

Use of Logistic Regression Approach to Determine the Effective Factors Causing Renal Failure Disease

Sami Jabbar Shappot

sami.jabbar.19800@gmail.com

Hazim Mansoor Gorgees

Hazim5656@yahoo.com

Department, of Mathematics, College of Education for Pure Science Ibn Al-Haitham, University of Baghdad, Baghdad, Iraq

Article history: Received 3 June 2018, Accepted 5 August 2018, Published December 2018

Abstract

The main goal of this research is to determine the impact of some variables that we believe that they are important to cause renal failure disease by using logistic regression approach. The study includes eight explanatory variables and the response variable represented by (Infected, uninfected). The statistical program SPSS is used to perform the required calculations.

Key word: Renal failure disease, logistic regression, Wald test, Hosmer and Lemeshow test.

1. Introduction

It is well known that the regression model is the most popular statistical models. One type of regression models is known as logistic regression. This type is suitable when the response variable is binary variable, while the explanatory variables may be categorical or continuous variables. Practically situations involving categorical outcomes are quite common. In a medical setting for example, an outcome might be presence or absence of a disease. Logistic regression is based on the logit transformation of dependent variable. This transformation is necessary since dichotomous dependent data violates least squares assumptions, furthermore, the error terms are not normally distributed which implies that all normality tests become invalid [1].

2. Assumptions of Logistic Regression

The main difference between the linear regression analysis and the logistic regression analysis is that the normally distributed dependent variable and homogeneity of the variance are not required in logistic regression analysis. The probabilities and the nature of log curve are the basis of the theory underlying logistic regression. The only assumptions of logistic regression are that the resulting logit transformation is linear, the resultant logarithmic curve does not include outliers, the dependent variable must be categorical and the categories have to be mutually exclusive so that a case can be only in one category and every case must be a member of one of the categories [2].

3. Types of Logistic Regression

Two types of logistic regression can be identified, namely, the binary and multinomial logistic regression. Binary logistic regression is a predictive model that can be used when the categorical response variable consists of two categories.

For example, live |die, presence absence of a disease. Multinomial logistic regression is an extension of binary type so that it allows the response variable to include more than two categories, however, the focus on this paper is on the binary logistic regression.

4. The Logit Model

Assuming that y_1, y_2, \dots, y_n denoted n independent observations on a binary response variable y : let π_i be the probability that $y_i=1$ then for p explanatory variables , the logit model is defined as :

$$\log \left[\frac{\pi_i}{1 - \pi_i} \right] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} \tag{1}$$

Unlike the usual linear regression model ,there is no random error term in the expression for logit model , this doesn't mean that the model is deterministic since there is still room for random variation represented by probabilistic relationship between π_i and y_i . We can solve the logit model for π_i to obtain

$$\begin{aligned} \pi_i &= \frac{\exp[\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}]}{1 + \exp[\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}]} \\ &= \frac{e^{\beta'x_i}}{1 + e^{\beta'x_i}} \end{aligned} \tag{2}$$

Where $\beta' = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)$ is a vector of unknown parameters ,
 $x' = (1, x_{i1}, \dots, x_{ip})$ is a vector of explanatory variables (1 is for the intercept)

Equation (2) can be simplified as follows

$$\pi_i = \frac{1}{1 + e^{-\beta'x_i}} \tag{3}$$

This equation has desired property that whatever we substitute for β' 's or x' 's , the value of π_i will always be a number between 0 and 1.

5. Maximum Likelihood Estimator

The data with binary response variable is one case that maximum likelihood method handles very nicely. The likelihood of observing the values of y for all of the n observations can be written as:

$$L = \prod_{i=1}^n pr(y_i) \tag{4}$$

Since $pr(y_i = 1) = \pi_i$ and $pr(y_i = 0) = 1 - \pi_i$

$$\text{Thus } f(y_i) = \pi_i^{y_i} (1 - \pi_i)^{1-y_i}, y_i = 0, 1, \dots, n \tag{5}$$

$$L = \prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{1-y_i} = \prod_{i=1}^n \left(\frac{\pi_i}{1 - \pi_i} \right)^{y_i} (1 - \pi_i)$$

Then we take the natural logarithm of both sides to get:

$$\ln L = \sum_{i=1}^n y_i \ln \frac{\pi_i}{1 - \pi_i} + \sum_{i=1}^n \ln(1 - \pi_i) \tag{6}$$

$$= \sum \beta' x_i y_i - \sum \ln(1 + e^{-\beta' x_i}) \tag{7}$$

Taking the derivative of $\ln L$ with respect to β and setting it equal to zero we get:

$$\begin{aligned} \frac{\partial \ln L}{\partial \beta} &= \sum x_i y_i - \sum x_i (1 + e^{-\beta' x_i})^{-1} \\ &= \sum x_i y_i - \sum x_i \hat{y}_i = 0 \end{aligned} \tag{8}$$

Where $\hat{y}_i = \frac{1}{1 + e^{-\beta' x_i}}$ is the predicted probability of y for a given value of x_i .

Actually, (8) is a system of $(k+1)$ equations one for each element of β .

There is no explicit solution to (8), instead, we must utilize iterative methods which yield successive approximations to the solution until the approximations converge to the correct value.

One of the most common iterative methods is referred to as Newton Raphson method which can be explained as follows [3]:

Let $u(\beta)$ be the vector of the first derivative of $\ln L$ with respect to β . That is :

$$u(\beta) = \frac{\partial \ln L}{\partial \beta} = \sum x_i y_i - \sum x_i \hat{y}_i$$

Let $I(\beta)$ be the matrix of second derivative, of $\ln L$ with respect to β

$$I(\beta) = \frac{\partial^2 \ln L}{\partial \beta \partial \beta'} = - \sum x_i x_i' (y_i - \hat{y}_i)$$

The Newton Raphson algorithm is then

$$\beta_{(j+1)} = \beta_{(j)} - I^{-1}[\beta_{(j)}] u[\beta_{(j)}] \tag{9}$$

Practically, we need a set of initial values β_0 , which can be started with all coefficients equal to zero.

These initial values are substituted into the right hand side of Equation (9) which give the result for the first iteration β_1 , again these values are substituted back into the right hand side of Equation (9) where the first and second derivatives are recomputed and the result is β_2 . This process is repeated until the maximum change in each parameter estimate from one step to the next is less than the value specified for tolerance on the logistic regression modeling.

6. Statistical Hypotheses and Model Fitting

In analyzing logistic regression two hypotheses are of interest specifically:

1-Null hypothesis which arises when all logistic regression coefficients are equal to zero i.e. there is no relationship between the binary response variable and the predictor variables.

2- Alternative hypothesis which arises when the logistic regression coefficients differ significantly from zero. i.e. there exists a significant effect of predictors on the binary response variable. There are two popular approaches for testing the null hypothesis. The first is performed by calculating the Wald statistic, which is similar to t test in multiple linear regression. This statistic is the division of the parameter estimate by the standard error of that

estimate $\left(\frac{\hat{\beta}_j}{s_{\hat{\beta}_j}} \right)^2$. For the large sample sizes the distribution of $\hat{\beta}_j$ approaches to normality this implies that the standard errors $s_{\hat{\beta}_j}$ are asymptotic, accordingly, the standard error should be

regarded approximate for small sample sizes..The wald statistic is more reliable for large sample sizes.

7. Hosmer and Lemeshow (H – L) Goodness of Fit Test

A popular test for the goodness of fit of the logistic regression model depending upon the Chi square statistic is known as Hosmer and Lemeshow test [4].

According to this test , the whole subject have been divided into ten ordered groups which are constructed on the basis of their estimated probability , those with estimated probability between 0 and 0.1 form one group , and so on , up to those with probability between 0.9 to 1 .The test statistic here is based on two components , namely , the observed component which does not depend on any theoretical distribution and expected component obtained from the estimated logistic model .A probability p value is computed from the Chi square distribution with 8 degrees of freedom to test the fit of the logistic model .We accept the null hypothesis if the H-L test statistic is greater than 0.05 .This means that there is no difference between the observed and model predicted values which implies that the model's estimate fit the data at an acceptable level.

8. Omnibus Test of Model Coefficients

This test is applied to determine whether the overall model with all predictors is significantly different from the model with only the intercept. The Omnibus test is interpreted as a test of the capability of all predictors in the model jointly to predict the response variable. The test is based on the difference between the log likelihood for the overall model and log likelihood of the model with only the constant term, such difference follows chi square distribution where the number of predictors represent the degrees of freedom [5].

9. The Practical Study

In our practical study, we attempted to assess the impact of some factors (explanatory variables) that we think they are important to cause the renal failure disease by employing logistic regression procedures.

The real data were collected from (60) real patients suffering renal failure and (55) people do not suffer from this disease from Al Kindi Hospital in Baghdad.

The logistic regression analysis was then performed with two groups (Infected and uninfected) and 8 predictor variables that we believe they cause the disease. The variables for each group are:

A. The dependent variable which is represent [1 for Infected (group 1)] and [0 for uninfected (group 2)]

B. Eight independent variables are described below:

- 1- Diabetes (0 uninfected,1 Infected) X_1
1. 2-Blood pressure (0 uninfected,1 Infected) X_2
- 2- (Glp) Glomerulonephritis (Renal syndrome) (0 uninfected,1 Infected) X_3
- 3- (UTI) Chronic urinary tract infection (0 uninfected,1 Infected) X_4
2. 5-Genetics (0 don't exist, 1 exist) X_5
3. 6-Age X_6
4. 7- Sex (1 for Male) (0 for Female) X_7
5. 8-Kidney stones (0 don't exist, 1 exist) X_8

The statistical program SPSS was used to perform the required calculations.

10. Logistic Regression Analysis

Employing the statistical program SPSS, the required results were obtained, and arranged in the following tables. **Table 1.** summarizes data enters to the analysis, the sample size studied and the missing data.

Table 1.Case Processing Summary

		N	Percent
Selected Cases	Included in Analysis	115	100.0
	Missing Cases	0	.0
	Total	115	100.0
Unselected Cases		0	.0
Total		115	100.0

The code (or symbol) of the dependent variable values is displayed in **Table 2.**

Table 2. Dependent Variable Encoding

Original Value	Internal Value
Uninfected	0
Infected	1

Table 3. includes the number of iterations for the derivatives of likelihood function in order to obtain the minimum value of -2log likelihood, that is required to get the optimal estimates for the model coefficients. The minimum value of -2log likelihood was obtained at the six iteration, it was equal to 81.207 the process was stopped at this iteration since the differences between the values of coefficients became very small (less than 0.001). In fact, the variation between the estimated coefficients became very small after the forth iteration as it is shown in **Table 3.** Then its estimated coefficients to be the best estimated coefficients that can be obtained.

Table 3. Iteration History

Iteration	-2 Log likelihood	Coefficients							
		Constant	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇
Step 11	62.354	-1.777	-0.005	-0.005	0.025	0.046	1.528	0.054	0.293
2	48.716	-2.998	-0.003	-0.003	0.090	0.132	2.445	0.090	0.632
3	44.753	-4.020	0.000	0.000	-0.302	0.226	3.210	0.157	0.949
4	43.919	-4.691	0.003	0.003	-0.492	0.307	3.735	0.239	1.140
5	43.851	-4.939	0.004	0.004	-0.566	0.346	3.933	0.281	1.197
6	43.850	-4.967	0.004	0.004	-0.574	0.350	3.955	0.287	1.203
7	43.850	-4.968	0.004	0.004	-0.574	0.350	3.955	0.287	1.203

Table 4. describes the estimator of the parameters of the optimal model obtained from the sixth iteration given in **Table 3**. All estimated coefficients ($\beta_0, \beta_1, \dots, \beta_8$) as well as the standard error and the Wald statistic for each estimated coefficient and the upper and lower bound for $\exp(\beta)$ are include

Table 4. Variables in the Equation

	β	S.E.	Wald	Df	Sig.	Exp(β)	95.0% C.I. for EXP(β)	
							Lower	Upper
Step1 X₁	2.415	0.742	10.596	1	0.001	11.195	2.614	47.934
X₂	3.168	0.798	15.767	1	0.000	23.749	4.973	113.417
X₃	4.725	0.955	24.464	1	0.000	112.780	17.339	733.588
X₄	3.258	0.854	14.561	1	0.000	26.005	4.878	138.637
X₅	0.728	0.861	0.714	1	0.398	2.070	0.383	11.196
X₆	-0.004	0.018	0.043	1	0.835	0.996	0.962	1.032
X₇	-1.027	0.598	2.942	1	0.086	0.358	0.111	1.158
X₈	1.863	0.748	6.200	1	0.013	6.446	1.487	27.945
Constant	-4.644	1.381	11.311	1	0.001	0.010		

Where Wald statistic = $(\frac{b}{s_b})^2$, and s_b is the standard error of b , the Wald statistic for the first estimate corresponding to X_1 is given as:

$$Wald = (\frac{2.415}{0.742})^2 = 10.596$$

From this table we conclude that the logistic regression equation is

$$\begin{aligned} \text{logit} \left(\frac{\hat{\pi}_i}{1 - \hat{\pi}_i} \right) &= -4.644 + 2.415x_1 + 3.168x_2 + 4.725x_3 + 3.258x_4 + 0.728x_5 \\ &\quad - 0.004x_6 - 1.027x_7 + 1.863x_8 \end{aligned} \tag{11}$$

To test the efficiency and the goodness of fit for the whole model we use the log likelihood ration, which follows the relationship

$$\chi^2 = 2(\log L_0 - \log L_1)$$

Where: L_1 is the value of likelihood function of the full model, L_0 is the value of likelihood function of the reduced model, the value of χ^2 was found to be 77.791 which is significant at the level α less than 0.001 and 8 degrees of freedom with sig=0 which ensure the significance of the whole fitted model as shown in **Table 5**.

Table 5. Omnibus Tests of Model Coefficients

	Chi-square	df	Sig.
Model	77.791	8	.000

Another test for the goodness of fit of the model depending upon the χ^2 statistic is presented in **Table 6**. The test statistic here is based on two components, as Hosmer and Lemeshow

suggested. Namely, the observed component which does not depend on any theoretical distribution and expected component obtained from the estimated logistic model.

Table 6. Contingency Table for Hosmer and Lemeshow Test

		group = uninfected		group = Infected		Total
		Observed	Expected	Observed	Expected	
Step 1	1	12	11.697	0	0.303	12
	2	9	11.163	3	0.837	12
	3	11	10.327	1	1.673	12
	4	8	8.474	4	3.526	12
	5	8	5.997	4	6.003	12
	6	5	3.572	7	8.428	12
	7	1	1.826	11	10.174	12
	8	0	0.725	12	11.275	12
	9	0	0.215	12	11.785	12
	1	0	0.003	7	6.997	7
	0					

The value of χ^2 was found to be 10.308 with 8 degrees of freedom and significant value 0.244 as shown in **Table 7.**, hence, we accept the null hypothesis that there is no difference between the actual and expected frequencies which implies that a good fitting for the whole model. From table the closeness of observed and expected values is clear.

Table 7. Hosmer and Lemeshow Test

Step	Chi-square	Df	Sig.
1	10.308	8	.244

The percentage of the correct classification is presented in **Table 8.** The overall percentage was found to be 83.5% calculated as $[(44+52)/115] \times 100\%$ while the percentage of not correct classification was found to be $[(9+10)/115] \times 100\% = 16.5\%$ and this is a good indicator that the model fits well the data.

Table 8. Classification table

	Observed		Predicted		
			Group		Percentage Correct
			uninfected	Infected	
Step 1	group	Uninfected	44	10	81.5
		Infected	9	52	85.5
	Overall Percentage				83.5

11. Conclusions

1) The best way to appreciate the Wald statistic is to consider its significance values, we reject the null hypothesis and accept the alternative hypothesis which implies that the variable under consideration does make a significant contribution. In our study, we noted that five

explanatory variables are contributed significantly to the prediction, namely, X_1 , X_2 , X_3 , X_4 and X_8 whose significant values are less than 0.05 as it is shown in **Table 4**.

2) Hosmer and Lemeshow (H – L) goodness of fit test is one of the most popular methods in logistic regression for testing the null hypothesis which assumes that there is no difference between the observed and model predicted values. If the value of the test statistic is more than 0.05 as it is desired for good fitting model, we accept the null hypothesis. In our practical study, the H – L statistic has 0.244 as it is shown in **Table 7**. This ensures that our model is quite of good fit, this preferable conclusion indicates that there are no significant differences between the observed and predicted values.

3) In addition to using a goodness of fit statistic, we often interest in looking at the proportion of outcomes we have managed to classify correctly. For this we need to look at the classification table. In our study, 81.5% were correctly classified for the uninfected group and 85.5% for the infected group. Overall 83.5% were correctly classified. This indicates a good performance of logistic regression model.

References

1. Agresti A. *An Introduction to Categorical Data Analyses*. NY: John Wiley & Sons, Inc. New York. 1996.
2. Chatterjee, S.; Hadi, A. *Regression Analysis by Example*. Fourth Edition, John Wiley and Sons, Inc. 2006.
3. Press, J.; Wilson, S. choosing between Logistic Regression and Discriminant Analysis. *Journal of the American Statistical Association*. **1978**, 73, 364, 99-705.
4. Ahmed, L.A. Using Logistic Regression in Determining the Effective Variables in Traffic Accidents. *Applied mathematical Sciences*. **2017**, 11, 42, 2047- 2058.
5. Kuha, J.; Mills, C. on group Comparisons with logistic regression models. A variable in LSE Research Online: **2017**.