American Journal of Applied Mathematics and Statistics, 2018, Vol. 6, No. 6, 262-265
Available online at http://pubs.sciepub.com/ajams/6/6/7

DOI:10.12691/ajams-6-6-7

Statistical Modelling of Categorical Outcome with More than Two Nominal Categories

Fatma D.M. Abdallah*

Department of Animal Wealth Development, Faculty of Veterinary Medicine, Zagazig University, Egypt
*Corresponding author: Nour_stat2013@yahoo.com

Received August 13, 2018; Revised October 04, 2018; Accepted December 04, 2018

Abstract This paper aims to explain and apply an important statistical method used for modelling categorical outcome variable with at least two unordered categories. Logistic regression model especially multinomial logistic type (MNL) model is the best choice to model unordered qualitative data. A simulation study was done to examine the efficiency of the model in representing categorical response variable. Three explanatory variables (age, species, and sex) are used for discrimination. While the outcome variable was Rose Bengal Plate Test (RBPT) results which has four outcome categories (negative, positive, false positive, and false negative). Therefore, logit model will be utilized to model this data. MNL models were fitted using SPSS packages and parameters estimated depending on maximum likelihood (MLE) by the Newton-Raphson algorithm. This model depends mainly on two estimates to interpret the results, they are the regression coefficient and the exponentiated coefficients which known as the odds ratio. This model was a good fitted for description the data of 500 values of Rose Bengal Plate Test results of Brucella in sheep and goat species. The results showed fitting of the model to the data with highly significant likelihood ratio statistic for the overall model (P value = 0.000**). Wald test was significant for all variables in positive category and this indicated that age, species and sex are good predictors for test results. The odds ratio in case of positive category for age, species and sex was 1.589, 0.214 and 0.133 respectively.

Keywords: *multinomial logistic regression, odds ratio, Rose Bengal Plate Test (RBPT), maximum likelihood and pseudo R^2*

Cite This Article: Fatma D.M. Abdallah, "Statistical Modelling of Categorical Outcome with More than Two Nominal Categories." *American Journal of Applied Mathematics and Statistics*, vol. 6, no. 6 (2018): 262-265. doi: 10.12691/ajams-6-6-7.

1. Introduction

Different statistical techniques for modelling of categorical data have increased in last years. Many traditional methods were used for this purpose such as Chi square and log-linear models. It is believed that the logistic models are convenient for representing and predicting categorical data. Multinomial regression logistic models are types of logistic regression models and considered an extension of binary logistic regression methods that can be used in this area of statistics.

In the health sciences, the model is known as polychotomous logistic regression and in econometrics as the discrete choice model [1].

It is a non-linear S-shaped distribution function, and can be utilized in many implementations [2]. So the estimated probabilities of the logit distribution are between 0.0 and 1.0. The variation between the logistic function and the linear regression function is that, the values of the first one falls between zero and one, whereas the second one values are absolute.

Multinomial regression models are used to analyze data where the outcome variable is categorical with at least two unordered categories while the predictor variables could be continuous, categorical variables, or both [1].

These models do not require the known assumptions of any model such as normality, linearity and the variances homogeneity for continuous data with exception of multicollinearity [3], so these models can be used in many fields with major advantage.

Fitting the model to the data depending on two types of measures: predictive power measures such as pseudo R^2 and goodness of fit tests such as deviance and Pearson chi-square. It's common for models with high R^2 to give acceptable goodness of fit tests. But, models with very small R^2, may fit the observed data very well.

In order to measure the action of any explanatory variable on the outcome one, the test statistic is the likelihood ratio (LR). If this probability ratio is significant for the overall model, this means that the explanatory variables have participated in the forecasting of the resultant one. If the outcome variable has K groups, then the coefficients (β_n) associated with each explanatory variable X_n are $K - 1$ [4].

The beta coefficient statistic and the odd ratios (the exponentiated coefficients) can explain the results of the model under study. Positive beta coefficients, increases in predictor values leads to an increase of probability in the higher-numbered response categories [5,6,7].

2. Materials and Methods

A simulation study was done to explore the model efficiency in predicting categorical nominal response variables.

The dependent or the outcome variable (RBPT results) contain four groups (negative, positive, false positive and false negative).It is unordered categories therefore the model which will be applied is the multinomial logistic regression (MNL). It is important to select what is called reference category (negative category is the reference one). Then the model compares all the groups of the outcome variable with this reference one. This category considered as the comparison point for all analyses [5].

RBPT is a serological test where the samples of serum and RBPT antigen were thoroughly mixed and reading was done within 4 minutes to detect the results.

The independent or explanatory variables are: age which is considered quantitative and species (sheep and goat), and sex. Species and sex are considered categorical.

2.1. The Model Assumptions

- The independent variable may either be quantitative or qualitative.
- The outcome variable is qualitative which divided into different unordered categories.
- The data do not require to have a normal distribution, no linear relationship and no equality of variances.

2.2. The Mathematical Model

The logistic function according to [2] is

$$\log it(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon \qquad (1)$$

where,

p: the chance of choosing a category of RBPT results,

X_1,…, X_k: the predictor variables k,

β_0: the intercept,

β_1, …, β_k: the parameter estimates on the explanatory variables X_k.

ε: error terms.

The model parameters produced depending on estimation of maximum likelihood (MLE) of the model. To get the maximum likelihood estimate of β, a Newton-Raphson iterative estimation method is used. SPSS packages version 20 [8] is used to execute the analysis process.

2.3. -2 log likelihood

The -2 log likelihood has a chi-square distribution, which applied to show whether the test result is significant.

$$LR = \left[-2LL\left(\text{of full model}\right)\right] - \left[-2LL\left(\text{of restricted model}\right)\right]. \qquad (2)$$

The LR statistic of the full model includes all of the explanatory factors and producing LR_F. The LR statistic of the reduced model excluded the explanatory factors x_n and producing LR_R.

2.4. R square Measures or Pseudo R^2

In the ordinary linear regression there is no pseudo R^2. The goodness of fit measure (R^2) is as follows:

$$r^2 = \frac{b^2\left(\Sigma x^2 - \left(\Sigma x\right)^2 / n\right)}{\Sigma y^2 - \left(\Sigma y\right)^2 / n} = \frac{SS_{exptained}}{SS_{total}} \qquad (3)$$

Where b is the regression coefficient, x is the independent variable and y is the dependent one.

In this model, several pseudo R^2 have been developed due to the lack of statistic equivalent to R^2 .They are called “pseudo” R-squares as their values, lying between 0 to 1, so they are similar to R^2 [9].

There are some of the common pseudo R^2:

1. McFadden’s R^2, which is shown as in the following equation [10].

$$R^2_{McF} = 1 - [\ln\left(L_M\right) / \ln\left(L_0\right) \qquad (4)$$

where L_0 means the likelihood function in a model without predictor variables, and L_M means the likelihood for the estimated model. $L_n(L_0)$ is similar to the residual sum of squares in linear regression model.

2. The Cox and Snell R^2 which is as in this equation [11].

$$R^2_{C\&S} = 1 - \left(L_0 / L_M\right)^{2/n} \qquad (5)$$

where n is the number of observations. This measure shows the efficiency of the full model over the intercept model.

3. The Nagelkerke R^2 is defined as [12].

$$R^2_{NK} = (1 - \left(L_0 / L_M\right)^{2/n}) / (1 - \left(L_0\right)^{2/n}). \qquad (6)$$

These tests are a part of goodness of fit statistics. Goodness of fit its meaning is that the model convenient for the data. Any model fits the data poorly if its residual variation is large [13].

2.5. Wald Test

It is a statistical measure to test if each variable in the model is significant or not.

2.6. Odds Ratio

The odds ratio is a method of comparison between two categories to show whether the probability of a certain event is similar between them [2] and [5]. It is used as a statistic in the logistic regression instead of regression coefficients of the linear one to explain the prediction results.

It is the exponentiation of the multinomial logit coefficients ($e^{(\beta)}$). The number of multinomial logistic regression models are (M − 1), where M is the number of groups of the outcome variable, and with consideration of the referent group.

β_1 is the probability of transformation in the reference category versus the transformation in the comparison category as the explanatory variable changes.

3. Results and Discussion

The result of -2 log likelihood chi-square of overall model was 325.318 showed P value = 0.000 meaning that the difference was highly significant, so the covariates have a significant effect on the RBPT results as shown in Table 1.

Table 1. Model fitting information

Model	Model fitting criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	Degrees of freedom	P-value
Intercept only	385.095			
Final	325.318	59.777	9	0.000

The final row showed that the added variables develop the model in comparison with to the intercept only (with no added variables). The significance column (p = 0.000) denotes that the full model more efficient in predicting the outcome variable than the intercept-only model alone.

Thus the null hypothesis which stated that the difference is absent between the model without explanatory variables and the model with explanatory variables was not accepted and this indicated that the relationship between the explanatory variables and the outcome variable is present, hence accepting the alternative hypothesis.

There are some pseudo R^2 measures in MNL model, as presented in Table 2.

Table 2. Pseudo R^2 measures as a predictive measures

Cox and Snell	0.113
Nagelkerke	0.170
McFadden	0.109

Cox & Snell R^2, the Nagelkerke R^2 and McFadden R^2 values indicate the amount of variation in the outcome variable and described as pseudo R^2. Their values are 0.113, 0.170 and 0.109 respectively, suggesting that 11.3%, 17 % and 10.9% of the variability is explained by this set of variables included in the model.

The relationship between predictor and outcome variables depends mainly on two types of tests. The first is likelihood ratio test which determines the overall relationship between an explanatory variable and outcomes. The second is the Wald test which determines whether or not the explanatory variable is statistically significant in differentiating between two categories in each of binary logistic comparisons. If an explanatory variable has an overall relationship to the outcome variable, it does not necessarily suggest statistical significance. In fact, it may or may not be significant in differentiating between pairs of categories defined by the outcome variable.

Table 3 shows which of the predictors are statistically significant. Age, species and sex were significant because their p values are (0.033, 0.000 and 0.000) respectively.

Table 3. Likelihood ratio tests

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	Degrees of freedom	P-value
Intercept	325.318a	0.000	0	
Age	334.074	8.756	3	0.033
Species	344.196	18.878	3	0.000
Sex	354.284	28.966	3	0.000

Table 4 shows the parameter estimates (the model coefficients). There is a regression coefficient for each explanatory variable for each RBPT outcome variable category. There were four categories of the outcome, so there are three sets of logistic regression model coefficients. The first group of coefficients are belonging to the "positive" cell (which showing the comparison of the positive result group to the reference group, (negative)). The second group of coefficients are belonging to the "false positive" row (which showing the comparison of the false positive result group to the reference (negative) group. The third group of coefficients are belonging to the "false negative" row (which showing the comparison of the false negative result group to the reference (negative) group).

It is noticed that age was statistically significant for positive category as its p value was 0.022 and it was non-significant in case of false positive and false negative categories as the p value was 0.815 and 0.153 respectively.

Species was significant for positive and false positive categories as the p value was 0.008 and 0.009 respectively. It was not significant in case of false negative category as p value was 0.338.

Sex was significant for positive and false negative categories as the p value was 0.000 and 0.001 respectively. It was not significant in case of false positive category as p value was 0.454.

As mentioned before that the coefficients are exponentiated "Exp (β)" to interpret of the results.

Higher odd ratios than 1.0, mean that the predictor variables are positively related with the resultant one and lower odd ratios than 1.0, then the predictor variables are negatively related with the resultant one.

In case of age, the odd ratio is larger than 1.0 (1.589), indicating the positive contributing of this predictor to the RBPT results at positive category and its effect could increase by a value of 1.589 while the other factors are constant.

For species, the odd ratio is less than 1.0 (0.214), which indicates the negative contributing of this predictor to the RBPT results at positive level, and its effect could be predicted to decline by a value of 0.214 with fixation of the rest factors.

For sex, the odd ratio is lower than 1.0 (0.133), which indicates the negative contributing of this predictor to the RBPT results at this level (positive), and its effect could be predicted to decrease by a value of 0.133 while the rest variables in the model are fixed as shown in Table 4.

Table 4. Summary of the final model of risk factors for RBPT results of Brucella disease in a sample of 500 animal (sheep and goat) (the parameter estimates table)

RBPT results		β	SE	Wald test	P-value	Exp(β)	95% CI for Exp(β)	
Positive	Intercept	-3.110	0.776	16.071	0.000			
	Age	0.463	0.202	5.259	0.022	1.589	1.070	2.360
	[Species = goat]	-1.542	0.583	6.989	0.008	0.214	0.068	0.671
	[Species = sheep]	0						
	[Sex = f]	-2.019	0.572	12.450	0.000	0.133	0.043	0.408
	[Sex = m]	0						
False positive	Intercept	-3.463	0.657	27.774	0.000			
	Age	-0.036	0.156	0.055	0.815	0.964	0.711	1.308
	[Species = goat]	1.356	0.515	6.925	0.009	3.882	1.414	10.661
	[Species = sheep]	0						
	[Sex = f]	-0.310	0.414	0.562	0.454	0.733	0.326	1.650
	[Sex = m]	0						
False negative	Intercept	-1.533	0.579	7.013	0.008			
	Age	-0.252	0.176	2.046	0.153	0.777	0.550	1.098
	[Species = goat]	-0.465	0.485	0.919	0.338	0.628	0.243	1.625
	[Species = sheep]	0						
	[Sex = f]	-1.885	0.582	10.508	0.001	0.152	0.049	0.475
	[Sex = m]	0						

In case of age the odd ratio is less than 1.0 (0.964), which indicates the negative contributing of this predictor to the RBPT result at false positive level. Its effect could be predicted to decline by a value of 0.964 while the other factors are held constant.

When checking species variable in the RBPT results of the level of false positive category, the odd ratio is higher than 1.0 (3.882), which indicates the positive contributing of this predictor to the RBPT results at this level, and its effect could be predicted to increase by a value of 3.882 with fixation the rest variables in the model.

For sex, the odd ratio is lower than 1.0 (0.733), which indicates the negative contributing of this predictor to the RBPT results at this level (false positive), and its effect would be predicted to decline by a factor of 0.733 with fixation the other factors of the model.

In case of age the odd ratio is less than 1.0 (0.777), which indicates the negative contributing of this predictor to the RBPT result at false negative level. Its effect could be expected to decline by a factor of 0.777 and the other factors are held constant.

When checking species predictor in the RBPT results of the level of false negative result, the odd ratio is less than 1.0 (0.628), which indicates the negative contributing of this predictor to the RBPT results at this level, its effect would be predicted to decrease by a factor of 0.628.

For sex predictor in the RBPT results of the level of false negative result, the odd ratio is smaller than 1.0 (0.152), which indicates the negative contributing of this predictor to the RBPT results at this level (false negative) as in Table 4.

4. Conclusion

This work used multinomial logistic regression (MNL) (logit type) to detect the associations between the Rose Bengal Plate Test results which have four categories and the independent variables. The categories of the resultant variable (negative, positive, false positive and false negative) are considered nominal (cannot be ordered). This study examined using of three risk factors in RBPT results modeling. The statistic which used here was the odds ratio as a powerful measure for interpreting the prediction results instead of regression coefficients. The pseudo R^2 is used as a powerful goodness of fits statistic instead of known measures of fitness such as the deviance measure.

References

[1] Hosmer, D.W. and Lemeshow, S, *Applied logistic regression*, Wiley-Interscience, US, 2000.

[2] Judge, G., Griffiths, W.E., Hill, R.C., Lutkepohl, H. and Lee, T.C, *The theory and practice of econometrics*, 2nd Edition, Wiley, New York, 1985.

[3] Tabachnick, B.G. and Fidell, L.S. and Osterlind, S.J, *Using multivariate statistics*, Allyn and Bacon Boston, US, 2001.

[4] Abdulhafedh A, "Incorporating the Multinomial Logistic Regression in Vehicle Crash Severity Modeling: A Detailed Overview," *Journal of Transportation Technologies*, 7:279-303. 2017.

[5] Greene, W, *Econometric analysis*, 7th Edition, Prentice Hall, Upper Saddle River, 2012.

[6] Baltagi, B.H, Econometrics, 5th Edition, Springer, Berlin, 2011.

[7] Kleinbaum, D.G. and Klein, M, *Logistic Regression: A Self-Learning Text*, 3rd Edition, Springer, New York, 2010.

[8] SPSS, "Statistical Package for Social Sciences," Release 20.0 versions. SPSS Inc. USA, 2006.

[9] Menard, S, *Applied Logistic Regression Analysis*, Sage Publications, Thousand Oaks, 2002.

[10] McFadden, D, *Conditional logit analysis of qualitative choice behavior*. Frontiers in econometrics, 1974.

[11] Cox, D.R. and Snell, E.J, *Analysis of Binary Data*, Chapman & Hall, London, 1989.

[12] Nagelkerke, N.J.D, "A Note on a General Definition of the Coefficient of Determination," *Biometrika*, 78: 691-692. 1991.

[13] Hosmer, D.W., hosmer, T., le Cessie, S., and Lemeshow, S, "A comparison of goodness-of-fit tests for logistic regression model. Statistics in medicine," 16 (9). 965-80. 1997.