Available online at sjuoz.uoz.edu.krd

journals.uoz.edu.krd

Science Journal of University of Zakho

Vol. 5, No. 4, pp. 330 –334, Dec.-2017

p-ISSN: 2410-7549
e-ISSN: 2414-6943

APPLYING THE BINARY LOGISTIC REGRESSION ANALYSIS ON THE MEDICAL DATA

Qais M. Abdulqader

Technical College of Petroleum and Mineral Sciences, Duhok Polytechnic University, Kurdistan Region, Iraq.
(qais.mustafa@dpu.edu.krd)

Received: Sep. 2017 / Accepted: Dec., 2017 / Published: Dec., 2017 *https://doi.org/10.25271/2017.5.4.388*

ABSTRACT:
In this paper, the Binary Logistic Regression Analysis BLRA technique has been used and applied for building the best model for Hepatitis disease data using best subsets regression and stepwise procedures and depending on some laboratory tests such as glutamate oxalate transaminase, glutamate pyruvate transaminase, alkaline phosphatase, and total serum bilirubin which represents explanatory variables. Also, the technique has used for classifying persons into two groups which are infected and non-infected with viral Hepatitis disease. A random sample size consists of 200 persons has been selected which represents 86 of uninfected and 114 of infected persons. The results of the analysis showed that first, the two procedures identified the same three explanatory variables out of four and they were statistically significant, and it has been reliable in building the logistic model. And second, the percentage of visible correct classification rate was about 98% which represents the high ability of the model for classification.

KEYWORDS: Logistic regression, Hosmer–Lemeshow test, Likelihood ratio test, Maximum likelihood estimation, Wald test.

1. INTRODUCTION

Binary Logistic Regression Analysis BLRA analyzes the relationship between multiple explanatory variables and a single binary response variable, a categorical variable with two categories, (Sweet and Martin, 2011). Many medical applications have been done in this area. (Lihui et al., 2001) compared both linear regression and a logistic regression model for biological percentage data using different methods for comparison. (Sarkar et al., 2010) have used logistic regression method for model selection. The study aimed to increase the power of prediction while reducing the number of covariates. The procedure was depending on the stepwise method and best subsets regression through applying on health survey data. (Javali and Pandit, 2012) used a model depending on multiple logistic regression to make risk factors prognostication of oral health infirmities. (Reeda and Wub, 2013) used the logistic regression method and applied for building models of the risk factor in the stammer studies. (Mythili et al., 2013) suggested a formula based model compare the accuracies of applying formulas to the separate outcomes of support vector apparatus, judgment trees, and LRA on the database of Cleveland Heart Disease to obtain a reliable model of heart disease prediction. (Amir et al., 2014) studied the relation of hypertension with risk factors affecting significantly the execution of Hypertension using logistic regression technique. (Qais, 2015) used LRA and discriminant analysis DA and applied on natural and Caesarean births data to show the performance of such techniques and the capability in classification the type of birth. (Vaitheeswaran et al., 2016) examined the importance of keeping the possession of the ordinal nature of the outcome variable while marking the risk factors related to diabetic problems related to loss of vision using traditional and Bayesian approaches of ordinal logistic regression models.(Junguk et al., 2017) used multinomial logistic regression analysis MLRA to examine the effect of an estrogen on the rate of reverse pregnancy results. The purpose of this paper is to find a best BLRA model for fitting line and for obtaining the best classification and predicting the group membership. The remainder of this paper is structured as follows: section 2, explains BLRA and methodology. Section 3 presents the application on real data and finally, in section4 conclusions are presented.

2. BLRA

Regression analysis presents the association between a response variable and one or more explanatory variables. It is often the situation that the outcome variable is discrete, assuming two or more potential values (Hosmer & Lemeshow, 2000). BLRA represents a special condition of linear regression analysis LRA used when the response is binary not continuous, and the explanatory variables are quantitative or qualitative variables (Hair et al., 2010). It was first suggested in the 1970s to overcome difficulties of ordinary least squares OLS regression in treating binary outcomes (Peng et al., 2002). Logistic regression LR uses the theory of binomial probability which represents having only two values to predict: that probability (p) is 1 instead of 0, i.e. the event belongs to one group instead of the other. LR presents the best fitting function depending on the maximum likelihood ML approach, which maximizes the distinguishing probability of the observed data into the suitable category given the coefficients of regression (Burns & Burns 2008).

2.1 Assumptions of BLRA

Logistic regression ignores a linear relationship between the response and explanatory variables. It is also supposed that the response variable must be a binary and the explanatory variables need not be an interval, the distribution is normal, the relationship is linear, nor of equality of variance within each group. Furthermore, the groups must be mutually exclusive and detailed; a case can only be in one group and every case must be a member of one of the groups. Finally, the sample size must be large than for LR because ML coefficients are large sample estimates. A minimum of 50 cases for each explanatory variable is needed (Hair et al., 2010), (Burns & Burns 2008), (Kleinbaum & Klein, 2010).

2.2 The Logistic Model

To explore the implied association between a response variable and one or more explanatory variables, the LRA is suitable for study. By taking the case of one explanatory variable X with one binary outcome variable Y, the logistic model predicts the logit of Y from X which represents a natural logarithm of odds of Y. The simple formula can be written as the following (Peng et al., 2002), (James et al., 2013):

$$\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x. \tag{1}$$

The left-hand side is called the log-odds or logit. The LR model has a logit that is linear in X. Hence:

$$\pi(x) = E(Y|X) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}\,, \tag{2}$$

Where π is the probability of the outcome of interest given that X=x, α is a parameter which represents the Y-intercept, and β is a parameter of the slope, X can be qualitative (categorical) or quantitative variable, and Y is always qualitative or categorical. The formula (1) can be expressed and extended from simple to multiple linear regression as follows:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_{k.} \tag{3}$$

Therefore,

$$\pi(x) = \frac{e^{\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}}{1+e^{\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}}\,, \tag{4}$$

Where π is the event probability, α is the Y-intercept, β_s are parameters of the slope, and X`s are combinations of explanatory variables. α and β_s are estimated by the maximum likelihood estimator MLE approach.

2.3 Goodness of Fit Test

It is also called Hosmer-Lemeshow test which represents a X^2 (Chi-square) test used for testing the adequacy of the model for fitting the data. The null hypothesis is that the model is adequate to fit the data and we will only reject this null hypothesis if there are sufficiently strong grounds to do so (traditionally if the p-value is less than 0.05). See (Hosmer & Lemeshow, 2000) for details.

2.4 Likelihood Ratio Test (LRT)

The test depends on –2log likelihood ratio. We use this test for checking the significance of the difference between the likelihood ratio for the reduced model with explanatory variables and the likelihood ratio for current model with only a constant in it. Significance at 0.05 level or less means the reduced model with the explanatory variables is significantly different from the one with the constant only (all ‘b’ coefficients being zero). It measures the enhancement in a fit that the explanatory variables make compared to the null model. Chi-square is used to evaluate the significance of this ratio. When probability unable to reach the 0.05 significance level, we do not reject the null hypothesis that knowing the explanatory variables has no more effects in predicting the response variable. See (Burns & Burns, 2008), (Bewick et al., 2005), (Bergerud, 1996). for details.

2.5 Measures of Goodness of Fit Test

In linear regression method and depending on OLS, we use the coefficient of determination R^2 as a measurement of goodness of fit, which represents the variation ratio which explained by the model. Using logistic regression, a similar statistic does not exist, and therefore several pseudo-R^2 statistics have been developed. In this paper, we will depend on three pseudo R^2 values: Cox and Snell R^2, Nagelkerke R^2, and McFadden R^2. See (Nagelkerke, 1991), (StatSoft, 2013) for details.

2.6 Statistical Significant Test

In linear regression, we want to know how the model overall fits the data but also to determine the contributions of the explanatory variables. In logistic regression, we use another tool called Wald statistic, which is similar to the t-test performed on the coefficients of regression in a linear regression to test whether the variable has a real contribution to the prediction of the outcome, specifically whether the coefficient of explanatory variables is significantly different from zero. To evaluate the fit of a logistic regression model, we use the area under the curve which ranges from 0.5 and 1.0 with larger values indicative of better fit. (Kleinbaum & Klein, 2010)

2.7 The Classification Table

A good way to summarize the results of a fitted logistic regression model is via classification table which represents the result of cross-classification of the response variable Y and a binary variable whose values are gained from the probabilities of estimated logistic (Hosmer & Lemeshow, 2000). The reclassification table shows the accuracy of the model. It shows the frequencies of the predicted and observed classification of cases and percentage of correct predictions depend on the logistic regression model. When the predicted probability is greater than 0.5 an observation is predicted as 1 else it is predicted as 0. (StatSoft, 2013). Several criteria are existing to evaluate a set of classification rule and one of the simplest criteria is misclassification rate (Abdullah & Majid, 2014). For two groups, among the n_1 observations in G_1, n_{11} are classified correctly into G_1, and n_{12} are classified incorrectly into G_2, where $n_1 = n_{11} + n_{12}$. Similarly, of the n_2 observations in G_2, n_{21} are classified incorrectly into G_1, and n_{22} are classified correctly into G_2, where $n_2 = n_{21} + n_{22}$. Thus, the visible error rate VER can be presented as (Rencher, 2002):

$$VER = \frac{n_{12}+n_{21}}{n_1+n_2} = \frac{n_{12}+n_{21}}{n_{11}+n_{12}+n_{21}+n_{22}}\,. \tag{5}$$

Similarly, we can present overall visible correct classification rate VCCR as:

$$VCCR = \frac{n_{11}+n_{22}}{n_1+n_2}\,. \tag{6}$$

2.8 Methodology

To obtain the best binary logistic regression model we first try to get a combination of models using best subset regression depending on Akaike Information Criterion AIC which represents a way of choosing a model from a combination of models. This statistic seeks for a model that has a goodness-of-fit with few parameters (Lawless &Singhal, 1987). It is defined as:

$$AIC = -2(\ln(\text{likelihood})) + 2K \tag{7}$$

where likelihood is the probability of the data given a model and K is the number of free parameters in the model. The procedure is to select best of the best subsets regression models with minimum AIC value. Second, we apply the binary logistic regression analysis on the data set using a stepwise procedure which works as a judge for an order of importance of the explanatory variable. Finally, represents obtaining the best logistic model. The methodology can be presented in figure1:

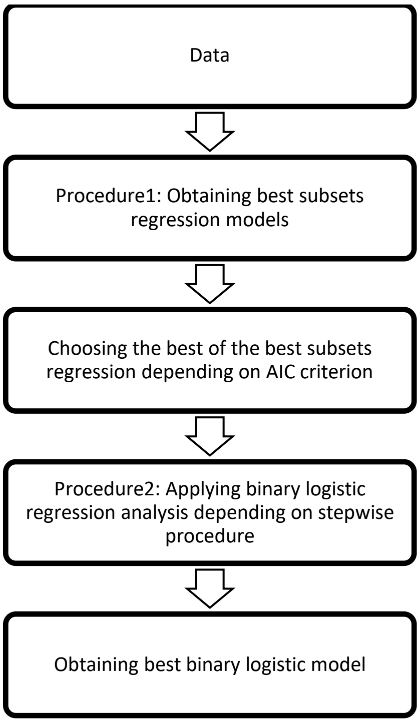
Data
Procedure1: Obtaining best subsets regression models
Choosing the best of the best subsets regression depending on AIC criterion
Procedure2: Applying binary logistic regression analysis depending on stepwise procedure
Obtaining best binary logistic model

Figure 1: Methodology of obtaining the best logistic model

3. APPLICATION ON REAL DATA

The application concerning the BLRA was performed using the Hepatitis data and working with SPSS statistical package program. The data were obtained from (Iehab & Sahar, 2013). The classification task consists of predicting whether a person would test positive for Hepatitis. The person's data were labeled, such that we put 1 for infected persons and 0 for non- infected persons or healthy persons. There are four laboratory tests which represent explanatory variables for 200 persons, and among them, 114 persons tested positive for infected. Table 1 presents the details about the frequency and percentage distribution of the groups.

Table 1. Distribution of the class of Hepatitis data

Class name	Class Size	Class Distribution
Infected persons or positive	114	57%
Non-infected persons or negative	86	43%

As presented in table1, the sample size of the data is 200 observations and the data set were classified into two groups such that, the first group with n_1=114 which represents the 57% of observations, and the second group with n_2=86 which represents the 43% 0f observations. Table 2 presents the information on statistical analysis which shows the mean and the standard deviation of the explanatory variables.

Table 2. Information of statistical analysis

Explanatory variables	Mean	Standard deviation
X_1	49.595	49.146
X_2	122.265	135.077
X_3	202.560	132.768
X_4	56.401	77.753

The response variable name is “Kind” which is the kind of test. The explanatory variables as shown in table2 are presented as the following: X1 for glutamate oxalate transaminase, X2 for glutamate pyruvate transaminase, X3 for alkaline phosphatase, and X4 for total serum bilirubin. We dealt with these explanatory variables as scale measure.

3.1 Getting Best Subsets Regression Models

Here we use the procedure representing best subsets of model selection using automatic selection which compares all possible models using a specified set of explanatory variables and displays the best-fitting models that contain one predictor or more. Table 3 shows the four best models selected depending on AIC criterion:

Table 3. Best models identified depending on AIC criterion

Model	Model covariates	AIC
1	X_1, X_2	44.40
2	X_2, X_4	39.85
3	X_1, X_2, X_4	**35.38**
4	X_1, X_2, X_3, X_4	36.87

From table3, four best subsets models specified and among them, third model has a minimum value of AIC which contains the three explanatory variables respectively: glutamate oxalate transaminase (X_1), glutamate pyruvate transaminase (X_2), and total serum bilirubin (X_4). The specified model represents the best of the best subsets model under study.

3.2 Applying BLRA on Real Data Using Stepwise Procedure

BLRA was applied to the Hepatitis data set to study the relationship between the response variable and combination of explanatory variables to find the most important predictors that discriminate the kind of test. Table4 gives the information of model fitting showing the statistical significance of the final x^2.

Table 4. Information of the Model fitting

Model	Model fitting criteria	Likelihood ratio tests		
	-2likelihood	x^2	D.F.	Sig.
Intercept only	273.326	245.949	3	0.000
Final	27.377			

It is clear from table 4 that when including only the intercept, the value of -2log likelihood of basic model was (273.326), and this value has decreased to (27.377) with the existence the set of explanatory variables in the model. The value of the $x^2$was (245.949) comparing with the probability (0.000) concluding that the model is significant and for this cause, we refuse the null hypothesis and take the alternative hypothesis which says that there is an essential relation between the explanatory variables and the response variable.

To evaluate the overall correlation between explanatory variables and the response variable we use the LRT for the coefficients of the logistic model. By using the stepwise procedure, three most important explanatory variables were selected respectively (X_4, X_2, X_1) and the explanatory variable X_3 were removed from the analysis because its contribution into the model was not significant for making discrimination the kind of test. The result of the tests presented in table 5.

Table 5. The result of likelihood ratio tests

Effect	Model fitting criteria	Likelihood ratio tests		
	-2likelihood of the reduced model	x^2	D.F.	Sig.
Intercept	156.966	129.589	1	0.000
X_4	38.399	11.022	1	0.001
X_2	58.187	30.810	1	0.000
X_1	33.850	6.473	1	0.011

The x^2 of Hosmer–Lemeshow test was 1.452 with P-value = 0.984 indicates that the numbers of infected persons of Hepatitis disease are not significantly different from those predicted by the model and that the overall model fit is good. Table 6 presents the three pseudo R^2 values.

Table 6. The pseudo R^2

Name of R^2	Value
Cox and Snell	0.708
Nagelkerke	0.950
McFadden	0.900

From table 6, the response variable defines 70.8% of the variance in explanatory variables according to Cox & Snell R^2 value, 95% according to Nagelkerke R^2 value which is the modified form of Cox & Snell coefficient, and 90% according to McFadden R^2.

3.2.1 Estimating the Logistic Regression Coefficients: The estimation of parameters of the logistic regression by using the MLE method depending on Wald statistic for the final model is showed in table 7 which gives the results of fitting the logistic regression model to Hepatitis data and showing coefficients, which are used in the equation for making the classifications in BLRA, much as. The constant term gathering with the sum of products of the coefficients with the observations gives the discriminant scores.

Table 7. Results of fitting the logistic regression model to Hepatitis data

Variable	β	Std. Error	Wald	D.F.	Sig.	Exp. (β)
Intercept	6.738	1.458	21.372	1	0.000	
X4	-0.046	0.018	6.589	1	0.010	0.955
X2	-0.128	0.044	8.541	1	0.003	0.880
X1	-0.050	.0220	5.452	1	0.020	0.951

From the above table, the estimated BLRA model can be presented by the following formula:

$$\ln\left(\frac{\pi}{1-\pi}\right) = 6.738 - 0.05(X_1) - 0.128(X_2) - 0.046(X_4) \qquad (8)$$

3.2.2 Evaluation of a Logistic Regression Model: For making an evaluation of the logistic model, we use the area under the curve as we mentioned in theoretical part. Table 8 shows the values of the area under the curve of the explanatory variables.

Table8.The values of the area under the curve of the explanatory variables

Test Result Variable (s)	Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
X1	0.984	0.011	0.000	0.963	1.000
X2	0.997	0.003	0.000	0.991	1.000
X4	0.990	0.009	0.000	0.972	1.000

From table 8, the area under the curve of the three explanatory variables are: 0.984, 0.997, and 0.990 respectively with 95% confidence interval (0.963, 1.000), (0.991,1.000) and (0.972,1.000) and they are all significant because p-value is equal to 0.000 for the three explanatory variables indicating that the logistic regression can classify the group significantly better than by chance.

3.2.3 Testing the Power of classification of the Logistic Model: To show the accuracy of the model, the two measures VER and VCCR as we referred in section 2.7 are performed here to evaluate the efficiency of the BLR model of the estimated function. Table 8 shows the final results of classification.

Table 9. Final classification results using binary logistic regression model

Kind of test	Non-infected	Infected	Total
Non-infected	84	2	86
Infected	2	112	114
VER		2%	
VCCR		98%	

From table 9, we can see that 84 of 86 persons from the first group were classified correctly, and 112 of 114 persons from the second group were classified correctly, we conclude that the LRA was able to classify 196 cases of persons out of 200 cases correctly. The VER was 2% and the VCCR was 98% indicating that the model has the high ability on classification.

4. CONCLUSION

In this paper, the BLRA has been applied on real data and two procedures have been used. The first procedure based on best subsets regression and depending on AIC criterion while the second procedure was depending on stepwise technique. The LRT has been performed for modeling, classifying, and selecting the most important explanatory variables. The accuracy of the model was depending on two statistical criteria: VER and VCCR. The results of the analysis showed that the performance of the BLRA gave the high ability of classification (VCCR =98%). In addition, the analysis showed that the two procedures have selected the same model consisting of three explanatory variables; X_4, X_2, and X_1 which represents the three tests respectively (total serum bilirubin, glutamate pyruvate transaminase, and glutamate oxalate transaminase) have contributed significantly to discriminate the kind of test and also both procedures excluded the remaining predictor; X_3 which represents alkaline phosphatase and cancelled from the analysis because it was unable to give a positive contribution when making discrimination. The best model obtained by using BLRA through looking at the value of the parameters of the logistic model and its signs in equation (8) it is observed that there is a negative correlation between the type of test and the specified explanatory variables.

REFERENCES

Abdullah E., &, Majid E. (2014). *A comparative study between Linear discriminant analysis and multinomial logistic regression*. An - Najah University Journal Research (Humanities), 28: 1525-1548.

Amir W, Mamat M, & Ali Z.(2014). *Association of hypertension with risk factors using logistic regression*. Applied Mathematical Sciences, 8, 2563 – 2572.

Bergerud W. (1996). *Introduction to logistic regression models: With worked forestry examples*. Biometrics Information Handbook, British Columbia, 1996.

Bewick V., Cheek L., & Ball J. (2005). *Statistics review 14: Logistic regression*. Critical Care, 9: 112-118.

Burns RB., & Burns RA. (2008). *Business research methods and statistics using SPSS*: Sage Publishing LTD.

Hair J., Black W., Babin B., & Anderson R. (2010). *Multivariate data analysis*, Seventh Edition.: Pearson Prentice Hall.

Hosmer D., & Lemeshow S. (2000). *Applied logistic regression*. Second Edition: John Wiley and Sons, Inc.

Iehab A. M, & Sahar H. A.(2013). *Choosing the best formula for the multiple linear regression model*. Magazine of College Administration and Economics for Economic, Babylon University, 242: 119-146.

James G., Witten D., Hastie T., & Tibshirani R. (2013). *An Introduction to statistical learning: with applications in R*: Springer.

Javali S., & Pandit P. (2012). *Multiple logistic regression model to predict risk factors of oral health diseases*. Romanian Statistical Review Journal, 73-86.

Junguk H., Eun-Hee C., Kwang-Hyun B., & Kyung J. L. (2017). *Prediction of gestational diabetes mellitus by unconjugated* *estriol levels in maternal serum*. International Journal of Medical Sciences. 14, 123-127.

Kleinbaum D., & Klein M. (2010). *Logistic regression: A self-learning text*. Third Edition.: Springer.

Lawless J. F., & Singhal K. (1987). *ISMOD: An all- subsets regression program for generalized linear models, I. Statistical and computational background*. Computer methods and programs in biomedicine, 24, 117-124.

Mythili T., Dev Mukherji M, Padalia N, & Naidu A. *A heart disease prediction model using SVM-decision trees-logistic regression (SDL)*. International Journal of Computer Applications, 68,11-14.

Nagelkerke N. (1991). *A note on a general definition of the coefficient of determination*. Biometrika, 78: 691-692.

Peng C., Lee K., & Ingersoll G. (2002). *An Introduction to logistic regression analysis and reporting*. The Journal of Educational Research, 96: 3-15.

Qais M. (2015). *Comparison of discriminant analysis and logistic regression analysis: An application on caesarean births and natural births data*. Journal of The Institute of Natural and Applied Sciences, 20, 34-46.

Reeda P., & Wub Y. (2013). *Logistic regression for risk factor modelling in stuttering research*. Journal of Fluency Disorders, 38, 88-101.

Rencher A. (2002). *Methods of multivariate analysis*. Second Edition: A John Wiley and Sons, Inc. Publication.

Sarkar S. K., Midi H., & Rana S. (2010). *Model selection in logistic regression and performance of its predictive ability*. Australian Journal of Basic and Applied Science, 12, 5813-5822.

StatSoft. STATISTICA.(2013). *Formula guide: Logistic Regression*, Version 1.1. www.statsoft.com.

Sweet S., & Martin K. (2011). *Data analysis with SPSS: A first course in applied statistics.* Fourth Edition.: Pearson publisher.

Vaitheeswaran K., Subbiah M., Ramakrishnan R., & Kannan T. (2016). *A comparison of ordinal logistic regression models using Classical and Bayesian approaches in an analysis of factors associated with diabetic retinopathy*. Journal of Applied Statistics, 43, 2254-2260.

Zhao, L., Chen, Y., & Schaffner, D.W. (2001). *Comparison of logistic regression and linear regression in modeling percentage data*. Applied and Environmental Microbiology, 67, 2129-2135.