Employing Logistic regression approach to determine impact of some variables on diabetes

Bushra Abdulrasool Ali

Ministry of Education

Abstract:-

Age, Sex, Uric Acid, Blood Urea, cholesterol, Serum Triglyceride, Smoking, Heredity and Alcohol are the nine variables that lead to get diabetes. logistic regression approach was employed using the statistical program SPSS to study the significant variables between them, the result's reveled that cholesterol, Serum Triglyceride, Heredity and age are the four effected parameters on getting the disease. The study based on a 200 population obtained from the National Center for Diabetes Treatment in Iraq. The true classification percentage is 91.5% and the total error is 8.5%.

Key words:-

Diabetes, Logistic regression, Logit model, Hosmer and Lemeshow test

توظيف الانحدار اللوجستي في تحديد اثر بعض المتغيرات المؤثرة على مرضى السكري

بشرى عبدالرسول علي

وزارة التربية

المستخلص:

العمر ، الجنس ،حامض اليوريك، املاح الدم، الدهون الثلاثية ،الوراثة والعمر هي تسع متغيرات تؤدي الى الاصابة بمرض السكري .تم توظيف أسلوب الانحدار اللوجستي واستخدام البرنامج الاحصائي (spss) اظهرت النتائج الى ان الكولسترول والدهون الثلاثية والوراثة والعمر هي اربع متغيرات ذات التاثير الاقوى من بين العوامل الانفة الذكر اعتمدت الدراسة على 200 شخص تم الحصول على البيانات من المركز الوطني لعلاج مرضى السكري في العراق وكانت نسبة التصنيف الصحيح 91.5 % ونسبة الخطأ % 8.5

Introduction:-

Patients who suffering from the diabetes were dramatically increased in the last decay [2]. This motivated the doctors and researchers to study the main reasons which cause the disease. Moreover that, diabetes has a negative consequences on other organs as heart, arteries, nerves, kidney and eye [6]. Diabetes is one of the most common diseases which is a widely spread over the world. It caused by increasing the sugar level in the blood due to the insulin reduction. The statistical and recent studies have shown that about 280 million person diagnosed with diabetes and it is expected that this number will be sorted to 420 million by 2030 [8]. Briefly, diabetes occur as a result of pancreas failure. it leads to insulin reduction within the body and less activity as well. It could cause the death if it not be correctly diagnosed and treated quickly and proficiently under monitoring of a doctor. It has a negative impact on the patient body according to the literature most of the diabetic patient suffering from the side effect of this disease as atherosclerosis and stroke [11]. 80-110 mg/dl is the normal blood sugar level range for a fasting person over night as it has been determined by the World Health Organization and the Union of diabetics in the United States [3]. This paper focus on the impact of the explanatory variables which are important for diabetes patients using logistic regression analyses and we have concluded that only four of these variables have a significant effect on the infection of the disease.

1- The logistic regression

The logistic regression is one of the regression model when the relation between the independent varables and the depended variable is nonelinear [4]. Logistic regression can be applied when the response variable is a categorical variable, while the independent variables may be a categorical or continuous [1]. It is a widely spread utilize a categorical outcomes in practical applications. For example, in education statistical program predications may be made for the dichotomous outcome of success/failure. Similarly, in a medical settings an outcome might be presence/absence of disease. The principle of the Logistic regression depends on a logit transformation, the logit transformation create a continuous logarithmic relation from non-continuous data which stimulate the development of the regression model. The model rely on the dependent variable data which can be employ in the outcome probability. The logit transformation is substantial since the dichotomous dependent data violates linear regression assumptions. An error terms are not normally distributed in dichotomous data is that the, thus all normality tests

are invalid. Linear regression is more flexible than Logistic regression. It does not required a normally distributed dependent data or homogeneity of variance. Predictions can be made by a linear regression based on the observed changes in the independent data [5]. While logistic regression is established on the log of the odds of a particular event occurring with a given set of observations. Logistic regression's underlying principles is founded on a probabilities and the nature of the log curve. The exclusive assumptions of logistic regression is that the results logit transformation is linear, the resultant logarithmic curve does not include outliers, the dependent variable is categorical, the categories must be mutually exclusive so that the case can only be in one category and every case must be a member of one of the categories. A minimum of 50 cases per predictor is recommended is needed for logistic regression which is not a condition for the linear regression because maximum likelihood coefficients are large sample estimates [12].

2- The Logit Model

The bounded probabilities between 0 and 1 is one of the limitation application in linear probability model, however that the linear functions is inherently unbounded. Transforming the probability to an odds removes the upper bound. If the natural logarithm of the odds was taken, we also remove the lower bound. Setting the result equal to a linear function of the explanatory variables, it can get the logit model. For k explanatory variables and i =1, 2, 3,…, n. Individuals, the model is [7]

$$\log\left[\frac{p_i}{1-p_i}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \qquad \ldots(1)$$

Where p_i is, as before, the probability that $y_i = 1$. The expression on the left hand side of the equation refer to logit or log odds.

Unlike the usual linear regression model there is no random disturbance term in the equation for the logit model. This does not mean that the model is deterministic because there is still room for the random variation in the probabilistic relationship between p_i and $\mathrm{y_i}$.

The logit equation for p_i can be solved to obtain [7]

$$p_i = \frac{\exp(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_k x_{ik})}{1+\exp(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_k x_{ik})} \quad \ldots(2)$$

It can simplify further by dividing both the numerator and denominator by the numerator itself:

$$p_i = \frac{1}{1+e^{-\beta_0-\beta_1 x_{i1}-\beta_2 x_{i2}-\ldots-\beta_k x_{ik}}} \quad \ldots(3)$$

This equation has desired property that no matter what values we substitutes for the β's and the $x's$, p_i will always be a number between 0 and 1.
If we have a single x variable with β_0=0 and β_1=1,

3- Measures and Significance Test

The simple logistic regression was firstly considered, the case in which a single predictor variable is included.

$g(x) = \beta_0 + \beta_1 x_i$ Where $g(x) = \ln \frac{p_i}{1-p_i}$

The null hypothesis H_0 of main interest when fitting this model is that β_1 equal to zero, i.e. there is no relationship between the binary response variable and the predictor variable. There are two common ways for testing this hypothesis, the first is calculate the wald statistic, which is the parameter estimate divided by its standard error. $\frac{b_i}{s_{bi}}$

We have to note that the standard error (s_{bi}) is asymptotic, which means that the distribution of b_1 approaches normality for the large sample sizes, so that the standard error should be considered approximate for small sample sizes.
The wald statistic is sometime called the wald t (or t ratio) statistic because of its similarity to t statistic [10]. The wald statistic is traditionally compared with the standard normal Z distribution. It is more reliable when the sample sizes are large. An alternative approach is more robust to small sample sizes, it is similar to what applied for linear regression models where we can compare full and reduced models. However that the use log likelihood as a measure of fit rather than least squares. To test the null hypothesis H_0 that β_1 equal to zero

for a simple logistic regression model with a single predictor, we compare the fit (the log likelihood) of the full model

$$g(x)=\beta_0+\beta_1 x_i \qquad \ldots (4)$$

To the fit of reduced model

$$g(x)=\beta_0 \qquad \ldots (5)$$

To compare likelihoods, we use a likelihood ratio statistic (Λ), which is the ratio of the log likelihood of reduced model to the log likelihood of full model. Larger log likelihoods mean a better fit, so if Λ is near one, then β_1 contributes little to the fit of the full model. To test H_0, the sampling distribution of Λ is needed when H_0 is true. The sampling distribution of Λ is messy so instead we calculate a G^2 statistic where

$$G^2=-2\ln(\Lambda) \qquad \ldots(6)$$

This also called the likelihood ratio χ^2 statistic. It can be simplified to

$$G^2=-2(\text{log likelihood reduced} - \text{log likelihood full}) \qquad \ldots(7)$$

If H_0 (β_1 equals zero) is true, the sampling distribution of G^2 is very close to a χ^2 distribution with one degrees of freedom. Therefore, we can test H_0 that β_1 equals zero with either the wald test or with G^2 test to compare the fit of reduced and full models. In contrast to least squares model fitting where the t test and the F test for testing β_1 equal to zero are identical for a simple linear regression, the wald and G^2 test are not the same in logistic regression. The wald test tends to be less reliable and lacks power for smaller sample sizes and the likelihood ratio statistic is recommended.

The model fitting procedures can be easily extended to situations with multiple predictor variables. The general multiple logistic regression model for p predictors can be represent by:

$$g(x)=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_p x_{ip} \qquad \ldots(8)$$

Where $g(x)$ is the natural log of the odds ratio, β_0 is the intercept or constant term, β_j, (j =1,2,…,p) is the partial regression coefficient for x_j holding the

remaining predictors constant. As multiple linear regression models, we can firstly test the significance of the overall regression model by comparing the log likelihood of the full model in eq (8) to the log likelihood of the reduced model (constant, or β_0 only).

G^2 statistic was calculated [-2(log likelihood reduced — log likelihood full)] to test H_0 that at least one of the regression coefficients equals zero. To test individual coefficients, Wald statistics can be measured, each one being the estimated regression coefficient divided by its standard error. These wald statistics are the equivalent of t tests for partial regression coefficients in multiple linear regression and can be compared to the standard normal (Z) distribution. Our reservations about wald tests (lack of power with small sample sizes) described for a simple logistic regression apply equally here. To achieved a advanced results can be achieved by fit a series of reduced models and compare their fit to the full model. To calculate H_0 that β_1 equals zero, we compare the fit of the full model given in equation (8) to the fit of a reduced model based on H_0 being true.

$$g(x) = \beta_0 + \beta_2 x_{i2} + ... + \beta_p x_{ip}$$

If the G^2 test is significant, it can be concluded that the inclusion of x_{i1} as a predictor makes the full model a best match to our data than the reduced model and therefore H_0 is rejected. A similar model can be done to comparison test for the other predictors.

4 - Other Methods for Significance Test

4-1 Hosmer and Lemeshow's Goodness of Fit Test

Hosmer and Lemeshow test is a subsitution which can be calculated by; divide subjects into 10 ordered groups of subjects, compares the number actually in the each group (observed) to number predicted by the logistic regression model (predicted). The 10 ordered groups are created based on their estimated probability, those with estimated probability below 0.1 form one group and so on up to those with probability 0.9 to 1.0. Each of these categories is further divided into two groups based on the actual observed outcome variable (success, failure). The expected frequencies for each of the cells are obtained from the 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. If the H-L goodness-of -fit test statistic is greater than 0.05 as

it should be for well-fitting models, we fail to reject the null hypothesis is failed to be rejected due to there is no difference between observed and model-predicted values, implying that the model's estimates fit the data at an acceptable level [13]. That is well-fitting models show non-significance on the H-L goodness-of-fit test. The desirable outcome of non-significance indicates that the model prediction does not significantly differ from the observed.

4-2 Omnibus Test of Model Coefficients

Omnibus test can be used as a substitute statistical to the Hosmer-Lemshow test. It can be employed whenever there is a significant differe between the model and predictors in the intercept. This test is expllicate as a test of the capability of all predictors in the model jointly to predict the response (dependent) variable. Significance corresponds to the conclusion that there is enough match between the data and the model, this mean that at least one of the predictors is significantly related to the response variable [9].

5- Description of Data

The data used in this study were collected from records of (100) patients suffering from the National Center for the treatment and research of diabetes, moreover, (100) healthy persons information where used to compare with the original data. The logistic regression analysis was then performed with two groups (patient, not patient) and nine predictor variables that we believe they cause the disease. The variables for each group are:

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

B. Nine independent variables described below:

1. Age (X1)
2. Sex (0 for Male) (1 for Female) (X2)
3. Uric Acid (X3)
4. Blood Urea (X4)
5. Cholesterol (X5)
6. Serum Triglyceride (S.T.G) (X6)
7. Smoking (0 for not smoker) (1 for smoker) (X4)
8. Heredity (X8) (0 for not Heredity) (1 for Heredity) (X9)
9. Alcohol (0 for not Alcohol) (1 for Alcohol) (X9)

The statistical program SPSS and Mat lab were employed to perform the required calculations.

Logistic Regression Analys

Employing the statistical program SPSS to elucidate the required results which obtained and arranged according to the following tables.
Table (1) summarized the data login to the analysis including the sample size studied and the missing data.

Table (1) Case Processing Summary

Unweight 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

The code (or symbol) of the dependent variable values is displayed in table (2)

Table (2) Dependent Variable Encoding

Original Value	Internal Value
Healthy	0
Patient	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 calculate the optimal estimates for the model coefficients. The minimum value of -2log likelihood was obtained at the eighth iteration, it was equal to **68.868** 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 become very slow after the fifth iteration as it shown in table (3), thus, we can say that 6, 7 and 8 are similar with very simple differences. Therefore, the process terminate at the eighth iteration and we regard its estimated coefficients to be the best estimated coefficients that can be obtained.

Table (3) Iteration History

Eration	-2 Log likelihood	Coefficients									
		Const.	SEX	AGE	B. UREA	S. CHOL	Serum Triglyceride S.T.G	S. URIC	Smoker	Hero	Cohol
Step 1	122.632	-8.116	-.203	.023	.025	.017	.007	.157	-.084	.578	.075
2	87.714	-13.463	-.505	.034	.033	.030	.013	.222	-.148	1.143	.141
3	73.939	-19.233	-.846	.042	.049	.042	.021	.286	-.215	1.704	.228
4	69.624	-24.765	-1.079	.048	.074	.054	.028	.344	-.252	2.066	.269
5	68.900	-28.236	-1.188	.049	.092	.061	.033	.377	-.251	2.204	.248
6	68.868	-29.165	-1.217	.049	.098	.063	.034	.386	-.244	2.230	.232
7	68.868	-29.217	-1.218	.049	.098	.063	.034	.387	-.243	2.231	.230
8	68.868	-29.217	-1.218	.049	.098	.063	.034	.387	-.243	2.231	.230

Table (4) elucidate the parameters of the optimal model obtained from the 8^{th} iteration given in table (3). All estimated coefficients $(\beta_0, \beta_1, \dots, \beta_9)$ as well as the standard error and the wald statistic for each estimated coefficient and the upper and lower bound for exp(β) are included

Table (4) Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp (B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1(a)	sex	-1.218	.996	1.495	1	.221	.296	.042	2.085
	AGE	.049	.024	4.112	1	.043	1.050	1.002	1.102
	B.UREA	.098		.536	1	.464	1.103	.849	1.433
	S.CHOL	.063	.014	21.108	1	.000	1.065	1.037	1.094
	S.T.G	.034	.011	9.714	1	.002	1.035	1.013	1.058
	S.URIC	.387	.290	1.778	1	.182	1.472	.834	2.600
	smoker	-.243	.671	.131	1	.717	.784	.211	2.922
	hero	2.231	1.007	4.908	1	.027	9.313	1.293	67.063
	khole	.230	.736	.098	1	.754	1.259	.297	5.332
	Constant	-29.217	6.822	18.340	1	.000	.000		

The logistic regression equation can be concluded from the above table as: *lo*

$$\text{Log } \frac{\hat{p}}{1-\hat{p}} = -29.217 - 1.218\ x_1 + 0.49x_2 + 0.98x_3 + 0.63x_4 + 0.34x_5 + 0.378x_6 + 0.243x_7 + 2.231x_8 + 0.230x_9 \quad \ldots\ldots (9)$$

To test the efficiency and the goodness of fit for the whole model we use the log likelihood ration can be used. Which follows the relationship

$$\chi^2 = 2(\text{Log } L_0 - \text{Log } L_1)$$

Where L_1 is the value of likelihood function with i variables, L_0 is the value of likelihood function with (i-1) variables. The value of χ^2 was found to be **208.391** which is significant at the level α less than 0.001 and 9 degrees of freedom with sig=0 which ensure the significant of the whole fitted model as shown in table (5)

Table (5) Omnibus Tests of Model Coefficients

		Chi-square	Df	Sig.
Step 1	Step	208.391	9	.000
	Block	208.391	9	.000
	Model	208.391	9	.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 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

		y = healthy		y = patient		Total
		Observed	Expected	Observed	Expected	
Step 1	1	20	19.998	0	.002	20
	2	20	19.985	0	.015	20
	3	20	19.844	0	.156	20
	4	18	18.960	2	1.040	20
	5	14	13.753	6	6.247	20
	6	7	5.901	13	14.099	20
	7	1	1.247	19	18.753	20
	8	0	.276	20	19.724	20
	9	0	.034	20	19.966	20
	10	0	.002	20	19.998	20

The value of χ^2 was found to be 1.781 with 8 degrees of freedom and significant value 0.987 as shown in table (7), hence, we accept the null hypothesis can be accepted due to 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	1.781	8	.987

The percentage of the correct classification is presented in table (8). The overall percentage was equal to 91%. It calculated as $\frac{91+92}{200} \times 100\%$ while the percentage of not correct classification was found to be $\frac{9+8}{200} \times 100\% = 8.5\%$ and this is an excellent indicator that the model represent the data very well.

Table (8) classification Table

	Observed		Predicted		
	Y				Percentage Correct
	Healthy		Patient		
Step 1	Y	Healthy	91	9	91.0
		Patient	8	92	92.0
	Overall Percentage				91.5

Conclusions

1- From our practical study we concluded that the important variables which is the main reasons to get the diabetes appeared in the order S.chol, followed by STG then Heredity and finally the age variable while other variables were not significant in influencing the dependent variable y.

2- The overall percentage of correct classification (91.5%) indicate that the logistic regression approach has a good performance for analyzing the data of our problem.

References

[1] Anderson, J.A. (1982), Logistic Discrimination, John Wiley, New York, London

[2] American Diabetes Association (2014),"Standards of medical care in diabetes" p.S14Jan;37 Suppl 1:S14-80. doi: 10.2337/dc14-S014.

[3] Brittany, N. (2015)."Fermentable Carbohydrate intake and differences in health Parameters in US College Students". M.SC., thesis University of Rhodeisland.

[4] David, W. Hosmer Jr., Stanley Lemeshow, Rodney X.(2013); "Applied Logistic Regression". Sturdivant chapter 9 Logistic Regression models for the analysis of correlated data.

[5] Draper, N.R. and Smith, H. (1966). Applied Regression Analysis, John Wiley, New York, London.

[6] Hu, F and Pro v Con Debate .(2013); "Role of sugar sweetened beverages in obesity Resolved: there is sufficient scientific evidence that decreasing sugar- sweetened beverage consumption will reduce the prevalence of obesity and obesity-related diseases" International Association for the Study of Obesity Vol 14 , pp:606–619

[7] Kleinbaum, D.G kupper,L.L.and muller.K.E.(1988) :Applied Regression Analysis and the other methods DWS-KENT publishing company, a division of wadsworth p 317.

[8] Li-Ying Huang, Wen-Yi Shau,Hseng-Long,Tsung-Tai Chen, Jun Yi Hsieh .(2015); "A model measuring therapeutic inertia and the associated factors among diabetes patients: A nationwide population-based study in Taiwan" Volume 55, Issue 1 January Pages 17–24 j of clinical pharmalogy

[9] Permpeh, E. A. (2009) “comparative study of the logistic regression analysis and the discriminant analysis" M.SC., thesis. University of cape coast.

[10] Pullinger, A.G., Seligner, D.A. & Gornbein, J.A. (1993). A multiple logistic regression analysis of the risk and relative odds temporamandibular disorder as function of common occlusal features. Journal of dental research, 72. (6) 968-979.

[11] Santiago Palacio, MD; Leslie A. McClure, Oscar R. Benavente, MD; Carlos Bazan, MD; Pablo Pergola, MD; Robert G. Hart .(2014) ; " Lacunar Strokes in Patients With Diabetes Mellitus: Risk Factors, Infarct Location, and Prognosis The Secondary Prevention of Small Subcortical Strokes Study",Stroke journal of the American Heart Association, September , pp.

[12] الراوي،خاشع محمود (١٩٨٧) ، "المدخل الى تحليل الانحدار"، جامعة الموصل.

[13] فهمي ،محمد شامل (٢٠٠٥)، " الاحصاء بلا معاناةالمفاهيم مع التطبيقات باستخدام spss الجزء الثاني ، مركز البحوث