IJSBAR
International Journal of Sciences: Basic and Applied Research

International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

A Probabilistic Model for Predicting Examination Performance: A Binary Time Series Regression Approach

I .U. Moffat, Ph.D*

Department of Mathematics & Statistics, University of Uyo, Uyo, 520231, Nigeria

Email: moffitto2011@gmail.com

Abstract

In any examination result, performance is a function of several variables, which could be linear or nonlinear in dimension. Under approximately normal condition, every examination result is a Bernoulli trial with two unique and independent outcomes: a success and a failure. In this work, we examine the goodness - of - fit of the ordinary least squares regression with binary dependent variables (linear probability model) and the logistic regression in modeling and predicting examination performance. The degree examination results of 2012/2013 graduating class of the Department of Statistics were considered having reflected all the categories of performance in our examination grading system [viz; First class, Second class (Upper & Lower) divisions, Third class, and Pass]. The analysis revealed that the binary logistic regression is a better approach for modeling and predicting examination performance since most examination conditions are abnormal and nonlinear in dimension.

***Keywords*:** Binary; binomial regression; logit; probability prediction; time series.

---

* Corresponding author.

E-mail address: moffitto2011@gmail.com.

1. Introduction

Studies involving examination results or performance of students have been in the increase recently. Various scholars and researchers have proposed different methods for modeling examination results.

This has been an important research area since effective modeling of examination results can lead to a better prediction of the performance of students in future, hence, providing the system an opportunity for evaluating the academic structures for optimal performance.

Recently, so many authors have considered researches involving categorical time series data of different dimensions and have showcased diverse challenges on the interpretation of the results [1,2,3]. This work, a probabilistic model is proposed for predicting examination performance, generated over time under an approximately normal condition, in linear and nonlinear dimensions. We also consider the materials and method in the next section followed by the empirical results and discussions.

2. Materials and Method

Degree examination results of 36 students in five different courses were considered for 2012/2013 graduating class. For the dependent binary variable (performance), a score of 50% and above was regarded as a pass and graded "1" while a score of below 50% was regarded as a failure and graded "0". The independent variables which were taken as the courses with grades A, B, C, D, E, and F were graded as 5, 4, 3,2, 1, and 0, respectively. The linear probability and binary logistic regression were separately applied to model the results based on their underlying assumptions.

2.1 Linear Probability Model

This is a special case of a binomial regression model where the observed variable for each observation takes values which are either 0 or 1. The probability of observing a 0 or 1 in any one case is treated as depending on one or more explanatory variables. For the "linear probability model", this relationship is particularly simple, and allows the model to be fitted by simple linear regression.

When Y is binary, the linear regression model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ is called the linear probability model. The predicted value is a probability given by: $E\{Y|X = x\} = \Pr(Y = 1|X = x)$ = Xβ.

Here, $\widehat{Y}$ = the predicted probability that $Y_{i} {}_{=1}$, given X.

β_1 = Change in probability that Y = 1 for a given change in X.

$$\beta_{1=} \frac{\Pr(Y=1|X=x+\Delta x)-\Pr(Y=1|X=x)}{\Delta x}$$

This model assumes that Y = Xβ + ε, such that the estimated coefficients are themselves the marginal effects.

With all exogenous regressors, $E(Y|X) = Pr[Y = 1|X] = X\beta$.

If on the other hand, some elements of X are endogenous, they will be correlated with ε. As the LPM with exogenous explanatory variables is based on standard regression, the zero conditional mean assumption $E(\varepsilon|X) = 0$ applies [4].

2.2 Logistic Regression

Logistic regression, or **logit regression**, as a type of probabilistic statistical classification model is also used to predict a binary response from a binary predictor. It helps in predicting the outcome of a categorical dependent variable (i.e., a class label) based on one or more predictor variables [5,6]. Thus, it is used in estimating the parameters of a qualitative response model. The probabilities describing the possible outcomes of a single trial are modeled, as a function of the explanatory (predictor) variables, using a logistic function. Frequently (and subsequently in this article) "logistic regression" is used to refer specifically to the problem in which the dependent variable is binary, that is, the number of available categories is two, while the problems with more than two categories are referred to as multinomial logistic regression or, if the multiple categories are ordered, as ordered logistic regression.

Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables, which are usually (but not necessarily) continuous, by using probability scores as the predicted values of the dependent variable. As such, it treats the same set of problems as does probit regression using similar techniques; the former assumed logistic function, the latter standard normal distribution function, although both yield similar results. Thus, logit regression models the probability of Y = 1 as the cumulative standard logistic distribution function, evaluated at: $Z = \beta_0 + \beta_1 X$.

This implies that: $Pr(Y = 1|X) = F(\beta_0 + \beta_1 X)$, where F is the cumulative logistic distribution function given by:

$$F(\beta_0 + \beta_1 X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Here, the measures of fit used are:

(i) The fraction correctly predicted = fraction of Y's for which predicted probability is > 50% (if $Y_i = 1$) or is < 50% (if $Y_i = 1$).

(ii) The pseudo-R^2 measure the fit using the likelihood function: measures the improvement in the value of the log likelihood, relative to having no X's, which simplifies to the R^2 in the linear model with normally distributed errors.

3. Results

Table 1 shows how the binary variables are introduced under the ordinary least squares procedure for the five Courses. From Table 2(a - e)**,** the predicted linear probability models are: Course A is $\hat{Y} = -0.034 + 0.237X$; Course B is $\hat{Y} = -0.093 + 0.28X$; Course C is $\hat{Y} = -0.017 + 0.235X$; Course D is $\hat{Y} = -0.006 + 0.219X$;

Course E is $\widehat{Y} = 0.012 + 0.22X$. On the proportion of the total variation accounted for by each of the linear models, about 86% is explained in Course A, 75% in Course B, 85% in Course C, 93% in Course D, and 98% in Course E. Using the fitted model to predict the probability of a student passing the course A, given the grade, it's seen that a student with an "F" grade, has a negative predicted probability value of -0.034, which is abnormal while the one with an "A" grade also has an abnormal probability value of 1.151 (which is 115.1%) of passing the course.

Table1: OLS Fit with Binary Variables

STUDENT	COURSE				
	A	B	C	D	E
1	$1 = \alpha_{0+\ \alpha_1\ +e_1}$	$1 = \alpha_{0} + 2\alpha_1 + e_1$	$1 = \alpha_{0} + 3\alpha_1 + e_1$	$1 = \alpha_{0} + 3\alpha_1 + e_1$	$1 = \alpha_{0} + 5\alpha_1 + e_1$
2	$1 = \alpha_{0} + 4\alpha_1 + e_2$	$1 = \alpha_{0} + 3\alpha_1 + e_2$	$1 = \alpha_{0} + 5\alpha_1 + e_2$	$1 = \alpha_{0} + 5\alpha_1 + e_2$	$1 = \alpha_{0} + 5\alpha_1 + e_2$
3	0	$0 = \alpha_{0} + 0\alpha_1 + e_3$	0	0	0
4	$1 = \alpha_{0} + 2\alpha_1 + e_4$	$1 = \alpha_{0} + 3\alpha_1 + e_4$	$1 = \alpha_{0} + 4\alpha_1 + e_4$	$1 = \alpha_{0} + 5\alpha_1 + e_4$	
5	$1 = \alpha_{0} + 5\alpha_1 + e_5$	$1 = \alpha_{0} + 2\alpha_1 + e_5$	$1 = \alpha_{0} + 5\alpha_1 + e_5$	$1 = \alpha_{0} + 5\alpha_1 + e_5$	$1 = \alpha_{0} + 5\alpha_1 + e_5$
6	$1 = \alpha_{0} + 5\alpha_1 + e_6$	$1 = \alpha_{0} + 4\alpha_1 + e_6$	$1 = \alpha_{0} + 3\alpha_1 + e_6$	$1 = \alpha_{0} + 4\alpha_1 + e_6$	$1 = \alpha_{0} + 4\alpha_1 + e_6$
7	$1 = \alpha_{0} + 4\alpha_1 + e_7$	$1 = \alpha_{0} + 2\alpha_1 + e_7$	$1 = \alpha_{0} + 5\alpha_1 + e_7$	$1 = \alpha_{0} + 5\alpha_1 + e_7$	$1 = \alpha_{0} + 4\alpha_1 + e_7$
8	$1 = \alpha_{0} + 4\alpha_1 + e_8$	$1 = \alpha_{0} + 3\alpha_1 + e_8$	$1 = \alpha_{0} + 4\alpha_1 + e_8$	$1 = \alpha_{0} + 4\alpha_1 + e_8$	$1 = \alpha_{0} + 5\alpha_1 + e_8$
9	$1 = \alpha_{0} + 5\alpha_1 + e_9$	$1 = \alpha_{0} + 3\alpha_1 + e_9$	$1 = \alpha_{0} + 3\alpha_1 + e_9$	$1 = \alpha_{0} + 3\alpha_1 + e_9$	$1 = \alpha_{0} + 4\alpha_1 + e_9$
10	$1 = \alpha_{0} + \alpha_1 + e_{10}$	$1 = \alpha_{0} + \alpha_1 + e_{10}$	$1 = \alpha_{0} + 4\alpha_1 + e_{10}$	$1 = \alpha_{0} + 4\alpha_1 + e_{10}$	$1 = \alpha_{0} + 5\alpha_1 + e_{10}$
11	0	$1 = \alpha_{0} + \alpha_1 + e_1$	$1 = \alpha_{0} + 2\alpha_1 + e_1$	0	$1 = \alpha_{0} + 4\alpha_1 + e_1$
12	$1 = \alpha_{0} + 4\alpha_1 + e_{12}$	$1 = \alpha_{0} + \alpha_1 + e_{12}$	$1 = \alpha_{0} + 3\alpha_1 + e_{12}$	$1 = \alpha_{0} + 4\alpha_1 + e_{12}$	$1 = \alpha_{0} + 5\alpha_1 + e_{12}$
13	0	0	0	0	0
14	$1 = \alpha_{0} + 5\alpha_1 + e_{14}$	$1 = \alpha_{0} + 4\alpha_1 + e_{14}$	$1 = \alpha_{0} + 5\alpha_1 + e_{14}$	$1 = \alpha_{0} + 5\alpha_1 + e_{14}$	$1 = \alpha_{0} + 4\alpha_1 + e_{14}$
15	$1 = \alpha_{0} + 2\alpha_1 + e_{15}$	0	0	0	0
16	$1 = \alpha_{0} + 4\alpha_1 + e_{16}$	$1 = \alpha_{0} + 3\alpha_1 + e_{16}$	$1 = \alpha_{0} + 3\alpha_1 + e_{16}$	$1 = \alpha_{0} + 5\alpha_1 + e_{16}$	$1 = \alpha_{0} + 4\alpha_1 + e_{16}$
17	$1 = \alpha_{0} + 3\alpha_1 + e_{17}$	$1 = \alpha_{0} + \alpha_1 + e_{17}$	$1 = \alpha_{0} + 4\alpha_1 + e_{17}$	$0 = \alpha_{0} + 0 + e_{17}$	$1 = \alpha_{0} + 4\alpha_1 + e_{17}$
18	0	0	0	0	0
19	0	0	0	0	0
20	0	0	0	0	0
21	0	$1 = \alpha_{0} + \alpha_1 + e_1$	0	0	$1 = \alpha_{0} + 4\alpha_1 + e_1$
22	$1 = \alpha_{0} + 3\alpha_1 + e_{22}$	$1 = \alpha_{0} + \alpha_1 + e_{22}$	0	0	0
23	0	0	0	0	0
24	0	0	0	0	0
25	0	0	0	0	0
26	$1 = \alpha_{0} + 3\alpha_1 + e_{26}$	$1 = \alpha_{0} + \alpha_1 + e_{26}$	$1 = \alpha_{0} + 2\alpha_1 + e_{26}$	$1 = \alpha_{0} + 2\alpha_1 + e_{26}$	$1 = \alpha_{0} + 4\alpha_1 + e_{26}$
27	0	0	0	0	0
28	$1 = \alpha_{0} + 3\alpha_1 + e_{28}$	$1 = \alpha_{0} + 3\alpha_1 + e_{28}$	$1 = \alpha_{0} + 4\alpha_1 + e_{28}$	$1 = \alpha_{0} + 5\alpha_1 + e_{28}$	$1 = \alpha_{0} + 5\alpha_1 + e_{28}$
29	$1 = \alpha_{0} + \alpha_1 + e_{29}$	0	0	0	0

30	$1 = \alpha_{0} + \alpha_{1} + e_{30}$	$1 = \alpha_{0} + 2\alpha_{1} + e_{30}$	$1 = \alpha_{0} + \alpha_{1} + e_{30}$	$1 = \alpha_{0} + 4\alpha_{1} + e_{30}$	0
31	0	0	0	0	0
32	0	0	0	0	0
33	0	0	0	0	0
34	0	0	0	0	0
35	0	0	0	0	0
36	0	$1 = \alpha_{0} + 3\alpha_{1} + e_{36}$	$1 = \alpha_{0} + 2\alpha_{1} + e_{36}$	0	$1 = \alpha_{0} + 4\alpha_{1} + e_{36}$

Table 2: Results of Model Fit with Ordinary Least Squares

Table 2a(i): ANOVA[b] (COURSE A)

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	7.188	1	7.188	218.768	.000[a]
	Residual	1.117	34	.033		
	Total	8.306	35			

a. Predictors: (Constant), SCORE

b. Dependent Variable: PERFORANCE

Table 2a(ii): Model Summary (COURSE A)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.930[a]	.865	.862	.18127
a. Predictors: (Constant), SCORE				

Table 2a(iii): Coefficients[a] (COURSE A)

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	T	Sig.
1	(Constant)	-.034	.040		-.840	.407
	SCORE	.237	.016	.930	14.791	.000

a. Dependent Variable: PERFORANCE

Table 2b(i): ANOVA[b] (COURSE B)

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	5.045	1	5.045	100.603	.000[a]
	Residual	1.705	34	.050		
	Total	6.750	35			

a. Predictors: (Constant), SCORE

b. Dependent Variable: PERFORMANCE

Table 2b(ii): Coefficients[a] (COURSE B)

Model		Unstandardized Coefficients		Standardized Coefficients	T	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.093	.051		-1.830	.076
	SCORE	.280	.028	.865	10.030	.000

a. Dependent Variable: PERFORMANCE

Table 2b(iii): Model Summary (COURSE B)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.865[a]	.747	.740	.22394

a. Predictors: (Constant), SCORE

Table 2c(i): Model Summary (COURSE C)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.922[a]	.850	.845	.19437

a. Predictors: (Constant), SCORE

Table 2c(ii): ANOVA[b] (COURSE C)

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	7.271	1	7.271	192.460	.000[a]
	Residual	1.285	34	.038		
	Total	8.556	35			

a. Predictors: (Constant), SCORE

b. Dependent Variable: PERFORMANCE

Table 2c(iii): Coefficients[a] (COURSE C)

Model		Unstandardized Coefficients		Standardized Coefficients	T	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.017	.044		-.378	.707
	SCORE	.235	.017	.922	13.873	.000

a. Dependent Variable: PERFORMANCE

Table 2d(i): Model Summary (COURSE D)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.966[a]	.934	.932	.12903

a. Predictors: (Constant), SCORE

Table 2d(ii): ANOVA[b] (COURSE D)

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	7.990	1	7.990	479.892	.000[a]
	Residual	.566	34	.017		
	Total	8.556	35			

a. Predictors: (Constant), SCORE

b. Dependent Variable: PERFORMANCE

Table 2d(iii): Coefficients[a] (COURSE D)

Model		Unstandardized Coefficients		Standardized Coefficients	T	Sig.
		B	Std. Error	Beta		
1	(Constant)	.006	.028		.210	.835
	SCORE	.219	.010	.966	21.906	.000

a. Dependent Variable: PERFORMANCE

Table 2e(i): Model Summary (COURSE E)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.988[a]	.976	.975	.07984

Table 2b(i): ANOVA[b] (COURSE B)

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	5.045	1	5.045	100.603	.000[a]
	Residual	1.705	34	.050		
	Total	6.750	35			

a. Predictors: (Constant), SCORE

a. Predictors: (Constant), SCORE

Table 2e(ii): ANOVA[b] (COURSE E)

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	8.783	1	8.783	1377.883	.000[a]
	Residual	.217	34	.006		
	Total	9.000	35			

a. Predictors: (Constant), SCORE

b. Dependent Variable: PERFORMANCE

Table 2e(iii): Coefficients[a] (COURSE E)

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	T	Sig.
1	(Constant)	.012	.019		.644	.524
	SCORE	.222	.006	.988	37.120	.000

a. Dependent Variable: PERFORMANCE

In reality, a student with grade F has a zero chance of passing the course while the one with grade A has a significant probability (not more than 1) of passing the course, as predicted by the logistic model. These further confirm that the Ordinary Least Squares is not appropriate for investigating dichotomous or otherwise “limited” dependent variables [5, 6, 7, 8].

In the same vein, if the performance is modeled as a function of the cumulative grade average points (CGPAs), the fitted model is; $E(Y) = -0.343 + 0.380X$.

Again, the problem here is the same as the LPM using the scores since a student with a CGPA that is below 1.0 will have a negative probability value while the one with a CGPA of 3.6 and above will have a probability value that is greater than one.

Furthermore, the prediction equation has not satisfied all the assumptions of OLS. Some of these are;

i. Constant Variance: It is expected that the variance of the residuals be the same for all Y's. In this case, it is obvious (from equation 2) that the variance is p(1-p), which means that the variance depends on the value of p, rather than be constant. The variance can only be constant if and only if the value of p is the same for all grades (which is not so). Hence, the assumption is violated.

ii. Normality: Also, OLS assumes that, for each set of values for the *k* independent variables, the residuals are normally distributed. This is equivalent to saying that, for any given value of performance, the residuals should be normally distributed. This assumption is also clearly violated as we can't have a normal distribution when the residuals are only free to take on two possible values.

4. Discussion

From the above results, we have further buttressed the fact that the linear probability model (LPM) is not a good estimator for modeling and predicting examination performance as it generates predicted probabilities that are too extreme even for moderate values of the grades obtained. These problems can be summarized thus: heteroskedasticity, unbounded predicted probabilities, non-normal errors, and the linear functional form. However, some researchers claim that although the predicted probabilities are flawed, their main interest lies in the models' marginal effects, and argue that it makes insignificant difference to use the LPM, with its constant marginal effects, rather than the more complex marginal effects derived from a proper estimated cumulative distribution functions (CDF) as demonstrated by the logistic model (see Table **3 (a - e) below)**. In all the cases, estimation procedure terminates at 20 iterations because a perfect fit is obtained in each logistic regression.

Table 3: LOGISTIC BINARY REGRESSION

Table 3a (i): Classification Table (a, b) for Course A

		Observed		Predicted		
				Y1		Percentage Correct
				.00	1.00	
Step 0	Y1	.00		20	0	100.0
		1.00		16	0	.0
	Overall Percentage					55.6

a. Constant is included in the model.

b. The cut value is .500

Table 3a (ii): Variables in the Equation

		B	S.E.	Wald	Df	Sig.	Exp(B)
Step 0	Constant	-.223	.335	.443	1	.506	.800

Table 3a (iii): Variables not in the Equation

			Score	Df	Sig.
Step 0	Variables	X1	31.895	1	.000
	Overall Statistics		31.895	1	.000

Table 3a (iv):Omnibus Tests of Model Coefficients

		Chi-square	Df	Sig.
Step 1	Step	49.461	1	.000
	Block	49.461	1	.000
	Model	49.461	1	.000

Table 3a (v): Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	.000(a)	.747	1.000

a. Estimation terminated at iteration number 20 because maximum iterations have been reached.

Table 3a (vi): Classification Table(a)

			Observed		
			Predicted		
			Y1		Percentage Correct
			.00	1.00	
Step 1	Y1	.00	20	0	100.0
		1.00	0	16	100.0
	Overall Percentage				100.0

a. The cut value is .500

Variables in the Equation

Table 3a(vii):

		B	S.E.	Wald	Df	Sig.	Exp(B)
Step 1(a)	X1	34.139	4820.852	.000	1	.994	670770709106501.000
	Constant	-84.463	12098.781	.000	1	.994	.000

a. Variable(s) entered on step 1: X1.

Table 3a(viii):Variables not in the Equation

			Score	Df	Sig.
Step 0	Variables	X1	36.000	5	.000
		X1(1)	20.571	1	.000
		X1(2)	3.600	1	.058
		X1(3)	.823	1	.364
		X1(4)	7.258	1	.007
		X1(5)	10.862	1	.001
		X2	20.925	4	.000
		X2(1)	14.400	1	.000
		X2(2)	.360	1	.549
		X2(3)	.090	1	.764
		X2(4)	9.600	1	.002
		X3	12.600	5	.027
		X3(1)	7.200	1	.007
		X3(2)	.823	1	.364
		X3(3)	.164	1	.686
		X3(4)	2.973	1	.085
		X3(5)	.569	1	.451
		X4	8.286	4	.082
		X4(1)	5.143	1	.023
		X4(2)	.823	1	.364
		X4(3)	.026	1	.871
		X4(4)	.569	1	.451
		X5	5.631	2	.060
		X5(1)	5.355	1	.021
		X5(2)	3.662	1	.056
	Overall Statistics		36.000	20	.015

Table 3b (i): Classification Table (a,b) for Course B

	Observed			Predicted		
	Y2		Percentage Correct			
	.00	1.00				
Step 0	Y2		.00	24	0	100.0
			1.00	12	0	.0
	Overall Percentage					66.7

a. Constant is included in the model.

b. The cut value is .500

Table 3b (ii): Variables in the Equation

	B	S.E.	Wald	Df	Sig.	Exp(B)
step 0 Constant	-.693	.354	3.844	1	.050	.500

Table 3b (iii): Variables not in the Equation

			Score	Df	Sig.
Step 0	Variables	X1	14.949	1	.000
	Overall Statistics		14.949	1	.000

Table 3b (iv):Omnibus Tests of Model Coefficients

		Chi-square	Df	Sig.
Step 1	Step	16.552	1	.000
	Block	16.552	1	.000
	Model	16.552	1	.000

Table 3b (v): Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	29.277(a)	.369	.512

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Table 3b (vi): Variables in the Equation

		B	S.E.	Wald	Df	Sig.	Exp(B)
Step	X1	.920	.288	10.184	1	.001	2.511
1(a)	Constant	-2.908	.945	9.474	1	.002	.055

a. Variable(s) entered on step 1: X1.

Table 3b (vii): Omnibus Tests of Model Coefficients

		Chi-square	Df	Sig.
Step 1	Step	29.277	1	.000
	Block	29.277	1	.000
	Model	45.829	2	.000

Table 3b (viii): Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	.000(a)	.720	1.000

a. Estimation terminated at iteration number 20 because maximum iterations has been reached.

Table 3b (viii): Classification Table (a)

	Observed			Predicted		
	Y2		Percentage Correct			
	.00	1.00				
Step 1	Y2		.00	24	0	100.0
			1.00	0	12	100.0
	Overall Percentage					100.0

a. The cut value is .500

Table 3b (ix): Variables in the Equation

		B	S.E.	Wald	Df	Sig.	Exp(B)
Step 1(a)	X1	-.007	1757.515	.000	1	1.000	.993
	X2	37.195	6001.843	.000	1	.995	14236737 14721980 0.000
	Constant	-92.616	14170.619	.000	1	.995	.000

a. Variable(s) entered on step 1: X2.

Table 3c (i): Classification Table(a,b) for Course C

			Predicted		
			Y3		Percentage Correct
Observed			.00	1.00	
Step 0	Y3	.00	22	0	100.0
		1.00	14	0	.0
	Overall Percentage				61.1

a. Constant is included in the model.

b. The cut value is .500

Table 3c (ii): Variables in the Equation

		B	S.E.	Wald	Df	Sig.	Exp(B)
Step 0	Constant	-.452	.342	1.748	1	.186	.636

Table 3c (iii): Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	X3	30.595	1	.000
	Overall Statistics		30.595	1	.000

Table 3c (iv): Classification Table(a)

	Observed			Predicted		
	Y3		Percentage Correct			
	.00	1.00				
Step 1	Y3		.00	22	0	100.0
			1.00	0	14	100.0
	Overall Percentage					100.0

a. The cut value is .500

Table 3c (v): Variables in the Equation

		B	S.E.	Wald	Df	Sig.	Exp(B)
Step 1(a)	X3	35.225	4800.310	.000	1	.994	1986529943907726.000
	Constant	-87.759	12124.286	.000	1	.994	.000

a. Variable (s) entered on step 1: X3.

Table 3c (vi): Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	48.114	20	.000
	Block	48.114	20	.000
	Model	48.114	20	.000

Table 3c (vi): Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	.000(a)	.737	1.000

a. Estimation terminated at iteration number 18 because a perfect fit is detected.

Table 3d (i): Classification Table (a,b) for Course D

	Observed			Predicted		
	Y4		Percentage Correct			
	.00	1.00				
Step 0	Y4		.00	22	0	100.0
			1.00	14	0	.0
	Overall Percentage					61.1

a. Constant is included in the model.

b. The cut value is .500

Table 3d (ii): Variables in the Equation

	B	S.E.	Wald	Df	Sig.	Exp(B)
Step 0 Constant	-.452	.342	1.748	1	.186	.636

Table 3d (iii): Variables not in the Equation

			Score	Df	Sig.
Step 0	Variables	X4	33.618	1	.000
	Overall Statistics		33.618	1	.000

Table 3d (iv): Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	48.114	1	.000
	Block	48.114	1	.000
	Model	48.114	1	.000

Table 3d (v): Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	.000(a)	.737	1.000

a. Estimation terminated at iteration number 20 because maximum iterations has been reached.

Table 3d (vi): Classification Table (a)

	Observed		Predicted		
			Y4		Percentage Correct
			.00	1.00	
Step 1	Y4	.00	22	0	100.0
		1.00	0	14	100.0
	Overall Percentage				100.0

a. The cut value is .500

Table 3d (vii): Variables in the Equation

		B	S.E.	Wald	Df	Sig.	Exp(B)
Step 1(a)	X4	32.871	4407.783	.000	1	.994	188685609292404.700
	Constant	-81.815	11202.959	.000	1	.994	.000

a. Variable(s) entered on step 1: X4.

Table 3e (i): Classification Table(a, b) for Course E

	Observed		Predicted		
			Y5		Percentage Correct
			.00	1.00	
Step 0	Y5	.00	19	0	100.0
		1.00	17	0	.0
	Overall Percentage				52.8

a. Constant is included in the model.

b. The cut value is .500

Table 3e (ii): Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-.111	.334	.111	1	.739	.895

Table 3e (iii): Variables not in the Equation

			Score	Df	Sig.
Step 0	Variables	X1	16.712	5	.005
		X1(1)	7.646	1	.006
		X1(2)	1.393	1	.238
		X1(3)	.920	1	.337
		X1(4)	1.726	1	.189
		X1(5)	5.166	1	.023
		X2	9.697	4	.046
		X2(1)	6.743	1	.009
		X2(2)	1.092	1	.296
		X2(3)	1.092	1	.296
		X2(4)	.037	1	.847
		X3	29.001	5	.000
		X3(1)	25.077	1	.000
		X3(2)	.920	1	.337
		X3(3)	3.658	1	.056
		X3(4)	6.490	1	.011
		X3(5)	2.503	1	.114
		X4	16.358	4	.003
		X4(1)	16.053	1	.000
		X4(2)	1.150	1	.284
		X4(3)	2.367	1	.124
		X4(4)	2.503	1	.114
		X5	36.000	2	.000
		X5(1)	36.000	1	.000
		X5(2)	15.475	1	.000
	Overall Statistics		36.000	20	.015

Table 3e (iv): Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	49.795	1	.000
	Block	49.795	1	.000
	Model	49.795	1	.000

Table 3e (v): Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	.000(a)	.749	1.000

a. Estimation terminated at iteration number 18 because a perfect fit is detected.

In this case, the predicted values for course A are all zero in the null model on the dependent variable with an overall percentage of 55.6% where B is the coefficient for the constants in each of the models. The Wald Chi-square test revealed that the null hypothesis of the constant being zero is not rejected since the p-value of 0.506 is greater the critical value of 0.05 (or 0.01). On the inclusion of X_1 (first predictor) in the model, the overall statistics shows that the test is significant at both 5% and 1% levels implying that the contribution of this course in predicting the overall performance of students is quite significant.

5. Discussion and Conclusions

From the predicted values of the dependent variable based on the model in Course A, 20 cases are observed to be 0 and are correctly predicted to be 0 while 16 cases are observed to be 1 and are correctly predicted to be 1. Also, no case is observed to be 0 and is incorrectly predicted to be 1, and vice versa. Thus the logistic regression describes the exact situation for predicting examination performance in this course. This is further confirmed by the overall percentage which increased from 55.6% (for the null model) to 100% (for the full model).

In Course B, 21 cases are observed to be 0 and are correctly predicted to be 0 while 8 cases are observed to be 1 and are correctly predicted to be 1. Also, no case was observed to be 0 but incorrectly predicted to be 1and vice versa, with an overall percentage of 61.7% (for the null model) and 100% (for the full model).

In Courses C and D, the situations were the same with 24 cases observed to be 0 and correctly predicted to be 0 while 14 cases were observed to be 1 and correctly predicted to be 1. However, no case was observed to be 0 or 1 but predicted otherwise, thereby recording an increase in the overall percentage from 61.1% (for the null model) to 100% (for the full model).

In Course E, 20 cases were observed to be 0 and correctly predicted to be 0 while 16 cases were observed to be 1and correctly predicted to be 1. Again, no case was observed to be 0 or 1 but incorrectly predicted. This generated an increase in significance from 52.8% (for the null model) to 100% (for the full model).

On the Omnibus tests for the significance of model coefficients, the Chi-square values for STEP, BLOCK and MODEL revealed that they are all significant at both 5% and 1% levels, thereby confirming the goodness - of - fit of the logistic regression in modeling and predicting examination performance with dichotomous variables.

Finally, some more predictive approaches based on these probable and other possible relationships (in linear and nonlinear dimensions) are being explored for subsequent research publications.

References

[1]. Agresti, A. An Introduction to categorical data analysis, 2007, 423, Wiley – Interscience.

[2]. Long, J. S. Regression models for Categorical and Limited Dependent Variables. Advanced Quantitative Techniques in the Social Sciences, 1997, 7. Sage Publications: Thousand Oaks, CA.

[3]. Powers, D., and Xie, Y. Statistical method for categorical data analysis Academic Press", 2000, San Diego, CA.

[4]. Melton, J. "Models for Binary dependent variables". IMT Institute of Advance studies Lecture Note, 2012.

[5]. Han, M., and Michael, A. G. Multiple Regression Analysis and Mass Assessment: A Review of the Issues. The Appraisal Journal, 2001, 1, pp. 89 - 109.

[6]. Hastie, T. and Tibshirani, R. Generalized additive models (2^{nd} Edition), 1990, London: Chapman & Hall.

[7]. Aldrich, J. H., and Nelson, F. D. Linear Probability, Logit, and Probit Models. John Herbert Aldrich Quantitative Applications in the Social Sciences, 1984, 45.

[8]. Strano, M., and Colosimo, B. M. Logistic regression analysis for experimental determination of forming limit diagrams. International Journal of Machine Tools and Manufacture, 2006, pp. 46 – 49.