USING MULTINOMIAL LOGIT MODEL FOR ANALYSIS OF EDUCATIONAL SURVEY DATA

Hairui Tang, College of Engineering, California State Polytechnic University, Pomona, CA 91768, 626-734-3498, hairuitang@cpp.edu
Wen Cheng, College of Engineering, California State Polytechnic University, Pomona, CA 91768 909-869-2957, wcheng@cpp.edu

ABSTRACT

This article develops a multinomial logit model to research and identify influential factors about learning outcomes. The model is based on a survey data of senior graduates at California State Polytechnic University Pomona in 2014 and contains information with two-part. The results show that the influence weight of each independent variable on a certain ability is disparate, and different competence is also significantly affected by different factors, which is detailed in part of the result. The generalized model will be dedicated to helping companies analyze and select talents.

Keywords: Multinomial logit model, Survey data, Student self-evaluation, Curriculum.

INTRODUCTION

Under the current social development, enterprises have higher requirements on the professional ability of new employees, and the cultivation of professional ability mainly comes from school period. Therefore, students' academic performance in school may become the key reference for enterprises to select talents. The original intention of this survey for Cal Poly Pomona is to find out the fields where students' abilities are relative low through the students' self-assessment, and based on clarifying the correlation between academic performance and outcomes, they can adjust the educational focus and improve the curriculum system.[3] It is related to the proposal of this research. This article is going to explore the influential factors of various professional and technical abilities by using a multinomial logit model, so that companies can also more accurately judge some competence of the graduates which they expect.

Multinomial Logit model is used to establish a model with multiple level's output variables, and these predictive variables become a final predictive variable through a linear combination. Multinomial logit model in the dependent variable can take categorical values. An example is the film classification, which is "interesting", "so-so" or "boring". This model is based on the multinomial logistic regression which has been introduced in many textbooks for applied statistical analysis. Combined with experimental data, each professional ability can be considered as a dependent variable, and the student's campus experience and academic performance are independent variables. Then the stepwise function can be used to determine multinomial logistic regressions. Finally, the generalized model will be chose statistically significant variables by minimizing AIC.

Usually, the multinomial Logit model based on minimizing AIC to select the optimization model, rather than an ordinary linear model regarding the R square or p value. Akaike's

information criterion (AIC) proposed by Akaike is widely used for selecting the best model among the candidate models[2]. A smaller AIC value indicates that the error of the competition model is lower and more reliable. In the multinomial logistic regression, the subset of explanatory variables in the optimization model is the best. Akaike's information criterion is increasingly being used in data analyses in the field of feature selection. This measure allows one to compare and rank multiple competing models and to estimate which of them best approximates the "true" process underlying the phenomenon under study.

DATA DESCRIPTION

The survey was collected from a group of CPP graduating seniors of civil engineering, including two parts. The first part is the experience and performance of students during their school years. There are 12 questions which are considered as input data in the multinomial logit model.

The second part represents the student's learning outcomes and covers 11 different abilities detailedly presented in Table 1. In order to reflect the difference in outcomes, each result was given five different levels of scores: (1) poor; (2) fair; (3) average; (4) good; (5) excellent, and a total of 507 students provided valid data for the survey. Table 2 summarizes the frequency of responses for each outcome student, which will be used to develop the initial multinomial logit model. The initial point of this study is going to select five lowest average outcome scores to make the targeted analysis which is calculated in table 2. Then, start to generalize the multinomial model for each outcome.

TABLE 1. Related Abilities

Student Outcomes ID	Description of Student Learning Outcomes
a	Ability to apply knowledge of mathematics, science, and engineering.
b	Ability to design and conduct civil engineering experiments, as well as to analyze and interpret data.
c	Ability to design a system, component, or process to meet desired needs within realistic constraints such as economic, environmental, social, political, ethical, health and safety, manufacturability, and sustainability.
d	Ability to function on multidisciplinary teams.
e	Ability to identify, formulate, and solve engineering problems.
f	Understanding of professional and ethical responsibility.
g	Ability to communicate effectively.
h	Understanding of the impact of engineering solutions in a global, economic, environmental, and societal context.
i	Recognition of the need for, and an ability to engage in life-long learning.
j	Knowledge of contemporary issues and their importance to engineering systems.
k	Ability to use the techniques, skills, and modern engineering tools necessary for engineering practice.

TABLE 2. Descriptive Statistics of Responses to Self-Evaluation of Student Outcomes

	Count of Responces	**Total**

Student Outcomes	Excellent (5)	Good (4)	Average (3)	Fair (2)	Poor (1)	Rating Average*	Response Count
(a)	232	238	31	4	2	4.37	507
(b)	173	262	63	7	2	4.18	507
(c)	143	237	113	13	1	4.00	507
(d)	271	199	31	5	1	4.45	507
(e)	217	243	41	5	1	4.32	507
(f)	320	167	14	5	1	4.57	507
(g)	249	207	40	10	1	4.36	507
(h)	228	228	43	7	1	4.33	507
(i)	302	174	28	2	1	4.53	507
(j)	184	243	69	9	2	4.16	507
(k)	197	260	43	7	0	4.27	507

METHODOLOGY OF MULTINOMIAL LOGIT MODEL

Despite each model used in the literature related to data analysis of categorical variables has its unique advantages, the MNL model seems to be the most common technique for identifying the relationship between dependent variables and independent variables. Here just use the "a" ability as an example, which is the ability to apply knowledge of mathematics, science, and engineering. As mentioned earlier, student learning outcomes are assigned to one of five discrete categories: 1. poor; 2. fair, 3. average, 4. good, 5. excellent. By importing these five discrete capability levels and other independent variables in the computer language R, a statistical model can be automatically derived, which can also be used to determine the probability of each categorical outcome happening to a certain student. The probability of each type of outcome for a student can be recorded as:

$$P_n(i) = P(U_{in} \geq U_{jn}) \; \forall j \neq i \tag{1}$$

Where $P_n(i)$ is the probability of having a learning outcome level i, and U_{in} is a function that determines the student's n-level of learning outcomes. To estimate this possibility, U_{in}'s linear function can be expressed as:

$$U_{in}=\beta_i X_n + \varepsilon_{in} \tag{2}$$

where X_n is a vector of explanatory variables that determine the level of student learning outcomes, β_i represents a vector of estimable coefficients for learning outcome level i, and ε_{in} is an error term that accounts for unobserved factors influencing evaluated level and is assumed to be identically and independently distributed. The term $\beta_i X_n$ in this equation is the observable component and ε_{in} is the unobserved portion.[1] Based on formular (1) and (2), the following equation can be written:

$$P_n(i) = P(\beta_i X_n - \beta_j X_n \geq \varepsilon_{jn} - \varepsilon_{in}) \; \forall j \neq i \tag{3}$$

Based on equation 3, assuming a distributional form for the error term can develop the model of estimating learning outcomes. Then, technically use generalized extreme value (GEV)

distribution to produce a closed form model that can be readily estimated using standard maximum likelihood methods by using following equation[5]:

$$P_n(i) = \exp(\beta_i X_n)/\sum_j \exp(\beta_i X_n) \quad (4)$$

METHODOLOGY OF CALCULATING AIC

AIC is calculated by using the number of fitted parameters, including the intercept, in the model (k), and either the maximum likelihood estimate for the model (L) or the residual sum of squares of the model (RSS), two measures that are also easily derived from the output of any statistics package. In the case of least-squares regression analyses, the value of k must be increased by 1 to reflect the variance estimate as an extra model parameter. [4] AIC is calculated as:

$$AIC = -2\ln(L)+2k \quad (5)$$

if using likelihood or

$$AIC= n[\ln(\frac{RSS}{n})]+2k \quad (6)$$

RESULT

This study selects five abilities that the students showed the lowest outcome grades overall to analyze related influential factors:

b. Ability to design and conduct civil engineering experiments, as well as to analyze and interpret data.

c. Ability to design a system, component, or process to meet desired needs within realistic constraints such as economic, environmental, social, political, ethical, health and safety, manufacturability, and sustainability.

e. Ability to identify, formulate, and solve engineering problems.

j. Recognition of the need for, and an ability to engage in life-long learning.

k. Ability to use the techniques, skills, and modern engineering tools necessary for engineering practice.

After the Multinomial logit model generated, AIC value will be the one and only reference to select competition model in this research. Actually, model selection is a process of choosing independent variables included in the model. Using different input variables in the model is going to develop a different model. The purpose of AIC here is that select an optimization model with the most influential variables related to the dependent variable. As mentioned earlier, this is a process of gradually reducing the AIC value. This process will constantly try to get the AIC value after subtracting each independent variable and compare with each other. The largest AIC value corresponding to the subtracted independent variable will be eliminated and the new model will be generated without this variable. Then continue this process until the AIC value obtained by subtracting any of the independent variables is not less than the AIC value of the current model. Finally, the AIC value of the current model will become the final AIC value, and the

variables still retained in this model are the most influential factors for the dependent variable processed by this process.

Based on AIC step by step process, influential factors of each ability is obtained presented from Table 3 to Table 7. The degree of effect of each factor on a possible outcome is different. As shown in the column of "Coefficients", there are some factors that have a greater impact weight, and some factors have negative impact weight.

TABLE 3. Influential Factors of Learning Outcome: b

Learning Out come	1	2	3	4	5	AIC Value without This Variable
Trait	Coefficients					
Q2	-	4.2867	4.2777	4.4024	5.0328	1067.520
Q7	-	5.9172	4.5564	5.4245	5.2074	1064.779
Q12	-	6.8721	6.3969	6.1953	6.2154	1067.520
Intercept	-	38.1320	38.0711	38.0669	38.0679	-
Final AIC	1063.689					

TABLE 4. Influential Factors of Learning Outcome: c

Learning Out come	1	2	3	4	5	AIC Value without This Variable
Trait	Coefficients					
Q4	-	85.52580	85.81096	86.70278	86.71185	1171.272
Q6	-	91.85218	92.29134	92.22124	93.06026	1165.882
Q8	-	-56.82973	-56.69983	-56.37731	-56.04308	1164.828
Q9	-	-50.54482	-50.34076	-50.72751	-50.86757	1162.804
Q12	-	-37.06390	-36.80031	-36.93123	-36.84164	1163.498
Intercept	-	196.7585	196.8121	196.8990	194.7223	-
Final AIC	1160.787					

TABLE 5. Influential Factors of Learning Outcome: e

Learning Out come	1	2	3	4	5	AIC Value without This Variable
Trait	Coefficients					
Q4	-	4.2867	4.2777	4.4024	5.0328	995.0973
Q7	-	5.9172	4.5564	5.4245	5.2074	986.2921
Intercept	-	9.4364	9.2790	9.2542	9.2553	-
Final AIC	986.1571					

TABLE 6. Influential Factors of Learning Outcome: j

Learning Out come	1	2	3	4	5	AIC Value without This Variable
Trait	Coefficients					
Q4	-	4.6952	5.7684	6.0314	6.5646	1103.903
Q10	-	18.6095	17.5167	18.7988	19.1424	1100.062
Intercept		10.8772	10.8088	10.7979	10.7989	-
Final AIC	1097.425					

TABLE 7. Influential Factors of Learning Outcome: k

Learning Out come	1	2	3	4	5	AIC Value without This Variable
Trait	Coefficients					
Q2	-	-	0.9275	0.8812	0.8823	993.3849
Q4	-	-	1.1873	1.1295	1.1309	992.706
Intercept	-	-	0.9879	2.6391	1.3211	-
Final AIC	989.5527					

CONCLUSION

The five abilities chose to study are relatively lacking for Cal Poly Pomona students. Modeling through the Multinomial logistic regression and using AIC to gradually select the variables contained in the model, the result shows influential factors of each learning outcome. From these factors, Cal Poly Pomona can find the mean point to adjust teaching focus that further improve the quality of education. These results are expected to take additional insights for department managers and administrators as they design courses that meet students' needs. As well enterprises can also refer to these factors of generally low ability before interviewing students, so that they can have a better understanding of the focus of recruitment.

In the field of statistical analysis, for the multiple categories data, multinomial logit model has been more and more popular. Because it does not assume normality, linearity, or homoscedasticity, and more data types are available. For this study, the deficiency is the accuracy of the original data, because this is a student self-evaluation survey, so it is hard to guarantee the objectivity and impartiality of the data. Some students may overestimate their abilities, while others may underestimate their abilities. If students' self-evaluation can be combined with teachers' evaluation of students, it is anticipated to yield more reliable findings.

REFERENCES

1. Zhen Chen; Wei(David) Fan (2019). “A multinomial logit model of pedestrian-vehicle crash severity in North Carolina.” International Journal of Transportation Science and Technology 8 (2019) 43–52.
2. Matthew R. E. Symonds; Adnan Moussalli (2011). “A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike’s information criterion.” Behav Ecol Sociobiol (2011) 65:13–21
3. Thurein Shwe; Wen Cheng. “Clustering Analysis of Student Learning Outcomes Based on Education Data.”
4. Hirokazu Yanagihara; Ken-ichi Kamo, Shinpei Imori, Kenichi Satoh (2012). “Bias-corrected AIC for selecting variables in multinomial logistic regression models.” Linear Algebra and its Applications 436 (2012) 4329–434.
5. Venkataraman Shankar; Fred Mannering (1996). “An Exploratory Multinomial Logit Analysis of Single-Vehicle Motorcycle Accident Severity.” Joumal of Safety Research, Vol. 27, No. 3. pp. 183-194.1996