Journal of Multidisciplinary Engineering Science Studies (JMESS)
ISSN: 2458-925X
Vol. 5 Issue 6, June - 2019

Logistic Regression Analysis and An Application Regarding Students' Success

Senol Celik
Bingöl University, Faculty of Agriculture,
Department of Animal Science, Biometry and Genetics
Bingol-Turkey
senolcelik@bingol.edu.tr

***Abstract*—"ALES (Academic Personnel and Postgraduate Education Entrance Exam), graduation grade, foreign language and science exam" data were studied to determine the key variables affecting student success in application for postgraduate study, with the aim to present a model implementation for logistic regression analysis.**
The aim of the study was to examine the logistic regression analysis being capable of defining the correlation between the clusters of independent variables and dependent variables consisting of dual results variable. The population of the study consisted 841 students who applied to Gazi University Institute of Natural Sciences in 2017-2018 academic year spring term in Turkey. Results of the logistic regression analysis showed that, "Exam" and "Graduate" were the factors having the highest effect on the students' level of success in the exam. As a result, logistic regression model was concluded to be an appropriate method in determining the students' probability of success.

Keywords—Logistic regression, success, exam, education.

I. INTRODUCTION

Logistic regression analysis is a method used to define the cause-result relationship between the categorical, double, triple and multiple categories dependent variable and independent variables [1]. The use of logistic regression analysis has some key advantages. It can be used in cases of the breach of certain hypotheses such as normality and possession of a covariance, it can be used when the dependent variable is a discrete variable with two or multiple categories and it can easily be interpreted in terms of mathematics [2, 3].
The purpose of the use of logistic regression analysis is to establish a model that defines the most congruent correlation between the dependent and independent variables by using the least amount of variables [4].
Logistic regression analysis is one of the three methods most commonly used in the group appointment of observations such as cluster analysis and discriminant analysis. Whereas the number of clusters to which observations will be appointed, is not exactly known in cluster analysis, the number of groups is known in discriminant and logistic regression analyses; a discrimination model is obtained by using the available data and the new observations added to the data cluster are appointed to groups [5]. In several fields there are numerous researches on logistic regression. Some of these are summarized below.
In agriculture; logistic regression model was used by [6] with regards to the factors affecting red meat preferences, by [7] with regards to farmer behaviors in struggle against stink bugs, by [8] with regards to the identification of the consumption behaviors regarding bee products. [9] brought the over-expansion observed in data set obtained in plant conservation area, under control. In transportation and traffic; logistic regression model was used by [10] with regards to tramway passenger satisfaction, by [11] with regards to the factors affecting local traffic accidents. [12] studied Dentistry Data to identify the risk factors affecting the use of prosthesis. Logistic regression analysis was used by [13] in mist event and aviation activities, and by [14] in genetics.
The purpose of this study is to study the factors affecting the success of students in postgraduate exam, by using the logistic regression model.

II. MATERIAL AND METHOD

A. Material

The study group consisted of 841 people who applied to Ankara Gazi University Institute of Natural Sciences for postgraduate study. Applications were made to departments of Biology, Forestry Products Engineering, Environmental Studies, Electrical-Electronic Engineering, Industrial Products Designing, Industrial Design Engineering, Energy Systems Engineering, Physics, Advanced Technologies, Production Engineering, Construction Engineering, Statistics, Environmental and Technical Analysis of Accidents, Chemistry, Chemistry Engineering, Mechanical Engineering, Mathematics, Metallurgy and Material Engineering, Architecture and Automotive Engineering. ALES (Academic Personnel and Postgraduate Education Entrance Exam), graduation grade, foreign language and science exam data were studied to assess success or failure. The level of success (successful 1, unsuccessful 0) was the dependent variable in logistic regression analysis and

was examined through ALES score, graduation grade, foreign language grade and science exam grade.

B. Method

Logistic regression model works similar to linear regression, but with a binomial response variable. Appropriate variables should be included in the model in logistic regression, and those that are not casually appropriate should not be included in the model. In general, 10 or more observations should be used for each variable in the model.

A logistic regression will model the chance of an outcome based on individual characteristics. Since chance is a ratio, what will be really modeled is the logarithm of the chance given by:

$$log\left(\frac{\hat{P}}{1-\hat{P}}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

where $\hat{P}$ indicates the probability of an event and β_i are the regression coefficients associated with the reference group and the x_i explanatory variables.

When there is one independent variable, Binary Logistic model

$$P(Y=1) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

Multiple Binary Logistic Regression model

$$P(Y=1) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k}} = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k)}}$$

Where, P: Observing probability of an analyzed event, β_0, β_1, β_2,...,β_k Regression coefficients of independent variables, X_1, X_2,...,X_k. k: The number of independent variables [15].

P represents the observing probability of an event analyzed in the logistic regression equation. The ratio of probability of an analyzed event to probability of the other events is Odds Value The ratio of the Odds values of two different analyzed events to each other is Odds Rate. Odds Rate is defined as exp(β) in the logistic regression. As Odds is the proportion of probability of an event happening to the probability of not happening, exp (βp) shows how many more folds in what percentage Y variable has observing probability with the effect of Xp variable [15].

III. RESULTS

In logistic analysis, the results involving the independent variables including ALES, graduation, foreign language and science exam scores affecting the response variable (0 and 1) where the success variable is categorical, are summarized below:

The model's "Omnibus test" among the significance tests, is equivalent to the multiple correlation coefficient test and is tested by χ2. According to Table 1, the model is significant (p<0.001).

TABLE 1. OMNIBUS TEST OF THE MODEL'S COEFFICIENTS

	Chi-square	df	Sig.
Step	590.48	4	0.001
Block	590.48	4	0.001
Model	590.48	4	0.001

df: Degrees freedom

As shown on the model's significance tests in Table 2, -2 Log Likelihood value was found as 554.271. As $\chi^2_{0.05;4} = 9.49 < 554.271$ for the model with significant variables, H_0 hypothesis was rejected and it was concluded that, at least one of the coefficients was different from zero (*α=0.05*).

TABLE 2. MODEL SUMMARY

In logistic regression analysis, the Cox and Snell R^2 value similar to the R^2 in the least squares method, was measured as 0.504. This ratio shows a 50.4% correlation between the dependent and independent variables. Nagelkerke R^2, which was developed to ensure that Cox and Snell R^2 statistics got a value between 0-1, was measured as 0.678. This value indicates a moderate level correlation between the dependent variable and independent variables. Hosmer-Lemeshow (H-L) test was performed considering the observed and expected frequency values for the goodness of fit test of the model obtained. According to Table 3 which presents these findings, H_0 hypothesis was rejected.

-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
554.271	0.504	0.678

TABLE 3. HOSMER AND LEMESHOW TEST

Chi-square	df	Sig.
80.583	8	0.001

With regards to the variables of the model and their significance levels (Table 4), parameters such as graduation grade and science exam score were observed to be significant (p<0.05 and p<0.01). Wald test was performed to test the significance of B parameters in logistic regression analysis. Positivity of the coefficients of the variables indicated an elevation in students' probability of success.

TABLE 4. VARIABLES IN THE EQUATION

	B	S. E.	Wald	df	Sig.	Exp(B)
X1: ALES	-0.024	0.019	1.507	1	0.22	0.977
X2: Graduate	0.023	0.013	2.866	1	0.09	1.023
X3: Language	-0.011	0.006	2.905	1	0.088	0.989
X4: Exam	0.082	0.005	247.173	1	0.000	1.085
Constant	-3.37	1.683	4.012	1	0.045	0.034

S. E.: standard error

The equation obtained for the model as the result of the analysis is as follows. Probability values for student success depending on the values obtained by the independent variables, can be calculated with the below logistic regression equation.

$$P(Y=1)=\frac{1}{1+e^{-(-3.37-0.024X1+0.023X2-0.011X3+0.082X4)}}$$

The coefficients in logistic regression analysis are interpreted through the odds ratio (superiority ratio-exp (B)) by taking the reciprocal of the natural logarithm of the coefficients. Further, natural logarithm of the superiority ratio is a linear function of the independent variable. The highest odds ratio (superiority ratio) in analysis belongs to "X4-science exam" variable. It was found to be elevating the probability for placement to a postgraduate program about 1.085 ($e^{0.082}$) times. Those with higher graduation scores had 1.023 times higher probability of success compared to those with lower graduation grades. In the model, "X1-ALES" variable had a lower odds ratio compared to others. Those with higher ALES scores had 0.977 times higher probability of success compared to those with lower scores. This means, as it is 1/0.977=1.023 times lower, it has no significant effect. Since the parameter estimations of X1-ALES and X3-Language variables are insignificant, their effect on success is also insignificant.

The right classification chart of the model is presented in Table 5. According to the classification chart, 296 of the "successful" students (83.6%) and 420 (86.2%) of the unsuccessful students are classified correctly. The model's likelihood of giving a correct estimation of students' level of postgraduate study success is, 85.1%.

TABLE 5. CLASSIFICATION TABLE

Observed		Predicted		
		Success		Percentage Correct
		0	1	
Success	0	420	67	86.2
	1	58	296	83.6
Overall Percentage				85.1

Exam success probabilities can be found by assigning different values to the independent variables with the logistic regression equation obtained in Table 4.

For example, for a student with ALES score (X1) of 73, graduation grade (X2) of 68, foreign language grade (X3) of 64 and science exam score (X4) of 80, probability of being successful in the postgraduate exam can be calculated as follows:

$$\hat{Y}=-3.37-0.024X1+0.023X2-0.011X3+0.082X4$$

$$\hat{Y}=-3.37-0.024(73)+0.023(68)-0.011(64)+0.082(80)=2.298$$

$$P(\hat{Y})=\frac{1}{1+e^{-(2.298)}}=\frac{1}{1+0.10046}=\frac{1}{1.10046}=0.9087$$

As the value calculated is P>0.50, the student has a high probability of passing the exam, which is 90.87%. Similarly, the probability of passing the exam is calculated as follows for a student with X1=55, X2=65, X3=38, X4=45.

$$\hat{Y}=-3.37-0.024X1+0.023X2-0.011X3+0.082X4$$

$$\hat{Y}=-3.37-0.024(55)+0.023(65)-0.011(38)+0.082(45)=0.077$$

$$P(\hat{Y})=\frac{1}{1+e^{-(0.077)}}=\frac{1}{1+0.92589}=\frac{1}{1.92589}=0.51924$$

The value obtained is P>0.50, with a probability of success of 0.51924, which corresponds to 51.92%. Estimated $P(\hat{Y})$ values for different values of X1, X2, X3 and X4, i.e., their probabilities of success are presented in Table 6.

TABLE 6. PROBABILITIES OF SUCCESS IN EXAMS

X1	X2	X3	X4	$P(\hat{Y})$
85	93	70	100	0.985
90	78	52	60	0.648
67	72	40	35	0.291
77	70	50	45	0.385
70	72	65	55	0.599

IV. DISCUSSION

One of the factors of the logistic regression model affecting university students' academic success has been class attendance. In respective study, the odds ratio related to the probability of success when the student does not attend the class, i.e., the probability of success in case of non-attendance, has been found as -0.84746. A 1-unit increase in non-attendance will reduce the student's probability of success 0.84746 times [16]. [17] used the logistic regression analysis to study the factors affecting university students' academic success. The rate of classification success obtained at the end of analysis was 66.10%.

In another logistic regression study on education, it was concluded that, the variables affecting students' attitudes towards Physics class were; "integrating the knowledge and skills learned in Physics class into daily life", "duration of weekly study" and "interest in the course" [18].

V. CONCLUSION

There are several statistical methods that are used to determine the students' level of success. Logistic regression method was used in this study. Logistic regression analysis was performed to classify the success grade variables of the students who had applied to Gazi University Institute of Natural

Sciences. The right classification probabilities of the data were found to be 85.1%. In conclusion, the science exam factor was found to be statistically significant in affecting the level of success.

REFERENCES

[1] K. Özdamar, "Paket Programlarla İstatistiksel Veri Analizi 1", Nisan Kitabevi, Eskişehir, Turkey, 2013.

[2] H. Tatlıdil, "Uygulamalı Çok Değişkenli İstatistiksel Analiz", Ziraat Matbaacılık, Ankara, Turkey, 2002.

[3] S. Lemeshow, D. Hosmer, "Applied Logistic Regression (Wiley Series in Probability and Statistics". Wiley-Interscience; 2 Sub edition, 2000.

[4] S. Coşkun, M. Kartal, A. Coşkun, H. Bircan, Lojistik Regresyon Analizinin İncelenmesi ve Diş Hekimliğinde bir Uygulaması. Cumhuriyet Üniversitesi Diş Hekimliği Fakültesi Dergisi, 2004, 7(1): 41-50.

[5] A. H. Elhan, Lojistik Regresyon Analizinin İncelenmesi ve Tıpta Bir Uygulaması, Ankara, 1997.

[6] F. Lorcu, V. A. Bolat, Edirne Merkez İlçede Tüketicilerin İthal Kırmızı Et Satın Alma Tercihlerini Etkileyen Faktörler. Academic Food Journal, 2011, 9(6): 38-45.

[7] M. Duman, C. Gözüaçık, V. Karaca, Ç. Mutlu, Farmer Behaviors in Sunnpest Struggle: A Case of Adıyaman, Diyarbakır, Mardin and Şanlıurfa. J. Agric. Fak. HR. U, 2008, 12(4): 65-71.

[8] R. I. Tunca, A. Taskin, U. Karadavut, "Determination of Bee Products Consumption Habits and Awareness Level in Some Provinces in Turkey. Türk Tarım – Gıda Bilim ve Teknoloji Dergisi, 2015, 3(7): 556-561.

[9] A. Yeşilova, I. Kasap, Investigation of Overdispersion in Logistic Regression. Yüzüncü Yıl Üniversitesi, Tarım Bilimleri Dergisi, 2008, 18(1): 21-25.

[10] N. Girginer, B. Cankuş, Measuring the Traveller Satisfaction of Tram Using Logistic Regression: Case Study of Estram. Yönetim ve Ekonomi, 2008, 15(1): 181-193.

[11] S. Bektaş, M. A. Hınıs, Investigation of Parameters Effecting Accidents in Urban Roads by Logistic Regression Modelling: Aksaray Case. J. Fac. Eng. Arch. Selcuk Univ., 2008, 23(3): 25-34.

[12] S. Coşkun, M. Kartal, A. Coşkun, H. Bircan, Lojistik Regresyon Analizinin İncelenmesi ve Diş Hekimliğinde bir Uygulaması. Cumhuriyet Üniversitesi Diş Hekimliği Fakültesi Dergisi, 2004, 7(1): 41-50.

[13] C. Aktaş, O. Erkuş, "Investigation of Fog Forecasting of Eskisehir Using Logistic Regression Analysis. İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 2009, 8(16): 47-59.

[14] M. Özkan, H. Çamdeviren, "Logistic Regression Analysis and an Application to Genetic Studies", Tarım Bilimleri Dergisi, 2000, 6(3): 42-48.

[15] D. N. Gujurati, "Basic Econometrics", McGraw-Hill, Inc., New York, 2003, p. 1002.

[16] Ş. Can, T. Özdil, C. Yılmaz, "Estimation of the Factors Affecting the Success of the University Students by Logistic Regression Analysis. International Review of Economics and Management, 2018, 6(1): 28-49.

[17] G. Çırak, Ö. Çokluk, "The Usage of Artificial Neural Network and Logistic Regression Methods in the Classification of Student Achievement in Higher Education. Mediterranean Journal of Humanities, 2013, 3(2): 71-79.

[18] M. Şata, M. Çakan, Comparison of Results of CHAID Analysis and Logistic Regression Analysis. Dicle University Journal of Ziya Gökalp Faculty of Education, 2018, 33: 48-56.