The Application of Logistic Regression Analysis to the Cummulative Grade Point Average of Graduating Students: A Case Study of Students' of Applied Science, Federal Polytechnic, Ilaro

FAGOYINBO, I.S. BSc (Mathematics), MSc(Statistics)
Department of Mathematics and Statistics, Federal Polytechnic, Ilaro, Nigeria
E-mail: ifagoyinbo@yahoo.com

AJIBODE, I.A. BSc (Mathematics), MSc(Statistics)
Department of Mathematics and Statistics, Federal Polytechnic, Ilaro Nigeria
E-mail: ileloveseun@yahoo.com

OLANIRAN, Y.O.A
Department of Marketing Federal Polytechnic, Ilaro Nigeria

Abstract
Logistic regression deals with the relationship existing between a dependent variable and one or more independent variables. It provides a method for modelling a binary response variable which takes values 1 and 0. In this study, 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 School of Applied Science, Federal Polytechnic, Ilaro, Ogun State. The study reveals that the final year grade point average (GPA) of the graduates has significant effect among other variables.

Keywords: Logistic Regression, Odd Ratio, Cgpa, Statistical Significance, Log Likelihood, Standard Error

INTRODUCTION

Logistic regression model introduced in late 1960's and early 1970's has in the early 1980's become routinely available in statistical packages. It has also found many applications in field like the social sciences (Chaung, 1977) and in educational research, especially in higher education (Austine et al, 1992). Logistic regression analysis extends the technique 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.

Situation involving categorical outcomes are quite common in practice. In educational program, predictions are made for the binary or success/failure in the same vein, operation units could be classified successful or not successful according to some objectives 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 2009/2010 graduates of the School of Applied Science, Federal Polytechnic, Ilaro. It drives motivation from the work of Peng et al (2002), supported by Karp (2007) who argued that logistic regression is an increasingly popular analytical 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 has been applied in a number of contexts; which includes: 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 as multivariate normal at each level of response variables). It has also been applied to study investigating the risk factors for low birth weight babies (Hosmer and Lemeshow, 1989). Other vital applications of 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 wherethe dependent variable is two cases knwon as multinomial or polytomous. Tshachride and Fids (1996) use the term polychotomous.

THEORETICAL FRAMEWORK OF LOGISTIC REGRESSION

Logistic regression analysis is part of a category of statistical model knwon as generalised 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).

Let X denotes the vector of predictors (X_1 , X_2 , X_3 , ..., X_k) and let the conditional probability that the outcome is present be denoted by the equation as:

$$P(Y = \frac{1}{X} = \pi(X)) \qquad (1)$$

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

$$\pi(X) = \frac{1}{1+ e^{X\beta}} \qquad (2)$$

$\pi(X)$ = The success probability of value X.
$X\beta$ = Stands for $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...+ \beta_n X_n$
e = exponent or the base of the system of natural logarithms.
Its transformation generates:

$$\text{Odd} = \frac{\pi}{1- \pi} \qquad (3)$$

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

$$\text{Logit } \{ \pi(X)\} = \log \{\frac{\pi(X)}{1- \pi(X)}\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...+ \beta_n X_n \qquad (4)$$

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

$$\pi = \text{P(Y = outcome of interest x: x is a specific value of x)} = \frac{e^{X\beta}}{1+ e^{X\beta}} \qquad (5)$$

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

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

The unknown parameter β_i in the logistic regression model are estimated by the method of maximum likelihood. Solving for logistic regression coefficient β_i 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 coefficient is evaluated using the WALD TEST:

$$Z = \frac{\hat{\beta}}{SE} \qquad (6)$$

Wald test statistics has a normal distribution when β =0. For the logistic regression model, the hypothesis H: β = 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, regression coefficient is 0, which is H: β=0.

Goodness of fit shows how effectively the model we have described the outcome variable. Selection is made to the available that it deems important in describing the dependent variable. Log likelihood is calculated for a candidate model based on summing the probabilities associated with the predicted an actual outcome of each case.

$$L(\beta) = \ln (\frac{1}{\beta}) = \sum(y ln(\pi(X) + (1 - y)\ln(1 - \pi(X)) \qquad (7)$$

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

$$D = -2\sum[y ln(\frac{\pi(X)}{y} + (1 - y))\ln(1 - \frac{\pi(X)}{1-y})] \qquad (8)$$

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 (Model without the variable – Model with the variable) (9)

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

The Cox and Snell's measures are based on log likelihood. Equation 10 provides the method of calculation for Cox and Snell's R^2.

$$R^2 = 1 - e^{(\frac{2}{n}(D\ (Model\ without\ the\ variable))} \qquad (10)$$

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

$R_N^2 = R_{CS}^2/R_{MAX}^2$ (11)

Where R_{MAX}^2 = 1 - $e^{(\frac{2}{N}(D\ (Model\ with\ the\ variable))}$ (12)

The MC Fadden's R^2 is a less common pseudo-R^2 variant, based on log-likelihood Kernels for the full versus the intercept only. Only models, Hosmer and lemeshow Chi-Square test of fit evaluates the goodness of fit by creating 10 ordered group of subjects then it compares the number actually in each group (observed) to the number predicted by the logistic regression model (MC Fadden, 1974). A good model fit is indicated by a non significant Chi Square value.

AIMS AND OBJECTIVES OF THE STUDY

Before the inception of the application of logistic regression model analysis, many institutions, organisation and companies only collate and stored data. Having no knowledge of detecting the relationship that exists between two or more explanatory variables. This research work is designed purposely to:

- ✓ Identify patterns and trends occurring in the Cumulative Grade Point Average (CGPA) frequently among the genders.
- ✓ The parameter estimation of patterns and trends in the CGPA.
- ✓ The diagnostic analysis checking.
- ✓ To evaluate and test whether the final year result of the students determines the final Cumulative Grade Point Average (CGPA)

DATA PRESENTATION AND ANALYSIS

The data tabulated below consists of the result of students both male and female in Applied Science and their cumulative Grade Point Average.

	GENDER		TOTAL
CGPA	MALE	FEMALE	
Graduating Students with CGPA<2.5	36	31	67
Graduating Student with CGPA>2.5	113	101	214
TOTAL	149	132	281

The characteristic of the data set are the dependent binary variable represented with 0 stands for graduating students with CGPA < 2.5 and 1 stands for graduating students with CGPA > 2.5 and male and female students are coded with 0 and 1 respectively.

From the table above, about 23.8% of the students had CGPA < 2.5 while 76.2% of students had CGPA > 2.5. And the females are 46.98% while males are 53.02%. A males odd of being graduated with CGPA<2.5 relative to females odd. This gives an odd odd ratio of 1:4 which suggests that males being graduated with CGPA < 2.5 are 4 times less than that of female.

The computations arre as follows:

The percentage of students:

With CGPA < 2.5: $\frac{67}{281}$ x $\frac{100}{1}$ = 23.8%

With CGPA > 2.5: $\frac{214}{281}$ x $\frac{100}{1}$ = 76.2%

Percentage of male students:

$\frac{149}{281}$ x $\frac{100}{1}$ = 53.02%

Percentage of female students:

$\frac{132}{281}$ x $\frac{100}{1}$ = 46.98%

The odd of male students:

$\frac{36}{113}$ = 0.319

The odd of female students:

$\frac{31}{103}$ = 0.306

The odd ratio of male students to female students is:

$\frac{0.314}{0.306}$ = 1.04 = 1:04

Taking the natural logarithm i.e. log of odd:

Log(1.04) = 0.017

Using GPA as the predicator, the logistic equation for log-odds in CGPA < 2.5 is obtained in SPSS logistic regression as:

$\text{Log}\left(\frac{\pi}{1-\pi}\right) =$ -0.150 + 0.037GPA

= -0.150 + 0.037(2.310)

= -0.150 + 0.0693

Log (odds) = -0.0807

The odds:

$$\frac{\pi}{1-\pi} = e^{-0.150+\,0.037\mathrm{GPA}} \qquad (13)$$

$= e^{-0.150+\,0.037(2.31)}$

$= e^{-0.0807}$

Odds = 0.922

Probability:

$$\pi = \frac{1}{1+e^{-(-0.150+\,0.037\mathrm{GPA})}} \qquad (14)$$

$\pi = \frac{1}{1+e^{0.0807}}$

$\pi = \frac{1}{1+1.084}$

$\pi = \frac{1}{2.084}$

$\pi = 0.48$

$\pi = 48\%$

And for GPA of 3.81, the log-odds and probability of obtaining CGPA > 2.5 are calculated below:

Log-odds:

= -0.150 + 0.037(3.81)

= -0.150 + 0.141

Log (odds) = -0.01

The odds:

$= e^{-0.150+\,0.037(3.81)}$

$= e^{-0.01}$

Odds = 0.99

Probability:

$\pi = \frac{1}{1+e^{0.009}}$

$\pi = \frac{1}{1+1.009}$

$\pi = \frac{1}{2.009}$

$\pi = 0.500$

At least 50%.

The logistic regression coefficient for GPA in final year is 0.037 and the exponent is $e^{0.037}$ = 1.038, while the standard error for β is 0.281 and the statistical significant is assessed by the Wald Chi-Square statistics as:

$Z = \frac{\hat{\beta}}{SE} = \frac{0.037}{0.281}$

= 1.32

$Z^2 = 1.32^2 = 1.74$

With the degree of freedom of 1, from the table it is significant at p-value of 0.00 for GPA in final year. Hence H_o is accepted, which supports the conclusion that GPA in the final year is a useful predictor of student performance upon graduation.

FINDINGS AND INTERPRETATION

From the data presentation analysis of CGPA of graduating students, the CGPA of students < 2.5 from the sample data are about 23.80% and 76.20% has CGPA > 2.5 and 53.02% of the male students while 46.98% of female students were present in the sample data. With the odd ratio of male students graduating with CGPA < 2.5 are 1:04 times that of female students.

This simply means that there are more female student graduating with CGPA<2.5 than male students. The result of the logit of the odds of male and female students allowed or shows a linear regression between the two dichotomous outcome variables (male and female) which are dependent variables because the outcome is 0.017 which is not equal to zero or undefined.

The latter part of the analysis reveals that there is a linear regression in students with lowest CGPA of 2.310 to graduate with CGPA >2.5 are -0.0807. The odd is 0.922 with a probability of 0.48 equivalent ton possibility of having chance of 48% in obtaining a CGPA > 2.5. However, for the CGPA of 3.81 having a linear regression of 0.01 i.e. logit = -0.01 to obtain a CGPA > 2.5 with odd 0.99 and probability of at least 0.5 equivalent to at least 50% of the students with GPA 3.81. The estimated logistic regression coefficient of GPA in final year is 0.037 and the exponent is $e^{0.037}$ = 1.038 which indicates that for an increase in GPA in final year the odd in favour of CGPA > 2.5 are estimated to be increased by a multiplicative factor of 1.038. The statistical significance is assessed by the Wald Chi-Square statistic with value 1.74 with degree of freedom of 1. It is significance at conventional level of significance of 0.05 with p-value of 0.00. This simply shows that the GPA in final year is a useful predictor of student's performance upon graduation.

CONCLUSION AND RECCOMMENDATION

This research work has been able to show that there is a linear regression between the entire compared dichotomous variable with the use of logit in the analysis. The study reveals that the factor that contributed to the students' success in the model is the final year GPA which is significant at conventional level.

- ✓ Moreover, from the findings, it was recommended that can employ the use of logistic regression so as to determine the CGPA of students and that student should take their final year GPA seriously.
- ✓ Government can make use of the logistic regression in the allocation of infrastructural facilities so as to bring development to various segments of the country.
- ✓ Banking system can also make use of logistic regression to know their customer's demand and how to effectively improve on their services.
- ✓ Finally, the research work is recommended for all bodies and institutions to determine job satisfaction, promotion levels as well as development of organisations.

REFERENCES

Agresti, A. (2002): Categorical data analysis. New York; Wiley interscience.

Agresti, A. (2007): Building and applying logistic regression models "An introduction to categorical data analysis. Hoboken, New Jersey; Wiley p. 138.

Ahan, A.E; & Okafor, R. (2010): Application of logistic regression model to graduating (CGPA of University Graduate-University of Lagos). Journal of Modern Mathematics and Statistics 2010, 2(2), pp. 58 – 62. Medwel publishing and scientific research publishing company.

Alemes, S., Jonasson, J.M., Genell, A., & Steinech, G. (2009): Bias in odd's ratio by logistic regression modelling and sample size. BMC Medical Research Methodology 9:56 Bromed Central Publishing.

Bolakrishnan, N. (1991): Handbook of the logistic distribution, Marcel Dekker.

Hilbe, J.M. (2009): Logistic Regression Models. Chapman and Hall/CRC press.

Hosmer, D.W. & Stanley, L. (2000): Applied Logistic Regression, 2nd Edition, New York; Chichiester Wiley.

Jonathan, M. & Michael, A. G. (2001): Multiple regression analysis and mass assessment: A review of the issues. The Appraisal Journal, pp. 89 – 109.

Mayers, J.H. & Forgy, E.W. (1963): The development of numerical credit evaluation systems. Journal of the American Statistics Association, 58(303), pp. 799 – 806.

Peduzzi, P., Concato, J., Kemper, E., Holford, T.R., & Feinsten, A.R. (1996): "A simulation study of the number of event per in a variable in logistic regression analysis". Journal of Epidemiology, 49(12), pp. 1371 – 1379.