Application of Multinomial Logistic Regression to analyze learning difficulties in statistics courses

A Abdillah[1], A Sutisna[2], I Tarjiah[3], D Fitria[1] and T Widiyarto[1]

[1] Economic Education Department, Indraprasta PGRI University
[2] Educational Research and Evaluation Department, Jakarta State University
[3] Special Education Department, Jakarta State University

e-mail: agus.abdillah@unindra.ac.id

Abstract. Every student basically wants to get the opportunity to achieve satisfactory academic performance. But from daily reality, it seems clear that students have differences in intellectual abilities, physical abilities, family background, habits and learning approaches that are sometimes very striking between a student and other students. The data used in this study consisted of primary data and secondary data. Primary data in the form of a questionnaire for the dependent variable and response variable. While secondary data contains general information on the number of students in the fourth semester of economic education courses obtained from the administration of the faculty of education, Unindra. Data analysis using multinomial logistic regression analysis. Based on the results of this study, independent variables that influence student learning difficulties in Statistics courses include; gender, regional origin, learning motivation, and learning resources. While the variables that did not have a significant effect on learning difficulties include; major at school, monthly tuition allowance, and study patterns. Furthermore, the value of the accuracy of the classification of students produced using multinomial logistic regression analysis of seventy-six point five percent.

1. Introduction

Every student basically wants to get the opportunity to achieve satisfactory academic performance. But from daily reality, it seems clear that students have differences in intellectual abilities, physical abilities, family background, habits and learning approaches that are sometimes very striking between a student and other students.

Meanwhile, the delivery of education in high schools, in universities, in general, is only shown to students with average abilities, so students who have more ability or who have less ability are neglected. Thus, students who are categorized as "out of average" do not get adequate opportunities to develop in accordance with their capacity. From this then arises what is called learning difficulties which not only afflict average students, but also those with low abilities and those with high abilities. Difficulty learning is a symptom that is seen in students which are characterized by a low learning achievement or below the established norms [1].

Based on these problems, it is no longer downsized that in education there are various kinds of difficulties caused by the situation or nature of the students themselves or by the environment and or by the educator himself [2].

One factor that can be categorized as an innate factor affecting learning difficulties is syndrome. Syndromes can be interpreted as a unit of symptoms that appear that indicate psychological abnormalities. These types of syndromes include; dyslexia namely the inability to learn to read, dysgraphia ie the inability to learn to write, and dyscalculia namely the inability to learn mathematics. However, students who experience the above syndromes in general actually have a normal IQ potential even among them having above-average intelligence. Therefore, the learning difficulties of students suffering from these syndromes may only be due to minimal brain dysfunction, which is a mild disturbance in the brain [3].

Furthermore, motivation to learn becomes one of the factors of learning difficulties, because motivation is recognized as the driving force that drives learning activities. If motivation is low on a learning subject, students will tend to ignore it and will lead to learning outcomes that are not optimal. Motivation also influences learning achievement, high and low motivation is always used as an indicator of the good and bad learning achievement of a student [4]. In addition, the factor of study habits is also a study in this study. Study habits of students who study only at a time determined by the campus and learning techniques that are used less precisely such as memorization. Though skills require a learning process based on practice and practice that is carried out continuously and steadily. And language skills certainly require mastery of the linguistic component and mastery of the content component. These two things are basic in speaking skills. As for mental attitude in the form of anxiety which can be in the form of lack of mastery of the linguistic component and mastery of the component content. Meanwhile, anxiety will cause learning difficulties [5].

Several previous studies have examined how to overcome learning difficulties, including research; [6], [7]. Most researchers analyze their research data using multiple regression analysis. Not too much considering that learning difficulties are the response variable or the dependent variable in the form of qualitative data. Therefore, the authors try to present a different analysis from previous studies. Multinomial logistic regression analysis the authors believe is the most suitable model. That is because the indicators on the learning difficulty variable are alternative answers that are categorical or qualitative. In accordance with the theme of this study, in essence, the writer wants to know and analyze the characteristics of students. So the data obtained are nominal data and also ordinal data. As for some relevant studies that use this analysis technique, among others; [8], [9] and [10]. Nevertheless, the focus of the problem faced is not the same as this research.

The general objective of this research is to find out and analyze the factors that influence students' learning difficulties in Statistics courses, economic education courses, University of Indraprasta PGRI. While the specific purpose of this study is to classify the characteristics of economic education students who have learning difficulties in the Statistics subject in terms of several aspects including; gender (X_1), regional origin (X_2), majors taken at school (X_3), monthly tuition allowance (X_4), learning motivation (X_5), learning resources (X_6), and learning patterns (X_7).

2. Data and samples
The population in this study were all students of semester 4 of Unindra's economic education study program, which included 1,388 people. By using the Slovin formula a sample of 311 people was obtained. Data consists of primary data and secondary data. Primary data referred to in the form of a questionnaire filled out by students. Both for dependent and independent variables. For the dependent variable (Y) in the form of student opinion about the difficulty of studying the Statistics course. Likewise for the independent variable, in the form of a questionnaire about the characteristics of students consisting of; gender (X_1), regional origin (X_2), majors taken at school (X_3), monthly tuition allowance (X_4), interest in learning (X_5), learning resources (X_6), and learning patterns (X_7).

3. Data analysis technique
Multinomial logistic regression analysis is used to find the best model to describe the relationship between the dependent variable and the independent variable [11]. Called logistic regression, because in this regression analysis the formation of the model is based on logistic curves. The resulting value

of the logistic regression equation is the chance of the event being used as a measure for classification [12]. Data analysis techniques using multinomial logistic regression performed with the following steps:

Estimating parameters, that is to estimate the logit model with a qualitative scale response variable, the parameter estimation technique that is feasible to use is the maximum likelihood method. It aims to get a model that will be used in classifying [13]. The maximum likelihood function is

$$l(\beta) = \prod \pi_0 (x_i)^{y0i} \pi_1(x_i)^{y1i} \pi_2(x_i)^{y2i} \text{ , with } \sum ij = 1 \qquad (1)$$

then the log function is a likelihood

$$l(\beta) = \sum y_{1ig1} (x_i) + y_{2ig2}(x_i) - \ln (1 + e^{g_1(x_i)} + e^{g_2(x_i)} \qquad (2)$$

Test the significance of parameters simultaneously with the test statistic G or likelihood ratio using a hypothesis

$H_0 : \beta_1 = \beta_1 = ….= \beta_p = 0$

H_a : there is a minimum of one $\beta_j \neq 0$, with $j = 1,2,3,.., p$

$$\text{So, test statistics } G = -2ln\left[\frac{L_0}{L_a}\right] \qquad (3)$$

Where L_0 is Likelihood without independent variables, and L_a is Likelihood with independent variables. Test criteria rejected H_0 if $G > x^2{}_{(df,\alpha\}}$ or p-value < α.

Therefore, partial test using a Wald test with a hypothesis

$H_0 : \beta_j = 0$

$H_a : \beta_j \neq 0$, with $j = 1,2,3,.., p$

$$\text{So, test statistics } W = \left(\frac{\widehat{\beta}_i}{SE(\widehat{\beta}_i)}\right)^2 \qquad (4)$$

Test criteria, rejected H_0 if $W > x^2{}_{(df,\alpha\}}$ or p-value < α.

The next step is to calculate the accuracy of the classification. The criteria for comparing classification techniques are based on their classification error, known as the Apparent Error Rate (APER). Then interpret the results and draw conclusions.

4. Results and discussion

The initial step taken in this analysis is parameter estimation. The method used in estimating parameters is a maximum likelihood. Multinomial logistic regression models obtained from the data are as follows:

$g_1 x = 20.518 - 21.882X_1 + 2.558X_2 + 1.664X_3 - 1.239X_4 - 5.663X_5 - 17.290X_6 - 0.443X_7$

$g_2 x = 18.424 - 23.601X_1 + 2.779X_2 + 2.109X_3 - 1.381X_4 - 4.687X_5 - 14.350X_6 - 0.404X_7$

$g_3 x = 18.018 - 7.078X_1 + 1.085X_2 + 1.628X_3 - 1.742X_4 - 2.194X_5 - 13.181X_6 - 1.385X_7$

4.1. Simultaneous test

The simultaneous test is a significance test used to find out the variables that simultaneously influence the formation of multinomial logistic regression models. This simultaneous test uses the likelihood ratio test which can be seen in the following table 1.

Table 1. Simultaneous test

	Model Fitting Criteria	Likelihood Ratio Tests			
Effect	-2 Log Likelihood of Reduced Model	Chi-square	df	Sig.	Information
Intercept	105.937	0.000	0		
Gender	286.908	180.970	3	0.000	Significant
Origin	117.856	11.919	3	0.008	Signifikan
Majors at school	108.760	2.823	3	0.420	Not significant
Monthly tuition allowance	109.624	3.687	3	0.297	Not significant
Learning motivation	141.993	36.055	3	0.000	Significant
Learning resources	149.806	43.869	3	0.000	Significant
Learning patterns	107.108	1.171	3	0.760	Not significant

The results obtained in table 1 show that the significant independent variables produce multinomial logistic regression models, among others; gender, regional origin, learning motivation, and learning resources. While the variables that did not significantly affect the model were the majors at school, monthly tuition allowance, and learning patterns.

4.2. Partial test
The partial test is a test of the significance of each parameter in the model using the Wald Test. This test is conducted to determine whether each independent variable can be relied on to build a model or not in the classification process of learning difficulties Statistics courses. The partial test is obtained in Table 2.

Table 2. Partial test

Difficulty Learning		Wald	Sig.	Exp(B)	Information
Easy	Intercept	18.722	0.000		
	Gender	0.000	0.994	3.139	Not significant
	Origin	3.182	0.074	12.904	Not significant
	Majors at school	1.413	0.235	5.281	Not significant
	Monthly tuition allowance	1.004	0.316	0.290	Not significant
	Learning motivation	4.345	0.037	0.003	Significant
	Learning resources	302.300	0.000	3.098	Significant
	Learning patterns	0.011	0.915	0.642	Not significant
Moderate	Intercept	15.628	0.000		
	Gender	0.000	0.994	5.626	Not significant
	Origin	6.916	0.009	16.428	Signifikan
	Majors at school	2.517	0.113	8.241	Not significant
	Monthly tuition allowance	1.376	0.241	0.251	Not significant
	Learning motivation	3.072	0.080	0.009	Not significant
	Learning resources	223.896	0.000	5.861	Significant
	Learning patterns	0.010	0.922	0.668	Not significant
Difficult	Intercept	15.544	0.000		
	Gender	27.399	0.000	0.001	Significant
	Origin	1.297	0.255	2.960	Not significant
	Majors at school	1.538	0.215	5.092	Not significant
	Monthly tuition allowance	2.340	0.126	0.175	Not significant
	Learning motivation	0.679	0.410	0.111	Not significant
	Learning resources	202.055	0.000	1.887	Significant
	Learning patterns	0.116	0.734	0.250	Not significant

Based on the results obtained in table 2, it can be shown that there are two independent variables that distinguish students who have difficulty learning Statistics in the easy and very difficult categories, namely the variable of learning motivation and learning resources. Student motivation and learning resources tend to be more likely to affect learning difficulties in the easy category than learning difficulties in a very difficult category. As for what distinguishes students who have difficulty learning Statistics courses in the medium and very difficult categories, namely the area of origin and learning resources. The area of origin and source of student learning in the medium category tend to be higher than the very difficult category. Then what distinguishes students who have difficulty learning Statistics in the difficult and very difficult categories is gender and learning resources. The sex and learning resources of students in the difficult category tend to be lower than the very difficult categor

Furthermore, the value of the accuracy of the classification of students produced using a multinomial logistic regression analysis of 76.5%. In table 3 shows that the percentage for guessing learning difficulties in Statistics courses with easy categories correctly was 83.3%, suspecting learning difficulties in Statistics courses with moderate categories correctly at 67.8%, and suspecting learning difficulties in Statistics courses with difficult categories correctly at 54.9%. While suspect learning difficulties in Statistics courses with very difficult categories correctly at 92.5%.

Table 3. Classification accuracy test

Observed	Predicted				Percent correct
	Easy	Moderate	Difficult	Very difficult	
Easy	75	14	1	0	83.3%
Moderate	13	61	16	0	67.8%
Difficult	2	20	28	1	54.9%
Very difficult	0	1	5	74	92.5%
Overall percentage	28.9%	30.9%	16.1%	24.1%	76.5%

5. Conclusions

Based on the results of research that has been done, it can be taken that the seven independent variables that influence learning difficulties in Statistics courses using multinomial logistic regression, obtained four independent variables that have a significant effect, including; gender, regional origin, learning motivation, and learning resources. The highest classification opportunity that classifies learning difficulties in Statistics is gender. The value of the accuracy of student classifications produced using a multinomial logistic regression analysis of 76.5%. Furthermore, the suggestion that can be given in connection with this research is that more research needs to be done on other variables that affect students' learning difficulties in the Statistics subject so that better results can be obtained. Male students are expected to be able to compete, must have a high level of struggle or motivation, and adequate learning resources so that their learning abilities are equal to female students. Then the origin of the student area for example from remote or rural areas, should not be used as a reason to lose academic achievement. It should instead be a trigger in order to become successful students, a matter of pride for the residents of their respective villages after graduating from college.

References

[1]   Sugihartiono, et al 2007 *Psikologi Pendidikan* (Yogyakarta: UNY Press) p 149

[2]   P Ngalim 2007 *Ilmu Pendidikan Teoretis dan Praktis* Bandung PT Remaja Rosdakarya Offset

[3]   M Syah 2013 *Psikologi Belajar* Jakarta PT Raja Grafindo Persada

[4]   S B Djamarah 2008 *Psikologi Belajar* Jakarta PT Rineka Cipta

[5]   O Hamalik 2009 *Psikologi Belajar dan Mengajar* Bandung Sinar Baru Algensindo

[6]   Darmawilis, et al 2013 Analisis Kesulitan Belajar Siswa Kelas X IPS Dalam Memahami Konsep Pokok BahasanFungi di SMA Negeri 11 *Jambi Jurnal Pendidikan*

[7]   H Prasetyono, A Abdillah, T Widiyarto and H Sriyono 2018 *Character-based economic learning implementation and teacher's reinforcement on student's affective competence in minimizing hoax* Cakrawala Pendidikan 37(3):426-435

[8]   Soubhi et al 2015 *Learning Difficulties Related of Health Status of Moroccan Students* (Morocco: Observatory of Research in Didactics and University Pedagogy and Universite Pierre Mendes France Grenoble 2)

[9]   J M Zulkifli 2014 *Pendekatan Regresi Logistik Multinomial Pada Klasifikasi Pemilihan Jurusan Siswa SMA Negei 5 Malang* Jurnal Mahasiswa Statistika 2(5): 349 – 352

[10]   W R A Ndangi, Resmawan, I Djakaria 2019 *Analisis Diskriminan dan Regresi Logistik Multinomial pada Proses Pengklasifikasian Siswa* Jambura Journal of Mathematics 1 (2): 54-63

[11]   Subekti 2014 *Model Regresi Logistik Multinomial Untuk Menentukan Pilihan Sekolah Lanjutan Tingkat Atas Pada Siswa SMP* (Malang:Universitas Brawijaya Malang)

[12]   Hosmer W and Lemeshow S 2002 *Applied Logistic Regression Second Edition* (United States of American: John Wiley & Sons Inc.)

[13]   Qudratullah M F 2011 *Misklasifikasi Mahasiswa Baru F Saintek UIN Sunan Kalijaga Jalur Tes Tertulis dengan Analisis Regresi Logistik* Jurnal Matematika Murni dan Aplikasi, 1(4)