World Applied Sciences Journal 9 (9): 1062-1066, 2010
ISSN 1818-4952
© IDOSI Publications, 2010

Binary Logistic Regression Analysis Technique Used in Analyzing the Categorical Data In Education Sciences: A Case Study of Terengganu State, Malaysia

[1]Wan Muhamad Amir W. Ahmad, [2]Nor Azlida Aleng and [3]Zalila Ali

[1,2]Jabatan Matematik, Fakulti Sains dan Teknologi, Malaysia, Universiti Malaysia Terengganu (UMT),
21030 Kuala Terengganu, Terengganu Malaysia
[3]Pusat Pengajian Sains Matematik,
Universiti Sains Malaysia (USM), 11800 Minden Pulau Pinang, Malaysia

Abstract: Multiple logistic regression is a technique for modeling and studying an association between several variables. It is observed that both of multiple linear regression and multiple logistic regression methods are frequently used in social sciences, medical sciences, education sciences and many more. Similar with the multiple linear regression cases, it's focused on the relationship between a dependent variable and one or more independent variables. However, to use of this method it's depends on some circumstances. The problems always arising at the stage of doing the analysis and interpreting the computer output. In this study, the practical applications of the above mentioned method are discussed. Then, an in-depth analysis follows on advantages and utilization of logistic regression, which is suggested as the technique that resolves the issues related to the statistical.

Key words: Multiple logistic regression • Categorical data variables • And Receiver Operating Characteristic (ROC)

INTRODUCTION

Data can be reached in various modes depending on the core principle of the research. For example, while in certain cases it is necessary to measure, elaborate or analyzes a feature; in other circumstances it is necessary to classify the data into categorical groups (or categories). On other occasions data is gathered in ordinal form. Statistical interpretation of these data requires different statistics methods[1].

According to Akturk [2] and Baspinar and Mendes [3], behavioral sciences, response variables are often consist of categorical or categorized data. Related analysis that widely used is Chi-Square analysis, Fischer's exact test, G-Statistics and Ratio Test (Z-test). However, the use of these statistical techniques depend on several assumption and sometimes these assumption cannot be satisfied to use these techniques or even if the assumption are met results of the analysis are too general for interpretation.

MATERIAL AND METHODS

Logistic regression is a type of predictive model that can be used when the target variable is a categorical variable with two categories for instance live or die, has cancer or no cancer, having coronary heart disease or not having coronary heart disease, patient survives or dies and many more.

In logistic regression, the dependent variable is binary or dichotomous; its only contains data coded as 1 or 0. The objective of logistic regression is to find the best fitting model to illustrate the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables.

Let us consider that, there are P is independent variables which will be denoted as vector $x^1=\{x_1, x_2, ... x_p\}$ We assume that each of these variables is at least interval scaled. Let the conditional probability that the outcome is present be denoted by:

Corresponding Author: Wan Muhamad Amir W. Ahmad, Jabatan Matematik, Fakulti Sains dan Teknologi, Malaysia, Universiti Malaysia Terengganu (UMT), 21030 Kuala Terengganu, Terengganu, Malaysia.

Table 1: Explanation of the variables

Variables	Code	Explanation of the variables	Categorical
y	Useful	eBook is very useful to help students	0 = disagree 1 = agree
X_2	Hard	Students more hardworking doing their homework using eBook	0 = disagree 1 = agree
X_1	Good	eBook designed to help student to have a good performance in education	0 = disagree 1 = agree

$$P(Y=1|x)=\pi(x) \qquad (1)$$

Then the logit of the multiple regression models is given by the formula as follows:

$$\text{logit } (\pi)=b_0+b_1x_1+b_2x_2+b_3x_3+\text{.......}b_px_p \qquad (2)$$

The specific form of the logistics regression model is given by:

$$\pi(x)=\frac{e^{g(x)}}{1+e^{g(x)}} \qquad (3)$$

Where $\pi(x)$ is the probability of occurrence of the characteristic of interest. Since the model produce by logistic regression is nonlinear, the equations used to describe the outcomes are slightly more complex than those for multiple regression. This linear regression equation creates the logit transformation. This transformation is defined, in terms of π(x), as:

$$ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)=\sum_{j=1}^{p}b_jx_{ij} \qquad (4)$$

The objective of this study is to discuss multiple logistic analyses, which can be used in the analysis of categorical or categorized data.

Material of this study is a hypothetical sample which is composed of seven variables. Namely variables are as in Table 1. Multiple logistic regression technique was used in the analysis of relationship between variables. Data of 70 respondents (teachers) were collected. The sample size calculations for the data are given by single proportion method.

Sample Size Required: The sample sizes required at analysis stage are as follows:

Anticipated population proportion (P) = 0.953
Level of significance= 5% (0.05)
Absolute precision (△) = ±5%

$$= \left(\frac{196}{005}\right)^2 0953(1-0953)$$

= 68.8 ≈70 respondents.

The sample of 70 respondents required at the analysis stage.

RESULTS AND DISCUSSION

Step 1 : Perform Univariable Analysis:

Table 1: Variables X_2 in the Equation

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
X_2	3.9444	1.1005	12.8457	1	0.0003	0.4124	51.6473
Constant	-0.0953	0.4369	.0476	1	0.8273		

We found that x_2 is an important factor (p-value < 0.001) at univariable analysis. The crude (unadjusted) OR is 51.6473. At univariable analysis, teachers with "agree status" (for the statement,x_2) has 52 times the odds (chance) than teachers with "disagree status" to have variable$y$ (eBook is very useful to help students).

We found that x_1 is an important factor (p-value < 0.001) at univariable analysis. The crude (unadjusted) OR is 27.778. At univariable analysis, teachers with "agree status" (for the statement, x_2) has 28 times the odds (chance) than teachers with "disagree status" to have

Table 2: Variables X_2 in the Equation

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
X_1	3.3242	0.7874	17.8235	1	0.0000	0.4868	27.7778
Constant	-.5108	0.5164	0.9785	1	0.3226		

Table 3: Variables X_1 and X_2 in the Equation

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
X_2	3.1463	1.1850	7.0494	1	0.0079	0.2823	23.2508
X_1	2.7359	1.0000	7.4849	1	0.0062	0.2942	15.4229
Constant	-1.3100	0.6739	3.7789	1	0.0519		

Table 4: Correlation Correlation between X_1 and X_2

Correlation between X_1 and X_2	
	X_1
X_2	r=-0.08500

Table 5: Variables X_1 and X_2 in the Equation

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
X_2	2.1972	1.3333	2.7156	1	0.0994	0.1063	9.0000
X_1	2.1972	1.0541	4.3450	1	0.0371	0.1924	9.0000
$X_1 \times X_2$	7.9070	40.8574	0.0375	1	0.8465	0.0000	2716.3
Constant	-1.0986	0.6667	2.7156	1	0.0994		

variable y. From the above analysis, we found that the variables x_1 and x_2 are important and we have to include them into the model. So, we run the model again and gained the results as follows:

Step 2: Perform Variable Selection: We review all the results in univariable analysis and select the variables based on their p value (<0.25)). In our cases, we select the two variables that's x_1 and x_2.

Step 3: Check Multicollinearity: In step 3, we check multicollinearity in order to assess which variables are correlated highly. There are two methods that can be used and they are as follows:

The first step is by referring to the Table 4, from that table, we found that correlation between x_1 and x_2 are correlated lowly with one another. The second step is by referring to the standard errors of the variables. Table 3 indicates that that the standard errors of the variables are small and this means that there are no multi-collinearity effects.

Step 4: Checking for the Interaction: We have checked all two way interaction between two important independent variables in the equation and the results shows that the interaction part is not significant (p = 0.8465). So we make decision not to include the interaction in the model while the others two variables alone are significant.

Step 5: Asses the Goodness-of-fit: In order to know the goodness-of-fit of the model, we tested the model with the three method. They are

- The classification Table
- Hosmer-Lemeshow test
- Area under the ROC curve

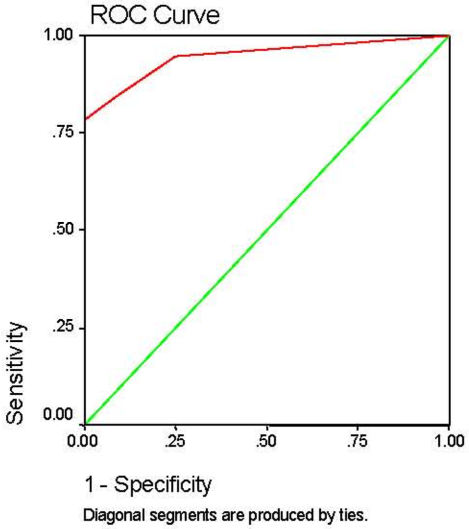

Fig. 1: Receiver Operating Characteristic (ROC) Curve

Table 6: Summary under the Curve

Area		Predicted Value	0.946
Std. Error		Predicted Value	0.026
Asymptotic Sig.		Predicted Value	0.000
Asymptotic 95%			
Confidence	Lower	Predicted Value	0.895
	Upper	Predicted Value	0.998

The area under the ROC curve is 0.946, it is significantly different from 0.5 (with p-value is less than 0.05). This information tells us that the model can discriminate 94.6% of the cases.

In order to assess the goodness-of-fit, we tested the hypothesis that data fit the model well versus data do not fit the model well. From the point of view of Hosmer and Lemeshow Test, we accept the hypothesis (with P=0.5642) that the model can predict well. p-value is greater than 0.05 indicated that the model is good enough to use.

Table 7: Summary of Hosmer and Lemeshow Test

	Chi-Square	df	Sig
Goodness-of-fit test	1.1448	2	0.5642

Table 8: Classification Table

		Predicted		
		Agree	Disagree	% Correct
Observed		1	0	
Agree	1	53	3	94.64%
Disagree	0	3	9	75.00%
Overall % correct	91.18			

The classification table in Table 8 shows that the overall percentage correct is very good (above 90%). 91.18% of all respondent said that “eBook is very useful to help students” can be predicted accurately by the model. Ideally, a valid test needs to be high sensitivity and high in specificity.

Step 6: Establish Final Model: Respondents with an agree opinion with the statement “eBook designed to help student to have a good performance in education” has 15 times the odds to agree with the opinion “eBook is very useful to help students”(Cl: 2.27 to 237.21, p-value <0.001). A respondent with an agree opinion with the statement “students more hardworking doing their homework using eBook” has 23 times the odds to agree with the opinion “eBook is very useful to help students”(Cl: 2.17 to 109.49, p-value <0.001).

Table 9: Sensitivity and Specificity Table

Sensitivity = 53/(53+3)% = 94.6%.	percentage of occurrences correctly predicted
Specificity = 9/ (3+9) % = 75%.	percentage of nonoccurrence correctly predicted
False Positive Rate =3/53% = 5.67%.	percentage of predicted occurrences which are incorrect
False Negative Rate =3/ (3+9) % = 25%.	percentage of predicted nonoccurrence which are incorrect

Table 10: Variables X_1 and X_2 in the Equation

					95% CI for Exp(B)	
Variable	B	Sig	R	Exp(B)	Lower	Upper
X_2	3.1463	0.0079	0.2823	23.2508	2.1725	109.4872
X_1	2.7359	0.0062	0.2942	15.4229	2.2790	237.2149
Constant	-1.3100	0.0519				

Results Presentation:

Table 11: Logistic Regression Analysis of 70 eBook Respondents

							95% CI for Exp(B)	
Predictors	B	S.E.	Sig	df	Wald	Exp(B)	Lower	Upper
Constant	-1.3100	0.6739	0.0519	1	3.7789			
X_2 *(1= Agree)(0= Disagree)*	3.1463	1.1850	0.0079	1	7.0494	23.2508	2.1725	109.4872
X_1 *(1= Agree)(0= Disagree)*	2.7359	1.0000	0.0062	1	7.4849	15.4229	2.2790	237.2149

Test	X_2	df	p
Overall model evaluation-2 Log Likelihood	28.503	2	0.000
Goodness-of-fit test			
Pearson	1.145	1	0.285
Deviance	1.511	1	0.219
Hosmer-Lemeshow	0.263	1	0.608

DISCUSSION AND CONCLUSION

In this paper, we show that logistic regression is a very powerful analytical technique when the outcome variable is dichotomous. In order to know more about effectiveness of the logistic model we used the guideline as follow:

- Checking significance test for each predictor
- Checking significance tests of the model against the null model
- Descriptive and inferential goodness-of-fit indices
- Predicted probabilities

We hoped that this article give some guidelines and illustration of how logistic regression is applied to a dataset. As a research finding, from this analysis, we found that respondents with an agree opinion with the statement "eBook designed to help student to have a good performance in education" has 15 times the odds to agree with the opinion "eBook is very useful to help students" and the same interpretation goes to the statement that "students more hardworking doing their homework using eBook" have 23 times the odds to agree with the opinion "eBook is very useful to help students".

ACKNOWLEDGEMENTS

We wish to thank Vice Chancellor of University Malaysia Terengganu (UMT) and their research team for providing real dataset for this analysis.

REFERENCES

1. Akturk, D.A., G. Sema and K. Taner, 2007. Multiple Correspondence Analysis Technique Used in Analyzing the Categorical Data in Social Sciences. J. App. Sci., 7: 585-588.
2. Akturk, D., 2004. Multiple Correspondence Analysis Technique: Its Application In Social Science Researches. J. Agri. Sci., 10: 218-221.
3. Baspinar, E. and M. Mendes, 2000. The Usage of Correspondence Analysis Technique At The Contingency Tables. J. Agric. Sci., 6: 98-106.