Journal of Modern Mathematics and Statistics 4 (2): 58-62, 2010
ISSN: 1994-5388

Application of Logistic Regression Model to Graduating CGPA of University Graduates

[1]E. Ahani, [2]O. Abass and [1]O. Okafor Ray
[1]Department of Mathematics, [2]Department of Computer Science, University of Lagos, Nigeria

Abstract: Logistic regression model deals with the relationship that exists between a dependent variable and one or more independent variables. It provides a method for modeling a binary response variable which takes values 1 and 0. In this stuty, a brief review of the underlying theory for the approach is presented and the Logistic regression model to estimate the graduating Cumulative Grade Point Average (CGPA) of graduates were fitted and tested. Data were collected from Faculty of Science, University of Lagos. The study reveals that the final year Grade Point Average (GPA) of the graduands has significant effect among all other variables.

Key words: Logistic regression, binary data, categorical variables, diagnostic test, grade point average, University of Lagos

INTRODUCTION

Logistic regression model, introduced in late 1960s and early 1970s has in the early 1980s become routinely available in statistical packages. It has also found many applications in fields like the social sciences (Chuang, 1997) and in educational research, especially in higher education (Austin *et al*., 1992). Logistic regression analysis extends the techniques of multiple regression analysis to research situations in which the outcome variable is categorical. There is a binary response of interest and the predictor variables are used to model the probability of that response.

Situations involving categorical outcomes are quite common in practice. In educational program, predictions are made for the binary outcome of success/failure. In the same vein, operation units could be classified as successful or not successful according to some objective criteria in industries.

The several characteristics of the units could be measured and logistic regression analysis could be used to determine which characteristics best predict success. Similarly, in a medical arena, an outcome might be due to the presence or absence of a particular disease. This research gives a brief review of the underlying theory of logistic regression with its application to Graduating Cumulative Grade Point Average (CGPA) of the 2007/2008 graduates of Faculty of Science, University of Lagos. It derives motivation from the study done by Peng *et al.* (2002). Further support is derived from Karp (2007) who argued that logistic regression is an increasingly popular analytic tool, used to predict the probability that the event of interest will occur as a linear function of one (or more) continuous and/or dichotomous independent variables. Logistic regression model have been applied in a number of contexts. Some examples include applications to adjust for bias in comparing two groups in observational studies (Rosenbaum and Rubin, 1983). Efron (1975) compared logistic regression to discriminant analysis (which assumes the explanatory variables are multivariate normal at each level of the response variable); it has also been applied to a study investigating the risk factors for low birth weight babies (Hosmer and Lemeshow, 1989). Other applications include using logistic regression analysis to determine the factors that affect green card usage for health services (Senol and Ulutagay, 2006). Applications of logistic regression have also been extended to cases where the dependent variable is >2 cases, known as multinomial or polytomous. Tabachnick and Fidell (1996) use the term polychotomous.

University of Lagos, Nigeria: The University of Lagos was established in 1962. It is made up of two campuses: the main campus located in Akoka, Yaba and the college of medicine in Idi-Araba, Surulere, Nigeria. The institution started with 131 student but today, it can boast of annual intake of 39,000 students.

In addition, it has a total staff strength of 3,365 administrative and technical staff (1,386), junior staff (1,164) academic staff (813). The university has nine faculties and a college of medicine. The faculties include: Arts, Business Administration, Education, Engineering, Environmental Sciences, Law, Pharmacy, Sciences and Social Sciences. These faculties offer a total of 117

Corresponding Author: E. Ahani, Department of Mathematics, University of Lagos, Nigeria

programmes. The university also offers Master's and Doctorate degree in most of its programmes. The distance learning institute of the university offers courses in Accounting, Business Administration, Science Education and Library/Information Sciences.

The vision of the university is to be a top-class institution for the pursuit of excellence in knowledge through learning and research as well as in character and service to humanity while the mission is to provide a conducive teaching, learning, research a nd development environment where staff and students can interact and compete effectively with their counterparts both nationally and internationally in terms of intellectual competence and zeal to add value to the world.

In the spirit of this vision and mission the university recently rewarded 19 of its researchers for their outstanding excellence in the 2005 Research Conference and Fair.

Logistic regression model and general theory: Logistic regression analysis is part of a category of statistical models known as generalized linear models which consist of fitting a logistic regression model to an observed proportion in order to measure the relationship between the response variable and set of explanatory variables (Lavange *et al*., 1986).

Letting X denote the vector of predictors $\{x_1, x_2, ... x_k\}$ and let the conditional probability that the outcome is present be denoted by the equation given as:

$$P(Y = 1 \mid x) = \pi(x) \quad (1)$$

The logistic regression model (Harrell, 2001) is given by:

$$\pi(x) = \frac{1}{1+e^{x\beta}} \quad (2)$$

$\pi(x)$ = The success probability at value x
$x\beta$ = Stands for $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$
e = The base of the system of natural logarithms

It can be transformed to give a new interpretation. Specifically, we define the odds as the following ratio:

$$odd = \frac{\pi}{1-\pi} \quad (3)$$

The logistic regression model has a linear form for the logit of this probability:

$$\begin{aligned} logit[\pi(x)] &= \log\left(\frac{\pi(x)}{1-\pi(x)}\right) \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \end{aligned} \quad (4)$$

Thus:

$$\log (odds) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

Equation 4 is in the same form as the multiple linear regression equation. The inverse transformation of Eq. 4 is the logistic function of the form:

$$\begin{aligned} \pi &= P(Y = \text{outcome of interest} \mid X \\ &= x, \text{a specific value of } X) = \frac{e^{x\beta}}{1+e^{x\beta}} \end{aligned} \quad (5)$$

With Eq. 5, one predicts the probability of the occurrence of the outcome of interest. According to Eq. 4, the relationship between log it (Y) and X is linear. Yet, according to Eq. 5, the relationship between the probability of Y and X is non-linear. Thus, the natural log transformation of the odds in Eq. 4 is necessary to make the relationship between a categorical outcome variable and its predictor(s) linear.

The value of the coefficient β determines the direction of the relationship between C and the logit of Y. When β is >0, larger (or smaller) X values are associated with larger (or smaller) logits of Y. Conversely, if β is <0, larger (or smaller) X values are associated with smaller (or larger) logits of Y.

Fitting the logistics regression model to data: The unknown parameters β_j in the logistic regression model are estimated by the method of maximum likelihood. Solving for logistic regression coefficients β_j and their standard errors involves calculus, in which values are found using maximum likelihood methods.

These values, in turn are used to evaluate the fit of one or more models. The statistical significance for individual logistic regression coefficients is evaluated using the Wald test:

$$z = \frac{\hat{\beta}}{SE} \quad (6)$$

Wald test statistics has a standard normal distribution when $\beta = 0$. For the logistic regression model, the hypothesis H:$\beta = 0$ states that the probability of success is independent of X.

The usefulness of the model (Dayton, 1992) as a whole can be assessed by testing the hypothesis that simultaneously all of the partial logistic regression coefficients are 0 that is H: $\beta = 0$.

Goodness of fit: Goodness of fit shows how effectively the model we have described the outcome variable. Selection is made to the available list of independent variable that it deems important in described the dependent variable. A log-likelihood is calculated for a candidate model-based on summing the probabilities associated with the predicted and actual outcomes for each case i:

$$L(\beta) = \ln[l(\beta)] = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1-y_i)\ln[1-\pi(x_i)] \quad (7)$$

The comparison of observed to predicted values using the likelihood function is based on the statistic, D known as deviance. The resulting deviance is:

$$D = -2\sum_{i=1}^{n} \left[y_i \ln\left(\frac{\hat{\pi}(x_i)}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\hat{\pi}(x_i)}{1-y_i}\right)\right] \quad (8)$$

The value of D is compared with and without the independent variable in the equation as given below which aids in the assessment of the significance of an independent variable:

$$G = D \text{ (model without the variable)} - D \text{ (model with the variable)} \quad (9)$$

This goodness of fit process evaluates predictors that are eliminated from the full model. In general, as predictors are added/deleted, log-likelihood decreases/ increases. The logistic regression in SPSS uses three R^2 like measures: Cox and Snell's, Nagelkerke's and McFadden's and then the Hosmer and Lemeshow Chi-square test of goodness of fit.

The Cox and Snell measure is based on log-likelihood. Equation 10 provides the method of calculation for Cox and Snell's R^2:

$$R^2 = 1 - e^{-\frac{2}{n}[D(\text{model without the variable}) - D(\text{model with the variable})]} \quad (10)$$

However, Cox and Snell's R^2 cannot achieve a maximum value of 1. The Nagelkerke's R^2 which stands as a modification of the Cox and Snell, assures that a value of 1 is achieved. In order to achieve a measure that ranges from 0-1, Nagelkerke's R^2 divides Cox and Snell's R^2 by its maximum. Equation 11 provides the measure for Nagelkerke R^2:

$$R_N^2 = \frac{R_{CS}^2}{R_{NAX}^2} \quad (11)$$

Where:

$$R_{MAX}^2 = 1 - e^{\left[\frac{2}{n} D(\text{model with the variable})\right]} \quad (12)$$

The McFadden's R^2 is a less common pseudo-R^2 variant, based on log-likelihood kernels for the full versus the intercept-only models (McFadden, 1974).

Hosmer and Lemeshow Chi-square test of goodness of fit evaluates the goodness of fit by creating 10 ordered groups of subjects. Then it compares the number actually in each group (observed) to the number predicted by the logistic regression model. A good model fit is indicated by a non-significant Chi-square value.

Application of logistic regression model: Logistic regression analysis was applied to the CGPA of the 2007/2008 graduates of Faculty of Science, University of Lagos (Table 1). The data set was obtained from the result record office of the Faculty which includes other pieces of information (e.g., age, gender and UME score) concerning the students that entered the Faculty of Science in 2003/2004 session and scheduled to graduate in 2007/2008 academic year. All analysis was performed initially using Microsoft Excel and this was loaded into SPSS.

The characteristics of the data set are as follows: the dependent binary variable Y represented with 1, stands for graduation of the students with CGPA >2.4 and 0, stands for graduation of the students with CGPA <2.4 is as follows:

$$Y = \begin{cases} 1, \text{ Graduating student with CGPA} > 2.4 \\ 0, \text{ Graduating student with CGPA} < 2.4 \end{cases}$$

The explanatory variables, used to predict whether or not an individual student graduated with CGPA above

Table 1: Logistic regression analysis of faculty of science students' CGPA

							95.0% C.I. for Exp(B)	
Parametters	B	S.E	Wald	df	Sig.	Exp(B)	Lower	Upper
GPA in final year	1.004	0.151	44.514	1	0.000	2.730	2.033	3.667
Umescore	0.006	0.004	1.920	1	0.166	1.006	0.998	1.014
Gender	0.289	0.283	1.043	1	0.307	1.336	0.767	2.327
Age	-0.028	0.021	1.835	1	0.176	0.972	0.933	1.013
Contant	-2.786	1.223	5.190	1	0.023	0.062	-	-

Table 2: Sample data for gender and graduation of the students with CGPA below or above 2.4

	Gender		
Data	Male	Female	Total
CGPA			
Graduating student with CGPA <2.4	49	34	83
Graduating student with CGPA >2.4	148	113	261
Total	197	147	344

or below 2.4 were the students' final year GPA, UME score, age, gender (0 = female student, 1 = male student). About 75.87% of the students (261 students) had CGPA >2.4 while 24.13% (83 students) <2.4 as shown in Table 2.

The gender predictor was coded as 0 = male and 1 = female with 57.27% (n =197) males and 42.73% (n = 147) females. Assessing a male's odd of being graduated with CGPA <2.4 relative to female's odds. The result is an odd ratio of 1.10 which suggests that males being graduated with CGPA <2.4 are 1:10 times that of female. The odd ratio is derived from two odds:

$$\left(\frac{149}{148}\text{for males and }\frac{34}{113}\text{for females}\right)$$

Its natural logarithm (that is $\log_e(1.10)$ is a logit equal to 0.04. Using GPA as the predictor, the logistic equation for log-odds in CGPA >2.4 is (the SPSS logistic regression is provided in Table 2).

$$\log_e\left[\frac{\pi}{1-\pi}\right] = -1.753 + 0.966\ \text{GPA}$$

The equation is exponentiated to estimate odds:

$$\frac{\pi}{1-\pi} = e^{-1.753+0.966\text{GPA}}$$

The probability is obtained by:

$$\pi = \frac{e^{-1.753+0.966\ \text{GPA}}}{1-e^{-1.753+0.966\ \text{GPA}}} = \frac{1}{1+e^{-1.753+0.966\ \text{GPA}}}$$

In the data, GPA in final year ranged from 1.49-4.95. Thus, for the lowest GPA recorded, 1.49, the log-odds, odds and the probability of CGPA >2.4 are -0.31366, 0.731 and 0.422, respectively. At the other extreme for GPA of 4.95, the log-odds of CGPA >2.4 are 3.0287, the odds are 20.6703 and the probability is 0.954. The relation between GPA and CGPA is showed in Fig. 1. Figure 1 was constructed by systematically varying CGPA from 1.00-5.00 (shown on the abscissa) and calculating the

Table 3: Log-odds and probabilities for various combinations of the predictor

Data	B	S.E	Wald	df	Sig.	Esp (B)
GPA in final year	0.966	0.146	43.606	1	0.000	2.627
Constant	-1.753	0.437	16.109	1	0.000	0.173

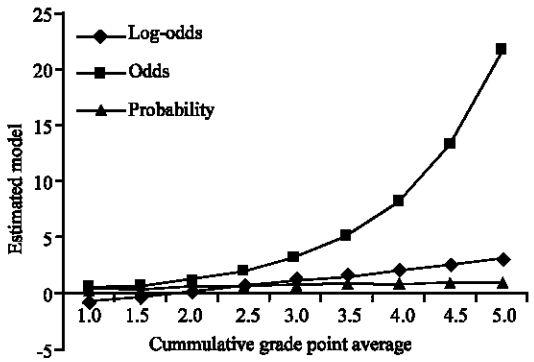

Fig. 1: Relation between GPA and CGPA

estimated probability of GPA (shown on the ordinate). It is evident from the graph that students with a CGPA over 3.5 have an estimated odd of 0.5 while for π is estimated to be equal 0.3 (Table 3). The estimated logistic regression coefficient for GPA in final year is 0.966 and the exponential of this value is $e^{0.966} = 2.63$. This indicates that for an increase in GPA in final year, the odds in favor of CGPA above 2.4 are estimated to be increased by a multiplicative factor of 2.63.

The reported standard error for b is 0.146 and statistical significance is assessed by the Wald Chi-squared statistic $(0.966/0.146)^2 = 43.78$ that with 1 degree of freedom, it is significant at conventional levels (the empirical two-tailed p-value is 0.0000 in Table 2). Thus, this study supports the conclusion that the GPA in final year is a useful predictor of student performance upon graduation.

As shown in Table 4, a classification table is constructed by predicting CGPA >2.4 or <2.4 of each student based on whether or not the odds for CGPA >2.4 are greater or less than 1.0 and comparing these predictions to the actual outcome for each student. The percents of correct decisions are 93.1 for students who graduated with CGPA >2.4, 28.9 for students who graduated with CGPA <2.4 and 77.6 overall. This overall result is compared with a rate of 24.1% that would be obtained by simply predicting students who graduated with CGPA <2.4 as the outcome for every student (i.e., since 83 of the 344 graduated with CGPA <2.4, this is the prediction with the greater likelihood of being correct). A four-predictor logistic model was fitted to the data. The result showed that the logistic regression equation for log-odds is estimated to be:

Table 4: The observed and perdicted fruquencies for graduating grades by Logistic regression

Observed	Predicted: Graduating student with CGPA <2.4	Predicted: Graduating student CGPA >2.4	Correct percentage
Step 1 CGPA			
Graduating student with CGPA <2.4	24	59	28.9
Graduating student CGPA >2.4	18	243	93.1
Overall percentage	-	-	77.6

$$\log_e\left[\frac{\pi}{1-\pi}\right] = -2.786 + 1.004\text{GPA} + 0.006\text{UMEscore} + 0.289\ \text{Gender} - 0.028\ \text{Age}$$

To estimate odd, the equation is exponentiated as:

$$\left[\frac{\pi}{1-\pi}\right] = e^{-2.786+1.004\text{GPA}+0.006\ \text{UMEscore}+0.289\ \text{Gender}-0.028\ \text{Age}}$$

The probability of success is obtained by applying the logistic transformation:

$$\pi = \frac{e^{-2.786+1.004\ \text{GPA}+0.006\ \text{UMEscore}+0.289\ \text{Gender}-0.028\ \text{Age}}}{1+e^{-2.786+1.004\ \text{GPA}+0.006\ \text{UMEscore}+0.289\ \text{Gender}-0.028\ \text{Age}}}$$

The Wald Chi-squared statistics are non-significant for UME score, gender and age (that is p-values of 0.166, 0.307 and 0.176, respectively) while the chi-squared value for GPA in final year is statistically significant at the 0.05 level (that is p-value of 0.000). Thus, given that the other predictors remain in the model, removing the GPA in final year as a predictor would result in significantly poorer predictive efficiency, although removing any of the other predictors does not have a significant impact (Table 3). The Hosmer-Lemeshow test

Model summary:	
Step 1	Values
-2 Log likelihood	0.14900
Cox and Snell R^2	324.600
Nagelkerke R^2	0.22300
Hosmer and Lemeshow test:	
Step 1	
χ^2	6.56900
df	8.00000
Sig.	0.58400

yielded a χ^2distribution with 8 degrees of freedom of 6.569 and was insignificant (p value = 0.584), indicating that the model fit is good. Measuring the usefulness of the model, the Cox and Snell and Nagelkerke R^2 are two such statistics. The values for the data are 0.149 and 0.223, respectively. In addition to these measures, includes a classification (Table 4) that documents the validity of predicted probabilities.

CONCLUSION

Logistic regression provides a useful means for modeling the dependence of a binary response variable on one or more explanatory variables where the latter can be either categorical or continuous. The model appears to suggest this conclusion: given that the other predictors remain in the model, removing the GPA in final year as a predictor would result in significantly poorer predictive efficiency, although removing any of the other predictors does not have a significant impact. The factor that contributed in the model is the final year GPA.

REFERENCES

Austin, J.T., R.A. Yaffee and D.E. Hinkle, 1992. Logistic regression for research in higher education. Higher Educ. Handbook Theory Res., 8: 379-410.

Chuang, H.L., 1997. High school youths dropout and re –enrollment behavior. Econ. Educ. Rev., 16: 171-186.

Dayton, C.M., 1992. Logistic regression analysis. http://bus.utk.edu/stat/datamining/Logistic%20Regression%20Analysis%20(Dayton).pdf.

Efron, B., 1975. The efficiency of logistic regression compared to normal discriminant analysis. J. Am. Statist. Assoc., 70: 892-898.

Harrell, F.E., 2001. Regression Modelling Strategies. 1st Edn., Springer-Verlag Inc., New York, ISBN: 978-0-387-95232-1, pp: 568.

Hosmer, D.W. and S. Lemeshow, 1989. Applied Logistic Regression. Wiley, New York.

Karp, A.H., 2007. Getting started with proc logistic. http://www2.sas.com/proceedings/sugi26/p248-26.pdf.

Lavange, L.M., V.G. Iannacchione and S.A. Garfinkel, 1986. An application of logistic regression methods to survey data: Predicting high cost users of medical care. pp: 270-275. http://www.amstat.org/sections/SRMS/Proceedings/papers/1986_049.pdf.

McFadden, D., 1974. Conditional Logit Analysis of Qualitative Choice Behavior. In: Frontiers in Econometrics, Zarembka, P. (Ed.). Academic Press, New York, pp: 105-142.

Peng, C.Y., K.L. Lee and G.M. Ingersoll, 2002. An introduction to logistic regression analysis and reporting. J. Educ. Res., 96: 3-14.

Rosenbaum, P.R. and D.B. Rubin, 1983. The central role of the propensity score in observational studies for causal effects. Biometrika, 70: 41-55.

Senol, S. and G. Ulutagay, 2006. Logistic regression analysis to determine the factors that affect green card usage for health services. JFS, 29: 18-26.

Tabachnick, B.G. and L.S. Fidell, 1996. Using Multivariate Statistics. 3rd Edn., Harper Collins College Publishers, New York.