International Journal of Multidisciplinary Research and Development
www.allsubjectjournal.com
Online ISSN: 2349-4182, Print ISSN: 2349-5979
Received: 19-01-2022, Accepted: 03-02-2022, Published: 19-02-2022
Volume 9, Issue 2, 2022, Page No. 68-73

Notes and applications in multinomial logistic regression model

AbuElgasim Abbas Abow Mohammed
College of Business and Economic, Qassim University, Buraydah, Saudi Arabia

Abstract
This study presented a notes and an application in Multinomial Logistic Regression model , according to assumption's model , such a model is very important deals with one nominal or ordinal response variable that has more than two categories, whether nominal or ordinal variable, this model has been applied in many areas , a real data of sample size 694 student according to their (status, grades and gender) in two semester year 2021 on education from College of Business and Economics Qassim University Kingdom of Saudi Arabia has been used in the study. In order to build a multinomial logistic regression model, student status is considered as dependent variable and grades and gender as independent variables, which tested through a set of statistical test. The study concluded that, the student grades is significant variable that affect the student status and it recommended that to use a multinomial logistic in a large different fields of educational.

Keywords: multinomial logistic regression, response variable, student status, student grades, gender

Introduction

In the multiple regression model, we assume that a linear relationship exists between some variable, Y which we call the dependent variable, and k independent variables, $X_1, X_2, \ldots . X_k$. The independent variables are sometimes referred to as explanatory variables, because of their use in explaining the variation in Y. In case that the outcome variable is discrete, taking on two or more possible values the method is called logistic regression (LR) means that the outcome variable is binary or dichotomous. Multinomial logistic regression (MLR) model is used widely in biometrics, econometrics, psychometric sociometric and many other fields, where it uses a set of explanatory variables to predict the probabilities of different possible outcomes of categorically distributed responses (Dakin *et al*., 2006 [7]; Millington *et al*; 2007 [13]; Briz and Wald, 2009 [4]; Choi *et al*.,2011 [6]; Dey, Olio *et al*.,2011 [8]. The goal of an analysis of each method is finding the best fitting model and testing for significance coefficient in the model Abow (2020) [2].

A common problem that is how to determine the appropriate model of regression in the case of the multiple response dependent variable, it is more general than logistic regression because the dependent variable is not restricted to two categories. These methods include regression models but in a way that suits the condition of the dependent variable, which is multiple response these include the multinomial logistic regression model Seber (1977) [15].

The aim of the study is to introduce the concepts of multinomial logistic regression and to explain its measures according to the study problem

The importance of this study stems from the use of multinomial logistic regression by applying it in a data collected from the Business and Economics College Qassim University Kingdom of Saudi Arabia, which is in the long run will enrich and pave the way towards scientific researches in the college, as well as expanding the circle of knowledge on how to apply such a model.

Literature Review

Abow (2020) [2] presented a comparison between a simple linear regression model and binary logistic regression model, used preparatory year student's data of Business and Economic College to know the effect of students grades and term level on student status, found that the grades have a significant effect on student status.

Lin, Deng, MA (2014) [10] provided a comparison of multinomial logistic regression and logistic regression: which is more efficient in allocating land uses, to determine the relationship between land use change and its driving factors. The comparison of two regression methods indicated that the proportion of correctly allocated pixels using multinomial logistic regression 92.98 % which was 8.47% higher than obtained using logistic regression, the results also showed that the pixels were more clearly distinguished by multinomial logistic regression than by logistic regression. Abdalla. ELhabil (2012) [2] presented an application of multinomial logistic regression, he used real data on physical violence against children, from a survey of youth 2003 which was conducted by Palestinian Central Bureau of Statistic. Segment of the population of children in the age group (10-14 years) for residents of Gaza governorate, he used the multinomial logistic regression for defining the relationship between the group of explanatory variables and the response variable, identify the effect of each of the variables, and predict the classification of any individual case.

Presentation of the data

Data for this study was provided by Business and Economic college of Qassim University, Saudi Arabia, it was taken from quantitative method unit which is the grades of student's course of Statistics for the Management & Economic (2) divided in two semester of year 2021.

Assignment of dependent variable

Two types of variables have been used in the study which were categorical and continuous variables.

Dependent variable: (student status) Categorical classified into five levels;

1. (i.e. from 90 to 100 degree)
2. B (i.e. from 80 to 89 degree)
3. C (i.e. from 70 to 79 degree)
4. D (i.e. from 60 to 69 degree)
5. F (i.e. less than 60 degrees - failure)

Independent variables

The independent variables of this study represented as followed

1. Students' grades
2. Gender (Dichotomous binary) in two semester year 2021.

Methodology

This study depends on the theoretical aspect that dealt with the multinomial logistic regression model supported the practical aspect that depend on course of Statistics for the Management & Economic (2) of the College of Business and Economics at Qassim University in the Kingdom of Saudi Arabia. The study used SPSS for analyzing data.

Multinomial logistic regression (**MLR**)

MNL model is used to predict a nominal dependent variable given one or more independent variable ,its sometimes considered an extension of binomial logistic regression to allow for a dependent variable with more than two categories, as with other types of regression, MLR model can have nominal / or continuous independent variables and can have interactions between independent variable, also MLR model that generalizes the Logistic Regression(LR) model by allowing for more than two discrete and unordered response by Luce (1959) [11]. Mc Fadden (1974) [12] provided a general procedure formulating the MLR model. Currently the MLR model is widely used in biometrics, econometrics, psychometrics, sociometric, and many other fields, where it uses in a set of explanatory variables to predict the probabilities of different possible outcomes of categorically distributed responses (Dakin *et al*.,2006, Millington *et al*.,2007) [7, 13].

Assumptions of multinomial logistic regression model

The multinomial logistic regression model assumes that:

1. Dependent variable should be measured at the nominal level with more than or equal to three values.
2. one or more independent variables that are continuous, ordinal or nominal (including dichotomous variables). However, ordinals independent variables must be treated as being either continuous or categorical.
3. Should have independence of observations and the independent variable should have mutually exclusive and categories
4. There should be no multicollinearity occurs when we have two or more independent variables that are highly correlated with each other
5. There needs to be a linear relationship between any continuous independent variables and the logit transformation of the dependent variable
6. There should be no outliers high leverage values or highly influential points for the scaled continuous variables.

Multinomial logistic regression model.

The logistic regression can be extending to models with multiple explanatory variables. Let k denotes number of predictors for a binary response Y by $x_1, x_2, \dots x_k$ the model for log odds is

Logit $[p(y=1)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k$ and the alternative formula, directly specifying $\pi(x)$, is

$$\pi(x) = \frac{\exp(\alpha+\beta_1 x_1+\beta_2 x_2+\cdots\beta_k x_k)}{1+\exp(\alpha+\beta_1 x_1+\beta_2 x_2+\cdots\beta_k x_k)} \quad (1)$$

The parameter β_i refers to the effect of X_i on the log odds that y=1, controlling other X_j, for instance, exp (β_i is the multiplicative effect on the odds of a one-unit increase in X_j at fixed levels of other X_j if we have n independent observations with p-explanatory variables, and qualitative response variable has k categories, to construct the logit in the multinomial case one of the categories must be considered the base level and all the logits are

constructed relative to it. Any category can be taken as the base level, so, we will take category k as the base level. Since there is no ordering.it is apparent that any category may be labeled k.

Let π_j denote the multinomial probability of an observation in the jth category to find the relationship between this probability and the p explanatory variable, $X_1, X_2, \ldots X_p$ the multiple logistic regression model then is

$$log\left[\frac{\pi_j(x_i)}{\pi_k(x_i)}\right] = \alpha_{0i} + \beta_{1j}x_{1i} + \beta_{2j}x_{2i} + \cdots \beta_{pj}x_{pi} \quad (2)$$

Where j=1, 2,…,(k-1), i =1,2,…n. since all the π's odd to unity, this reduces to

$$\log(\pi_j(x_i)) = \frac{\exp(\alpha_{0i}+\beta_{1j}x_{1i}+\beta_{2j}x_{2i}+\cdots\beta_{pj}x_{pi})}{1+\sum_{j=1}^{k=1}\exp(\alpha_{0i}+\beta_{1j}x_{1i}+\beta_{2j}x_{2i}+\cdots\beta_{pj}x_{pi})} \quad (3)$$

Where j=1,2,…,(k-1), the model parameters are estimated by the method of ML practically, we use statistical software to do this fitting, Chatterjee and Hadi (2006) [5]

Baseline-category logistic

Suppose that $x_i = (y_{i1}, y_{i2}, \ldots, y_{ij})^T$ has a multinomial distribution with index $n_i = \sum_{j=1}^{k} y_{ij}$ and parameter, $j = 1,2, \ldots k-1$

$\pi_i = (\pi_{i1}, \pi_{i2}, \ldots \pi_{ij})$ When the response categories 1, 2,…,k are unordered, the most popular way to relate π_j to covariates is thought a set of j-1 baseline category logit. Taking J as the baseline category,

$$\log\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) = x_i{}^T\beta_j, \; j \neq J \quad (4)$$

If x_i has length p, then this model has $(j-1)\times p$ parameters, which we can arrange as a matrix or a vector Schafer (2006) [14]

Where j= 1, 2, …(J-1), simultaneously describes the effect of X on these (J-1) logits, the effects vary according to the response paired with baseline, these (J-1) determine parameters for logits with other pairs of response categories since

$$\log\frac{\pi_a(x)}{\pi_b(x)} = \log\frac{\pi_a(x)}{\pi_J(x)} - \log\frac{\pi_b(x)}{\pi_J(x)} \quad (5)$$

With categorical predictors person chi-square statistic χ^2 and likelihood ratio chi-square G^2 goodness fit statistic provides a model check when the data are not sparse Agresti (2000) [3]

Goodness of fit model

If the estimated expected counts $\widehat{\mu_{ij}} = n_i\widehat{\pi_{ij}}$ are large enough, we can test the fit of our model versus a saturated model that estimates π independently for i=1…N.the deviance for comparing this model to a saturated one is

$$G^2 = 2\sum_{i=1}^{N}\sum_{j=1}^{k} y_{ij} = \log\frac{y_{ij}}{\mu_{ij}} \quad (6)$$

The saturated method has N (k-1) free parameters p is the length of x_i, so the degrees of freedom are df = (N-p)(k-1) Schafer (2006) [14].

Person Statistic

The corresponding Person statistic is

$$\chi^2 = \sum_{i=1}^{N}\sum_{j=1}^{k} k_{ij}{}^2 \quad (7)$$

Where $k_{ij} = \frac{y_{ij}-\widehat{\mu_{ij}}}{\sqrt{\widehat{\mu_{ij}}}}$

Results and Discussion

The first four assumptions of multinomial logistic regression assumptions, have been validated, the others will be validated with application to the data, regarding the multicollinearity, when multicollinearity is present the regression coefficients and statistical significance become unstable and less trust worthy though it doesn't affect how well the model fit the data parse. The able (1) below observe that tolerance for the variables grades students,

gender and term level is equal 1, similarly the variance inflation factor (VIF) corresponding the explanatory variables is 1, when VIF is less than 2.5 indicate that there is no collinearity problem. Midi, Sarkar, Rana (2013).

Table 1: Collinearity statistic

Model		Tolerance	VIF
1	Students' grades	1	1
1	gender	1	1
1	term level	1	1

Dependent Variable: Student Status.

Table (2) presented the valid observation used in the study is distributed among five categories, Column (N) provides the number of observations fitting the description in the first column. The marginal percentage column lists the proportion of valid observation found in each of the response variable groups, 14.4% of F, and 13.7% of D, and 15.4% of C, and 21.2% of B, and 35.3% of A.

Table 2: Case Processing Summary

Variables with classifications		N	Marginal Percentage
Student Status	F	100	14.40%
	D	95	13.70%
	C	107	15.40%
	B	147	21.20%
	A	245	35.30%
gender	male	340	49.00%
	female	354	51.00%
Term level	semester one	406	58.50%
	semester two	288	41.50%
Valid		694	100.00%
Missing		0	
Total		694	

The below table (3) showed the chi-square value of the model, which is significance because the p-value was (0.000), less than the level of significance 0.05, the interaction effect is contributing significantly to the full model and should be retained, so, the null hypothesis has been rejected indicates that the explanatory variables affect the response variable.

Table 3: Model Fitting Information

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	2110.415			
Final	286.049	1824.366	12	.000

Good ness of-fit table (4) provides labelled person chi-square statistic, which is statistically significant (p=0.000, is less than 0.05) indicates that the model was well fitted the data, and the value of deviance is significant as well, however, the model is considered to be fully fit the data.

Table 4: Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	3220925177	844	.000
Deviance	270.297	844	1

Table (5) represent pseudo R-square. There are three R-square in logistic regression, which were not equivalent to R-square in ordinary least square regression coefficient determination but use as indicator for model fit. A model with largest pseudo R-square statistic is best according to the measures; however, classification coefficient as overall affect size measures are preferred over pseudo R-square measures as they have some severe limitation for this purpose Garson (2009), according to value of R-square statistic (Cox and Snell is 0.928) and (Nagelkerke is 0.973) and (McFadden is 0.856), the model is fully fit the data.

Table 5: Pseudo R-Square

Cox and Snell	0.928
Nagelkerke	0.973
McFadden	0.856

Table (6). represents the likelihood ratio tests which reflects the overall effect of a nominal variable, the results showed that the independent variables were not all statistically significant (i.e. students grades row) was statistically significant because p = 0.00 $<$ 0.05, but the (gender row) and the (Term level row) are not statistically significant. because the p-values of these rows are greater than 0.05, There is not usually any interest in the model intercept (i.e. the "Intercept" row).

Table 6: Likelihood Ratio Tests

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	286.049[a]	0	0	.
Students' grades	1978.01	1691.961	4	.000
gender	295.069	9.02	4	0.061
Term level	286.804	0.755	4	0.944

The likelihood ratio test is mostly useful for nominal independent variables because it is only test that considers the overall effect of a nominal variable unlike the parameter estimates test (i.e. table (7)), the table presents the parameter estimate (also known as the coefficients of the model) as there were five categories of dependent variable and three of independent variable of coefficients which is students' grades, gender and term level. From the table it's obvious each student status has a coefficient equivalent to student grade which were all statistically significant (p = 0.00), meanwhile, the related classification male, is not statistically significant for both F and A (p = 0.934, p =0.149 respectively) and statistically significant for C and B (p=0.024, p=0.016). Regarding the term level 1, it is clear that they are all not statistically significant (p=0.886, p=0.662, p=0.894, p=0.891) for F, C, B, and A respectively. In respect to the value of coefficients β , if we increase student grade score by one point the multinomial log-odds of preferring D to F would be expected to decrease by 0.683 unit while holding all other variables in the model constant, and if we increase students grades score by one point the multinomial log-odds of preferring C to D would be expected to increase by 1.087 unit while holding all other variables in the model constant, and if we increase students grades score by one point the multinomial log-odds of B preferring to D would be expected to increase by 2.063 unit while holding all other variables in the model constant, and if we increase students grades score by one point the multinomial log-odds of preferring A to D would be expected to increase by 3.023 unit while holding all other variables in the model constant.

The "exp (B)" column in the table label for odds ratio of the explanatory variables with the response variable, it is predicted change in odds for a unit increase in the corresponding explanatory variable. Odds ratios less than 1 correspond to decreases and odds ratio more than 1 correspond to increases. Odds ratios close to 1.0 indicates that unit changes in that explanatory variable does not affect the response variable.

Table 7: Parameter Estimates

Student status	(B)	Std Error	Wald	df	sig	Exp (B)	95% confidence Interval	
							Lower	Upper
F intercept	40.398	7.514	28.908	1	0			
Students' grade	-0.683	0.125	30.023	1	0	0.505	0.396	0.645
Male	-0.061	0.736	0.007	1	0.934	0.941	0.223	3.977
Female	0			0				
Term1	-0.111	0.774	0.02	1	0.886	0.895	0.196	4.082
Term2	0			0				
C intercept	-76.809	13.582	31.98	1	0			
Student grades	1.087	0.193	31.815	1	0	2.966	2.033	4.327
Male	1.64	0.726	5.102	1	0.024	5.156	1.242	21.402
Female	0			0				
Term1	0.326	0.746	0.191	1	0.662	0.722	0.167	3.117
Term2	0			0				
B intercept	-154.78	17.963	74.244	1	0			
Student grades	2.063	0.243	72.305	1	0	7.868	4.891	12.658
Male	2.306	0.957	5.809	1	0.016	10.032	1.533	65.414
Female	0	0.971		0				
Term1	0.13		0.018	1	0.894	1.138	0.17	7.641
Term2	0			0				
A intercept	-240.16	21.149	128.951	1	0			
Student grades	3.023	0.273	122.903	1	0	20.551	12.043	35.07
Male	1.566	1.086	2.079	1	0.149	4.786	0.57	40.213
Female	0			0				

Term1	0.151	1.102	0.019	1	0.891	1.163	0.134	10.089
Term2	0			0				

The reference category is (D)

Conclusion and Recommendation

From the results of the study and application in the multinomial logistic regression, we conclude the following notes

1. From the application of the method of multinomial logistic regression on the collected data to find out the main factors that have real impact on the dependent variable, the study showed that, the tests that had been carried out in the analysis emplies that the model is well fitted the data
2. The likelihood ratio test with chi-square = (1824.366, i.e. 2110.415-286.049) indicated that the existence of a relationship between the independent variables and the dependent variable was supported.
3. the student's grades variable is the only explanatory variable that has significant impact on the student status while other explanatory variables are not.
4. The MLR model is a suitable model to many types of data when the response variable is more than two categories.
5. The MLR model indicates the effect of each explanatory variables as well as its additive effect by used in the analysis simultaneously, which we are looking for in such a case of study. Thus, the paper recommends that both educationalists and researchers take advantage of the applications of multinomial logistic regression in different educational fields to solve some problems related to the relationships between different educational variables

References

1. Abdalla. EL-HABlL, An Application on Multinomial Logistic Regression Model, pak.j.stat.oper.res, 2012, 271-291.
2. Abow AA. Inferences About the Use of Linear regression and Logistic regression, International Journal of Recent Scientific Research,2020:11(8):39547-39552.
3. Agrestic A. Categorical Data Analysis, John Wiley and Sons, Inc, 2000.
4. Briz and Wald. Consumer awareness of Organic Products in Spain: An application of Multinomial Logit Models. Food Policy,2009:34(3):295-304.
5. Chatterjee S, Hadi A. Regression Analysis by Example. John Wiley & Sons, 2006.
6. Choi *et al*. An assessment of influence of bioenergy and marketed and amenity values on land uses in the mid-western Us Ecol Econ,2011:70(4):713-720.
7. Dakin *et al*. H A, Devlin NJ, Odeyemi IAO "yes" "No" or "Yes but"? Multinomial Modelling of Nice Decision-Making. Health Polices,2006:77(3):352-367.
8. Dey Olio *et al*. The quality of service desired by public Transport Users Transport Policy,2011:18(1):217-227
9. Garson D. Logistic Regression with SPSS, North Carolina State University, Public Administration Program, 2009.
10. Lin, Deng MA. A comparison of Multinomial Logistic Regression and Logistic Regression, 2014.
11. Luce R D. Individual Choice Behavior. A theoretical Analysis, Yew York: John Wiley and Sons, 1959, 139-141.
12. Mc Fadden D. Conditional Logit Analysis of Qualitative Choice Behavior. In: Zarembka P, ed Frontiers in Econometrics. New York: Academic press, 1974, 105-142.
13. Millington *et al*. Regression techniques for examine Land use/ cover change: case study of a Mediterranean Landscape. Eco system (N.Y),2007:10(4):562-578.
14. Schafer JL. Multinomial Logistic Regression Model, Stat-544 lecture 19, 2006.
15. Seber GAF. Linear Regression Analysis, John Wiley and Sons, New York, 1977, 75.