Supply and demand analysis for flood insurance by using logistic regression model: case study at Citarum watershed in South Bandung, West Java, Indonesia

P. Sidi[1*], M. Mamat[2], Sukono[3*], S. Supian[4]

[1*]Department of Mathematics, Universitas Terbuka, INDONESIA
[2]Graduate School, Universiti Sultan Zainal Abidin, MALAYSIA
[3,4]Department of Mathematics, Universitas Padjadjaran, INDONESIA

E-mails: pram@ecampus.ut.ac.id; sukono@unpad.ac.id; sudradjat@unpad.ac.id

Abstract. Floods have always occurred in the Citarum river basin. The adverse effects caused by floods can cover all their property, including the destruction of houses. The impact due to damage to residential buildings is usually not small. Indeed, each of flooding, the government and several social organizations providing funds to repair the building. But the donations are given very limited, so it cannot cover the entire cost of repair was necessary. The presence of insurance products for property damage caused by the floods is considered very important. However, if its presence is also considered necessary by the public or not? In this paper, the factors that affect the supply and demand of insurance product for damaged building due to floods are analyzed. The method used in this analysis is the ordinal logistic regression. Based on the analysis that the factors that affect the supply and demand of insurance product for damaged building due to floods, it is included: age, economic circumstances, family situations, insurance motivations, and lifestyle. Simultaneously that the factors affecting supply and demand of insurance product for damaged building due to floods mounted to 65.7%.

1. Introduction

Flooding has always been a natural disaster that always happens every year in the Citarum river basin in the South Bandung West Java. The disadvantage of this disaster is diverse, ranging from the loss of: property, a car, people, until the damage of houses [4]. Various flood disaster management efforts have been made by the government, such as evacuating victims and provide financial aid for home repairs. However, the donation cannot cover the losses suffered by the community. Therefore, public participation needs to be improved so that they participate in preparation for the floods that continue to occur. Insurance damage of houses is a form of risk management that is effective in overcoming the loss [2].

Flood insurance is a form of cooperation between people who want to minimize the risk of damage uncertain. Flood insurance provides protection and guarantees if there is a risk of damage due to flood. Because of the risk of damage due to flood, it raised awareness of a person to bestow such risks to the

insurance company by filing a flood insurance demand [5]. The demand for flood insurance encourages insurance companies to make efforts in attracting customers. Various insurance companies race to create a services product that can meet the needs of customers [10]. Understanding of customer behavior is an important task for the company to determine the factors that encourage customers to follow the life insurance. In order to know these factors, then the method can be used is the logistic regression analysis [1], [3].

This paper intends to conduct research, with the aim of analyzing the supply and demand of insurance caused by flood damage of houses Citarum river. The analyzed data is obtained by distributing questionnaires to gather public perception in the Citarum river basin in South Bandung regency. The analysis is done in stages as follows: (i) to analyze deals flood insurance products, (ii) the analysis of demand for insurance products, and (iii) analyze the factors that influence the demand for insurance products flooding.

2. Logistic Regression Model

Logistic regression is a model used to describe the relationship between independent variables and response variables with ordinal scale. In this logistic regression, the variable response with ordinal scale more than two categories, and each category can be rated. The model for logistic regression is *cumulative logit models*. According Fahrmeir [6], the cumulative probability model for k categories is:

$$P\left(Y \le r \mid x\right) = \frac{\exp\left(\theta_r + x'\beta\right)}{1+\exp\left(\theta_r + x'\beta\right)} \tag{1}$$

while the cumulative logit model as follows:

$$\log\left[\frac{P\left(Y \le r \mid x\right)}{P\left(Y > r \mid x\right)}\right] = \log\exp\left(\theta_r + \boldsymbol{x'\beta}\right) = \theta_r + \boldsymbol{x'\beta} \tag{2}$$

where $r = 1,\ldots,k-1$, $\theta_r =$ constants to -*r*, $\boldsymbol{\beta} = \left(\beta_1, \beta_2, \beta_3, \ldots, \beta_p\right)$ is vector coefficients of free variable, and $\boldsymbol{x'} =$ vector of free variable. Method of estimation parameters used in the logistic regression model is the Maximum Likelihood Estimation (MLE). This parameter estimator is obtained by maximizing the likelihood function respect to parameters. The likelihood functions for three categories as follows [7], [12]:

$$l\left(\boldsymbol{\beta}\right) = \prod_{i=1}^{n}\left\{\left[\frac{e^{\theta_1+\boldsymbol{x'\beta}}}{1+e^{\theta_1+\boldsymbol{x'\beta}}}\right]^{y_{1i}}\left[\frac{e^{\theta_2+\boldsymbol{x'\beta}}-e^{\theta_1+\boldsymbol{x'\beta}}}{\left(1+e^{\theta_2+\boldsymbol{x'\beta}}\right)\left(1+e^{\theta_1+\boldsymbol{x'\beta}}\right)}\right]^{y_{2i}}\left[\frac{1}{1+e^{\theta_2+\boldsymbol{x'\beta}}}\right]^{y_{3i}}\right\} \tag{3}$$

In a logistic regression model, is necessary to test the significance of the regression coefficients simultaneously and partially. Testing the significance of the regression coefficients simultaneously is performed by using a likelihood estimation ratio test symbolized by *G*. According to Hosmer and Lemeshow [7], [9], this is defined as follows:

$$G = -2\ln\left[\frac{L_0}{L_1}\right] \tag{4}$$

with $L_0 =$likelihood in the condition H_0, $L_1 =$likelihood in the condition H_1. The hypothesis of this test is: $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ against alternative H_1: at least there is one $\beta_p \neq 0$. Criteria test the null hypothesis is rejected if $P\left[\chi^2(v) > G\right] < \alpha$, the level of significance was set.

Statistics *G* will follow Chi-Square distribution with degrees of freedom *v*. While testing the significance of regression coefficient partially is done by using the *Wald* test is defined as:

$$W = \frac{\hat{\beta}_l}{SE\left(\hat{\beta}_l\right)} \tag{5}$$

The hypothesis of this test is: $H_0 : \beta_i = 0$ against alternative H_1: $\beta_i \neq 0, 1 = 1, 2, \ldots, p$. Criteria test the null hypothesis is rejected if $P[|Z| > W] < \alpha$, the level of significance was set. Where, *Z* is a random variable that follows a standard normal distribution [7].

In addition, it is also testing the suitability of the model used to check the model obtained is appropriate or not in accordance with the observed data. The test is performed using Pearson test and Deviance test [6], [8]. Pearson statistic is defined as:

$$\chi^2 = \sum_{i=1}^{g} \chi^2{}_p\left(y_i, \hat{\pi}_i\right) \text{ with } \chi^2{}_p\left(y_i, \hat{\pi}_i\right) = n_i \sum_{j=i}^{k} \frac{\left(y_{ij} - \hat{\pi}_{ij}\right)^2}{\hat{\pi}_{ij}} \tag{6}$$

Deviance statistic is defined as:

$$D = 2\sum_{i=1}^{g} \chi_D^2\left(y_i, \hat{\pi}_i\right) \text{ with } \chi_D^2\left(y_i, \hat{\pi}_i\right) = n_i \sum_{j=1}^{k} y_{ij} \log\left(\frac{y_{ij}}{\hat{\pi}_{ij}}\right) \tag{7}$$

The hypothesis of this test is H_0: model is appropriate, against the alternative H_1: model is not appropriate. Pearson and Deviance statistics was spread by Chi-Square distribution with degrees of freedom $g(k-1) - p$. Decisions reject the null hypothesis. Criteria test the null hypothesis is rejected if: χ^2 and $D > \chi^2_{(g(k-1)-p)}$ or $P\left[\chi_v^2 > \chi^2_{pearson}\right]$ and $P\left[\chi_v^2 > D\right]$ less than α, the level of significance was set. If a null hypothesis is rejected, then the conclusion is that the model obtained is not appropriate. Conversely, if the opportunities generated is greater than opportunities desired or alpha, then H_0 accepted so that the model is appropriate [11], [12].

3. Methodology

The data used in this study are primary data that obtained by distributing questionnaires to customers of an insurance company at branch office, Bandung on 15-26 February 2016 as many as 98 people. Response variable used in this study is a request of flood insurance, while the independent variables were age (X_1), the economy condition (X_2), the family situation (X_3), insurance motivation (X_4), and lifestyle (X_5). The steps are performed in this study: the first is to test the assumption of multicollinearity. This test uses correlation matrix. Furthermore, the univariable of logistic regression model is analyzed. This analysis is used to sort out the variables that will be used in a multivariable of logistic regression

model, and to generate a parameter estimator. After that, in this analysis was also performed coefficient test of logistic regression simultaneously by using a likelihood ratio test. While, the coefficient test of logistic regression partially was done by using *Wald* test. After getting the best logistic regression model, further it is performed testing for the suitability of the model.

4. Results and Discussion

From the results of instrument test showed that variable X_2, X_3, X_4, and X_5 are valid and reliable. Therefore, all of the study variables are eligible for use in this research. Before performing univariable logistic regression analysis, performed first assumption test of multicollinearity. Based on the examination results of multikolinieritas is obtained that correlation values between X_4 with X_3 for 0.627 is greater than the correlation between X_4 and Y is 0.374. It indicates the occurrence of multicollinearity, but it is not excluded for allegedly influencing the dependent variable. The next step is to perform univariable logistic regression analysis. In this analysis, first performed is the likelihood ratio test for univariable logistic regression model. Tests carried out by using Equation (4), and the results are given in Table 1.

Table 1. Results of Likelihood Ratio Test for Univariable Logistic Regression Model

Variables	Statistic of *G*	*P*-Value
X_1	0.573	0.574
X_2	12.453	0.000
X_3	53.363	0.000
X_4	25.421	0.000
X_5	57.244	0.000

Based on the results in Table 1, it can be seen that the p-value of the estimation ratio likelihood *G* of four independent variables, namely the economy condition (X_2), the family situation (X_3), insurance motivation (X_4), and lifestyle (X_5) is less than = 5 %. Therefore it can be concluded that the four significant variables that are not all the parameters in the model is zero. After that, the partial test of the value of the parameter estimator for univariable logistic regression model. Testing is done by using equation (5), and the results are given in Table 2.

Table 2. Result of Parameter Estimation, Value-Z of Univariable Logistic Regression Model

Variables	Predictors	Coefficients	Statistic of *Z*	*P*-Values
Age	C_1	-3.23312	-2.67321	0.00511
	C_2	-2.41325	-1.34320	0.13124
	C_3	0.67924	0.75627	0.34152
	X_1	0.02115	0.56489	0.48636
Economy condition	C_1	2.67413	2.11631	0.03547
	C_2	4.12536	3.22953	0.00251
	C_3	5.98525	4.77351	0.00001
	X_2	-0.43574	-3.76835	0.00000
Family situation	C_1	2.32356	3.24362	0.01784
	C_2	3.24798	4.85796	0.00002
	C_3	5.67595	7.00345	0.00000
	X_3	-1.38462	-5.17364	0.00000
Insurance motivation	C_1	2.47463	2.63352	0.02301
	C_2	4.26738	4.01732	0.00002
	C_3	5.98584	4.87642	0.00000
	X_4	-0.70185	-5.21364	0.00000
Lifestyle	C_1	4.32911	4.53722	0.00002
	C_2	5.88678	5.75691	0.00000
	C_3	10.1325	9.00231	0.00001
	X_5	-0.89150	-5.43510	0.00000

Based on Table 2, it is known that the statistical results of the value of *Z* and *P*-value on the age variable is not significant in the model, because $P[|Z| > 0.56489] = 0.48636$ more than $\alpha = 5\%$. The economy condition variable is significant in the model, because $P[|Z| > 3.76835] = 0.00000$ less than $\alpha = 5\%$. The family situation variable is significant in the model, because $P[|Z| > 5.17364] = 0.00000$ less than $\alpha = 5\%$. The insurance motivation variable is significant in the model, because $P[|Z| > 5.21364] = 0.00000$ less than $\alpha = 5\%$. The lifestyle variable is significant in the model, because $P[|Z| > 5.43510] = 0.00000$ less than $\alpha = 5\%$. Therefore, the independent variables that should be in the logistic regression models is the economy condition (X_2), the family situation (X_3), insurance motivation (X_4), and lifestyle (X_5).

Furthermore, logistic regression analysis will be done by using X_2, X_3, X_4, and X_5 so that it obtained results in Table 3.

Table 3. Results of Parameter Estimation, Standard Errors of Logistic Regression Model

Predictors	Coefficients	*SE*-Coefficients
C_1	8.97139	2.13838
C_2	13.27131	2.28379
C_3	15.57133	2.73726
X_2	-0.426424	0.13617
X_3	-0.724789	0.28345
X_4	-0.089617	0.15675
X_5	-0.827016	0.25379

Based on the estimation results of logistic regression in Table 3, it is obtained three logit models, namely:

Logit regression model 1:

$$\log[P(Y \le 1)] = 8.97139 - 0.348356X_2 - 0.724789X_3 - 0.089617X_4 - 0.827016X_5;$$

Logit regression model 2:

$$\log[P(Y \le 2)] = 13.27131 - 0.348356X_2 - 0.724789X_3 - 0.089617X_4 - 0.827016X_5;$$

Logit regression model 3:

$$\log[P(Y \le 3)] = 15.57133 - 0.348356X_2 - 0.724789X_3 - 0.089617X_4 - 0.827016X_5.$$

After that, testing the logistic regression coefficient simultaneously by using estimation ratio test likelihood and it is obtained the value $G = 74\,543$ with a *P*-value $= 0.000 < \alpha = 5\%$. It can be concluded that at least one of the independent variables significantly influence the response variable. Then, the logistic regression coefficient test partially using the *Wald* test, refer to the equation (5), and the results are given in Table 4.

Table 4. Results of Wald Test

Variables	Statistic of *Z*	*P*-Value
X_2	-3.74	0.004
X_3	-4.37	0.002
X_4	-1.68	0.437
X_5	-5.64	0.000

On the results of the Wald test is known that there are three independent variables are variables X_2, X_3 and X_5 which has a smaller chance of α. This means that the independent variable is significant at the level 5%. In addition, there is one independent variable is the variable X_4 which have greater opportunities than the level 5%. It can be concluded that the factor of economic conditions, the family situation, and lifestyle have a significant influence on demand for flood insurance. Furthermore, the

suitability tests of the model by using Pearson and Deviance test. Tests were done according to Equation (6) and (7). Based on the results of testing the suitability of the model obtained *p-value* for the Pearson method was not significant because $P[\chi^2_{216} > 216.455]$ =0.437 more than $\alpha = 5\%$. In addition, *p-value* for the method of Deviance is not significant because $P[\chi^2_{216} > 148.132]$= 1.000 more than $\alpha = 5\%$. This means accepting H_0, so it can be concluded that the model obtained in accordance with observed data. Based on the results of *Pseudo R-Square* is obtained Nagelkerke value of 65.7%. This means that 65.7% variability of the dependent variable is able to be explained by variable X_2, X_3, X_4, and X_5 while the remaining 34.3% is explained by other variables.

Thus the interpretation of the best logistic regression for logit model 1, it can be concluded that the chances of economic condition variable (X_2) affect demand for flood insurance for the first category were lower by 0.426424 as compared to the second category. Chances of the family situation variable (X_3) affect demand for flood insurance for the first category were lower by 0.724789 as compared to the second category. Chances of insurance motivation variable (X_4) affect demand for flood insurance for the first category were lower by 0.089617 as compared to the second category. Chances of lifestyle variable (X_5) affect demand for flood insurance for the first category were lower by 0.827016 as compared to the second category.

Interpretation of the logit models 2, it can be concluded that the chances of economic condition variable (X_2) affect demand variables flood insurance for second categories was lower by 0.426424 compared to the third category. Chances of the family situation variable (X_3) affect demand for flood insurance for both categories was lower by 0.724789 compared to the third category. Chances of insurance motivation variable (X_4) affect demand for life insurance for both categories was lower by 0.089617 compared to the third category. Chances of lifestyle variable (X_5) affect insurance demand for both categories was lower by 0.827016 compared to the third category.

Interpretation of the logit models 3, it can be concluded that the chances of economic condition variable (X_2) affect demand for the third category of flood insurance was lower by 0.426424 compared to the fourth category. Chances of the family situation variable (X_3) affect demand for flood insurance for the third category was lower by 0.724789 compared to the fourth category. Chances of insurance motivation variable (X_4) affect demand for life insurance for the third category was lower by 0.089617 compared to the fourth category. Chances of lifestyle variable (X_5) affect insurance demand for the third category was lower by 0.827016 compared to the fourth category.

5. Conclusion

In this paper has carried out analysis of the factors affecting demand and supply flood insurance, with a case study in the Citarum river basin Bandung regency, West Java. Results of analysis showed that the supply and demand for insurance of damage to buildings caused by floods, influenced by factors: age, economy condition, family situation, insurance motivation, and lifestyle. From the results of data analysis that has been done, there are three best ordinal logistic regression models on the factors that affect the supply and demand for insurance of damage to buildings caused by floods significantly. Based on Pseudo R-Square, insurance demand of damage to buildings caused by floods able to be explained by the logistic regression model amounted to 65.7%, while the rest of 34.3% is explained by other factors. After knowing the factors that significantly influence demand for flood insurance, it is expected that an insurance company branch office in Bandung can further increase the sales of products, especially products of flood insurance.

Acknowledgment

We would like thanks the academic leadership grant (ALG), Faculty of Mathematics and Natural Sciences, University Padjadjran, which has provided facilities to conduct research and publication.

References

[1] R. Baker, C. Weinand, J. Jeng, H. Hoeksema, S. Monstrey, S. Pae, R. Spence and D. Wilson. 2009. Using Ordinal Logistic Regression to Evaluate the Performance of Laser-Doppler Predictions of Burn-Haling Time. *BMC Medical Research Methodology,* (online), 9-11, (www.biomedcentral.com), diakses 5 Mei 2013.

[2] J. Czajkowski, H. Kunreuther, and E. Michel-Kerjan. 2013. Catastrophe Model Based Quantification of Riverine and Storm Surge Flood Risk in Texas. *Working Paper* # 2013-01. The Wharton School, University of Pennsylvania.

[3] S. Das and R. Rahman. 2011. Application of Ordinal Logistic Regression Analysis in Determining Risk Factors of Child Malnutrition in Bangladesh. *Nutrition Journal, (online),* 10:124, (www.nutritionj.com), diakses 5 Mei 2013.

[4] V.A. Dei-Tutu. 2002. Flood Hazards, Insurance, and House Prices-A Hedonic Property Price Analysis. *M.S. Research Paper*. Department of Economics, College of Arts and Sciences, East Carolina University.

[5] R.E. Dumm, C. Nyce, D.L. Eckles, and J. Volkman-Wise. 2015. Demand for Catastrophe Insurance and the Representative Heuristic. *Working Paper.* Florida State University.

[6] L. Fahrmeir and G. Tutz. 1994. *Multivariate Statistical Modelling Based on Generalized Linear Models*. New York: Springer Verlag.

[7] D.W. Hosmer and S. Lemeshow. 2000. *Applied Logistic Regression.* Second Edition. Canada: John Wiley & Sons, Inc.

[8] S. Javali and P. Pandit. 2010. A Comparison of Ordinal Regression Models in an Analysis of Factors Associated with Periodontal Disease. *Journal of Indian Society of Periodontology,* (online), 14(3):155-159, (www.jisponline.com), diakses 5 Mei 2013.

[9] A. Luca. 2011. Ordinal Logistic Regression for the Estimate of the Response Functions in the Conjoint Analysis. *iBusiness,* (online), 3:383-389, (http://www.SciRP.org/journal/ib), diakses 2 Mei 2013.

[10] Y. Paudel, W. J. W. Botzen, and J. C. J. H. Aerts. Estimation of Insurance Premiums for Coverage Against Natural Disaster Risk: an Application of Bayesian Inference. Nat. Hazards Earth Syst. Sci., 13, 737–754, 2013. Published by Copernicus Publications on behalf of the European Geosciences Union.

[11] R.T. Ramadhayanti, I.N. Parta, and H. Permadi. 1912. Implementasi Regresi Logistik Ordinal pada Faktor-Faktor yang Mempengaruhi Permintaan Asuransi Jiwa: (Studi Kasus di AJB Bumiputera 1912 Cabang Malang Dieng). *Paper*. Universitas Negeri Malang.

[12] C. Trihendradi. 2007. *Kupas Tuntas Analisis Regresi, Strategi Jitu Melakukan Analisis Hubungan Causal.* Yogyakarta: ANDI.