International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

A Comparison Study between Regression Models for Analyzing Anemia Diseases

Taghreed Al-Said PhD [a]*, Sanaa Al-Marzouki PhD [b], Mona Adham MD. S[c]

[a]*A lecturer of Statistics at AL AZHAR University, Faculty of Commerce, Department of Statistics, Cairo, Egypt, Assistant professor at King Abdul-Aziz University, Faculty of Science, Department of Statistics*

[b]*Assistant professor of Statistics at King Abdul-Aziz University, Faculty of Science, Department of Statistics*

[c]*A student of Master Degree at King Abdul-Aziz University, Faculty of Science, Department of Statistics*

[a]*Email: taghreed_minna@yahoo.com*

[b]*Email: sanaa_no_1@hotmail.com*

[c]*Email: Mony.walid@gmail.com*

Abstract

Regression models are the suitable statistical techniques for drawing inferences about relationships among interrelated variables. These models are applicable in many fields, such as the social field, physical field, biological sciences, business and medical fields. Regression models are perhaps the most used of all data analysis methods. This research interests in comparing regression models and applying these models in analyzing two real data sets of anemia diseases. Also, many evaluating methods are applied in the research to choose between models, determining variables that effective the anemia diseases. The analysis of the results detects the best variables, the suitable model and the best criterion can be used with the medical data.

Keywords: logistic regression models; anemia diseases; Iterative weighted least square methods; r-squared measure; Hosmer-Lemeshow test.

---

* Corresponding author.

1. Introduction

Regression models are the widely statistical used models for analyzing the relationship between dependent and independent variables. The ordinary regression models is the suitable regression models when the outcome variable, Y is a continuous variable. The regression context assumed that there are a set of predictor variables $\boldsymbol{X_1, X_2, \ldots X_p}$ that related with the response variable Y and provide additional information for predicting Y [3]. The ordinary regression models are not appropriate for situations in which Y is categorical such as logistic regression models. Section (2) has details of ordinary and logistic regression models. Section (3) has the estimation and evaluating methods of ordinary and logistic models. Two real data sets are used to apply and compare between the two models. The application will be stated in section (4). The analysis of the results and conclusion are included in section (5). The recommendations is in section (6). Finally, the references will be stated at the end of the research

2. The Ordinary and Logistic Regression Models

Regression analysis is an important statistical tool to analyze data sets in all fields. It enables researcher to identify and characterize the relationships among multiple factors. It also identifies the prognostic relevant risk factors and calculates risk scores [1]. This section introduces the classical statistical regression model that are defined by the ordinary regression models either the simple regression models or the multiple regression models. Also, the details of the logistic regression models are introduced.

2.1. The Ordinary regression models

The regression analysis is the widely used techniques for analyzing multifactor data. It expresses the relationship between a variable of interest, the response variable Y and a set of related predictor variables [9].

There are many types of regression models each one can be applied and suitable in a situation on the basis of some conditions. The simple linear regression model is the simplest model that has only one independent variable and a continuous response variable. The model states the true mean of the dependent variable changes at a constant rate as the value of the independent variable increases or decreases. The functional relationship between the true mean of Y_i, denoted by $E(Y_i)$, and the independent variable has the following equation:

$$E(Y_i) = \beta_0 + \beta_1 X_1 \qquad (1)$$

Where the intercept is β_0, and the rate of change in $E(Y_i)$ per unit change is β_1[1].

The deviation of an observation y_i from its population mean $E(Y_i)$ is taken into account by adding a random error i to form the simple model as follows:

$$E(y_i) = \beta_0 + \beta_1 X_i + \epsilon_i \qquad (2)$$

The subscript i indicates the particular observational unit, i = 1, 2... n. The X_i are the n observations of the independent variable and are assumed to be measured without error. The observed values of X assumed to be a set of known constants. The Y_i and X_i are paired observations measured on every observational unit. The random errors are normal distributed with zero mean and assumed to have common variance σ^2 and is pair wise independent [12].

A multiple regression model describes the relationship between two variables, the continuous dependent variable and many explanatory variables. The multiple model with two independent variables can be defined as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \quad (3)$$

Where the unknown parameters of the model are β_0 , β_1 and β_2. The response variable y. is related with the predictor variables in a linear link. If there are not only two independent variables, then the model for k independent variables can be defined as follows

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \quad (4)$$

The parameters are β_j , $j = 0,1, \dots, k$, that are called the regression coefficients. This model describes a hyperplane in the k - dimensional space of the regression variablesx_j. The parameter β_j represents the expected change in the response variable y per unit change in the explanatory variable x_j when all of the remaining regressor variables x_j $(i \neq j)$ are held constant. For this reason, the parameters β_j , $j = 1,2, \dots, k$, are the partial regression coefficients [9].

2.2. The logistic regression models

The use of logistic regression model back to 1958 by statistician David Cox in 1958 and is appeared in mathematical studies since that time [14]. The logistic regression is the increasingly and popular statistical technique used to model the probability of discrete outcomes. When the logistic regression analyses applied, it yields very powerful insights about what variables are more or less likely to predict event outcome in the population of interest [8]. The logistic model is the suitable used model to explain the relationship between a set of variables and the probability of an event. He focuses on the binary logistic regression and considers the binomial case [4].

There are two types of logistic regression models, the binary logistic model and the multinomial logistic model. In the binary logistic model and for each observation, the response Y can take only one of two possible values which are denoted by 0 for failure and 1 for success. The relation can be described as follows:

$$\text{logit}[\pi(x)] = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = a + \beta x \quad (5)$$

Where the constant of the equation is a , and the regression parameter is β . There is a single explanatory variable X, which is quantitative or qualitative variable. The response variable Y, has the probability of success

at value x denoted by $\pi(x)$. This probability is the parameter for the binomial distribution. The logistic regression model has the linear form for logit of this probability. This formula implies that $\pi(x)$ increases or decrease as an S-shaped function of x. The form of logistic regression model with multiple explanatory variables will be defined as follows:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta^T x_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \quad (6)$$

The parameters β_i refers to the effect of x_i on the log odds, Y=1. The component structural of the logistic model sets the logit link between the probability of success and the linear combination [11].

In a medical study about the risk factors for the carcinogenesis of oral sub-mucous fibrosis in mainland China. The data is related to risk factors and collected using a short structured questionnaire. Results are associated significantly with increased risk for the malignant transformation of oral sub mucous fibrosis [17]. The complemented of diagnostic the breast cancer from the mammogram uses the logistic regression model. The results using logistic regression cross tabulation is to obtain the significant value between the breast cancer factors. The classification table for 130 samples shows that the percentage of correct classification for mammogram is 91.5%. The accuracy is compared with validated samples which are 46 samples where the percentage of correct classification is 67.4% [15].

The nonlinear logistic regression model kernel logistic regression model that based on kernel density estimation is suitable in application for the classification proposes. A suitable comparison between logistic model and the other nonlinear model is made. This approach is important for clinical applications. Results of real datasets reveals that this approach not only achieves superior classification accuracy, but also reduces the computing time as compared to other methods [2].

3. The Estimation and Evaluating methods of the Regression models

There are two estimation methods for estimating the parameters of the regression models, the maximum likelihood method, and the least square method. Also, there are many evaluating measures to test the effect of the variables, the efficiency of the estimation methods, and the best fitted model. This section introduces the suitable estimating and evaluating methods to the proposed model.

3.1. The Estimation methods of the Regression models

The maximum likelihood method for the simple linear regression models is derived by using the joint probability density functions of $Y_1, Y_2, \ldots, Y_n$. It is given as follows under the normal errors assumption:

$$f(Y_1, Y_2, \ldots, Y_n \backslash \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} f(Y_i \backslash \beta_0, \beta_1, \sigma^2)$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2\right\} \quad (7)$$

The log-likelihood function is defined as follows:

$$log\ L(\beta_0, \beta_1, \sigma^2 \backslash Y_1, Y_2, \dots, Y_n)$$

$$= -\frac{n}{2} log(2\pi) - \frac{n}{2} log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \qquad (8)$$

Taking the first partial derivatives of the log-likelihood function, the estimators of the parameters are obtained as follows:

$$\hat{\beta}_{1,ML} = {S_{xy}}/{S_{xx}} \qquad (9)$$

$$\hat{\beta}_{0,ML} = \bar{Y} - \hat{\beta}_{1,ML} \bar{X} \qquad (10)$$

Where the estimated maximum likelihood of the slop is $\hat{\beta}_{1,ML}$, and the estimated value of the intercept is β_0 ,[3].

The maximum likelihood for the multiple linear regression analysis of the parameters' vector $\underline{\beta}$ are the same as the least square estimator. The errors are independently identically distribution with N (0,$\underline{\sigma}^2$) and hence the probability density function of Y is defined as follows:

$$f\left(\underline{Y} \backslash \underline{\beta}, \underline{\sigma}^2\right) = \frac{1}{\left(2\pi \underline{\sigma}^2\right)^{n/2}} exp\left\{-\frac{1}{2\underline{\sigma}^2} \left(\underline{Y} - \underline{X\beta}\right)^T \left(\underline{Y} - \underline{X\beta}\right)\right\} \qquad (11)$$

The log-likelihood function of $\underline{Y}$ can be defined as follows:

$$log\ L\left(\underline{\beta}, \underline{\sigma}^2 \backslash \underline{Y}\right) = -\frac{n}{2} log(2\pi) - \frac{n}{2} log\left(\underline{\sigma}^2\right) - \frac{1}{2\underline{\sigma}^2} \left(\underline{Y} - \underline{X\beta}\right)^T \left(\underline{Y} - \underline{X\beta}\right) \qquad (12)$$

The partial derivatives of the log-likelihood function yield the estimated estimator of β as follows:

$$\underline{\hat{\beta}}_{ML} = \left(\underline{X}^T \underline{X}\right)^{-1} \underline{X}^T \underline{Y} \qquad (13)$$

Where the vector $\underline{\hat{\beta}}_{ML}$ are the estimated vector of the parameters [3, 6].

The maximum likelihood for the logistic regression models is different from the multiple regression models. The likelihood of the sample data is defined as product across all the sampled cases of the probabilities for success as follows:

$$L = \prod_{i=1}^{n} P\left(Y_i \middle| X_{i1}, \dots, X_{ip}\right) = \prod_{i=1}^{n} \left[\frac{e^{\alpha + \sum_{j=1}^{p} \beta_j X_j}}{1 + e^{\alpha + \sum_{j=1}^{p} \beta_j X_j}}\right]^{Y_i} \times \left[\frac{1}{1 + e^{\alpha + \sum_{j=1}^{p} \beta_j X_j}}\right]^{1 - Y_i} \qquad (14)$$

where Y is the outcome variable for the i $^{\underline{th}}$ case (i=0,1), and the values of the predictor variables based on a sample of n cases are $X_{i1}, \ldots, X_{ip}$. The use of Y_i and $1 - Y_i$ as exponents in the equation includes that the likelihood is an appropriate probability dependent on whether $Y_i = 1$ or $Y_i = 0$. Using the methods of calculus, a set of values for α, and the β_j can be calculated as maximizing L, and these resulting values that are known as the maximization process. It is more complicated than the multiple regression analysis for finding estimates. It involves initial guesses for the unknown parameters. This iterative solution procedure is available in popular statistical procedures such as the SPSS and SAS packages [3].

3.2. The least square estimating method

The least squares method minimizes the sum of squares of the vertical distances from each point to the fitted line. The vertical distances represent the errors in the response variable. These errors can be defined as follows:

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i \qquad i=1, 2,\ldots, n \qquad (15)$$

The values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of squares are defined as follows:

$$\hat{\beta}_1 = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} \qquad (16)$$

$$\widehat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (17)$$

The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are the intercept β_0 and the slope β_1 of the line respectively. The least square estimator of β for the multiple regression models is $\hat{\beta} = \left(\underline{X}^T \underline{X}\right)^{-1} \underline{X}^T \underline{Y}$ [13].

For the binary logistic regression where (y=0 or y=1), the iterative reweighted least squares is equivalent minimizing the log - likelihood of the Bernoulli distributed process using the Newton's method. If the problem is written in vector matrix form with parameters$w^T = \beta_0, \beta_1, \beta_3, \ldots]$, and with explanatory variables x_i = *[1, x_1 (i), x_2 (i), …]*, the parameters can be found using the following iterative algorithm:

$$w_{k+1} = (X^T S_K X)^{-1} X^T (S_K\, X\, w_k + y - \mu_k) \qquad (18)$$

Where a diagonal weighting matrix S=diag (μ(i) (1- μ(i))), and μ= [μ (1) , μ (2),…] is the vector of the expected values [10].

3.3. The Evaluating Measures

3.3.1. The Wald test statistic

The Wald statistic used to assess the contribution of individual predictors or the significance of individual coefficients in a given model. The Wald statistic is the ratio of the square of the regression coefficient to the square of the standard error of the coefficient. The Wald statistic asymptotically distributed as a Chi-square

distribution. It can be defined as follows:

$$W_j = \frac{\beta_j^2}{SE_{\beta_j}^2} \quad (19)$$

Each Wald statistic is a compared with a Chi-square with 1 degree of freedom. Wald statistics are easy to calculate, but their reliability is questionable [1].

3.3.2. The Hosmer-Lemeshow test

The Hosmer–Lemeshow test is used to examine whether the observed proportions of events are similar to the predicted probabilities of occurrence in subgroups of the model population. It performed by dividing the predicted probabilities into deciles and then computing a Pearson Chi-square that compares the predicted to the observed frequencies in a 2_x10 - table. The value of the test statistics is defined as follows:

$$H = \sum_{g=1}^{10} \frac{(O_g - E_g)^2}{E_g} \quad (20)$$

Where O_g and E_g denote the observed and the expected events for the g^{th} risk decile group. The test statistic is asymptotically χ^2 distribution with 8 degrees of freedom. Small values (with large p-value closer to 1) indicate a good fit to the data, good overall model fit. Large values with ($p < .05$) indicate a poor fit to the data [7].

3.3.3 The R-squared and the adjusted R-squared measures

The R-squared is the most widely measure for detecting goodness of fit of a model. The used symbol is symbolled by R^2. There are several definitions of $R2$. It is also known as the square of the coefficient of correlation (Pearson's R) between x and y for a set of n points (x_i, y_i). It is also the proportion of the variance in the dependent variable that is predictable from the independent variable(s). It can be defined as follows:

$$R = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 (y_i - \bar{y})^2}} = \frac{S_{XY}}{\sqrt{S_{XX}\, x\, S_{YY}}} \quad (21)$$

where overbars designate averages. The R is also given by $R = (b\, \acute{b})^{1/2}$, where b and b' are the least-squares slopes for linear regression of y upon x and x upon y respectively. It is also given by:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \hat{y})^2}$$

$$= 1 - \frac{SSE}{S_{YY}} \quad (22)$$

Where $\hat{y}_i$ is the calculated value of y at x_i from the regression of y upon x. The second term is recognized as the ratio of the residual sum of squares to the total sum of squares, and it expresses the extent to the fit model accounts for the variability in y. The R^2 represents the efficiency of the least square fit. A perfect fit means that

R^2= 1. If there is a fit to a true model, R^2 can be expressed as follows:

$$R^2 = 1 - \frac{(n-p)s^2}{(n-1)S^2} \quad (23)$$

Where the number of adjustable parameters is p, and an estimate of σ^2 is s^2. The S^2 is the total variance in y, equivalent to the estimated variance for a fit to the model y=a [16]

The Adjusted R^2 does not have the same interpretation as R^2, it is instead a comparative measure of suitability of alternative nested sets of explanatory variables. The adjusted R^2 is particularly useful in the feature selection stage of model building. It can be defined as follows:

$$R^2_{adj} = 1 - \frac{(n-1)SSE}{(n-p)S_{yy}} = 1 - \frac{s^2}{S^2} \quad (24)$$

where *p* is the total number of explanatory variables in the model (not including the constant term), and n is the sample size. The degrees of freedoms are n– 1[10].

4. Applications of the Anemia Diseases Data Sets

Anemia is a serious public health problem that affects populations in many countries. It arises when a person has lower number of red blood cells than the normal number; the amount of hemoglobin in the red blood cells drops below the normal level. The red cells prevent the body parts by the oxygen, where the body contains about 5-6 quarts of blood, and the heart are then constantly pumped it throughout all the body parts. The blood carries oxygen, nutrients, and other essential compounds. When something goes wrong in the blood, it makes a big problem and impact on the health and the life of the person. There are many types of anemia like aplastic anemia, iron-deficiency anemia, hemolytic anemia and pernicious anemia [18].

The anemia affects 1.62 (1.50–1.74) billion people that corresponding %24.8 of the population. The highest prevalence is in the pre-school children that prevalence about %47.4 (%45.7–%49.1), and the lowest prevalence is in men %12.7 (%8.6–%16.9). The greatest number of individuals affected is non-pregnant women 468.4 (446.2 –490.6) million [19].

A study of anemia diseases in Philippine and results reveal negatively effects of the pre-school children. It damages the hindered physical, cognitive development and leads to a weaken immune system. The hemoglobin level is analyzed by the ordinal logistic regression models, using data set from National Nutrition Survey in 2008. The results showed children aged between 6-11 months required more attention[5].

This section introduces two applications using two real data sets. The first data set are chosen from the Maternity and Children Hospital in Jeddah and the second data set are chosen from King Fahad Hospital in Jeddah. The SPSS program is used to descriptive and apply regression (ordinary and logistic) models.

4.1. The application of the children data set

The data set of children is for children who aged from (12-168) months. There are 92 cases suffering from the anemia diseases. There are five available independent variables, the age/ month, sex, weight, HB (hemoglobin) and WBC (white blood cell) that affect the dependent variable (the anemia level). The following Table 1 describes the data set, the minimum and the maximum values in the data, the mean for all used variables, the standard deviation and the variance:

Table 1: The Descriptive statistics of variables

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
Age/.month	92	12	168	86.87	38.573	1487.851
Sex	92	1	2	1.40	.493	.243
Weight	92	3.35	54.30	18.6510	7.93988	63.042
Anemia level	92	0	1	.25	.435	.190
HB	92	5.80	10.90	8.4511	1.27749	1.632
WBC	92	4.06	30.16	11.9408	5.16562	26.684

The multiple regression model is used at first to fit and analysis the relationship between the dependent variable anemia level and all the independent variables age/.month, weight, sex, HB, and WBC. The correlation R, the multiple correlation of determinationR^2, adjusted R square and the standard error of the model estimate for the regression model is summarized in the following Table 2:

Table 2: The summary of evaluating measures of multiple regression model for children

R	R - square	Adjusted R - square	Std. Error of the estimate
.671	.450	.418	.332

The ANOVA table for the multiple regression model is stated in Table 3 as follows:

Table 3: The ANOVA for the multiple regression for children

Sum of squares		df	Mean square	F	Sig.
Regression	7.757	5	1.551	14.054	.000
Residual	9.493	86	.110		
Total	17.250	91			

The coefficients of the model are stated in Table 4 as follows:

Table 4: The coefficient of the multiple regression model

Unstandardized coefficients		Standardized coefficients		t	Sig.
B	Std. error	Beta			
(Constant)	2.017	.314		6.417	.000
Age/month	.004	.001	.315	2.924	.004
sex	.001	.073	.002	.020	.984
weight	-.006	.006	-.117	-1.113	.269
HB	-.226	.029	-.664	-7.782	.000
WBC	-.004	.008	-.047	-.519	.605

The binary logistic regression model is also used to fit the data set where the dependent variable is divided into two levels, the moderate anemia level and the severe anemia level. The beginning classifications are in Table 5 as follows:

Table 5: The Classifications of the binary logistic regression model

Observed		Predicted		
		Anemia .level		
		Moderate anemia	Severe anemia	Correct percentage
Anemia level	Moderate anemia	69	0	100.0
	Severe anemia	23	0	.0
Overall percentage				75.0

Table 6 reveals no variable in the equation of the logistic model:

Table 6: Variables in the logistic model

	B	S.E.	Wald	df	Sig.	Exp (B)
Constant	-1.099	.241	20.820	1	.000	.333

The summary of evaluating measures of the logistic model, and the classification tables are stated in Tables 7 and 8 respectively as follows:

Table 7: The summary of evaluating measures of the logistic model

Step	-2 Log likelihood	Cox & Snell Rs quare	Nagelkerke R square
1	49.232	.445	.660

Table 8: The classification of the logistic model

Observed			Predicted		
			Anemia level		
			Moderate anemia	Severe anemia	Correct percentage
Step 1	Anemia level	Moderate anemia	65	4	94.2
		Severe anemia	7	16	69.6
	Overall percentage				88.0

The cut value is .500, while Table 9 reveals the variables in the equation as follows:

Table 9: The Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	Age/month	.045	.018	6.118	1	.013	1.046
	sex	.366	.809	.205	1	.651	1.442
	weight	-.102	.085	1.422	1	.233	.903
	HB	-2.354	.547	18.518	1	.000	.095
	WBC	.004	.071	.003	1	.958	1.004
	Constant	14.801	4.156	12.680	1	.000	2678505.532

4.2. The application of the adults' data set

The second data set is for adults who suffering from anemia diseases. There are 55 cases. The adults are aged from 16-80 years and having anemia on the bases of the laboratory analysis.

There are five available independent variables, the age/year, sex, weight, HB (hemoglobin) and WBC (white blood cell) variables and the dependent variable is the anemia level. The description of the data is in the following Table 10:

Table 10: The descriptive statistics of variables

	N	Minimum	Maximum	Mean	Std. deviation	Variance
Age/years	55	16	80	34.05	14.123	199.460
sex	55	1	2	1.75	.440	.193
weight	55	30.00	105.00	54.9545	14.41651	207.836
Anemia/.level	55	0	1	.29	.458	.210
HB	55	5.70	10.90	8.6109	1.29524	1.678
WBC	55	2.68	668.00	21.3087	88.93254	7908.996
Valid N (list wise)	55					

The multiple regression model is used to fit and analysis the relationship between the dependent and independent variables.

The independent variables are entered together in the analysis. The correlation R, the multiple correlation of determinationR^2, adjusted R square and the standard error of the estimate for the multiple model is stated in the following Table11:

Table 11: The summary of evaluating measures of multiple regression model for adults

Model	R	R square	Adjusted R square	Std. error of the estimate
1	.830[a]	.689	.657	.269

The ANOVA table for the adults' model is stated in Table 12 as follows:

Table 12: The ANOVA for the multiple regression for adults

Model		Sum of squares	df	Mean square	F	Sig.
1	Regression	7.812	5	1.562	21.666	.000[a]
	Residual	3.533	49	.072		
	Total	11.345	54			

The regression coefficients for the adults model is stated in the following Table 13:

Table 13: The Coefficient of the Multiple Regression Model

Model		Unstandardized coefficients		Standardized coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.854	.347		8.222	.000
	Age/years	.002	.003	.073	.836	.407
	sex	-.120	.090	-.115	-1.332	.189
	weight	.001	.003	.024	.281	.780
	HB	-.286	.028	-.809	-10.072	.000
	WBC	-.001	.000	-.099	-1.139	.260

The binary logistic regression model is also used to fit the data set where the dependent variable is the two levels of anemia, the moderate anemia level and the severe anemia level. The classification table of the logistic model when there are no predictor variables is stated in table (14) as follows:

Table 14: The classification of the logistic model

Observed			Predicted		
			Anemia/level		Correct percentage
			Moderate anemia	Severe anemia	
Step 0	Anemia/level	Moderate anemia	39	0	100.0
		Severe anemia	16	0	.0
	Overall percentage				70.9

Table 15 reveals the significance of the logistic model and there are no independent variables in the model as follows:

Table 15: The variables in the equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-.891	.297	9.006	1	.003	.410

Tables 16 and 17 have the summary of the model and the classification results respectively after entering the independent variables:

Table 16: The Summary of evaluating measures of the logistic model

Step	-2 Log likelihood	Cox & Snell R square	Nagelkerke R square
1	14.193	.612	.874

Table 17: The classification of the logistic model

Observed			Predicted		
			Anemia/level		Correct percentage
			Moderate anemia	Severe anemia	
Step 1	Anemia/level	Moderat anemia	37	2	94.9
		Severe anemia	2	14	87.5
	Overall percentage				92.7
The cut value is .500					

5. The Analysis of the Results and the Conclusions

5.1. The analysis of the results

For the children data, all variables entered the analysis; the R-squared for the regression model is 0.45 whereas for the logistic model is 0.66 with correct classification is % 88. The variables are tested to determine which of them more impact the level of anemia and the correct classifications. There are just two variables, the WBC, and HB, and the R- squared for using the regression model reduced to 0.39 , whereas for the logistic model the resulted R-squared is approximately 0.6 with correct classification approximately %84 with slowly decrease.

For the adult data, the R-squared for the regression model is 0.689 whereas for the logistic model is 0.874 for the case of all independent variables with %92.7 correct classification. By using only the variables WBC and HB, R-squared for the regression model is 0.67 while for the logistic model is .862 with correct classification %92.7.

5.2. The conclusions

In analyzing the two data sets, the regression model sometimes produces right results although the outcome variable, the anemia / level is binary. This is a wrong technique in analyzing such data sets especially medicine data sets where the outcomes in the two data sets are not continuous. The results are suspected and researchers have to be aware to the conditions each method based on. For children data set the logistic model reveals % 88 correct classifications, while for the adult's data set the logistic model reveals %92.7.

6. Recommendations

In the medicine applications, the logistic models provide excellent results and does not required any conditions except the categorical outcomes. There are also many other models such as the discriminant models, the operations research techniques, and the classification methods can be used if the conditions of each method exist. If the researcher has not experience it is good advice to use the logistic model either the binary or the multinomial model. Measures of evaluating the variables, and the models are available in the packages that analysis the data and reveals the contributions of each variable and the strength of the model. If there is an experience there are also available measures to evaluate the models and to select suitable variables explain the relationship between the outcome and independent variables

References

[1] A. Agresti. Categorical Data Analysis. New Jersey: John Wiley & Sons, 2007.

[2] W. Chen, Y. Chen, and Y.M.B. Guo. "Density-based logistic regression". The National Science Foundation of USA, 2013.

[3] C.M. Dayton. “Logistic regression analysis. Statistics & evaluation linear programming for resolving classification problem”. International Mathematical Forum, pp. 3125-3141, 1992.

[4] G. Gregoire. “Logistic regression”. EAS Publication series, 66, pp. 89-120, 2014.

[5] J. Gorospe, S. Bismonte, R. Areilla, and G. Gironella. "Ordinal logistic regression analyses on anemia for children aged 6 months to 5 years old in the Philippines". Presented at the De La Salle University Research Congress, 2014.

[6] J.F. Hair, R.F. Anderson, R.L. Tatham, and W.C. Black. Multivariate Data Analysis. Internationals edition, New York, USA, Maxwell, Macmillan**,** 1992.

[7] D.W. Hosmer and S. Lemeshow. Applied Logistic Regression. John Wiley & Sons Inc., 2000.

[8] A. Krap. “Using logistic regression to predict customer retention”. Sierra Information Service, Inc., 1998.

[9] D. Montgomery, E. Peck, and G.Vining. Introduction to Linear Regression Analysis. Fifth edition. John Wiley & Sons, Inc., 2012.

[10] K.P. Murphy. Machine Learning-A Probabilistic Perspective. The Massachusetts Institute of Technology Press, pp. 245-259, 2012.

[11] M. Pohar, M. Blas, S. Turk (2004)."Comparison of logistic regression and linear discriminant analysis: A Simulation Study". Metodoloski zvezki.1 (1), pp. 143 – 161, 2004.

[12] J. Rawlings, S. Pantula, and D. Dickey. Applied Regression Analysis: A Research Tool. Second Edition, 1932.

[13] R.L.Smith, and J.C. Naylor. “A comparison of maximum likelihood and Bayesian estimators for the three-parameter Weibull distribution”. Applied Statistics, 36, pp. 358-369, 1987.

[14] S.H. Walker, and D.B. Duncan. "Estimation of the probability of an event as a function of several independent variables". Biometrika. 54, pp. 167–178, 1967.

[15] H. Yusuff, N. Mohamad, U.K. Ngah, and A. Yahya.. "Brest cancer analysis using logistic regression". International Journal of Research in Engineering and Technology. 10, 1, pp. 14-22, 2012.

[16] J. Tellinghuisen, and C.H. Bolster. “Using $\mathbf{R^2}$ to compare least-squares fit models: when it must fail”. Chemometrics and intelligent laboratory systems. 105, pp. 220-222, 2011.

[17] S. Zhou, F. Guo, L. Li, Y. Zhou, Y. Lei, Y. Hu, H. Su, X. Chen., P. Yin, and X. Jian. "Multiple logistic regression analysis of risk factors for carcinogenesis of oral submucous fibrosis in mainland China". International Journal of Oral Maxillofacial Surgery. 37, pp. 1094–1098, 2008.

Sites:

[18] The National Heart, Lung, and Blood Institute https://www.nhlbi.nih.gov/files/docs/public/blood/anemia-inbrief_yg.pdf , 2011.

[19] B. De- Benoist, E. McLean, I Egli, and C. Cogswell (1993-2005)

http://www.who.int/ vmnis/anaemia/prevalence/summary/anaemia_data_status_t2/en/ , 1993-2005.