Journal of Mathematical Sciences: Advances and Applications
Volume 22, 2013, Pages 1-17

CLASSIFICATION OF HIGHER EDUCATION LOANS USING MULTINOMIAL LOGISTIC REGRESSION MODEL

DENNIS K. MURIITHI[1], GLADYS G. NJOROGE[2], MARK O. OKONGO[2] and ELIZABETH W. NJOROGE[1]

[1]Department of Business Administration
Chuka University
P. O. Box 109-60400
Chuka
Kenya
e-mail: kamuriithi@yahoo.com

[2]Department of Physical Sciences
Chuka University
P. O. Box 109-60400
Chuka
Kenya

Abstract

This paper looks into the allocations made by the Higher Education Loans Board (HELB) relative to the economic status of the student. In this paper, we modelled HELB loan application data from three public universities to determine whether the loan was allocated based on the needs of the respective applicants. The data was classified to consider the amounts awarded by the HELB. This was possible since we observed that HELB loans were awarded in distinct categories (Kshs 35,000, Kshs 40,000, Kshs 45,000, Kshs 50,000, Kshs 55,000, and Kshs 60,000). In this paper, we used multinomial logistic regression in classifying the applicants into the identified categories. The models were

2010 Mathematics Subject Classification: 62J15.

Keywords and phrases: regression, logistic, multinomial, higher education, loan.

Received June 7, 2013

generated that included all predictor variables that were useful in predicting the response variable. The study revealed that the loans were not awarded based on the need of respective applicants. This has led to mis-classification when allocating loan. This study revealed that wealth, house worth, and amount of fees paid for siblings were other factors that could be considered to identify needy students. This results show that multinomial regression model gives accurate estimates that can enable HELB make a viable awarding decision thus minimizing the number of mis-classifications when awarding HELB loan, if any, although further studies may be commissioned to confirm or disapprove our findings.

1. Introduction

The Higher Education Loans Board was established by an Act of Parliament and as one of its roles was to grant loans. The statute known as The Higher Education Loans Board Act, 1995 was legally established as Act number 3 of 1995. It came into existence on the 21st day of July 1995 through Kenya Gazette Supplement (Cap 213A). The board applies a means testing instrument in order to identify deserving students, Web site [1].

Joint Admission Board (JAB) is an organization in Kenya that manages admission of government sponsored students. The students apply for the HELB loan upon receiving admission at the university. Once the student applies for the loan, HELB goes through a process of checking the forms filled and depending on several factors they are able to determine who will be awarded a loan and what amount will be awarded. In the case of the HELB loans, we can classify the loans into two natural categories of those not allocated the loan (0) and those allocated the loan (1), Chacha [2]. We could as well classify further to consider the amounts awarded by the HELB.

2. Literature Review

Multinomial logistic regression

Multinomial logistic regression is used to analyze relationships between a non metric dependent variable and metric or dichotomous independent variables. Multinomial logistic regression compares multiple

groups through a combination of binary logistic regressions. The group comparisons are equivalent to the comparisons for a dummy-coded dependent variable, with the group with the highest numeric score used as the reference group. For example, if the students wanted to study differences in BSc, MSc, and PhD students using multinomial logistic regression, the analysis would compare BSc students to PhD students and MSc students to PhD students. For each independent variable, there would be two comparisons, Wedel [9]. Multinomial logistic regression provides a set of coefficients for each of the two comparisons. The coefficients for the reference group are all zeros, similar to the coefficients for the reference group for a dummy-coded variable, thus, there are three equations, one for each of the groups defined by the dependent variable. The three equations can be used to compute the probability that a subject is a member of each of the three groups. A case is predicted to belong to the group associated with the highest probability. Predicted group membership can be compared to actual group membership to obtain a measure of classification accuracy.

In the study of the primary food choices of alligators in four Florida lakes by Delany and Moore [3]. The study classified the stomach contents of 219 captured alligators into five categories: Fish (the most common primary food choice); Invertebrate (snails, insects, crayfish, etc.); Reptile (turtles, alligators); Bird and other (amphibians, plants, household pets, stones, and other debris). These data was described by a baseline-category model, with primary food choice as the outcome and lake, sex, and size as covariates. In the study, the overall test of relationship among the independent variables and groups defined by the dependent variables is based on the reduction in the likelihood values for a model, which does not contain any independent variables and the model that contains the independent variables. This difference in likelihood follows a Chi-square distribution, and is referred to as the model Chi-square. The significance test for the final model Chi-square (after the independent variables have been added) is our statistical evidence of the presence of a relationship between the dependent variable and the combination of the independent variables.

Modelling the relationship between explanatory and response variables is a fundamental activity encountered in statistics. Multinomial logistic regression is used to predict a categorical variable from a set of predictor variables. Multinomial logistic regression models are used to model relationships between a polychotomous response variable and a set of regressor variables. These polychotomous response models can be classified into distinct types, depending on whether the response variable has an ordered of unordered structure.

In an ordered model, the response of an individual unit is restricted to one of m ordered values. For example, the severity of a medical condition may be; none, mild, and severe. The cumulative logit model assumes that the ordinal nature of the observed response is due to methodological limitations in collecting the data that result in lumping together values of an otherwise continuous responses variable by McKelvey and Zavoina [7]. In an unordered model, the polytomous response variable does not have an ordered structure. Two classes of models, the generalized logit models and the conditional logit models, can be used with nominal response data. The generalized logit model consists of a combination of several binary logits estimated simultaneously. For example, the response variable of interest is the occurrence of non-occurrence of infection after a caesarean section with two types (I, II) of infection. Two binary logits are considered: one for type I infection versus no infection and the other for type II infection versus no infection. The conditional logit model has been used in biomedical research to estimate relative risks in matched case-control studies. The nuisance parameters that correspond to the matched sets in an unconditional analysis are eliminated by using a conditional likelihood that contains only the relative risk parameter.

A study by Muriithi et al. [4], used ordinal logistic regression and multiple binary logistic regression to classify applicants for loans and predict their appropriate allocations. Three predictors were used for analysis in this study, which gave fairly good predicted probabilities.

They however expressed need to come up with a better predictive model, which would consider more factors and give results with more precision and accuracy. They recommended the use of multinomial logistic regression model (MLRM), which allows for consideration of more than two categories of the outcome variable. The review of the existing literature shows little application of multinomial logistic regression as a predictive tool. Woo-Yong et al. [5], for example, used MLRM to ascertain and predict the extent to which saltwater anglers were willing to substitute fishing at one location for fishing at another location. They confirmed that MLRM provided more satisfactory results compared to other analysis techniques because it not only requires strict assumptions but also enables a direct interpretation of the relationship between independent variables and the dependent variables. Chao-Ying et al. [6] used MLRM to predict adolescent risk behaviours based on several personal as well as family characteristics. Another example is that of Yun-Wang [8], who applied MLRM to predict multi-attack types of anomaly intrusion and profile use behaviour in computers. He used bootstrap simulation method.

3. Methodology

Multinomial logistic regression

Let us consider a situation where an individual faces K choices and a set of variables characterizes the individual. Let i-th individual be characterized by the vector $z_i = (z_{i1}, \ldots, z_{im})$ contains variable like age, sex, and income. Consequently, U_{ir} will denote the utility of the r-th category for individual i. Y_i will denote the categorical response variable. A simple linear model for U_{ir} is given by

$$U_{ir} = \beta_{r0} + z_i'\beta_r, \tag{1}$$

where $\beta_r = (\beta_{r0}, \ldots, \beta_{rm})$ is a parameter vector. That means the preference of the r-th alternative for i-th individual is determined by z_i

and a parameter β_r that depends on category. Therefore, a parameter that depends on the category will be called *specific*, the variable z_i are called *global*.

Assuming only global variable z_i the multinomial logistic regression model is given by

$$P(Y_i = r) = \frac{\exp(\beta_{r0} + z_i'\beta_r)}{1 + \sum_{s=1}^{m} \exp(\beta_{s0} + z_i'\beta_s)}, \tag{2}$$

which can be written equivalently as

$$\text{Log}\,\frac{P(Y_i = r)}{P(Y_i = k)} = \beta_{r0} + z_i'\beta_r. \tag{3}$$

Hence z_i is the vector of covariables determining the log odds for category r with respect to the category k. We fit a multi-categorical logit model to the HELB data taking Kshs 60,000 as reference category of Y.

For the reference category

$$P(Y_i = k) = \frac{1}{1 + \sum_{s=1}^{m} \exp(\beta_{s0} + z_i'\beta_s)}. \tag{4}$$

According to the equation above, the models can be written as

$$\text{Log}\,\frac{P(35{,}000)}{P(60{,}000)} = \beta_{10} + z_i'\beta_1; \tag{5}$$

$$\text{Log}\,\frac{P(40{,}000)}{P(60{,}000)} = \beta_{20} + z_i'\beta_2. \tag{6}$$

In general can be written as

$$\log\frac{P(Y_i = r)}{P(Y_i = k)} = \beta_{r0} + z_i'\beta_r. \tag{7}$$

The response variable is

$$Y_i = \begin{cases} 1 & \text{if awarded Kshs 35,000,} \\ 2 & \text{if awarded Kshs 40,000,} \\ 3 & \text{if awarded Kshs 45,000,} \\ 4 & \text{if awarded Kshs 50,000,} \\ 5 & \text{if awarded Kshs 55,000,} \\ 6 & \text{if awarded Kshs 60,000.} \end{cases}$$

In fitting such a model, the study estimate m–1 sets of regression coefficients. The regression coefficients should give the significance of each independent variable to the outcome Y_i. The estimated regression coefficients, forming the model, will be used to classify the remaining part of the data into either of five groups. The outcome of the classification will be compared with already known outcome (model without independent variables). Finally, the percentage of the correctly classified data will be obtained. The percentage performance will be used for comparison.

4. Data Collection

Stratified random sampling

This method of sampling is applied where the population embraces a number of distinct categories. The frame can be organized by these categories into separate strata. Each stratum is then sampled as an independent sub-population, out of which individual elements can be randomly selected. This sampling method is used because sampling problems may differ remarkably in different parts of the population. In this paper, strata were HELB loan applicants from Kenyatta University (KU), Jomo Kenyatta University of Agriculture and Technology (JKUAT), and Kimathi University College of Technology (KUCT). Each stratum

was treated as an independent population and proportional sampling approaches were applied to different strata. The stratum population size was the number of students graduating or admitted per academic year to the university.

The data that was captured by questionnaire included: gender, year of birth, if one has both parents, fathers year of birth and his occupation; mothers year of birth and her occupation, guardians year of birth and their occupation; how many siblings one has, year of birth of the first and last born; the level of education of the siblings, category of high school attended, the amount of fees paid per year and who exactly did the payment. Also captured: the year one was admitted to the university; the year of study they are in; what amount is spent on fees, food, clothes, rent, medical bills, and travel per year; what amount was awarded by HELB and in which years did he receive the loan; what is HELB in the context of transparency, fairness, courtesy, competence, and if it works as a team. Also, it captured: what the parents own in terms of motor vehicles and domestic animals; what land acreage they have; what are the sources of income of the family; what kind of house they are living in and whether rented or owned. For the owned houses, information was recorded on the roofing materials, wall materials, floor materials, and how many rooms there are in total. For the ones in rented accommodation, interest is on how much they pay as rent.

Finally, the data was analyzed by using statistical package for social scientist (SPSS) and R-program. This information was used to come up with a conclusion and probably a recommendation.

5. Empirical Results

The questionnaire was administered physically. The data has the consideration that all the respondents were awarded or not awarded a loan on application.

Data coded

0-No

1-Yes

In this analysis, the following variables were under consideration:

x_1 -House worth (computed from the total materials of the house they live in). In the survey, questionnaire applicants were to indicate what kind of house they live in, whether in rented or their own. We also inquired on what materials were used and we had roof, floor, walls, and how many rooms. We valued all the materials and computed how many are required for each room. Then, the total value of the house was the sum total of the materials by the number of rooms.

x_2 -Wealth (sum of property owned by the parents of an applicant). This was computed from lorry/bus, Matatu, motorbike, bicycle, cars, cows, goats, poultry, land, and posh mill. Each was allocated a fixed value and this was used to multiply with the number of units of the property.

x_3 -Amount of fees (the cost of education of the siblings of an applicant each year). This was done for those in secondary school and college. We inquired on the number of siblings the respondents had and who were still in school. This was intended to know what other expenses the parents had. A constant value for those in secondary school and college per year was given.

5.1. Multinomial logistic regression results

This section presents the results of multinomial logistic regression model.

Table 1. Model fitting information

	Model fitting criteria		Likelihood ratio test	
	– 2Log likelihood	Chi-square	df	Sig
Intercept only	1077.825			
Final	989.808	88.016	15	0.000

The presence of a relationship between the dependent variable and combination of independent variables is based on the statistical significant

of the final model Chi-square value. In this case, the probability of the model Chi-square (88.016) was 0.00, less than significance level of 0.05. The null hypothesis that there was no difference between the model without independent variables and model with independent variables is rejected. Thus, there exists sufficient evidence that a relationship between the independent variable and the dependent variable was statistically significant at 5% significance level.

Table 2. Likelihood ratio tests

	Model fitting criteria		Likelihood ratio test	
Effect	– 2Log likelihood	Chi-square	df	Sig
Intercept	1030.196	40.388	5	0.000
x_1	1018.948	29.140	5	0.000
x_2	1009.203	19.395	5	0.002
x_3	1023.853	34.045	5	0.000

The Chi-square statistic is the difference in – 2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0. In this case, there exist a statistically significant relationship between independent variables (house worth, wealth, and amount of fees) and the dependent variable (loan awarded) since the significance value were all less than 0.05. This implies that the independent variables (house worth, wealth, and amount of fees) were highly statistically significant hence considerable in multinomial logistic regression model.

Table 3. Model results

Loan		B	S. E.	Wald	df	Sig	Exp (B)
	Intercept	– 1.888	0.569	11.031	1	0.001	
1	x_1	0.047	0.033	1.957	1	0.162	1.048
	x_2	0.001	0.002	0.355	1	0.551	1.001
	x_3	0.073	0.039	3.465	1	0.063	1.075
	Intercept	– 1.118	0.390	8.199	1	0.004	
2	x_1	0.055	0.029	3.634	1	0.057	1.057
	x_2	0.004	0.001	11.085	1	0.001	1.004
	x_3	0.097	0.028	11.832	1	0.001	1.102
	Intercept	0.030	0.342	0.008	1	0.930	
3	x_1	0.040	0.029	1.924	1	0.165	1.040
	x_2	– 0.004	0.001	11.318	1	0.001	0.996
	x_3	0.060	0.027	5.004	1	0.025	1.062
	Intercept	– 0.263	0.437	0.361	1	0.548	
4	x_1	– 0.057	0.047	1.461	1	0.227	0.945
	x_2	– 0.002	0.001	4.164	1	0.041	0.998
	x_3	0.011	0.037	0.097	1	0.756	1.011
	Intercept	– 1.733	0.536	10.452	1	0.001	
5	x_1	– 0.069	0.029	5.649	1	0.017	0.933
	x_2	– 0.004	0.001	8.889	1	0.003	0.996
	x_3	0.075	0.056	1.816	1	0.178	1.078

Table 3 shows the multinomial logistic regression coefficients, Wald test, and odds ratio for each of the predictors in all the five categories under study. Employing a 0.05 criterion of statistical significance, all the variables had no significant effects when a loan of Kshs 35,000 is awarded. However, the amount of fees is marginally significant with a significance level of $0.063 > 0.05$. The exponentiated coefficients in the last column of the output are interpretable as multiplicative effects on

loan. Thus, for example, holding all other variables constant, an additional unit of amount of fees increases the likelihood of being awarded a loan (Kshs 35,000) by a factor of 1.075 on average, i.e., an increases by 7.5%. The Wald statistic tests the unique contribution of each predictor in the context of the other predictors in each of the category. For instance, in the first category (loan of Kshs 35,000), amount of fees paid has higher contribution with Wald statistic value of 3.465. The multinomial logistic regression model in this paper is thus

$$\log \frac{P(Y_i = r)}{P(Y_i = k)} = \beta_{r0} + z_i'\beta_r, \tag{8}$$

where $P(Y_i = r)$ is the probability of being allocated the loan in r-th category with reference to k-th category. In the first category, we develop the following model although none of the variable is statistically significant at 5% significance level:

$$\text{Log} \frac{P(35{,}000)}{P(60{,}000)} = -1.888 + 0.073x_3. \tag{9}$$

In the second category (loan of Kshs 40,000), we observed that the significance values of the last two variables (x_2 and x_3) were less than 0.05 ($0.001 < 0.05$ and $0.001 < 0.05$, respectively) meaning they are statistically significant. When holding all other variables constant, an additional unit of amount of fees increases the likelihood of being awarded a loan (Kshs 40,000) by a factor of 1.102 on average, i.e., an increases by 10.2%. In this case, the Wald statistic tests (11.832) of the amount of fees paid has higher unique contribution than other predictors. The model equation in this category is thus

$$\text{Log} \frac{P(40{,}000)}{P(60{,}000)} = -1.118 + 0.055x_1 + 0.004x_2 + 0.097x_3. \tag{10}$$

In the third category (loan of Kshs 45,000), we found that the (x_2 and x_3) had a significant effect in awarding a loan of Kshs 45,000. Their

significance levels were less than 5% $(0.001 < 0.05$ and $0.025 < 0.05)$. In terms of contribution in the model, wealth has higher contribution than others with Wald statistic value of 11.318. The exponentiated coefficients, indicate that holding other variables constant, an additional unit of wealth decreases the likelihood of been awarded a loan by a factor of 0.996 (0.4%) on average. The model

$$\text{Log}\frac{P(45{,}000)}{P(60{,}000)} = -0.004x_2 + 0.060x_3. \tag{11}$$

In the fourth category (loan of Kshs 50,000), we found that the wealth has a significant effect on awarding a loan of Kshs 50,000. The exponentiated coefficient, indicate that holding other variables constant, an additional unit of wealth decreases the likelihood of been awarded a loan (Kshs 50,000) by a factor of 0.998 (0.2%) on average. Wealth was the only variable statistically significant at 5% significance level $(0.041 < 0.05)$. The model equation in this category is thus

$$\text{Log}\frac{P(50{,}000)}{P(60{,}000)} = -0.002x_2. \tag{12}$$

In the fifth category (loan of Kshs 55,000), we observed that the wealth and house worth had significant effect on awarding a loan of Kshs 55,000. In fact, wealth has a higher contribution (Wald statistic = 8.889) than other variable in awarding the loan. It was found that, holding all other variables constant, an additional unit of wealth decreases the likelihood of being awarded a loan (Kshs 55,000) by a factor of 0.996 on average, i.e., a decrease by 0.4%. Also, when holding all other variables constant, an additional unit of house worth decreases the likelihood of being awarded a loan (Kshs 55,000) by a factor of 0.933 (6.7%) on average. The model equation in this category is thus

$$\text{Log}\frac{P(55{,}000)}{P(60{,}000)} = -0.733 - 0.069x_1 + 0.004x_2. \tag{13}$$

5.2. Computing probabilities

We fit a logits model for each category (r = 1, 2, 3, 4, 5). The logit estimates for one unit each for wealth, house worth, and amount of fees paid for siblings were estimated. The probability of the response in category r can be calculated with Kshs 60,000 (maximum loan) as reference category as follows:

$$P(Y = 1) = \frac{\exp(-1.825)}{1 + \exp(-1.825) + \exp(-0.962) + \exp(0.056) + \exp(-0.002) + \exp(-1.804)} = 0.043; \quad (14)$$

$$P(Y = 2) = \frac{\exp(-0.962)}{1 + \exp(-1.825) + \exp(-0.962) + \exp(0.056) + \exp(-0.002) + \exp(-1.804)} = 0.102; \quad (15)$$

$$P(Y = 3) = \frac{\exp(0.056)}{1 + \exp(-1.825) + \exp(-0.962) + \exp(0.056) + \exp(-0.002) + \exp(-1.804)} = 0.281; \quad (16)$$

$$P(Y = 4) = \frac{\exp(-0.002)}{1 + \exp(-1.825) + \exp(-0.962) + \exp(0.056) + \exp(-0.002) + \exp(-1.804)} = 0.265; \quad (17)$$

$$P(Y = 5) = \frac{\exp(-1.804)}{1 + \exp(-1.825) + \exp(-0.962) + \exp(0.056) + \exp(-0.002) + \exp(-1.804)} = 0.044; \quad (18)$$

$$P(Y = 6) = \frac{1}{1 + \exp(-1.825) + \exp(-0.962) + \exp(0.056) + \exp(-0.002) + \exp(-1.804)} = 0.266. \quad (19)$$

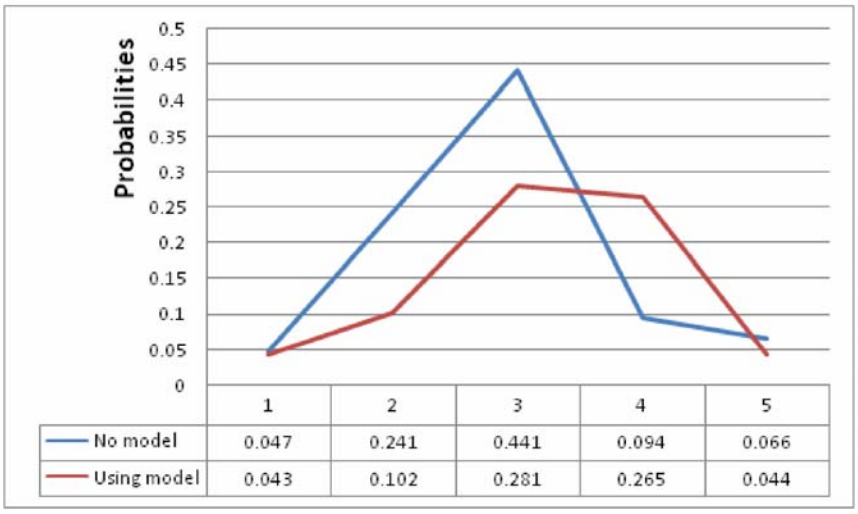

Figure 1. Distribution of amount awarded.

As shown in Figure 1, we observe that the likelihood of being correctly allocated Kshs 35,000 depending on the variable used in Equation (14) is less than 5%. This implies that for applicant from wealth back ground, wealth will reduce the chances of getting maximum loan. Also, we observe that probability of being allocated Kshs 55,000 using model is 4.4% as shown in the Figure 1 above. In general, it was revealed that the chances of been allocated loan on both extreme ends is very minimal with higher chances on middle categories. Also, we observed that, some of the probabilities in both case were very close to each other. Therefore, mean square error (MSE) computed to determine the error associated with model for recommendation in this study.

6. Conclusion and Recommendations

This study was aimed at investigating whether the HELB loan is allocated based on the need of the respective applicants. The study found that the loan was not awarded based on the need of respective applicant. This has led to mis-classification when allocating loans. The study revealed that wealth, house worth, and amount of fees paid for siblings were other factors that could be considered to identify needy student.

The multinomial logistic regression model discussed above has the significance being less than 0.05, hence it is statistically significant. We observed that wealth and amount of fees paid for siblings were statistically significant in this model. The above model can classify about 79.8% of the cases under study. The results show fairly good predicted probabilities that can be used for loan allocation or classification.

In this case, the probability of the model chi-square was 0.00, less than significance level of 0.05. The null hypothesis that there was no difference between the model without independent variables and model with independent variables was rejected. Thus, there exists sufficient evidence that a relationship between the independent variable and the dependent variable was statistically significant at 5% significance level.

Using mathematical tool, the mean square error for the multinomial regression model was 0.0012304 hence an appropriate model than no model at all. In conclusion, we recommend use of multinomial regression model for HELB loan allocation to determine the amount of loan, if any, to be awarded to an applicant. This will minimize the number of mis-classification when awarding HELB loan.

References

[1] http://helb.co.ke.

[2] P. Chacha, Logistic Regression Versus Neural Networks in Classification of Binary Data, Master's Thesis, Jomo Kenyatta University of Agriculture and Technology, 2007.

[3] Delany and Moore, Classification of Primary Food Choices of Alligators in a Lake, Master's Thesis, Florida Lake, 1987.

[4] D. K. Muriithi, J. Waititu and A. Waititu, Ordinal logistic regression versus multiple binary logistic regression model for predicting student loan allocation, Journal of Agriculture, Science and Technology 14(1) (2012).

[5] W.-Y. H. and D. R. B., Using multinomial logistic regression analysis to understand anglers willingness to substitute other fishing locations, Proceedings of the Northeastern Research Symposium (2006), GTR-NRS-P-14.

[6] J. P., C.-Y. and N. R. N., Using multinomial logistic models to predict adolescent behavioural risk, Journal of Modern Applied Statistical Methods 2(1) (2003).

[7] R. McKelvey and W. Zavoina, A statistical model for the analysis of ordinal level variable, Journal of Mathematical Sociology 4 (1975), 103-120.

[8] Y. Wang, A multinomial logistic regression modelling approach for anomaly intrusion detection, Computers Security 24 (2005), 662-674.

[9] M. Wedel, Concomitant variables in finite mixture models, Statistica Neerlandica 56 (2002).

■