Current Journal of Applied Science and Technology

40(20): 1-6, 2021; Article no.CJAST.71482
ISSN: 2457-1024
(Past name: British Journal of Applied Science & Technology, Past ISSN: 2231-0843, NLM ID: 101664541)

Binary Logistic Regression Analysis on Predicting Academics Performance

E. K. Akinyemi[1*], O. A. Ogunleye[1], H.O Olaoye[1] and J. Brakoru[1]

[1]Department of Statistics, Federal School of Statistics, Ibadan, Oyo State, Nigeria.

Authors' contributions

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/CJAST/2021/v40i2031458
Editor(s):
(1) Dr. Vitaly Kober, CICESE, Mexico.
Reviewers:
(1) Kayondo Wasswa Hassan, Makerere University, Uganda.
(2) N. Varathan, University of Jaffna, Sri Lanka.
Complete Peer review History: https://www.sdiarticle4.com/review-history/71482

Original Research Article

Received 22 May 2021
Accepted 28 July 2021
Published 21 August 2021

ABSTRACT

This paper considers the application of logistic regression model to predict academics performance of students. The choice of this model becomes imperative as a result of dichotomous relationship existing in the model (either pass or fail). 100 students from the four department where engaged in the study. Statistical package for social scientist (SPSS) was used for the analysis. The results show that monthly allowance of students, and study time of the students were significant predictors. While gender and educational level of parent were insignificant predictors. The fitness of the model was assessed using Hosmer and Lemeshow test, split-sample approach and other supplementary indices to validate the model. The fitted model indicated that fitted binary logistic regression model could be used to predict the future performance of students.

Keywords: Logistic regression; hosmer–lemeshow test; likelihood ratio test; maximum likelihood estimation; wald test.

1. INTRODUCTION

Many research problem calls for the analysis and prediction of a dichotomous outcome whether a student will succeed in college, whether a child should be classified as learning disable, whether a teenager is prone to engage in risky behavior, and so on. Traditionally, these research

**Corresponding author: E-mail: emmak106@gmail.com;*

questions were addressed by either ordinary least squares (OLS) regression or linear discriminate function analysis. Both techniques were subsequently found to be less than ideal for handling dichotomous outcomes due to their assumption that is linearity, normality, and continuity for OLS regression and multivariate normality with equal variances and covariance for discriminate analysis. Generally,Binary Logistic Regression Analysis (BLRA)study the relationship between multiple explanatory variables and a single binary response variable, a categorical variable with two categories, [1].

2. EMPIRICAL REVIEW

Many applications of binary logistics have been. (Lihui *et al*., 2001) compared both linear regression and logistic regression model for biological percentage data using different methods for comparison. [2] have used logistic regression method for model selection. The study aimed to increase the power of prediction while reducing the number of covariates. The procedure was depending on the stepwise method and best subsets regression through applying on academic data.

Javali and Pandit [3] used a model depending on multiple logistic regression to make risk factors prognostication of oral health infirmities. [4] used the logistic regression method and applied for building models of the risk factor in the stammer studies. [5] suggested a formula based model to compare the accuracy of applying formulas to separate outcomes of support vector apparatus, judgment trees, and LRA on the database of Cleveland Heart Disease to obtain a reliable model of heart disease prediction [6]. studied the relation of hypertension with risk factors affecting significantly the execution of Hypertension using logistic regression technique.

Qais [7] used LRA and discriminant analysis DA and applied on natural and Caesarean births data to show the performance of such techniques and the capability in classification the type of birth [8]. examined the importance of keeping the possession of the ordinal nature of the outcome variable while marking the risk factors related to diabetic problems related to loss of vision using traditional and Bayesian approaches of ordinal logistic regression models. [9] used multinomial logistic regression analysis MLRA to examine the effect of an estrogen on the rate of reverse pregnancy results. The purpose of this paper is to find a best BLRA model for fitting line and for obtaining the best classification and predicting the group membership. The remainder of this paper is structured as follows: section 2, explains BLRA and methodology. Section 3 presents the application on real data and finally, in section4 conclusions are presented.

2.1 Binary Logistic Regression Analysis (BLRA)

Regression analysis presents the association between a response variable and one or more explanatory variables. It is often the situation that the outcome variable is discrete, assuming two or more potential values [10]. BLRA represents a special condition of linear regression analysis LRA used when the response is binary not continuous and the explanatory variables are quantitative or qualitative variables [11]. It was first suggested in the 1970s to overcome difficulties of ordinary least squares OLS regression in treating binary outcomes [12]. Logistic regression LR uses the theory of binomial probability which represents having only two values to predict: that probability (p) is 1 instead of 0, i.e. the event belongs to one group instead of the other. LR presents the best fitting function depending on the maximum likelihood ML approach, which maximizes the distinguishing probability of the observed data into the suitable category given the coefficients of regression [13].

2.2 Assumptions of Binary Logistic Regression Analysis (BLRA)

Logistic regression ignores a linear relationship between the response and explanatory variables. It supposed that the response variable must be a binary and the explanatory variables,need not be an interval, the distribution is normal, the relationship is linear, nor of equality of variance within each group. Furthermore, the groups must be mutually exclusive and detailed; a case can only be in one group and every case must be a member of one of the groups. Finally, the sample size must be large than that of LR because ML coefficients are large sample estimates. A minimum of 50 cases for each explanatory variable is needed [3,13], (Kleinbaum & Klein, 2010).

2.3 The Logistic Model

To explore the implied association between a response variable and one or more explanatory variables, the LRA is suitable for study. By taking the case of one explanatory variable X with one binary outcome variable Y, the logistic model

predicts the Logit of Y from X which represents a natural logarithm of odds of Y. The simple formula can be written as the following (Peng et al., 2002), (James et al., 2013):

$$In\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x \ldots\ldots\ldots\ldots\ldots\ldots \text{(i)}$$

The left-hand side is called the log-odds or logit. The LR model has a logit that is linear in X. Hence:

$$\pi(x) = E(Y/X) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} \quad \text{(ii)}$$

Where π is the probability of the outcome of interest given that X=x, α is a parameter which represents the Y-intercept, and β is a parameter of the slope, X can be qualitative (categorical) or quantitative variable, and Y is always qualitative or categorical. The formula (1) can be expressed and extended from simple to multiple linear regressions as follows:

$$In\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x \quad \text{(iii)}$$

3. RESEARCH METHODOLOGY AND MATERIALS

This work is an empirical survey study conducted on students of federal school of statistics, Ibadan Oyo, Nigeria. The research used a non-random sampling technique, quota sampling method to select respondents for the study, due to lack of control over the students initially selected based on stratified random sampling method. There were a total of four (4) departments from which hundred (100) questionnaires were administered to students willing to participate in the study after brief enlightenment address on the objective of the study. A 15 item, structured questionnaires, administered on one hundred (100) students across the four departments. An IBM statistical package for social science (SPSS, version 23.0) was used for the statistical data analysis conducted (Binary Logistics Regression). An association test between academic performance and certain factors were conducted and factors found significant in association with formulated regression model. A binary logistic regression model of the academic performance of the students is formulated such as:

$$Li = In\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x1 + \beta x2 + \beta x3 + \beta x4 + \cup i \quad \text{(iv)}$$

Where, Li =1 if the mark is from40-70
0 if the mark is from0 to 39

The outcomes are coded in binary response: Y = 1 if the student academics performance falls within pass mark and Y = 0 if the student academics performance falls within fail mark as shown below:

$$Y = \begin{cases} 1, \text{Good academics performance} \\ \\ 0, \text{Bad academics performance} \end{cases}$$

Pi = the probability of the student having good academic performance.
X1 rep. Gender
X2 rep. Weekly Allowance
X3 rep. Study Time
X4 rep. Mother Educational level
X4 rep. Father Educational level

3.1 Data Analysis

Stepwise logistic regression analysis was used to reduce number of covariates results as presented in Table 1 below.

Table 1. Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Gender	-.305	.660	.214	1	.644	.737	.202	2.685
	allowance	1.278	.439	8.473	1	.004	3.589	1.518	8.486
	time	2.686	.609	19.451	1	.000	14.671	4.447	48.398
	education	-.484	.380	1.629	1	.202	.616	.293	1.296
	Constant	-4.130	1.635	6.380	1	.012	.016		

Variable(s) entered on step 1: gender, allowance, time, education.

4. RESULTS

It is noted that the covariates (monthly allowance of the students and study time of the students) of the students are statistically significant; while the covariates (gender and parentals educational level) are statistically non-significant.

The Wald test is obtained by comparing the maximum likelihood estimate of the beta's β, to its standard error. The resulting ratio, under the hypothesis that βi = 0 are given in Table 1. But the covariate type of allowance and time is statistically non-significant and statistically significant using Wald test and likelihood ratio test respectively. However, Hauk and Donner (1977) and Jennings (1986) examined the performance of the Wald test and found that the test often failed to reject the null hypothesis when the coefficient was significant. They recommended that the likelihood ratio test to be used. Therefore, it is evident that the covariate (time and allowance) is statistically significant. And the logit is:

$\boldsymbol{Y}$ = $\boldsymbol{logi}$t (**performance**) = $-4.130 - 0.305x1 + 1.278x2 + 2.686x3 - 0.484x4 + \cup i$

Here, the relationship between logit (performance) and X_1, X_2, X_3&X_4 is linear.

Hence,

The (Y) above indicates that: students academics performance are less possibility to be pass their course; students based on gender are less possibility to pass their course; a students with allowance have high likelihood than a student with high parental academic qualification. Firstly, time spent to study increases the chance of having a good academic performance, in other words, the higher the time spent to study, the higher the chances that a student will have a good academics performance. Secondly, monthly allowance of the students increases the chance of having a good academic performance, in other words, the higher the monthly allowance, the higher the chances that a student will have a good academics performance.

Moreso, the "B" values are the logistics coefficients that can be used to create a predictive equation. In this research:

$(Y|X1 = x1, X2 = x2, X3 = x3, X4 = x4)$

$$= \frac{e^{\alpha+\beta x1+\beta x2+\beta x3+\beta x4}}{1+e^{\alpha+\beta x1+\beta x2+\beta x3+\beta x4}}$$
$$= \frac{1}{1+e^{-(\alpha+\beta x1+\beta x2+\beta x3+\beta x4)}}$$

Here, the relationship between the outcome and the predictors is non-linear.
(Performance) =

$$\frac{e^{-4.130-0.305x1+1.278x2+2.686x3-0.484x4}}{1+e^{(-4.130-0.305x1+1.278x2+2.686x3-0.484x)}}$$
$$= \frac{1}{1+e^{-(-4.130-0.305x1+1.278x2+2.686x3-0.484x)}}$$

The exponent (Exp (B)) in Table 1 above is the odds ratio, thus:

- The odds ratio for gender to pass academically is 0.757
- The odds ratio for student's monthly allowance is 3.589
- The odds ratio for students study time is 14.671
- The odds ratio for student's mother academics qualification is 0.616

Table 2 shows the classification table. Using the obtained Y function observations which are classified as follows, using a prior probability of 0.50.

(a) 66.7% of all students offering statistics that fail in their course were correctly classified and 33.3% were incorrectly classified.
(b) 92.0% from all students offering statistics that pass in their course were correctly classified, 8.0% were incorrectly classified.
(c) The overall correct percentage was 85.9%, which reflects the model's overall explanatory strength.

From Table 3, Cox & Snell R-Square indicates that 38.3% of the variation in the independent variable is explained by the logistic model. Nagelkerke R Square indicates a moderately strong relationship of 57.2% between the predictors and the prediction.

From Table 4, The value of the Hosmer Lemeshow goodness-of-fit statistic for the full model was Chi-square = 13.268 and the corresponding p-value from the chi-square distribution with 8 degree of freedom is 0.103 which means that it is not statistically significant and therefore our model is quite good.

Table 2. Classification Table[a]

Observed			Predicted		
			Academics Performance		Percentage Correct
			Fail	Pass	
Step 1	Academics Performance	Fail	16	8	66.7
		Pass	6	69	92.0
	Overall Percentage				85.9

a. The cut value is .500

Table 3. Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	61.835[a]	.383	.572

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.p-value

Table 4. Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	13.268	8	.103

5. DISCUSSION OF FINDINGS

Significance testing for the logistic coefficients using likelihood ratio and stepwise regression method show that, at 0.05 level of significance, monthly allowance of students, and study time of the students were significant predictors. While gender and educational level of parent were insignificant predictors. The odds ratio for gender of the students ranges between 0.202 times to 2.685 times with confidence of 95%. The odds ratio for monthly allowance of the students ranges between 1.518 times to 8.486 times with confidence 95%. The odds ratio for study time of the students ranges between 4.447 times to 48.398 times with confidence of 95%. The odds (likelihood) associated with mother education level between 2.93 times to 1.296 times with confidence 95%. To assess the fitness of the model the maximum likelihood test and Hosmer Lemeshow goodness-of-fit test suggest that the fitted logistic regression model has significant predictive ability for future subjects. Prediction error rate for validation of the model is not so high. The asymptotic significance of the model is less than 0.005 which indicates that the predictive ability of the fitted model is good. Thus, different summary measures of goodness-of-fit and others supplementary indices of predictive ability of the fitted model indicate that the fitted binary logistic regression model can be used to predict the performance of students offering statistics.

6. CONCLUSION

Based on the study done, we conclude based on the model used that students have less possibility to perform in their course of study. In other words, the higher the monthly income of the students, the high likely performance of the students and the more time they spend studying the high likely performance of the students. However, parental educational level and gender cannot predict academics performance of students.

7. RECOMMENDATION

Based on the findings of this paper we recommend that the same study to all other higher institutions with increase sample size be carried out. Develop a logistic regression model that contains repeated measures. Apply Classification and Regression Tree (CART) and compare the result with binary logistic regression model.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Sweet S, Martin K. Data analysis with SPSS: A first course in applied statistics. Fourth Edition.: Pearson Publisher; 2011.

2. Sarkar SK, Midi H, Rana S. Model selection in logistic regression and performance of its predictive ability. Australian Journal of Basic and Applied Science. 2010;12:5813-5822.
3. Javali S, Pandit P. Multiple logistic regression model to predict risk factors of oral health diseases. Romanian Statistical Review Journal. 2012;73-86.
4. Reeda P, Wub Y. Logistic regression for risk factor modelling in stuttering research. Journal of Fluency Disorders. 2013;38: 88- 101.
5. Mythili T, Dev Mukherji M, Padalia N, Naidu A. A heart disease prediction model using SVM-decision trees-logistic regression (SDL). International Journal of Computer Applications. 2013;68:11-14.
6. Amir W, Mamat M, Ali Z. Association of hypertension with risk factors using logistic regression. Applied Mathematical Sciences. 2014;8:2563–2572.
7. Qais M. Comparison of discriminant analysis and logistic regression analysis: An application on caesarean births and natural births data. Journal of The Institute of Natural and Applied Sciences. 2015; 20:34-46.
8. Vaitheeswaran K, Subbiah M, Ramakrishnan R, Kannan T. A comparison of ordinal logistic regression models using Classical and Bayesian approaches in an analysis of factors associated with diabetic retinopathy. Journal of Applied Statistics. 2016;43:2254-2260.
9. Junguk H, Eun-Hee C, Kwang-Hyun B, Kyung JL. Prediction of gestational diabetes mellitus by unconjugated estriol levels in maternal serum. International Journal of Medical Sciences. 2017;14:123-127.
10. Hosmer D, Lemeshow S. Applied logistic regression. Second Edition: John Wiley and Sons, Inc; 2000.
11. Hair J, Black W, Babin B, Anderson R. Multivariate data analysis, Seventh Edition.: Pearson Prentice Hall; 2010.
12. Peng C, Lee K, Ingersoll G. An Introduction to logistic regression analysis and reporting. The Journal of Educational Research. 2002;96: 3-15.
13. Burns RB, Burns RA. Business research methods and statistics using SPSS: Sage Publishing LTD; 2008.

Peer-review history:
The peer review history for this paper can be accessed here:
https://www.sdiarticle4.com/review-history/71482