تحديد اهم العوامل المؤثرة على الإصابة بمرض القلب باستخدام الانحدار اللوجستي
(دراسة تطبيقية في مستشفى اربيل)

Determine of the Most Important Factors that Affect the Incidence of Heart Disease Using Logistic Regression Model (Applied Study in Erbil Hospital)

Mohammed A. Mohammed
Department of Accounting Techniques
Al-Dewanyia Technical Institute
Al-Furat Al-awsat Technical University

Saif Hosam Raheem
Department of Statistics
Faculty of Administration and Economics
University of Al Qadisiyah

Mohammed A. Mohammed **Saif Hosam Raheem**

الملخص :

مرض القلب هو مصطلح يشير إلى مجموعة متنوعة من الأمراض المختلفة التي تصيب القلب. في عام ٢٠٠٧ ؛كانت أمراض القلب في مقدمة أسباب الوفاة في كل من الولايات المتحدة والمملكة المتحدة وكندا. في الولايات المتحدة هناك ما يقارب ما نسبته ٢٥.٤٠٪ من جميع الوفيات تعزى إلى هذه الأمراض . في عام ٢٠١٥ حوالي ١٧.٧ مليون حالة وفاة بسبب أمراض القلب، وهو ما يمثل ٣١ ٪ من جميع الوفيات في العالم في نفس العام.

وفقًا لمنظمة الصحة العالمية ، فإن أكثر من ثلاثة أرباع الوفيات الناجمة عن أمراض القلب والأوعية الدموية تحدث في البلدان المنخفضة والمتوسطة الدخل مثل سوريا واليمن وبلدنا ؛ العراق. في هذه الدراسة ، تم استخدم نموذج الانحدار اللوجستي لوصف العلاقة بين متغير الاستجابة الثنائية (مصاب ، غير مصاب) ومجموعة المتغيرات المستقلة (ثمانية متغيرات ثنائية) لتحديد أهم العوامل التي تؤثر على أمراض القلب.حيث تم تطبيق اسلوب الانحدار اللوجستي على بيانات تمثل عينة من مجموعة من المرضى في مستشفى الجراحة المتخصصة في أربيل حجمها (٢٠٠ مريض). شخصت الدراسة بعض العوامل المهمة التي تؤثر على أمراض القلب مثل التدخين والسمنة وإدمان الكحول.

Abstract

Heart disease is a term that refers to a variety of different types of diseases that infect the heart. In 2007, heart disease is at the top of the causes of death in the United States, United Kingdom and Canada. In the United States, there are around 25.40% of all deaths are attributed to these diseases (Wikipedia). In 2015, about 17.7 million people deaths due to heart disease, accounting for 31% of all deaths in the world in the same year.

According of World Health Organization, more than three-quarters of deaths from cardiovascular disease occur in low- and middle-income countries such as Syria, Yemen and our country ;Iraq. In this study, a logistic regression model is used to describe the relationship between the binary dependent response variable (infected, uninfected) and set of independent variables (eight binary variables) to identify the most important factors affecting heart disease. We applied the logistic regression on the data represent a sample of a group of patients at the specialized surgery hospital in Erbil (200 patients).

The application study indicated some interesting factors that effect on the heart disease such as smoking, obesity and alcoholism.

Keywords: Logistic regression, heart disease, likelihood ratio test, Wald test, Hosmer & Lemeshow test

1 Introduction

The term of heart disease (HD) refers to several diseases (vascular diseases, birth defects, irregular heartbeat). All these diseases lead to heart attack and angina. As for the symptoms of HD, it begins with chest pain accompanied by shortness of breath as well as numbness and coldness in the limbs, and sometimes there are pain in the neck and jaw. It is interesting to see that there is a three quarters of cardiovascular deaths in the world occurred in low income countries (see [1], [2]). People in poor countries such as Iraqi peoples are often do not have benefit from integrated primary health care programs which aimed for early identify and therapy of people exposed to risk factors as compared with those in developed countries.

A lot of studies that dealt with HD physiologically, however there is no significant works for detection the factors affecting the incidence of HD statistically (for more details, see [1], [2]). The aim of this article is to diagnose the most important factors affecting heart disease such as diabetes, blood pressure, obesity, smoking, family history of heart disease, alcoholic beverages, social status, and psychological stress. The purpose of identify the important factors is to avoid them to reduce the risk of HD. In addition, the following methodology was used to achieve the research objective.

- Adopting the descriptive analytical approach in characterizing the logistic regression model and estimating its parameters.
- Analysis the collected data based on logistic regression model using statistical programs.

This study is introduced as follows; the logistic regression model is briefly displayed in Section 2. Parameters estimation of logistic regression is introduced in Section 3. In sections 4 and 5, the Wald test and Hosmer and Lemeshow test are given. In Section 6, the variables of study are described. The analyses of data are discussed in Section 7. Finally, the conclusion and some important recommendation are introduced in sections 8 and 9.

2 Logistic Regression model

Logistic regression (LR) is a mathematical approach that used to analyze data in which the dependent variable follows Bernoulli distribution, whereas, binary logistic regression is a type of regression analysis where the dependent variable is a dummy variable: coded 0 (uninfected) or 1(uninfected). Assumed the following model [1], [5];

$$y_i = X_i' \beta + \varepsilon_i \quad \dots (1)$$

where,

$X_i = [1, X_{i1}, X_{i2}, \dots, X_{in}$ represent a vector of independent variables,
$\beta' = [\beta_0, \beta_1, \dots, \beta_n]$ represent a vector of unknown parameters, and
$\varepsilon_i \sim N(0, \sigma^2)$ is a vector of random error term and $E(\varepsilon_i) = 0$.

The dependent variable y_i follows the Bernoulli distribution with binary outcome 0 and 1. The probability function of logistic regression can express as [10];

$$P(y_i = 1) = \pi_i$$
$$P(y_i = 0) = 1 - \pi_i \quad \dots (2)$$

The expected value of independent variables is given by;

$$E(y_i) = 1(\pi_i) + 0(1 - \pi_i) = \pi_i$$
$$= X_i' \beta = \pi_i \quad \dots (3)$$

This is lead to, $0 \leq E(y_i) = \pi_i \leq 1$

Therefore, the expected of dependent variable represents the probability function of the dependent variable when it takes value 1.

There are essential problems in the regression model in Equation (1) with a binary data. Where, the random error term is not normally distributed due to it takes only two values, shown as follows [10], [11];

$$\varepsilon_i = \begin{cases} 1 - X_i' \beta & \text{when } y_i = 1 \\ X_i' \beta & \text{when } y_i = 0 \end{cases} \quad \dots (4)$$

and the variance is not constant, given by;

$$\sigma_{yi}^2 = E\{y_i - E(y_i)\}^2 = (1 - \pi_i)^2 \pi_i + (0 - \pi_i)^2 (1 - \pi_i) = \pi_i (1 - \pi_i)$$
$$= E(y_i)[1 - E(y_i)] \quad \dots (5)$$

From Equation (5), we can notice how the variance could be potentially different for each y_i, unlike for normal linear regression. In general word, with binary dependent variable follows

Bernoulli distribution we cannot apply linear regression model [5]. Figure (1) shows comparing between linear regression model and logistic regression model for binary data. In order to solve this problem, many mathematical transformation approaches are available such as logistic response function, expressed as follows (see [6], [10], [11]);

$$E(y) = \frac{exp\,(X'\beta)}{1+exp\,(X'\beta)} \quad \ldots (6)$$

Let θ represent the linear predictor, given as;

$$\theta = X'\beta \quad \ldots (7)$$

By applied the logit transformation,

$$\theta = ln\ \frac{\pi}{1-\pi} \quad \ldots (8)$$

where, $\frac{\pi}{1-\pi}$ is called a log odds ratio.

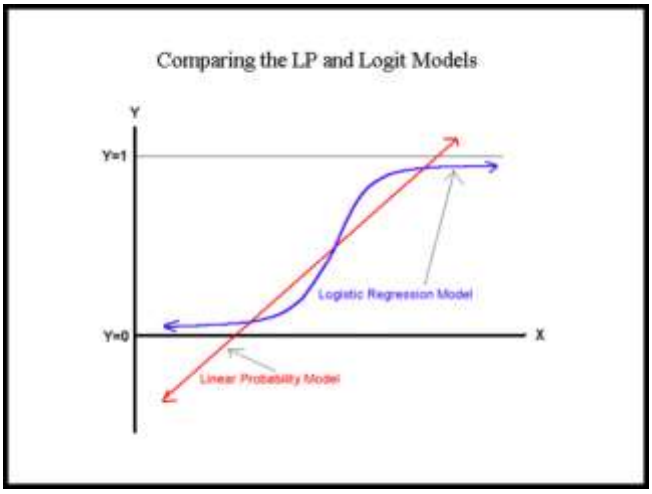

Figure (1): Comparing between linear regression and logistic regression for binary data

3 Parameters Estimation of Logistic Regression

Maximum likelihood estimation (MLE) is commonly used to estimate the parameters of the model. The likelihood function for "*n"* observation is [3], [4]

$L(y_1, y_2, \ldots., y_n, \beta) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$ … (9)

where

$$f_i(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}, \ \ i = 1,2, \ldots n \quad \ldots (10)$$

$$\pi_i = E(y_i) = \frac{\exp(X'\beta)}{1+\exp(X'\beta)} \quad \ldots (11)$$

Now as we can find the log likelihood functions as:

$ln\ L(y_1, y_2, \ldots., y_n, \beta) = \ln \prod_{i=1}^{n} f_i(y_i) = \sum_{i=1}^{n}[y_i \ln\left(\frac{\pi}{1-\pi_i}\right)] + \sum_{i=1}^{n} \ln(1-\pi_i)$ … (12)

in the above expression. Thus, we can write

$$\ln L(y,\beta) = \sum_{i=1}^{n} y_i\, X_i'\beta - \sum_{i=1}^{n} \ln[1+\exp(X_i'\beta)]$$

$$= \sum_{i=1}^{n} y_i \ln(\pi_i) + \sum_{i=1}^{n} n_i \ln(1-\pi_i) - \sum_{i=1}^{n} y_i \ln(1-\pi_i)$$

$$= \sum_{i=1}^{n} y_i \ln(\pi_i) + \sum^{ni=1}(n_i - y_i)\ln(1-\pi_i)$$

where,

$$1-\pi_i = [1+\exp(X_i'\beta)]^{-1}\ ,\ n_i = \ln\left[\frac{\pi_i}{1-\pi_i}\right] = X_i'\beta$$

Taking derivatives with respect to unknown parameters $\beta_1, \beta_2, \dots, \beta_p$, setting them equal to 0, then, solve it leads to the maximum likelihood estimates (MLE). These parameter estimates are denoted by $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$. Corresponding estimates of π are [6], [7];

$$\hat{\pi} = \frac{exp\,(X'\,\hat{\beta}\,)}{1+exp\,(X'\,\hat{\beta}\,)} \dots (13)$$

The MLE can be found by using iterative numerical approaches. Most software packages such as SPSS, SAS and R language use an iteratively reweighted least squares procedure (IRLS) to find the MLE for logistic regression models (see [9], [12], [13]).

4 Wald Test

It is a commonly test which used to test of significance for individual coefficient in LR (recall that we use t-tests in LR). This test is calculated by dividing the value of the coefficient by the standard error σ. This test is following normal distribution *N* (0,1) or *Z* distribution in case of large samples.

The Wald statistic is defined as [2], [4], [8],

$$W = \left[\frac{\hat{\beta}_j}{se(\hat{\beta}_j)}\right]^2 \sim x^2 \text{ with 1 } df \dots (14)$$

$$W = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim N(0,1) \dots (15)$$

where;

$\hat{\beta}_j$ is an estimated coefficient ,

$se(\hat{\beta}_j)$ is an estimated standard error

The hypothesis test of Wald test is,

$$H_0: \beta_j = 0$$
$$vs \quad H_1: \beta_j \neq 0 \qquad \text{for } j = 1,2,\dots,n \dots (16)$$

5 Hosmer and Lemeshow Test

The Hosmer and Lemeshow (H&L) test is a statistical test for goodness of fit for LR model. It is commonly used in risk forecasting models. The test evaluate whether or not the observed values rates match expected values rates in subgroups of the model population. The H&L test follows chi-square distribution with one degree of freedom [2], [4], [8],

The hypothesis test of Hosmer and Lemeshow test is,

H_0: There are no significant differences between observed and expected values

H_1: There are significant differences between observed and expected values

The null hypothesis is accepted when the probabilistic value of the H & L test is greater than 0.05, meaning that there are no significant differences between observed and expected values.

6 Variables of Study

The variables of study included one dependent binary variable (y_i) and eight independent variables ($x_{i1}, x_{i2}, \dots, x_{ip}$) , $i = 1,2,\dots,200$, $p = 8$.

The data were collected through the questionnaire distributed to 200 patients, who visit the specialized surgery center in Erbil in 2018,

The variables are described as following,

y: Patient has a heart disease (1 = infected, 0 = uninfected)
x_1: Smoking (1 = smoker, 0 = non-smoker)
x_2: Obesity (1 = overweight, 0 = normal weight)
x_3: Diabetes (1 = infected, 0 = uninfected)
x_4: Family register (1 = has infected, 0 = there is no one infected)
x_5: Congenital defects (1 = have congenital heart defects, 0 = no congenital heart defects)
x_6: Drinking alcohol (1 = drink, 0 = not drink)
x_7: Marital status (1 = Married, 0 = single)
x_8: Psychological pressure (1 = always has a stress, 0 = has no stress)

7 Analyses of Data and Discussion

The statistical program SPSS was used to analyze the questionnaire data which were distributed to the patients of the specialized surgery center in Erbil. The results of the analysis are demonstrated in Table (1)

Table (1): Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	200	100.0
	Missing Cases	0	.0
	Total	200	100.0
Unselected Cases		0	.0
Total		200	100.0

a. If weight is in effect, see classification table for the total number of cases.

Table (1) shows the number of questionnaire (200 forms), the number of lost forms (0), and the total sample size is (200).

Table (2): Dependent Variable Encoding

Original Value	Internal Value
Uninfected	0
Infected	1

Table (2) shows the values of the dependent variable (binary response), where the incidence of heart disease was defined as (1) for affected and (0) for not affected.

Table (3): Iteration History a, b, c, d

Iteration		-2 Log likelihood	Coefficients								
			Constant	x1(1)	x2(1)	x3(1)	x4(1)	x5(1)	x6(1)	x7(1)	x8(1)
Step 1	1	269.047	.380	-.348-	-.184-	-.441-	-.072-	.454	-.155-	-.003-	-.164-
	2	269.037	.398	-.363-	-.190-	-.457-	-.076-	.469	-.163-	-.004-	-.170-
	3	269.037	.398	-.363-	-.191-	-.457-	-.076-	.469	-.163-	-.004-	-.170-

a. Method: Enter
b. Constant is included in the model.
c. Initial -2 Log Likelihood: 277.079

d. Estimation terminated at iteration number 3 because parameter estimates changed by less than .001.

Table (3) demonstrates the number of iteration cycles to obtain the lowest negative value of the (-2 log likelihood) function. This was done in the third cycle with a value of (269.037), while the initial value in the case of the model includes the constant term was (277,079). In the same table above (paragraph d), shows that the estimate was completed at the third iteration because the parameter estimates changed by less than (0.001). This indicates that the estimates obtained in the third cycle are the best estimate of the parameters of the logistic regression model.

Table (4): Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	8.042	8	.000
	Block	8.042	8	.000
	Model	8.042	8	.000

Table (4) shows the quality and efficiency of the model through the use of the likelihood ratio test which follows chi-square distribution with 8 degrees of freedom. The likelihood ratio test is computed as;

$$x^2 = -2\log L_0 - (-2\log L_1) \ldots (17)$$

where,

$-2\log L_1$ is computed for full model, whereas, $-2\log L_0$ is computed for only constant. Then,

$$x^2 = 277.079 - 269.037 = 8.042$$

Also, from the above table, the value of x^2 is significant due to the value of sig. is (0.000), this confirms that the significance and efficiency of the model.

Table (5): Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	269.037[a]	.539	.753

a. Estimation terminated at iteration number 3 because parameter estimates changed by less than .001.

From Table (5), the value of Cox & Snell R Square and Nagelkerke R Square are (0.539) and (0.753) respectively. The Cox & Snell R Square value represents the change in the dependent variable which was explained by the logistic regression model is 53.9 %. Whereas, the value of Nagelkerke R Square shows that 75.3% was explained to the dependent variable by the logistic regression model.

Table (6): Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	12.799	8	.119

From Table (6), the Hosmer and Lemeshow Test value is (12.799), and the significant value of the test is (sig = 0.119), which is greater than (0.05). Therefore, we reject the null hypotheses (there are no significant differences between observed and expected values).The observed and expected values are displayed in Table (7)

Table (7): Contingency Table for Hosmer and Lemeshow Test

		y = Uninfected		y = Infected		
		Observed	Expected	Observed	Expected	Total
Step 1	1	14	13.742	6	6.258	20
	2	14	12.467	6	7.533	20
	3	9	11.122	10	7.878	19
	4	15	11.173	5	8.827	20
	5	6	10.602	14	9.398	20
	6	9	10.075	11	9.925	20
	7	13	9.465	7	10.535	20
	8	9	9.319	12	11.681	21
	9	6	8.219	14	11.781	20
	10	8	6.817	12	13.183	20

Table (8): Classification

Observed			Predicted		
			y		Percentage Correct
			Uninfected	Infected	
Step 1	y	Uninfected	66	27	70.96
		Infected	29	58	74.35
	Overall Percentage				62.0

Table (8) shows the correct percentages of Predicted, where the percentage of non-infected people is (70.96) and the proportion of people infected is (74.35) and the overall percentage (62.0). Those results indicate that the model represents the data well.

Table (9): Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	x1(1)	3.235	.295	1.513	1	.019	25.406
	x2(1)	2.587	.299	.407	1	.024	13.290
	x3(1)	-.457-	.296	2.391	1	.122	.633
	x4(1)	1.547	.291	.068	1	.015	4.697
	x5(1)	.469	.294	2.547	1	.111	1.598
	x6(1)	1.985	.291	.313	1	.046	7.279
	x7(1)	-.004-	.294	.000	1	.988	.996
	x8(1)	-.170-	.295	.331	1	.565	.844
	Constant	.398	.445	.800	1	.371	1.489

a. Variable(s) entered on step 1: x1, x2, x3, x4, x5, x6, x7, x8.

Table (9) displays the estimated coefficients values and standard errors values (S.E). The logistic regression equation will be as following form,

$\hat{y} = 0.398 + 3.235\, x_1 + 2.587x_2 - 0.457x_3 + 1.547 + 0.469x_5 + 1.985x_6 - 0.004x_7 - 0.170x_8$

The last column in Table (9) shows the values of Exp.(B), which represents the Odd ratio. The Odd ratio represents the amount of change in the proportion of the likelihood that the person is infected when the variable independent is change, which is associated with parameter β.

The following are the most important factors that lead to the incidence of heart disease and its likelihood ratio.

x_1: Smoking:

$$Exp(B) = Exp(3.235) = e^{3.235} = 25.406$$

The Exp (B) value is (25.406) indicate that smokers are more likely to infect heart disease than (25.406 %) of non-smokers.

x_2: Obesity:

$$Exp(B) = Exp(2.587) = e^{2.587} = 13.290$$

Here, we find that the fat person is more likely to infect heart disease by (13,290%) than normal person.

x_3: Diabetes

$$Exp(B) = Exp(-0.457) = e^{-0.457} = 0.633$$

x_4: Family register

$$Exp(B) = Exp(\ 1.547) = e^{1.547} = 4.697$$

x_5: Congenital defects

$$Exp(B) = Exp(\ 0.469) = e^{0.469} = 1.598$$

x_6: Drinking alcohol

$$Exp(B) = Exp(\ 1.985) = e^{1.985} = 7.279$$

x_7: Marital status

$$Exp(B) = Exp(-0.004) = e^{-0.004} = 0.996$$

x_8: Psychological pressure

$$Exp(B) = Exp(-0.17) = e^{-0.17} = 0.844$$

8 Conclusions

From the results of data analysis and discussion, it is easy to see there is a considerable relationship between the heart disease and the independent variables of the study, where the Nagelkerke R Square value was 0.753. In addition, the Smoking variable has a major cause of heart disease, where the value of Wald test was significance and equal to (sig = 0.019) and it has the highest value of Exp.(B), which equal to 25.406. The second important variable around all of other variables is the Obesity variable with Exp.(B) value equal to and 13.290 and (sig = 0.024). The third significant variable that has high effect on heart disease is the Drinking alcohol with (sig.= 0.046) and Exp.(B)=7.279. The Family register has a lowest significant impact on the HD with sig.= 0.15 and Exp.(B) = 4.697. It is interesting to see the rest variables such as Diabetes, Congenital defects, Marital status and Psychological pressure are not significant variables in this study due to the Wald test values were not significant.

9 Recommendations

1- The study showed that the highest effect of heart disease is smoking followed by obesity and drinking alcohol, so we should wary about the negative effects of these things to avoid diseases especially heart disease. We advise the heart disease patients to give up smoking and alcohol and do exercises to avoid obesity.

2- The concerned health institutions, especially in the Iraq's cities, should take the responsibility of educating people about the risks of this common disease and alert them to avoid negative habits that increase the probability of heart disease.

3- The study indicated that there are a set of variables are not significant and may be due to the data of the sample, so we recommend conducting studies at a high level to reach results may be more accurate in the diagnosis of factors affecting heart disease

10 References

1- Go, A.S., Hylek, E.M., Phillips, K.A., Chang, Y., Henault, L.E., Selby, J.V. and Singer, D.E., 2001. Prevalence of diagnosed atrial fibrillation in adults: national implications for rhythm management and stroke prevention: the AnTicoagulation and Risk Factors in Atrial Fibrillation (ATRIA) Study. *Jama*, *285*(18), pp.2370-2375.

2- Hayes, A.F. and Matthes, J., 2009. Computational procedures for probing interactions in OLS and logistic regression: SPSS and SAS implementations. *Behavior research methods*, *41*(3), pp.924-936.

3- Hosmer Jr, D.W., Lemeshow, S. and Sturdivant, R.X., 2013. *Applied logistic regression* (Vol. 398). John Wiley & Sons.

4- Hu, F.B., Stampfer, M.J., Haffner, S.M., Solomon, C.G., Willett, W.C. and Manson, J.E., 2002. Elevated risk of cardiovascular disease prior to clinical diagnosis of type 2 diabetes. *Diabetes care*, *25*(7), pp.1129-1134.

5- Menard, S., 2002. *Applied logistic regression analysis* (Vol. 106). Sage.

6- Montgomery, D.C., Peck, E.A. and Vining, G.G., 2012. *Introduction to linear regression analysis* (Vol. 821). John Wiley & Sons.

7- Osmer Jr, D.W., Lemeshow, S. and Sturdivant, R.X., 2013. *Applied logistic regression* (Vol. 398). John Wiley & Sons.

8- Rasha A. S., 2015. Using the Logistic Regression Model in Studding the Assistant Factors to Diagnose Bladder Cancer. *Journal of Economics and Administrative Sciences*, *21*(83), 341.

9- Sweet, S.A. and Grace-Martin, K., 1999. *Data analysis with SPSS* (Vol. 1). Boston, MA: Allyn & Bacon.

10- Wasserman, L., 2013. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

11- Wooff, D., 2004. Logistic Regression: A Self-Learning Text. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *167*(1), pp.192-194.

12- Wuensch, K.L., 2014. Binary logistic regression with SPSS. *Retrieved March*, *18*, p.2015.

13- Yusuff, H., Mohamad, N., Ngah, U. and Yahaya, A., 2012. Breast cancer analysis using logistic regression. *IJRRAS*, *10*(1), pp.14-22.