Journal of Modern Applied Statistical Methods

Volume 7 | Issue 2 Article 25

11-1-2008

Construction of Insurance Scoring System using Regression Models

Noriszura Ismail
Universiti Kebangsaan, Malaysia, ni@ukm.my

Abdul Aziz Jemain
Universiti Kebangsaan Malaysia, azizj@ukm.my

 Part of the Applied Statistics Commons, Social and Behavioral Sciences Commons, and the Statistical Theory Commons

Recommended Citation
Ismail, Noriszura and Jemain, Abdul Aziz (2008) "Construction of Insurance Scoring System using Regression Models," *Journal of Modern Applied Statistical Methods*: Vol. 7 : Iss. 2 , Article 25.
DOI: 10.22237/jmasm/1225513440

Journal of Modern Applied Statistical Methods
November, 2008, Vol. 7, No. 2, 607-615
Copyright © 2008 JMASM, Inc.
1538 – 9472/08/$95.00

Construction of Insurance Scoring System using Regression Models

Noriszura Ismail
Universiti Kebangsaan

Abdul Aziz Jemain
Malaysia, Malaysia

This study suggests the regression models of Lognormal, Normal and Gamma for constructing insurance scoring system. The main advantage of a scoring system is that it can be used by insurers to differentiate between high and low risks insureds, thus allowing the profitability of insureds to be predicted.

Key words: Scoring system, insurance risks, regression model.

Introduction

One of the most recent developments in the U.S. and the European insurance industry is the rapidly growing use of a scoring system in pricing, underwriting and marketing of high volume and low premium insurance policies. In the Asian market, scoring system is still considered as relatively new, although several markets in the region have started utilizing the system especially in its rating of motor insurance premium. In Singapore for example, in 1992, the biggest private car insurer, NTUC Income, announced that it was changing from a tariff system to a scoring system, whereby the owners of newer cars and more expensive models would probably pay lower premiums (Lawrence 1996).

There are several advantages of utilizing scoring system in pricing, underwriting and marketing of insurance. The main advantage is that the scores may be used by insurers to differentiate between good and bad insureds, thus allowing the profitability of insureds to be predicted by using a specified list of rating factors such as driver's experience, vehicle's characteristics and scope of coverage.

Noriszura Ismail is an Associate Professor in Actuarial Science Department. Her research areas are actuarial and statistical modeling in non-life insurance. Email: ni@ukm.my. Abdul Aziz Jemain is a Professor in Statistics Department. His research areas are climate modeling, actuarial modeling, medical statistics and social statistics. Email azizj@ukm.my.

In addition to distinguishing the risks of insureds, insurers may also employ the scores to determine the amount of premium to be charged on each customer.

Several studies on scoring system have been carried out in the actuarial and insurance literatures. For example, Coutts (1984) proposed the Orthogonal Weighted Least Squares (OWLS) to convert premiums into scores; he examined the impact of changing several input assumptions such as inflation rates, base periods of TPBI claims, expenses and weights on the structure of scores. Brockman & Wright (1992) suggested Gamma regression model to convert premiums into scores, rationalizing that the variance of Gamma depends on the weights or exposures, and not on the magnitude of premiums.

In recent years, Miller & Smith (2003) analyzed the relationship between credit-based insurance scores and propensity of loss for private passenger automobile insurance, and found that insurance scores were correlated with propensity of loss due to the correlation between insurance scores and claim frequency rather than average claim severities. Anderson et al. (2004) suggested Generalized Linear Modeling (GLM) for deriving scores, and proposed the fitting of frequency and severity separately for each claim type as starting point. The expected claim costs resulting from frequency and severity fitting were then divided by the premiums to yield the expected loss ratios, and the profitability scores were derived by rescaling the loss ratios. Wu & Lucker (2004) reviewed the basic structure of several insurance credit scoring models in the

U.S. by dividing scoring algorithms into two main categories; the rule-based approach which assigns scores directly to each rating factor, and the formula approach which determines scores using mathematical formulas. The minimum bias and GLM were suggested for the rule-based approach, whereas the Neural Networks (NN) and Multivariate Adaptive Regression Splines (MARS) were suggested for the formula approach. Wu & Guszcza (2004) studied the relationship between credit scores and insurance losses using data mining methodology along with several predictive modeling techniques such as NN, GLM, Classification and Regression Trees (CART) and MARS. Vojtek & Kocenda (2006) reviewed several methods of credit scoring employed by banks such as linear discriminant analysis (LDA), logit analysis, *k*-nearest neighbor classifier (*k*-NN) and NN to evaluate the applications of loans in Czech and Slovak Republics. Their results showed that the logit analysis and LDA methods were mainly used, the CART and NN methods were used only as supporting tools, and the *k*-NN method was rarely used in the process of selecting variables and evaluating the quality of credit scoring models.

The objective of this article is to suggest the Lognormal, Normal and Gamma regression models for the construction of insurance scoring system. Even though several actuarial studies have been carried out on the construction of scoring system, the detailed procedures of these methods have not been provided, with the exception of Coutts (1984) who proposed the use of Orthogonal Weighted Least Squares (OWLS) to convert premiums into scores. Although the Lognormal model proposed in this study is similar to the OWLS method proposed by Coutts (1984), the fitting procedure slightly differs. The OWLS method assumed that the weights were possible to be factorized and the fitted values were calculated using the estimated weights, whereas in this study, the fitting procedure does not require the weights to be factorized and the weights were not replaced by the estimated weights. This study also compares the Lognormal, Normal and Gamma regression models whereby the comparisons were centered upon three main elements; fitting procedures, parameter estimates and structure of scores.

Methodology

The response variable, independent variables and weight for the regression models are the premiums, rating factors and exposures respectively. The datasets are (g_i, e_i), where g_i and e_i respectively denote the premiums and the exposure in the i-th rating class, $i = 1,2,...,n$.

Appendix A shows a sample of rating factors, premiums and exposures for the data set. The premiums were written in Ringgit Malaysia (RM) currency based on motor insurance claims experience provided by an insurance company in Malaysia. The exposures were written in number of vehicle years, and the rating factors considered were scope of coverage (comprehensive, non-comprehensive), vehicle make (local, foreign), use-gender (private-male, private-female, business), vehicle year (0-1, 2-3, 4-5, 6+) and location (Central, North, East, South, East Malaysia).

Lognormal Model

Let the relationship between premiums, g_i and scores, s_i, be written as,

$$g_i = b^{s_i}, \tag{1}$$

or,

$$\log_b g_i = s_i. \tag{2}$$

In this study, the value of $b = 1.1$ was chosen for Equation (1) to accommodate the conversion of premiums ranging from RM30 to RM3,000 into scores ranging from 0 to 100. For example, the score corresponding to the premium amount of RM3,000 is equal to 84.

If the premium, G_i, is distributed as Lognormal with parameters s_i and $e_i^{-1}\sigma^2$, then $\log_{1.1} G_i$ is distributed as normal with mean s_i and variance $e_i^{-1}\sigma^2$, where the density is,

$$f(\log g_i; s_i) = \frac{1}{\sqrt{2\pi\sigma^2 e_i^{-1}}} \exp\left(-\frac{e_i(\log g_i - s_i)^2}{2\sigma^2} \right).$$

The relationship between scores, s_i, and rating factors, x_{ij}, may be written in a linear function as,

$$s_i = \mathbf{x_i^T}\boldsymbol{\beta} = \sum_{j=1}^{p} \beta_j x_{ij} \,, \tag{3}$$

where $\mathbf{x_i}$ denotes the vector of explanatory variables or rating factors, and $\boldsymbol{\beta}$ the vector of regression parameters. In other words, $\beta_j, j = 1,2,...,p$, represents the individual score of each rating factor, and s_i represents the total scores of all rating factors.

The first derivatives of Equation (3) may be simplified into,

$$\frac{\partial s_i}{\partial \beta_j} = x_{ij} \,. \tag{4}$$

Therefore, the solution for $\boldsymbol{\beta}$ may be obtained from the maximum likelihood equation,

$$\frac{\partial \ell}{\partial \beta_j} = \sum_i e_i (\log g_i - s_i) x_{ij} = 0, \quad j = 1,2,...,p. \tag{5}$$

Since the maximum likelihood equation is also equivalent to the normal equation in standard weighted linear regression, $\boldsymbol{\beta}$ may be solved by using normal equation.

Normal Model

Assume that the premium, G_i, is distributed as normal with mean δ_i and variance $e_i^{-1}\sigma^2$, where the density function is,

$$f(g_i;\delta_i) = \frac{1}{\sqrt{2\pi\sigma^2 e_i^{-1}}} \exp\left(-\frac{e_i(g_i-\delta_i)^2}{2\sigma^2}\right).$$

The conversion of premiums into scores may be implemented by letting the relationship between scores (s_i) and fitted premium (δ_i) to be written in a log-linear function or multiplicative form. If the base value is equal to 1.1, the fitted premium is,

$$\delta_i = (1.1)^{s_i} \,, \tag{6}$$

where

$$s_i = \mathbf{x_i^T}\boldsymbol{\beta} = \sum_{j=1}^{p} \beta_j x_{ij} \,.$$

The first derivative of Equation (6) is,

$$\frac{\partial \delta_i}{\partial \beta_j} = \log(1.1)\delta_i x_{ij} \,, \tag{7}$$

and the solution for $\boldsymbol{\beta}$ may be obtained from the maximum likelihood equation.

$$\frac{\partial \ell}{\partial \beta_j} = \sum_i e_i (g_i - \delta_i)\delta_i x_{ij} = 0, \quad j = 1,2,...,p. \tag{8}$$

The maximum likelihood equation shown by Equation (8) is not as straightforward to be solved compared to the normal equation shown in Equation (5). However, since Equation (8) is equivalent to the weighted least squares, the fitting procedure may be carried out by using an iterative method of weighted least squares (see McCullagh & Nelder, 1989; Mildenhall, 1999; Dobson, 2002; Ismail & Jemain, 2005; Ismail & Jemain, 2007). In this study, the iterative weighted least squares procedure was performed using SPLUS programming.

Gamma Model

The construction of a scoring system based on the Gamma Model is also similar to the Normal Model. Assume that the premium, G_i, is distributed as Gamma with mean δ_i and variance $v^{-1}\delta_i^2$, where the density function is,

$$f(g_i;\delta_i) = \frac{1}{g_i\Gamma(v)}\left(\frac{vg_i}{\delta_i}\right)^v \exp\left(-\frac{vg_i}{\delta_i}\right),$$

and v denotes the index parameter.

The conversion of premiums into scores may also be implemented by letting the relationship between scores (s_i) and fitted premiums (δ_i) to be written in a log-linear function or multiplicative form. Therefore, the first derivative is the same as Equation (7).

Assume that the index parameter, v, varies within classes, and can be written as $v_i = e_i\sigma^{-2}$. Therefore, the variance of the response variable is equal to $\sigma^2\delta_i^2 e_i^{-1}$ and the solution for $\boldsymbol{\beta}$ may be obtained through the maximum likelihood equation,

$$\frac{\partial \ell}{\partial \beta_j} = \sum_i \frac{e_i (g_i - \delta_i) x_{ij}}{\delta_i}, \qquad j = 1,2,..., p \,. \quad (9)$$

The maximum likelihood equation shown by Equation (9) is not as straightforward to be solved compared to the normal equation shown by Equation (5), and the fitting procedure may be carried out using an iterative method of weighted least squares.

Results

Scoring System based on Lognormal Model

The best model for lognormal regression may be determined by using standard analysis of variance. Based on the ANOVA results, all rating factors are significant, and 89.3% of the model's variations ($R^2 = 0.893$) can be explained by using the same rating factors.

The parameter estimates for the best regression model are shown in Table 1. The class for 2-3 year old vehicles is combined with 0-1 year old vehicles (intercept), and the classes for East and South locations are combined with Central location (intercept) to provide significant effects on all individual regression parameters.

The negative estimates are converted into positive values using the following procedure. First, the smallest negative estimate of each rating factor is transformed into zero by adding an appropriate positive value.

Then, the same positive value is added to other estimates categorized under the same rating factor. Finally, the intercept is deducted by the total positive values which are added to all estimates. The final scores are then rounded into whole numbers to provide easier premium calculation and risk interpretation. The original estimates, modified estimates and final scores are shown in Table 2.

The final scores shown in Table 2 specify and summarize the degree of relative risks associated with each rating factor. For instance, the risks for foreign vehicles are relatively higher by four points compared to local vehicles, and the risks for male and female drivers who used their cars for private purposes are relatively higher by nine and five points compared to drivers who used their cars for business purposes. The goodness-of-fit of the scores in Table 2 may be tested by using two methods; (1) comparing the ratio of fitted over actual premium income, and (2) comparing the difference between fitted and actual premium income.

Table 3 shows the total difference of premium income and the overall ratio of premium income. The total income of fitted premiums is understated by RM560,380 or 0.2% of the total income of actual premiums.

Table 1: Parameter estimates for Lognormal Model

Parameters		Estimates	Std.dev.	*p*-values
β_1	Intercept	78.81	0.26	0.00
β_2	Non-comprehensive	-14.52	0.43	0.00
β_3	Foreign	4.23	0.26	0.00
β_4	Female	-4.30	0.28	0.00
β_5	Business	-9.25	0.53	0.00
β_6	4-5 years	-1.17	0.33	0.02
β_7	6+ years	-1.56	0.30	0.01
β_8	North	0.84	0.29	0.04
β_9	East Malaysia	-4.18	0.45	0.00

Table 2: Original estimates, modified estimates and final scores

Parameters	Original Estimates	Modified Estimates	Final Scores
Intercept (Minimum score)	78.81	49.30	49
Coverage:			
Comprehensive	0.00	14.52	15
Non-comprehensive	-14.52	0.00	0
Vehicle make:			
Local	0.00	0.00	0
Foreign	4.23	4.23	4
Use-gender:			
Private-male	0.00	9.25	9
Private-female	-4.30	4.95	5
Business	-9.25	0.00	0
Vehicle year:			
0-1 year & 2-3 years	0.00	1.56	2
4-5 years	-1.17	0.39	0
6+ years	-1.56	0.00	0
Vehicle location:			
Central, East & South	0.00	4.18	4
North	0.84	5.02	5
East Malaysia	-4.18	0.00	0

Table 3: Total premium income difference and overall premium income ratio

		Value
Total number of businesses/policies/exposures	$\sum_{i=1}^{240} e_i$	170,749
Total income from fitted premiums	$\sum_{i=1}^{240} e_i \hat{g}_i$	RM 275,269,816
Total income from actual premiums	$\sum_{i=1}^{240} e_i g_i$	RM 275,830,196
Total premium income difference	$\sum_{i=1}^{240} e_i (\hat{g}_i - g_i)$	- RM 560,380
Overall premium income ratio	$\frac{\sum_{i=1}^{240} e_i \hat{g}_i}{\sum_{i=1}^{240} e_i g_i}$	0.998

Therefore, the fitted premiums for all classes are suggested to be multiplied by a correction factor of 1.002 to match their values with the actual premiums.

Apart from differentiating between high and low risk insureds, a scoring system may also be used by insurers to calculate the amount of premium to be charged on each client. The procedure for converting scores into premium amounts involved two basic steps.

First, the scores for each rating factor are recorded and aggregated; then, the aggregate scores are converted into premium amount by using a scoring conversion table (a table listing the aggregate scores with associated monetary values). Table 4 shows a scoring conversion table, which is constructed using Equation (1).

Comparison of Scoring System based on Lognormal, Normal and Gamma Models

Comparison of parameter estimates resulted from Lognormal, Normal and Gamma regression models are shown in Table 5. The parameter estimates for Lognormal, Normal and Gamma models provided similar values, except for β_2 and β_5 which produced larger values in Normal and Gamma models compared to Lognormal model.

Table 4: Scoring conversion table

Aggregate Scores	Premium Amounts (RM)	Aggregate Scores	Premium Amounts (RM)
49	107	67	595
50	118	68	654
51	129	69	719
52	142	70	791
53	157	71	870
54	172	72	958
55	189	73	1053
56	208	74	1159
57	229	75	1274
58	252	76	1402
59	277	77	1542
60	305	78	1696
61	336	79	1866
62	369	80	2052
63	406	81	2258
64	447	82	2484
65	491	83	2732
66	540	84	3005

Table 5: Estimates for Lognormal, Normal and Gamma regression models

Parameters		Lognormal			Normal			Gamma		
		Est.	Std. Error	*p*-value	Est.	Std. Error	*p*-value	Est.	Std. Error	*p*-value
β_1	Intercept	78.81	0.26	0.00	79.02	0.01	0.00	78.89	0.02	0.00
β_2	Non-comp	-14.52	0.43	0.00	-12.79	0.05	0.00	-13.71	0.03	0.00
β_3	Foreign	4.23	0.26	0.00	4.02	0.01	0.00	4.19	0.02	0.00
β_4	Female	-4.30	0.28	0.00	-4.03	0.01	0.00	-4.25	0.02	0.00
β_5	Business	-9.25	0.53	0.00	-7.40	0.03	0.00	-8.55	0.04	0.00
β_6	4-5 years	-1.17	0.33	0.02	-1.17	0.01	0.00	-1.17	0.02	0.00
β_7	6+ years	-1.56	0.30	0.01	-2.10	0.01	0.00	-1.73	0.02	0.00
β_8	North	0.84	0.29	0.04	0.49	0.01	0.00	0.81	0.02	0.00
β_9	East M'sia	-4.18	0.45	0.00	-4.01	0.03	0.00	-4.21	0.03	0.00

Comparison of scoring system resulted from Lognormal, Normal and Gamma regression models are shown in Table 6. The scores for Lognormal range from 49 to 84, the scores for Normal range from 53 to 84, and the scores for Gamma range from 51 to 85. In terms of risk relativities, both Lognormal and Gamma models resulted in a relatively higher score for male driver, female driver and comprehensive coverage. Therefore, if an insurer is interested in charging higher premiums for male driver, female driver and comprehensive coverage, both Lognormal and Gamma models may be suitable for fulfilling this strategy. However, the difference between Lognormal and Gamma model is that the scores for low risk classes provided by Gamma are slightly higher compared to Lognormal.

Conclusion

This article shows the procedure for constructing insurance scoring systems using three different regression models; Lognormal, Normal and Gamma. The main advantage of a scoring system is that it may be used by insurers to differentiate between "good" and "bad" insureds, thus allowing the profitability of insureds to be predicted. In addition, the scoring system has an operational advantage of reducing premium calculations and can be treated as a more sophisticated device for customers to assess their individual risks.

Table 6: Scoring system for Lognormal, Normal and Gamma regression models

Rating factors	Scores		
	Lognormal	Normal	Gamma
Minimum scores	49	53	51
Coverage:			
Comprehensive	15	13	14
Non-comprehensive	0	0	0
Vehicle make:			
Local	0	0	0
Foreign	4	4	4
Use-gender:			
Private-male	9	7	9
Private-female	5	3	4
Business	0	0	0
Vehicle year:			
0-1 year	2	2	2
2-3 years	2	2	2
4-5 years	0	1	1
6+ years	0	0	0
Location:			
Central	4	4	4
North	5	5	5
East	4	4	4
South	4	4	4
East Malaysia	0	0	0

The relationship between aggregate scores and rating factors in Lognormal model was suggested as linear function or additive form, whereas the relationship between aggregate scores and rating factors in Normal and Gamma models were proposed as log-linear function or multiplicative form.

The best regression model for Lognormal model was selected by implementing the standard analysis of variance. The goodness-of-fit of scores estimates were tested by comparing the ratio of fitted over actual premium income and by comparing the difference between fitted and actual premium income.

Acknowledgements

The authors gratefully acknowledge the financial support received in the form of a research grant (IRPA RMK8: 09-02-02-0112-EA274) from the Ministry of Science, Technology and Innovation (MOSTI), Malaysia. The authors are also pleased to thank Insurance Services Malaysia (ISM) for supplying the data.

References

Anderson, D., Feldblum, S., Modlin, C., Schirmacher, D., Schirmacher, E., & Thandi, N. (2004). A practitioner's guide to generalized linear models. *Casualty Actuarial Society Discussion Paper Program*, 1-115.

Brockman, M. H., & Wright, T. S. (1992). Statistical motor rating: Making effective use of your data. *Journal of the Institute of Actuaries*. *119*(3), 457-543.

Coutts, S. M. (1984). Motor insurance rating, an actuarial approach. *Journal of the Institute of Actuaries*, *111*, 87-148. Dobson, A. J. (2002). *An introduction to Generalized Linear Models (second edition)*. NY: Chapman & Hall.

Ismail, N., & Jemain, A. A. (2005). Bridging minimum bias and maximum likelihood methods through weighted equation. *Casualty Actuarial Society Forum*, *Spring*, 367-394.

Ismail, N., & Jemain, A. A. (2007). Handling overdispersion with Negative Binomial and Generalized Poisson regression models. *Casualty Actuarial Society Forum*, *Winter*: 103-158.

Lawrence, B. (1996). Motor insurance in Singapore. In Low Chan Kee Ed., *Actuarial and insurance practices in Singapore*. 191-216. Addison-Wesley: Singapore.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Model.* (2nd ed.) London, UK: Chapman & Hall.

Mildenhall, S. J. (1999). A systematic relationship between minimum bias and generalized linear models. *Proceedings of the Casualty Actuarial Society*, *86*(164), 93-487.

Miller, M. J., & Smith, R. A. (2003). The relationship of credit-based insurance scores to private passenger automobile insurance loss propensity. *Presentation to NAIC*. July, 2003.

Vojtek, M., & Kocenda, E. (2006). Credit scoring models. *Czech Journal of Economics and Finance*, *56*(3-4), 152-167.

Wu, C. P., & Lucker, J. R. (2004, Winter). A view inside the "Black Box": A review and analysis of personal lines insurance credit scoring models filed in the state of Virginia. *Casualty Actuarial Society Forum*, 251-290.

Wu, C. P., & Guszcza, J. C. (2004, Winter). Does credit score really explain in losses? Multivariate analysis from a data mining point of view. *Casualty Actuarial Society Forum*, 113-138.

Appendix A: Rating factors, exposures and premium amounts for Malaysian data

Rating factors					Exposure (vehicle-year)	Premium amount (RM)
Coverage	Vehicle make	Use-gender	Vehicle year	Location		
Comprehensive	Local	Private-male	0-1 year	Central	4243	1811
				North	2567	2012
				East	598	1927
				South	1281	1869
				East Malaysia	219	983
			2-3 years	Central	6926	1704
				North	4896	1919
				East	1123	1854
				South	2865	1794
				East Malaysia	679	1301
			4-5 years	Central	6286	1613
				North	4125	1840
				East	1152	1770
				South	2675	1687
				East Malaysia	700	1162
			6+ years	Central	6905	1524
				North	5784	1790
				East	2156	1734
				South	3310	1633
				East Malaysia	1406	1144
		Private-female	0-1 year	Central	2025	1256
				North	1635	1343
				East	301	1396
				South	608	1289
				East Malaysia	126	787
			2-3 years	Central	3661	1210
				North	2619	1298
				East	527	1255
				South	1192	1212
				East Malaysia	359	942
			4-5 years	Central	2939	1139
				North	1927	1243
				East	439	1125
				South	959	1176
				East Malaysia	376	652
			6+ years	Central	2215	1072
				North	1989	1215
				East	581	1219
				South	937	1112
				East Malaysia	589	623
		Business	0-1 year	Central	290	722
				North	66	547
				East	24	107
				South	52	685
				East Malaysia	6	107
			2-3 years	Central	572	731
				North	148	630
				East	40	107
				South	91	657
				East Malaysia	17	107
			4-5 years	Central	487	654
				North	100	549
				East	40	540
				South	59	571
				East Malaysia	22	493
			6+ years	Central	468	567
				North	93	518
				East	33	562
				South	77	515
				East Malaysia	25	402