Utilize Logistic Regression Model to Classify Children with Intestinal Parasites

Altaiyb Omer Ahmed Mohmmed[1], Lana Salah Abdelati Hamdnalah[2], Haider Jamel Allah Mohmmed[3], Elsamoual Mohammed Kurtokeila Mamour[4*]

[1,2,3,4]Department of Statistics, College of science
Sudan University of Science and Technology - Sudan
*elsamoual@sustech.edu

Abstract: *This paper aims to predict the probability of infection of children infected with intestinal parasites, and determining the most important factors affecting the disease and their strength and impact by using logistic regression model. Data were collected from Bahri Teaching Hospital, Internal Medicine Department, from the record of reviews of injured children with sample size of 300 individuals, 150 were infected. The data focused on the variables of gender, age, family history disease, drinking contaminated water, fever, anaemia, diarrhoea, abdominal pain and vomiting. The result of the study that there is a significant effect of vomiting, abdominal pain, diarrhoea and fever on the infection of intestinal parasites, the model used has a high accuracy, as it has a high classification ability (70%) and the probability of error in classification was very small, as it reached (30%). This indicates the ability of model to correctly classify the sample members.*

Keyword: Discriminant Analysis, Logistic Regression, Classification, Intestinal Parasites, Infection

1. INTRODUCTION

Logistic regression method is the one of important methods that is used in many applied aspects such as the medical, social and economic. The logistic regression model is one of the important statistical models used in multivariate analysis when predicting the values of a qualitative dependent variable whose data are measured nominally or ordinally when the values of the different independent variables (quantitative or qualitative) are known. Although there are many multivariate statistical analysis methods that are used to study and analyze the relationship between dependent variables and a set of independent variables, there are some cases in which it is difficult to use these methods. The multiple regression analysis model assumes that the dependent variable is quantitative, and this is not achieved in many cases where the dependent variable is qualitative, and therefore it is difficult to apply the multiple regression analysis model. Using the discriminant analysis which allows predicting the groups of the dependent variable to which the individual belongs according to the odds ratio, this method requires some conditions or special assumptions such as assuming a multivariate normality for independent variables and also equals the variance co-variance matrices of the independent variables of the groups and this condition is not fulfilled in cases involving qualitative independent variables [13][16]. In such cases recommended another appropriate method, as the discriminant analysis method results in biased and inconsistent estimators [7].

Cox proposed a logistic regression method by which to predict the presence or absence of an event, in other words, studying the effect of a group of independent variables in the case that the dependent variable is a qualitative variable with two or more categories [3]. It is distinguished from the discriminant analysis method in that it requires fewer assumptions or conditions for its application, and even with the availability of the required conditions in the discriminant analysis model, it gives better results [9]. This model is also characterized by the possibility of explaining the relationships between variables with what is known as the Odds Ratio.

2. MATERIALS AND METHODS

Data collection: Data were collected from Bahri Teaching Hospital, Internal Medicine Department, from the record of reviews of injured children with sample size of 300 individuals, 150 were infected. The data focused on the variables of gender, age, family history disease, drinking contaminated water, fever, anaemia, diarrhoea, abdominal pain and vomiting.

Intestinal parasite: An intestinal parasite infection is a condition in which a parasite infects the gastro-intestinal tract of humans and other animals. Such parasites can live anywhere in the body, but most prefer the intestinal wall. Routes of exposure and infection include ingestion of undercooked meat, drinking infected water, fecal-oral transmission and skin absorption. Signs and symptoms depend on the type of infection. Intestinal parasites produce a variety of symptoms in those affected, most of which manifest themselves in gastrointestinal complications and general weakness [17]. Gastrointestinal conditions include inflammation of the small and/or large intestine, diarrhea/dysentery, abdominal pains, and nausea/vomiting. These symptoms negatively impact nutritional status, including decreased absorption of micronutrients, loss of appetite, weight loss, and intestinal blood loss that can often result in anemia. It may also cause physical and mental disabilities, delayed growth in children [2].

Logistic Regression model: Logistic regression is considered one of the general non-linear models because it is considered optimal in the analysis of phenomena with two possibilities. It is a statistical technique that has found multiple applications in many fields, because it represents an alternative to linear regression when the dependent variable is qualitative and some independent variables are also qualitative. This type of regression model depends on converting qualitative variables into dummy variables. There are several methods for this conversion, the most prominent of which is what is known as binary conversion. Logistic regression can also be used to distinguish (classify) between two or more groups as an alternative to analyzing the characteristic function when there are some qualitative independent variables or when there are some variables that do not normaly distributed[15].

Binary Logistics Model: Suppose that (y) the dependent variable takes the value (one) if a certain event occurs and the value (zero) if that event does not occur, that is, when the dependent variable has only two values and that (x) is a quantitative or qualitative variable called the binary logistic regression model. If we draw the curve that represents the relationship between (y) as a dependent variable and (x) as an independent variable, we will find that the function will take the form of (S-Shape Curve) where the values of (E(y)) are limited between (zero and one) and thus accumulate between these two values, also the shape of the incremental and decreasing function depends on the β_j sign, and the dependent variable in the estimated model is a Bernoulli variable that takes one of the two values (0, 1).

The function of this curve is the logistic function, so when (y) is a two-valued variable, we find that:

$$E(y/x)=\frac{e^{(\beta_0+\beta_1 x)}}{1+e^{(\beta_0+\beta_1 x)}} \quad \_\_\_\_(1)$$

Equation (1) is called binary logistic function; and one of the characteristics of the logistic function is that it can be converted to a linear function with an appropriate transformation. If we assume:

$$E(y/x)=\pi(x) \quad \_\_\_\_(2)$$

And we used the conversion:

$$g(x)=Ln\left[\frac{\pi(x)}{1-\pi(x)}\right] \quad \_\_\_\_(3)$$

Where

$\pi(x)$ ≡ the probability of phenomenon.

[1- $\pi(x)$] ≡ the probability of the absence of the phenomenon.

Where we find that:

$$g(x)=\beta_0+\beta_1 x \quad \_\_\_\_(4)$$

$g(x)$ Called the logit function, and the logistic model is important because the expression (e^{β_1}) gives the odds ratio.

To clarify the idea of the odds ratio, suppose that (y) takes the value (one) if the person has a specific disease and the value (zero) if he is not infected, also suppose that (x) takes the value (one) if the person applies to him a certain characteristic, and the value (zero) if he does not apply this characteristic[12].

The risk of a person having the trait (x=1) known as:

$$Odds(1)=\frac{The\ number\ of\ people\ who\ have\ the\ disease\ among\ those\ who\ have\ the\ trait}{The\ number\ of\ people\ who\ do\ not\ have\ the\ disease\ among\ those\ who\ have\ the\ trait}$$

The risk of a person being infected by a person who does not have the trait (x=0) known as:

$$Odds(0)=\frac{The\ number\ of\ people\ who\ have\ the\ disease\ among\ those\ who\ do\ not\ have\ the\ trait}{The\ number\ of\ people\ who\ do\ not\ have\ the\ disease\ among\ those\ who\ do\ not\ have\ the\ trait}$$

The odds ratio is defined as the ratio between those who have the trait and those who don't, known as:

$$OR=\frac{Odds(1)}{Odds(0)}$$

This ratio represents the risk of disease for people who have the trait compared to people who do not have the trait [11].

Multiple Logistics Model: If we have a binary dependent variable (Y) that takes the values (zero and one) for n number of independent variables

$$X=(X_1,X_2,\ldots,X_n) \quad \_\_\_\_(5)$$

The model in equation (1) can be generalized to take the following form:

$$P=E(y/x)=\frac{e^{(\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n)}}{1+e^{(\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n)}} \quad \_\_\_\_(6)$$

Equation (6) called the multiple logistic model[11], and the logit function takes the following form:

$$g(x_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \quad \_\_\_\_\_\_ (7)$$

Estimation of multiple logistic regression parameters: Kleinbaum and Klein stated that maximum likelihood is often used for the estimation of a parameter of either a linear or a nonlinear model [5]. To clarify the application of this method, we first notice that the form of the maximum likelihood function for (x_i, y_i) is:

$$\pi(x_i)^{y_i}\left[1-\pi(x_i)^{1-y_i}\right] \quad \_\_\_\_\_\_ (8)$$

By assuming that the observations are independent, the function for a sample of size n is as follows:

$$L(\beta) = \prod_{i=1}^{n}\left(\pi(x_i)^{y_i}\left[1-\pi(x_i)^{1-y_i}\right]\right) \quad \_\_\_\_\_\_ (9)$$

Test the significance of model parameters: Maximum likelihood estimation is used to compute logistic model estimates. The iterative process finds the minimal discrepancy between the observed response, Y, and the predicted response, Yˆ (see the handout "Maximum Likelihood Estimation"). The resulting summary measure of this discrepancy is the -2 loglikelihood or -2LL, known as the deviance [14]. The now familiar likelihood ratio test is used to compare the deviances of the two models (the null model, L_0 and the full model, L_1) [8].

$$\begin{aligned} G^2 &= Deviance_0 - Deviance_1 \\ &= \left[-2\log(L_0)\right] - \left[-2\log(L_1)\right] = -2\ln\left(\frac{L_0}{L_1}\right) \_\_\_ (10) \end{aligned}$$

The estimated value of G^2 is distributed as a chi-squared value with degree freedom equal to the number of predictors added to the model.

Wald Test: Wald statistics are used to test the significance of individual coefficients in the model and also the value of Wald used to measures the order of relative importance of the affecting factors [4][10], and is calculated as follows:

$$Wald = \left(\frac{\beta}{SE_\beta}\right)^2 \quad \_\_\_\_\_\_ (11)$$

Where:
β is the value of the coefficient of the predictive variable.
SE_β is the standard error of β.

3. RESULTS AND DISCUSSION

The dependent variable (y) is infection where it takes the value (1) when a person has intestinal parasite disease, and takes the value (0) when he is not infected.

The independent variables are: gender (X_1), age (X_2), family history disease(X_3), drinking contaminated water(X_4), fever(X_5), anaemia(X_6), diarrhoea(X_7), abdominal pain(X_8) and vomiting(X_9).

Steps to include affecting factors in a binary logistic regression model:

There are many criteria and measures for survival, elimination and affecting including the probability of F, the level of significance, and the value of the logarithm of maximum (-2 Log likelihood). The steps are:

Step zero: The model step contains only the constant term:

In this step, the model that contains the fixed term only is analyzed, and the value of the log-likelihood of this model (-2 Log Likelihood=415.888), and the results were as follows:

Table 1:

Step		Predicted		
0		Infection (Y)		Percentage correct
		Uninfected	Infected	
	Uninfected	0	150	0%
	Infected	0	150	100%

	Overall Percentage	50%

Prepared by the researcher from the study data using SPSS, 2021

Table (1) shows that 150 children have been infected with intestinal parasites, while 150 are not infected with this disease. Therefore, if the model predicts that all children have been infected, this prediction will be correct 150 times out of 300 by 50%, but if it is expected that children will not be infected, this prediction will be correct 150 times by 50%. Also we find that 0% of the accuracy of the sample who responded and were not infected with intestinal parasites, and 100% of the accuracy of the sample of those who responded and were infected. Overall, the model succeeded in classifying 50% of patients.

Table 2:

Step	β	S.E	Wald	df	Sig.	Exp(β)
0 Constant	0.00	0.112	0.000	1	1.000	1.000

Prepared by the researcher from the study data using SPSS, 2021

For table (2) this step shows the value of the constant only when there are no variables entered as an initial step where we note that the value of the constant limit is zero.

Table 3:

Step	Variable	Score	df	sig
0	X_1	0.013	1	0.908
	X_2	0.464	1	0.496
	X_3	1.405	1	0.236
	χ^2-value	1.754	3	0.625

Prepared by the researcher from the study data using SPSS, 2021

From Table (3), we find that the chi-square statistic for residuals is (1.754), which is a value that is not significant (P-value= 0.625> 0.05), and this statistic indicates that the coefficients of the variables not found in the model are equal to zero, and this means that adding one or more of these predictive variables to the model does not significantly affect its predictive ability. We also find that the Rao's efficient statistic values (Score value) for the excluded variables are not significant. Therefore, adding the variables X_1, X_2, and X_3 may not contribute to the predictive ability of the model at this stage.

Step 1: Add the uncontrollable risk factors (gender, age, family history disease):

In this step, the variables "uncontrollable risk factors: sex, age and family history disease " represented in X_1, X_2 and X_3 respectively are entered into the model, and the results are as follows:

Table 4:

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	414.117	0.006	0.008

Prepared by the researcher from the study data using SPSS, 2021

Table (4) shows the information about the model after adding the variables X_1, X_2, and X_3, where the value of (-2 Log likelihood) decreased to 414.117, which means a change of 1.771. Also the variables (X_1, X_2, and X_3) contribute to the prediction only 0.6% (Cox & Snell R^2 value) and the (Nagelkerke R^2=0.008), this means that only 0.8% of the changes that occur in infection are caused by gender, age and family history disease, which is a very weak and almost non-existent contribution.

Table 5:

Step	Variable	Chi-square	df	sig
1	Step	1.771	3	0.621
	Block	1.771	3	0.621
	Model	1.771	3	0.621

Prepared by the researcher from the study data using SPSS, 2021

From Table (5), we find that the value of the chi-square is 1.771, and this value indicates the model in the first step, while the Block1 indicates the extent of the improvement that occurred in the model since the previous block at step (0). The change in the amount of information justified by the model is not significant because (P-value=0.621>0.05), so entering the variables (X_1, X_2, and X_3) as predictive variables may not improve the ability to predict the probability of infection with intestinal parasites .

Table 6:

Step		Predicted		
		Infection (Y)		Percentage correct
		Uninfected	Infected	
1	Uninfected	104	46	69.3%
	Infected	58	92	38.7%
	Overall Percentage			54%

Prepared by the researcher from the study data using SPSS, 2021

From table (6) the model that contains gender, age, and family history to predict the outcome variable, it correctly classifies 104 children who are not infected with intestinal parasites, and incorrectly classifies 46 uninfected children, meaning that it classifies 69.3% of the cases correctly, as for children infected the model correctly classifies 92 of them, and classifies 58 of them incorrectly, meaning that it correctly classifies 38.7% of cases. The overall accuracy of the classification is 54%. When the model included only the fixed term, it classified 50% of the cases correctly. Now, after entering the variables sex, age, and family history as predictive factors, this value has increased to 54%.

Step 2: Adding the risk factors (drinking contaminated water, fever, anaemia) to step 1:

In this step, new predictive variables are entered into the model, which are physiological risk factors that can be controlled drinking contaminated water, fever and anaemia, represented by X_4, X_5 and X_6 respectively in block 2. The step 2 describes the model in step 1after adding the new predictive variables (X_4, X_5 and X_6).

Table 7:

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
2	409.857	0.020	0.027

Prepared by the researcher from the study data using SPSS, 2021

From Table (7), we find that the effect of adding these new predictive variables to the model is to reduce the value of the -2 Log likelihood to 409.857 (a reduction of 6.032) from the original model, the new predictive variables drinking contaminated water, fever and anaemia (X_4, X_5, and X_6) contribute to the prediction only 2% (Cox & Snell R^2) and the (Nagelkerke R^2) only 2.7% of the changes that occur in infection are caused by gender, age, family history disease, drinking contaminated water, fever and anaemia.

Table 8:

Step	Variable	Chi-square	df	sig
2	Step	4.261	3	0.235
	Block	4.261	3	0.235
	Model	6.032	3	0.420

Prepared by the researcher from the study data using SPSS, 2021

From Table (8), we find that the additional improvement by block 2 is not important because the chi-square value of block 2 is not significant and is equal to 4.261 with (P-value=0.420>0.05) and this is evidence that including the new predictive variables (X_4, X_5, and X_6) in the model may not significantly improve the ability to prediction and classification on the model.

Table 9:

Step		Predicted		
		Infection (Y)		Percentage correct
		Uninfected	Infected	
2	Uninfected	86	64	57.3%
	Infected	73	77	51.3%
	Overall Percentage			54.3%

Prepared by the researcher from the study data using SPSS, 2021

From table (9) the model that includes the factors X_1, X_2, X_3, X_4, X_5 and X_6 correctly classifies 86 children who were not infected with intestinal parasites, and incorrect classification of 64 uninfected children, meaning that 57.3% of cases were correctly classified. As for the children who were infected the model correctly classifies 77 of them, and classifies 73 of them incorrectly, meaning that it correctly classifies 51.3% of cases. The overall accuracy of the classification is 54.3%.

When the model contains risk factors that cannot be controlled, it classified 54% of the cases correctly, after entering the variables drinking contaminated water, fever and anaemia as predictive factors, the overall accuracy of classification increased to 54.3%.

Table 10:

Variable	β	S.E	Wald	df	Sig.	Exp(β)	95% C.I for Exp(β)	
							Lower	Upper
X_1	0.024	0.234	0.010	1	0.920	1.024	0.647	1.621
X_2	0.025	0.039	0.419	1	0.518	1.025	0.951	1.106
X_3	-0.574	0.496	1.337	1	0.248	0.563	0.213	1.490
X_4	0.083	0.655	0.016	1	0.899	1.086	0.301	3.920
X_5	**-0.479**	**0.237**	**4.087**	**1**	**0.043**	**0.620**	**0.389**	**0.986**
X_6	-0.240	0.541	0.196	1	0.658	0.787	0.272	2.273
Constant	0.842	0.799	1.110	1	0.292	2.321		

Prepared by the researcher from the study data using SPSS, 2021

Table (10) includes the predictive variables (Affecting factors) uncontrollable risk factors and physiological risk factors, and they are not significant, except for variable X_5 (fever) which is significant.

From this table, the value of Exp(β) for the variable X_5 (fever), which is equal to 0.620 indicates that if the value of the variable X_5 decreased by one unit, that is, it changed from fever to non-fever (the sign is negative), likelihood of infection of intestinal parasite disease decreases (because the value of Exp (β) is less than 1) with confidence interval between (0.389 and 0.986) at 95%.

Step 3: Adding the symptomatic risk factors (diarrhoea, abdominal pain, vomiting) to the second step:

In this step, new predictive variables are entered into the model, which are symptomatic risk factors that can be changed diarrhoea, abdominal pain, and vomiting, represented by X_7, X_8 and X_9 respectively. The step 3 describes the model in step 2 after adding the new predictive variables (X_7, X_8 and X_9).

Table 11:

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
3	332.037	0.244	0.325

Prepared by the researcher from the study data using SPSS, 2021

From Table (11) we find that the effect of adding these new predictive variables to the model is to reduce the value of the -2 Log likelihood to 409.857 (a reduction of 78.893) from the original model. the new predictive variables diarrhoea, abdominal pain, and vomiting (X_7, X_8 and X_9) contribute to the prediction by 24.4% (Cox & Snell R^2) and the (Nagelkerke R^2=0.325) this means that 32.5% of the changes that occur in infection are caused by gender, age, family history disease, drinking contaminated water, fever, anaemia, diarrhoea, abdominal pain and vomiting.

Table 12:

Step	Variable	Chi-square	df	sig
3	Step	77.82	3	0.000
	Block	77.82	3	0.000
	Model	83.852	9	0.000

Prepared by the researcher from the study data using SPSS, 2021

From Table (12), we find that the additional improvement by block 3 is important because the chi-square value of block 3 is significant and is equal to 77.82 with (P-value=0.000<0.05) and this is evidence that including the new predictive variables (X_7, X_8 and X_9) in the model has significantly improved the ability to prediction and classification.

Table 13:

Step		Predicted		
		Infection (Y)		Percentage correct
		Uninfected	Infected	
3	Uninfected	100	50	66.7%
	Infected	40	110	73.3%
	Overall Percentage			70.0%

Prepared by the researcher from the study data using SPSS, 2021

From table (13) the model that includes the factors X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8 and X_9 correctly classifies 100 children who were not infected with intestinal parasites, and incorrect classification of 50 uninfected children, meaning that 66.7% of cases were correctly classified. As for the children who were infected the model correctly classifies 110 of them, and classifies 40 of them incorrectly, meaning that it correctly classifies 73.3% of cases. The overall accuracy of the classification is 70%.

When the model contains uncontrollable risk factors and the controllable risk factors, it classified 54.3% of the cases correctly. Now after entering the variables diarrhoea, abdominal pain and vomiting as predictive factors, the overall accuracy of classification increased to 70%.

Table 14:

Variable	β	S.E	Wald	df	Sig.	$Exp(\beta)$	95% C.I for Exp(β)	
							Lower	Upper
X_1	-0.139	0.269	0.265	1	0.607	0.871	0.514	1.475
X_2	0.067	0.044	2.295	1	0.130	1.069	0.980	1.166
X_3	-0.267	0.564	0.224	1	0.636	0.766	0.253	2.313
X_4	-0.429	0.752	0.326	1	0.568	0.651	0.149	2.842
X_5	**-0.564**	**0.273**	**4.271**	**1**	**0.039**	**0.569**	**0.333**	**0.971**
X_6	-0.432	0.637	0.460	1	0.498	0.650	0.187	2.262

X_7	0.626	0.270	5.358	1	0.021	1.870	1.101	3.178
X_8	1.245	0.295	17.804	1	0.000	3.472	1.947	6.190
X_9	1.955	0.314	38.868	1	0.000	7.061	3.820	13.054
Constant	-1.516	0.969	2.447	1	0.118	0.220		

Prepared by the researcher from the study data using SPSS, 2021

Table (14) includes the predictive variables (Affecting factors) uncontrollable risk factors, physiological risk factors and symptomatic risk factors X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8 and X_9.

The uncontrollable risk factors gender, age and family history disease (X_1, X_2 and X_3) are not important for classification and prediction of infection because (all P-values >0.05), and the controllable risk factors drinking contaminated water, fever and anaemia (X_4, X_5 and X_6) we find that only fever (X_5) is important for classification and prediction of infection because (P-value=0.039 <0.05), while the symptomatic risk factors diarrhoea, abdominal pain and vomiting (X_7, X_8 and X_9) are very important for classification and prediction of infection with intestinal parasites because (all P-values <0.05).

Therefore, we conclude that gender, age, family history disease, drinking contaminated water, and anemia have no effect on infection with intestinal parasites, which means that the child will infected regardless of his gender, age and family history disease if there is vomiting, abdominal pain, diarrhea and fever. This is evidence that the symptoms are more effective and classified to predict the infection.

From this table, the value of Exp(β) for the variable X_7 (diarrhoea), which is equal to 1.870 indicates that if the value of the variable X_7 changed from(0 ≡ no diarrhea) to (1≡yes diarrhea) the likelihood of infection of intestinal parasite disease will increases because the value of Exp(β) is greater than 1 with confidence interval between (1.101 and 3.178) at 95%. Since the two values are greater than 1, we are confident that the relationship between the diarrhoea and infection with intestinal parasites in this sample is a valid relationship for the entire study population. This means that the likelihood of infection with intestinal parasites in those who have diarrhoea is two times greater.

the value of Exp(β) for the variable X_8 (abdominal pain), which is equal to 3.472 indicates that if the value of the variable X_8 changed from(0 ≡ no abdominal pain) to (1≡yes abdominal pain) the likelihood of infection of intestinal parasite disease will increases because the value of Exp(β) is greater than 1 with confidence interval between (1.947 and 6.190) at 95%. Since the two values are greater than 1, we are confident that the relationship between abdominal pain and infection with intestinal parasites in this sample is a valid relationship for the entire study population. This means that the likelihood of infection with intestinal parasites in those who have abdominal pain is three times greater.

the value of Exp(β) for the variable X_9 (vomiting), which is equal to 7.061 indicates that if the value of the variable X_9 changed from(0 ≡ no vomiting) to (1≡yes vomiting) the likelihood of infection of intestinal parasite disease will increases because the value of Exp(β) is greater than 1 with confidence interval between (3.820 and 13.054) at 95%. Since the two values are greater than 1, we are confident that the relationship between vomiting and infection with intestinal parasites in this sample is a valid relationship for the entire study population. This means that the likelihood of infection with intestinal parasites in those who have vomiting is seven times greater.

Contribution of the affecting factors in classification and prediction in the logistic model:

We find that the most important factors affecting infection are fever, diarrhoea, abdominal pain and vomiting X_5, X_7, X_8 and X_9 respectively because (all P-values <0.05). Wald's value indicates the variable's contribution to classification and prediction of infection, the large value of Wald means that the percentage of the variable contribution is greater, and accordingly, vomiting (X_9) contributes a greater percentage in classifying the two groups as uninfected and infected with a percentage of 38.9%, followed by abdominal pain (X_8) 17.8%, then diarrhoea (X_7) by 5.4%, and finally fever (X_5) by 4.3%.

Estimated binary logistic regression model

We conclude through these three steps is to determine the equation of the binary logistic regression model in terms of the affecting factors on infection of intestinal parasites according to health indicators, then the equation takes the following form:

$$P = E\left(y/x\right) = \frac{e^{\left(\beta_0+\beta_9 X_9+\beta_8 X_8+\beta_7 X_7+\beta_5 X_5\right)}}{1+e^{\left(\beta_0+\beta_9 X_9+\beta_8 X_8+\beta_7 X_7+\beta_5 X_5\right)}} \_\_\_\_\_\_ (12)$$

The affecting factors included in this model, according to the statistical criteria that are related to the (dependent) infection variable that contributes to the classification by 24.4% (Cox & Snell R^2) and they explain 32.5% of the change that occurs to the infection variable (Nagelkerke R^2), this factors are X_5, X_7, X_8 and X_9. And by substituting the values of (X_5, X_7, X_8 and X_9) β's into equation (12) from table (14), then the model is:

$$P = E\left(y/x\right) = \frac{e^{\left(-1.516+1.955X_9+1.245X_8+0.626X_7-0.564X_5\right)}}{1+e^{\left(-1.516+1.955X_9+1.245X_8+0.626X_7-0.564X_5\right)}} \_\_\_\_\_\_ (13)$$

This model is able to classify patients as infected or non-infected and predict by 70%.

Classification and prediction of infection using the estimated binary logistic model:

If it gives us data taken from a child who has vomiting (X_9=1) and does not have abdominal pain (X_8=0), not have diarrhoea (X_7=0) and does not have a fever (X_5=0), then by substituting these data in the prediction equation (13), we get the following:

$$P=E\left(y/x\right)=\frac{e^{\left(-1.516+1.955(1)+1.245(0)+0.626(0)-0.564(0)\right)}}{1+e^{\left(-1.516+1.955(1)+1.245(0)+0.626(0)-0.564(0)\right)}}=\frac{e^{0.439}}{1+e^{0.439}}=\frac{1.55}{2.55}=0.608$$

There is a 60.8% chance that this child will be classified as infected with intestinal parasites if he has vomiting. If he does not have vomiting (X_9=0) with the values of the rest of the factors remaining as they were, the equation will be:

$$P=E\left(y/x\right)=\frac{e^{\left(-1.516+1.955(0)+1.245(0)+0.626(0)-0.564(0)\right)}}{1+e^{\left(-1.516+1.955(0)+1.245(0)+0.626(0)-0.564(0)\right)}}=\frac{e^{-1.516}}{1+e^{-1.516}}=\frac{0.22}{1.22}=0.1803$$

There is an 18.03% chance that this child will be classified as infected with intestinal parasites if he does not have vomiting. By using this information, it is reasonable to predict that a child will have the above data in the case of has vomiting, he will be infected with intestinal parasites with an expected probability (0.608), and he will be uninfected, with an expected probability (1-0.608=0.392). And in the case of does not have vomiting, he will be infected with intestinal parasites with an expected probability (0.1803), and he will be uninfected, with an expected probability (1-0.1803=0.8197).

By using equation (13), we can classify when there are data taken from a child who has vomiting (X_9=1) and have abdominal pain (X_8=1), have diarrhoea (X_7=1) and have a fever (X_5=1), then by substituting these data in the prediction equation (13), we get the following:

$$P=E\left(y/x\right)=\frac{e^{\left(-1.516+1.955(1)+1.245(1)+0.626(1)-0.564(1)\right)}}{1+e^{\left(-1.516+1.955(1)+1.245(1)+0.626(1)-0.564(1)\right)}}=\frac{e^{1.746}}{1+e^{1.746}}=\frac{5.732}{6.732}=0.878$$

There is an 87.8% chance that this child will be classified as infected with intestinal parasites if he has fever. If he does not have fever (X_5=0) with the values of the rest of the factors remaining as they were, the equation will be:

$$P=E\left(y/x\right)=\frac{e^{\left(-1.516+1.955(1)+1.245(1)+0.626(1)-0.564(0)\right)}}{1+e^{\left(-1.516+1.955(1)+1.245(1)+0.626(1)-0.564(0)\right)}}=\frac{e^{2.31}}{1+e^{2.31}}=\frac{10.07}{11.07}=0.9097$$

There is an 90.97% chance that this child will be classified as infected with intestinal parasites if he does not have fever. By using this information, it is reasonable to predict that a child will have the above data in the case of have fever, he will be infected with intestinal parasites with an expected probability (0.878), and he will be uninfected, with an expected probability (1-0.878=0.122). And in the case of does not have fever, he will be infected with intestinal parasites with an expected probability (0.9097), and he will be uninfected, with an expected probability (1-0.9097=0.0903).

4. CONCLUSION

From foregoing, we find that there are some factors that have no effect on infection with intestinal parasites; these factors are gender, age, family history disease, drinking contaminated water and anaemia, and we find that the family history disease is not significant and has no effect on the infection with intestinal parasites, which confirms that this disease is not hereditary. Also there are some factors that have effect on infection with intestinal parasites; these factors are fever, diarrhoea, abdominal pain and vomiting, tthese factors are important in terms of effect on the infection, where vomiting is most important, followed by abdominal pain, then diarrhoea, and finally fever.

The logistic regression model with these affecting factors has the ability to classify and predict with 70%. The chance of classifying a child with intestinal parasites in the absence of any of the affecting factors (vomiting, abdominal pain, diarrhoea and fever) is 18% and the chance of being classified as infected with intestinal parasites in case of having these factors is 80%.

REFERENCES

[1] Agresti, Alan (2007). Categorical Data Analysis. Second edition. New York: Johnson Wiley & Sons, Inc.

[2] *Ashtiani, M. T. H.; Monajemzadeh, M.; Saghi, B.; Shams, S.; Mortazavi, S. H.; Khaki, S.; Mohseni, N.; Kashi, L.;* Nikmanesh, B. (2011).

[3] Cox, D.R. and Snell, E. J. (1989). The Analysis of Binary Data,London: Chapman and Hall.

[4] Cragg, J.G. & Uhler, R.S. (1970). "The demand for automobiles." The Canadian Journal of Economics, 3: 386-406.

[5] D.G. Kleinbaum and M. Klein (2011), Survival Analysis, Springer, Atlanta.

[6] David.W. Hosmer Jr., Stanley Lemeshow, Rodney X. Sturdivant (2013), Applied Logistic Regression, Third Edition, John Wiley & Sons, Inc.

[7] Halperin, M., Blackwelder, W.C.., and J.I. (1971). "Estimation of the Multivariate Logistic Risk Function: A Comparison of the Discriminant Function and Maximum Likelihood Approaches", Journal of Chronic Diseases. Vol 24, PP. 125-158.

[8] Hauck Jr, W. W., & Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. Journal of the american statistical association, 72(360a), 851-853.

[9] Hosmer, David W., and Stanley Lemeshow (1989). "Applied Logistic Regression" New York: John Wiley & Sons.

[10] Ennings, D. E. (1986). Judging inference adequacy in logistic regression. Journal of the American Statistical Association, 81(394), 471-476.

[11] Klein Baum, D.G. and Kipper, L.L. (1978). "Applied Regression and Other Multivariable Methods ", Duxbury Press North Situate, Mass.

[12] Kleinbaum, D.G ; Kipper, L.L.; and Muller K.E. (1988). Applied Regression Analysis and Other Multivariable Methods, Second edition. Pws-Kent Publishing Company. Boston.

[13] Koneke, J. D (1982). "Discriminant Analysis with Discrete and Continuous Variable". Biometric, Vol 38.PP. 191 – 200.

[14] McCullough and Nelder. (1989). "Generalized Liner Models", London, Chapman Hall, p.175.

[15] Sanforn Weisberg (2005), applied linear regression, third edition, Canada, John Wiley & Sons.

[16] Sayed-Ahmed, M. M. (1994). "A Comparison of the Discriminant Analysis and Logistic Regression Approaches". PH. D. Institute of Statistics & operational Research. Cairo University.

[17] *World Health Organization*. Retrieved (2017).