209254	1.127879	1.102567	0.871397	0.965814	0.976279	0.961449
	1.292142	1.209284	1.127265	1.102012	0.870950	0.965623	0.976815	0.962458
	1.292200	1.209285	1.127250	1.101997	0.870939	0.965618	0.976829	0.962484
Model 3	1.309298	1.219359	1.162829	1.142823	0.899809	0.991953	0.96854	0.939983
	1.329106	1.249564	1.154683	1.115882	0.894589	0.974865	0.987277	0.969541
	1.329725	1.249977	1.154582	1.115483	0.89444	0.974629	0.987536	0.970037
	1.329735	1.249983	1.154581	1.115477	0.894438	0.974626	0.98754	0.970045
Model 4	1.203546	1.207631	1.15438	1.123666	0.89052	0.970975	0.953408	0.921826
	1.239827	1.234401	1.144724	1.097248	0.88339	0.955742	0.969952	0.948634
	1.240569	1.234749	1.144644	1.096886	0.883228	0.955536	0.970163	0.949075
	1.24058	1.234754	1.144643	1.09688	0.883225	0.955533	0.970166	0.949082

Source: own calculations.

Table 3. Relativities for the variable *Y*

	y_1	y_2	y_3	y_4
Model 1	0.854618	0.887850	1.071906	1.390519
	0.850754	0.886292	1.073622	1.396364
	0.850680	0.886265	1.073654	1.396471
	0.850678	0.886264	1.073654	1.396473
Model 2	0.854744	0.888520	1.070131	1.386918
	0.850158	0.885728	1.071178	1.394628
	0.850034	0.885659	1.071205	1.394823
	0.850031	0.885657	1.071205	1.394828
Model 3	0.842359	0.874037	1.056052	1.376687
	0.838920	0.872723	1.057486	1.381844
	0.838863	0.872704	1.057509	1.381921
	0.838862	0.872703	1.057510	1.381922
Model 4	0.854173	0.887450	1.073693	1.393434
	0.850980	0.886537	1.075486	1.398901
	0.850929	0.886525	1.075513	1.398980
	0.850928	0.886525	1.075513	1.989810

Source: own calculations.

In every model that was built we observe that young drivers and drivers in their middle age are associated with the higher average value of claim compared to the base. It means that the premium has to be adjusted (increased) adequately. We have the similar situation for the car use variable, where the increase should be applied to the clients using the car on business and driving to work over 10 miles. For all the other cases there should be a discount in the base rate.

In order to choose the best model, taking as the criterion the minimization of *d*, we need to estimate the r_{ij} values and then compare the estimated values with the real values. The estimator has the general form:

$$\hat{r}_{ij} = Bx_i y_j.$$

Table 4. Estimators of values of the individual claims $\hat{r}_{ij}$

Model 1	17-20	21-24	25-29	30-34	35-39	40-49	50-59	60+
1	2	3	4	5	6	7	8	9
Not driving to work	258.88	251.20	233.44	225.83	180.34	197.10	199.86	196.20
Driving to work < 10 miles	269.70	261.71	243.20	235.28	187.88	205.35	208.22	204.41
Driving to work > 10 miles	326.73	317.04	294.63	285.02	227.61	248.77	252.25	247.63
Driving mostly on business	424.97	412.37	383.21	370.72	296.04	323.56	328.09	322.08
Model 2								
Not driving to work	265.22	248.21	231.37	226.18	178.76	198.19	200.49	197.55
Driving to work < 10 miles	276.34	258.61	241.06	235.66	186.25	206.50	208.90	205.83
Driving to work > 10 miles	334.23	312.79	291.57	285.04	225.27	249.76	252.66	248.95
Driving mostly on business	435.21	407.28	379.65	371.15	293.33	325.22	328.99	324.16

Table 4, cont.

1	2	3	4	5	6	7	8	9
Model 3								
Not driving to work	269.34	253.19	233.86	225.94	181.17	197.41	200.03	196.49
Driving to work < 10 miles	280.21	263.40	243.30	235.06	188.48	205.38	208.10	204.41
Driving to work > 10 miles	339.54	319.18	294.82	284.83	228.39	248.87	252.17	247.70
Driving mostly on business	443.71	417.09	385.26	372.21	298.46	325.21	329.52	323.68
Model 4								
Not driving to work	254.90	253.70	235.19	225.37	181.47	196.33	199.34	195.00
Driving to work < 10 miles	265.56	264.31	245.02	234.80	189.06	204.54	207.67	203.16
Driving to work > 10 miles	322.17	320.66	297.26	284.85	229.37	248.15	251.95	246.47
Driving mostly on business	419.07	417.10	386.66	370.53	298.35	322.78	327.72	320.60

Source: own calculations.

In the example we observe that the estimated values do not vary significantly when comparing different models. The estimation error d was computed to support the choice of the optimal model (Table 5).

Table 5. Weighted absolute percentage bias

	Model 1	Model 2	Model 3	Model 4
$d \rightarrow \min$	4.4537%	4.7045%	4.4229%	4.2584%

Source: own calculations.

In every analyzed model the weighted absolute percentage bias is a little bigger than 4%, however, the smallest value was obtained in the fourth model which is therefore taken as the optimal model for the given example. The above analysis can be extended by including other rate making methods and by using the weighted Pearson Chi-square statistics [3].

4. Monte Carlo simulation for the minimum bias procedure

The accuracy of the relativities depends on the claims' behaviour in the future. Thus, the simulation of the future claims is a useful tool illustrating changes in the relativities for different scenarios. We propose a Monte Carlo simulation procedure, where one iteration consists of five stages:

Stage 1. Assume the distribution for the average value of claims (independent variable).

Stage 2. Generate the data matrix (with the average values of claims).

Stage 3. Compute the relativities.

Stage 4. Estimate the theoretical average values of claims.

Stage 5. Compute the model estimation error.

In our simulation we assumed the Gamma distribution with the parameters α=1 and θ=r_{ij} (data in Table 1) for the average values of claims [3]. We used the fourth model (the optimal model with the smallest error d) for estimating the relativities. After iterating the procedure 1000 times we obtained the following variables distributions (Tables 6-8).

Table 6. Distributions of the relativities for the variable X

Deciles	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Min	0.8596	1.0000	1.0000	0.9912	0.7885	0.8928	0.8996	0.8715
10%	1.0665	1.1426	1.0907	1.0596	0.8453	0.9340	0.9413	0.9171
20%	1.1331	1.1683	1.1080	1.0742	0.8563	0.9424	0.9522	0.9283
30%	1.1736	1.1907	1.1172	1.0850	0.8646	0.9493	0.9585	0.9362
40%	1.2095	1.2047	1.1272	1.0937	0.8718	0.9549	0.9657	0.9434
50%	1.2447	1.2191	1.1363	1.1022	0.8807	0.9604	0.9716	0.9508
60%	1.2806	1.2369	1.1466	1.1096	0.8876	0.9656	0.9772	0.9592
70%	1.3224	1.2562	1.1577	1.1193	0.8938	0.9709	0.9831	0.9668
80%	1.3794	1.2796	1.1696	1.1297	0.9017	0.9763	0.9909	0.9749
90%	1.4536	1.3049	1.1870	1.1446	0.9134	0.9854	1.0017	0.9867
Max	1.8228	1.4535	1.2500	1.2144	1.0000	1.0391	1.0451	1.0386
Standard deviation	0.1466	0.0659	0.0376	0.0336	0.0265	0.0203	0.0235	0.0275

Source: own calculations.

Table 7. Distributions of the relativities for the variable Y

Deciles	y_1	y_2	y_3	y_4
Min	0.7793	0.8274	0.9806	1.0000
10%	0.8170	0.8631	1.0365	1.3262
20%	0.8273	0.8692	1.0447	1.3433
30%	0.8355	0.8737	1.0513	1.3565
40%	0.8416	0.8767	1.0567	1.3678
50%	0.8482	0.8801	1.0622	1.3777
60%	0.8540	0.8837	1.0671	1.3877
70%	0.8612	0.8873	1.0726	1.3991
80%	0.8689	0.8932	1.0798	1.4126
90%	0.8794	0.8998	1.0888	1.4282
Max	1.0000	1.0000	1.1230	1.5047
Standard deviation	0.0251	0.0147	0.0207	0.0415

Source: own calculations.

Table 8. Distribution of the error *d*

Deciles	*d*
Min	3.40%
10%	4.60%
20%	4.89%
30%	5.12%
40%	5.36%
50%	5.54%
60%	5.76%
70%	5.99%
80%	6.28%
90%	6.61%
Max	100.00%
Standard deviation	0.0309

Source: own calculations.

The simulation is also useful to compare the rate making results with the different average value of claims distribution, like inverse Gaussian or log-normal.

5. Summary

In the article we presented four approaches used in the estimation of the relativities in *a priori* rate making. All these approaches are based on the minimum bias procedure and have iterative structure, which means in practice the need of writing the computer implementation of the algorithm (it is not the case for, e.g., GLM models which are already implemented in the most widely used statistical software). When the distribution of the variable describing the claims is known, we are able to extend the iterative algorithms to the Monte Carlo simulation. As the result we obtain the distribution of the relativities, which makes us feasible to conduct the analysis of changes in the relativities for different scenarios.

The simulation can be applied even if it is hard to estimate the distribution of the claims based on the historical data, e.g., when the insurance company does not have the complete information about the claims experience in a given portfolio, in particular when the available data sets lack or do not isolate certain rating variables. The weakness of the simulation might appear when introducing too many rating variables to the model, because large number of obtained results may cause the analysis hard to conduct.

Bibliography

[1] Anderson D., Feldblum S., Modlin C., Schirmacher D., Schirmacher E., Thandi N., *A Practitioner's Guide to Generalized Linear Models*, A foundation for theory, interpretation and application, www.watsonwyatt.com, 2007.
[2] Bailey R.A., LeRoy J.S., "Two studies in automobile insurance ratemaking", *PCAS* 1960, vol. XLVII, pp. 1-19.
[3] Fu L., Moncher R.B., *Severity Distributions for GLMs: Gamma or Lognormal? Evidence from Monte Carlo Simulations*, Casualty Actuarial Society Discussion Paper Program Casualty Actuarial Society – Arlington, Virginia, 2004.
[4] Fu L., Wu Ch.P., "General iteration algorithms for classification ratemaking", *Variance Journal* 2007, vol. 1, no. 2.
[5] McCullagh P., Nelder J.A., *Generalized Linear Models*, Chapman & Hall/CRC, New York 1999.
[6] Mildenhall S., "A systematic relationship between minimum bias and generalized linear models", *PCAS* 1999, vol. LXXXVI, pp. 394-487.
[7] Ostasiewicz W. (Ed.), *Premiums and Insurance Risk*. *Stochastic Modeling*, Wrocław University of Economics Publishing House, Wrocław 2004 [in Polish].
[8] Śliwiński A., *Insurance Risk – Premium's Calculation and Optimization*, Poltext, Warszawa 2002 [in Polish].

PROCEDURY MINIMALNEGO OBCIĄŻENIA W TARYFIKACJI *A PRIORI*

Streszczenie

Zakłady ubezpieczeń działające w grupie ubezpieczeń majątkowych tworzą własne systemy taryfikacyjne w celu ustalenia sprawiedliwego poziomu składki dla każdego ryzyka w różnego rodzaju portfelach ubezpieczeniowych. Tworzenie systemu taryfikacyjnego jest oparte głównie na analizie danych dotyczących szkodowości oraz przebiegu ubezpieczenia jednostek bądź grup (klas) ubezpieczeniowych w danym portfelu ubezpieczeń. Ustalenie składki dla konkretnego ryzyka na podstawie danego systemu taryfikacyjnego jest dwuetapowe: taryfikacja *a priori* oraz taryfikacja *a posteriori*. W pracy analizowany jest proces taryfikacji *a priori* z wykorzystaniem tzw. procedur minimalnego obciążenia służących do szacowania poziomu zmiennych taryfikacyjnych wpływających na poziom szkodowości.