Journal of Statistics Applications & Probability Letters
An International Journal

http://dx.doi.org/10.18576/jsapl/050104

Efficiency of Multivariate Logistic Regression Analysis And Discriminant Analysis In Classification Of Rich Countries According To Human Development Index

Mohammed Mohammed Ahmed Al Mazah[1,2]

[1] Department of Mathematics, College of sciences and Arts (Muhyil), King Khalid University, KSA
[2] Ibb University, Yemen

Received: 15 Nov. 2017, Revised: 25 Dec. 2017, Accepted: 27 Dec. 2017
Published online: 1 Jan. 2018

Abstract: In this paper, we used the multivariate logistic regression method to classify countries and to arrive at a linear model for distinguishing countries into rich and very rich, according to Human Development Index based on Human Development Report 2016. To illustrate the importance of the logistic regression model, it has been compared with the discriminating function, where they both classify the value to their correct society. We found through the results, that the method of logistic regression is more efficient than the discriminatory function in the classification of the countries under study. The logistic regression has been classified 92 countries out of 96 with a probability of 95.8% classification. The discriminating function classified 89 countries with a probability of 92.7%. , the comparison was relied on statistical criteria (apparent error rate, and apparent correct classification rate. The ordering of the influential indicators was significant on the classification of countries according to their relative importance in the case of the regression model (mean years of schooling, life expectancy at birth, and gross national income(GNI) per capita).In the case of the discriminant function (GNI per capita, maternal mortality, and life expectancy Birth). Based on this, we suggest that countries should be classified according to the United Nations Human Development Report and data processing using multi-response logistic regression models.

Keywords: Linear discriminant analysis, Multivariate logistic regression, Apparent error rate, Apparent correct classification rate, Human Development Index

1 Introduction

Human development (HD) indicators play a prominent role in the classification of countries to countries with HD (very high, high, medium, low). In this paper, multivariate statistical analysis (MSA) methods will be widely applied to the classification of countries. There are two models known as multivariate logistic regression(MLR), whose importance is more powerful because it provides a test of the significance of transactions [1]. The linear discriminant analysis (LDA) is one of the most important methods of MSA, by which a set of variables is used to distinguish between two or more of the discriminant functions on which researcher can rely on to predict, and to determine the independent variables that contribute significantly to the distinction between groups, Groups with the lowest possible classification error [2].

Many researchers compared the two methods in terms of relative importance, efficiency, and ability to the classification for example, [3] compared the two methods and concluded that there were minor differences between them and the differences are in favor of logistic regression(LR). Similarly [4,5,6,7,8,9,10,11] found that the two methods have the same ability to classify and predict the same effectiveness. They concluded, however, that the method of LR gives better results than the Discriminant function (DF).The importance of this research is to classify countries into rich and very rich countries by using multivariate methods of analysis, as well as determining the influential and the most influential variables.

The aim of the research is to classify countries according to LR technique and DF, it aims to reach a linear model to distinguish countries into rich and very rich. The two methods were compared based on statistical criteria such as apparent error rate(AER) and apparent correct classification rate(ACCR). The research is based on the analytical

* Corresponding author e-mail: dalmazah2013@hotmail.com

descriptive approach through a brief review of the research literature on the theoretical side. On the applied side, the method of logistic regression analysis(LRA), and DA was used on data for 96 countries classified in the HD Report 2016 as countries with very high human development and countries with high HD, after excluding countries which have no complete data.

2 MATERIALS AND METHODS

2.1 Logistic Regression

It is well known in social and economic sciences that the dependent variable is a qualitative variable that takes binary value or more values. This is a challenge for researchers when they attempt to employ simple and multiple linear regression analysis, which is somewhat useful by requiring that the dependent variable should be a continuous quantity variable rather than a variable Descriptively. [12] recommended the use of LR because it has several characteristics that make it more appropriate in these cases. Gebotys [13] explained the importance of LRA when compared with the analysis of DF by saying that LR is a more powerful tool because it provides a test for the significance of the parameters as it gives the researcher an idea of the effect of the independent variable on the binary response variable as it arranges the effect of independent variables It also helps the researcher to conclude that a variable is stronger than the other variable in understanding the emergence of the desired result [14].

Walker [15] noted that LRA is less sensitive to deviating from normal distribution of variables under study. When compared to other statistical methods, such as DA, as well as the effect of the interaction between independent variables in the binary value variable, and can overcome many of the restricted assumptions for the use of least squares in linear regression. This makes LR the best method in the case of the binary dependent variable [16]. LR is defined as a statistical method to examine the relationship between the qualitative Y variable and one or more independent variables of any kind. It is also known as one of the regression methods used to predict the values of a qualitative dependent variable based on a set of mixed independent variables [16].

The LR model is based on the assumption that the Y variable of two values takes value (1) with probability P and value (0) with probability $(1-p)$, i.e., when the dependent variable has only two values $(0,1)$ and when the independent variables X_i are quantitative or qualitative variable, in this case, it is called Binary Logistic Regression Model. LR is used to predict the values of nominal (binary) variables based on the values of a mixed set of independent variables. It is known that the mean of the actual Y values at a certain value of variable X is $E(Y)$, and that the random error $E = Y - \hat{Y}$ so the model can be written as $E(Y/X_i) = \hat{\beta}_0 + \hat{\beta}_i X_i$, and the right side of the regression model takes values from $(-\infty)$ to $(+\infty)$, but if there are two variables, one of which is the binary Y, then the linear regression, as many researchers see, is inappropriate because: $E(Y/X) = P(Y=) = p$.Thus, the value of the right side confined between $(0,1)$. Hence the model is not applicable in the case of the binary dependent variable, value for the following reasons: first the Variance of the dependent variable Y changes by changing the values of the independent variable X, second the error variation is not distributed according to normal distribution. Finally, the estimated values cannot be interpreted as possibilities because their values do not range from $(0,1)$. One way to solve the problem is to introduce a mathematical transformation, It is known that $0 \leq p \leq 1$, i.e., $\frac{p}{1-p}$ is a positive number confined between $(0,\infty)$, i.e. $0 \leq \frac{p}{1-p} \leq \infty$ and by taking the natural logarithm $-\infty \leq Log_e\left(\frac{p}{1-p}\right) \leq \infty$, thus $Log_e\left(\frac{p}{1-p}\right) = \hat{\beta}_0 + \hat{\beta} X_i$. In other words:

$$\frac{p}{1-p} = e^{\hat{\beta}_0 + \hat{\beta} X_i} \Longrightarrow p = \frac{e^{\hat{\beta}_0 + \hat{\beta} X_i}}{1 + e^{\hat{\beta}_0 + \hat{\beta} X_i}}. \tag{1}$$

When there are k of independent variables, the equation is written as:

$$Log_e\left(\frac{p}{1-p}\right) = \hat{\beta}_0 + \sum \hat{\beta} X_i. \tag{2}$$

This model is called the logistic regression model and $Log_e\left(\frac{p}{1-p}\right)$ is called Logit transformation whereas:

$$Logit = Log_e(adds) = Log_e\left(\frac{p}{1-p}\right) = \hat{\beta}_0 + \sum \hat{\beta} X_i. \tag{3}$$

And that the logistic function is important because it takes input from $-\infty$ to $+\infty$, but outputs are always between 0, 1.

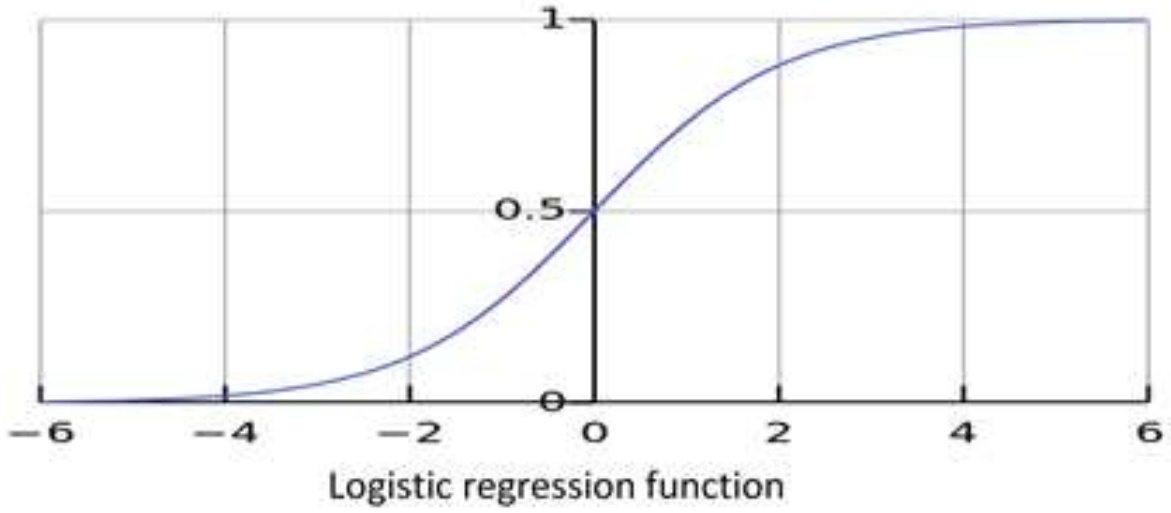

Fig. 1: Logistic regression function

The main function of the Logit function is to allow the application of linear regression when analysis relationships of binary dependent variables [17]. It is a similar function when the right side of this function is equal to zero and is called $\frac{p}{1-p}$ the ratio of preference or preference ratio of the desired event, and $Log_e\left(\frac{p}{1-p}\right)$ is called the logarithm of preference ratio or Logit [18].

For comparison between LR and the regression of the Least squares method, both methods are from the case of application is Symmetric [19]. However, they differ statistically as the mathematical formulas and detailed calculations for both are quite different, as the researchers suggest that the method of presentation, selection and examination of the model, assumptions of analysis, estimation of model parameters and interpretation of results are quite different in both analysis [20, 21].

The parameters of the LR model are estimated using the Maximum Likelihood method, which is one of the most popular estimation methods for all linear and nonlinear models. It has been considered better than the smaller squares method because it does not assume many constraints such as linearity and stability of variance. It measures the probabilities of viewing for the number of independent variables in the sample. i.e. , it is an iterative method that depends on repeating the calculations several times until to obtain the best estimate of the coefficients. The Maximum Likelihood method of calculating logit coefficients is used in LR to maximize Log Likelihood [16].

To explain LR coefficients we use Logit Coefficient. Which is called the non-standard LR coefficient, is used to estimate logarithm of log odds that the dependent variable is equal to one unit per unit change in the independent variable, and the LR calculates the amount of change in the logarithm of log odds of the dependent variable's weighting factor rather than the change in the dependent variable itself as it is in linear regression [22]. Thus LR coefficients by logit provide an explanation that is consistent with linear regression. The only difference is that the variable units in the LR represent logarithmic probability coefficients).

2.2 Discriminate Analysis

Discriminate Analysis is a multivariate statistical analysis technique that is concerned with building a base for redistribution or classification of societies that common characteristics. In other words, a discriminatory analysis is a technique used to classify observations into one communities or more through the use of the discriminatory function, which is a linear composition from the independent variables, which is used classification of observations.

And The Discriminate Analysis differs from the regression analysis in terms of the dependent variable. The dependent variable in the discriminate analysis is a nominal variable, which is a qualitative variable, whereas the dependent variable in the regression analysis is usually a continuous quantitative variable [23]. The classification process comes after the formation of a DF, where it depends on it in the process of predicting and classifying the new observations of one of

the groups under study with the least possible classification error. As well as their use in the knowledge of variables that contribute to the classification, which is defined as the collection of similar vocabulary in their characteristics or relations between them in certain categories.

2.2.1 Discriminate Function

It is a technique used to build a model for the predict, to be used in determine or distinguish observation to its correct group which is supposed to belong to it according to the criteria or measurements obtained from the known observations. There are two main types of the discriminatory function, the first linear type, called Fisher function, which is used when the relationship between the variables is linear and the nonlinear second is used when the relationship between the variables is not linear. The linear DF is used when the communities under study have a multivariate natural distribution with different averages, and the matrix of variance and common variance are equal, that is, the linear discrimination function is a linear formula of independent variables, written as the form:

$$L = a_1X_1 + a_2X_2 + \cdots + a_kX_k \tag{4}$$

where: $a_1, a_2, \ldots, a_k$ are the coefficients of the linear DF, and are chosen to give the highest percentage distinction between the two group , $x_1, X_2, \ldots, X_n$ denoted to the explanatory variables of the DF, and calculate the percentage of difference within the two groups from $\gamma = \frac{\text{BetweenGroup}}{\text{WithinGroup}}$ and chose the transactions that make the ratio of difference γ as large as possible [24]. To estimate the parameters of the discriminating function in the case there are two groups, as shown in following steps:

1. Finding the mean of each variable in each group, and finding the difference between the two means of each variable in the two groups from the formula:
$$d_i = \overline{X}_{k(1)} - \overline{X}_{k(2)} \tag{5}$$
2. Finding the matrix of variance and Co-variance for the two groups:
$$(n_j - 1)S_j = \left[X_{i(j)} - \overline{X}_{(j)}\right]' \left[X_{i(j)} - \overline{\overline{X}}_{(j)}\right], \quad j = 1,2 \tag{6}$$
3. Find the matrix of variance and Co-variance inside the groups
4. Calculate the parameter values of the linear discriminant function α from the formula:
$$\alpha = V^{-1}(\overline{X}_1 - \overline{X}_2) \tag{7}$$
5. Order the independent variables according to their relative importance in determining the variables of the affecting the process of discrimination from the formula:
$$\alpha_j^* = \alpha_j \sqrt{V_{ii}}, \quad i = 1,2,\ldots,k. \tag{8}$$

Where V_{ii}: represents the variance extracted from the elements of the diameter of the common variance matrix S_p^2 , by the Comparing the resulting values from the calculating a_j^* therefore the biggest value from all values of the variable, is the most important variable that has the ability to distinguish between the two groups, it follows, the second largest value which has the ability to discriminate [25].

2.2.2 Cutoff Point

It is the point that represents the Critical boundary between two groups, and is used when we want to classify a value or new value so that if the discriminating value of the value increases, from the cutoff point, the value is classified to the other group. If the discriminating value of that observation is equal to the value of the cutoff point, the classified is randomly assigned to either group. The cutoff point is calculated from the formula:

$$\overline{\overline{L}} = \frac{1}{2}\left(\overline{L}_{(1)} - \overline{L}_{(2)}\right) \tag{9}$$

Where $\overline{\overline{L}}$: represent cutoff point , $\overline{L}_{(1)}$: represents the mean of the discriminatory values of the first group and $\overline{L}_{(2)}$: represents the mean of the discriminatory values of the second group. The classification process is the subsequent process of forming the discriminant function and testing its ability to discriminate using the cutoff point and the following rule can be followed for the classification process.

1.If $\overline{L}_{(1)} > \overline{L}_{(2)}$ the observation belongs to the first group if the discriminate value of this observation is $L > \overline{\overline{L}}$ and the observation belongs to the second group if the discriminate value $L < \overline{\overline{L}}$

2.If $\overline{L}_{(1)} < \overline{L}_{(2)}$ the observation belongs to the first group if the discriminate value of this observation is $L < \overline{\overline{L}}$ and the observation belongs to the second group if the discriminate value $L > \overline{\overline{L}}$ [25].

3 Result And Discussion

3.1 Research Methodology

The research adopted the analytical descriptive approach to the styles of discriminant analysis and logistic regression, after looking in the scientific references and previous research related to the subject of research and applying on it, conducting to statistical operations and graphical presentations. We adopter on the United Nations data published in the Human Development Report 2016, where the countries of high and very high HD were chosen. The data was denoted as follows: High HD (rich = 0) and very high human development (very rich = 1), It represents the dependent variable Y.

The study included a number of independent variables and it is : X_1: Gross national income (GNI) per capita , X_2 : Life expectancy at birth , X_3 : Mean years of schooling, X_4 : Mortality Under-five, X_5 : Maternal birth ratio, X_6 :Mortality rate infant and X_7: Continental Location. The sample included 96 countries distributed on continents according to table (1)

Table 1: Distribution of countries by continental location

The continent	Europe	Asia	North America	South America	Africa	Australia
Number	37	26	9	17	6	1
Rate%	38.5	27.1	9.4	17.7	6.3	1.0

3.2 Results of data analysis using proposed models

3.2.1 Results of logistic regression

Table 2: the Calculation means and inflation factor index

Variables	X_1	X_2	X_3	X_4	X_5	X_6	X_7
Mean	27020.67	77.12	10.44	9.84	23.59	8.55	2.29
VIF	1.49	2.26	1.68	5.65	2.44	5.91	1.68

To detect the linear interference phenomenon, the VIF index, which represents a standard for detecting multiple linear correlation [26] stated that if VIF > 10 is an index of a multiple linear correlation between the dependent variable and the independent variables. Table (2) shows that VIF < 10 for all variables means that there is no linear interference between independent variables, which means continuing analysis. Hosmer and Lemshow [27] suggests that once the logistic regression model has been reconciled (identification of study variables), the model evaluation process begins in two ways:

1.Verify the suitability of the model as a whole.
2.Examining the statistical significance of each independent variable on Separately.

There are many important measures that help evaluate the final model that is created these include: statistic R, Hosmer-Lemeshow's test for goodness of fit, the maximum Likelihood ratio test, and classification tables. To test the suitability of the model as a whole and its Goodness of fit, Log Likelihood Ratio was used, which follows distribution χ^2 according to the following formula: $\chi^2 = 2\{LogD_0 - LogD_1\}$,
Where: D_0 :The maximum Likelihood function value that contains a variable $(i-1)$,
D_1 : The maximum Likelihood function value that contains one variable.

Table 3: Explanation of variables in the model

Step	$-2LogL$	Cox & Snell Square R^2	Nagelkerke Square R^2
1	15.820	0.705	0.94

Where the value of $\chi^2 = 117.23$, freedom degree 7 and significant level $(p < 0.05)$. This indicates that the statistical model that has been created is statistically significant, indicating that the variables in the model at step 12 are of great importance and statistically significant effect in the classification of countries into very rich and rich countries

Table3 represents the quasi-Coefficients of determination that express the explanatory power of the model. That is mean, the variables in the model at step 12 has been interpreted about 94% by Nagelkerke, and, it is high explanatory force, also 0.71% by Cox & Snell, and, it is Medium explanatory force. This shows that there is still a percentage of changes in the dependent variable due to other factors not included in the model.

To test the null hypothesis: the created model is suitable for the data. Versus alternative hypothesis: The model that has been created is not suitable for data.

The results of the analysis showed that the value $\chi^2 = 0.230$ and degree of freedom 8 and significant level $p-$value$= 0.973 > 0.05$, this means that there is no insufficient evidence to reject the null hypothesis; therefore, the model is considered to be suitable for the data is well, and hence indicates to there are complete vindication for the parameters of the model.

Table 4: Number of iteration cycles for the derivative maximum Likelihood function

Iteration		$-2LL$ Liklih00d	Coefficients							
			Constant	X_1	X_2	X_3	X_4	X_5	X_6	X_7
Steep1	1	53.78	-13.12	0.00	0.125	0.43	-0.005	-0.008	-0.046	-0.096
	2	39.956	-23.317	0.00	0.214	0.63	-0.007	-0.002	-0.072	-0.198
	3	28.381	39.943	0.00	0.36	1.06	-0.018	0.019	0.063-	-0.383
	4	22.792	-66.431	0.00	0.601	1.71	-0.088	0.053	0.027	-0.714
	5	18.900	-104.849	0.00	0.953	2.66	-0.470	0.101	0.453	-1.23
	6	16.723	-156.426	0.00	1.430	3.91	-0.877	0.166	0.895	-1.95
	7	15.964	205.643	0.00	1.884	5.10	-1.23	0.232	1.256	-2.61
	8	15.826	236.038	0.01	2.160	5.86	-1.46	0.274	1.484	-2.99
	9	15.820	-244.271	0.01	2.234	6.07	-1.52	0.285	1.545	-3.088
	10	15.820	-241.759	0.01	2.238	6.08	-1.52	0.286	1.549	-3.094
	11	15.820	-244.761	0.01	2.238	6.08	-1.52	0.286	1.549	-3.094
	12	15.820	-244.761	0.01	2.238	6.08	-1.52	0.286	1.549	-3.094

Table 4 includes The number of iteration cycles of derivatives of the maximum Likelihood function to obtain the lowest value equals twice the logarithm of the maximum Likelihood function to obtain the optimal estimation of the parameters of the derived model where parameters of the model was stabilized at step 12 with the lowest value (15.820). And we stopped at this iteration because the parameter estimates changed by less than 0.001 and in fact the change in the estimated transactions $\hat{\beta}_1, \hat{\beta}_2, \ldots$ became weak after the seventh iteration. It is also noted that the estimators of parameters of table (4), which in the iterations (8, 9, 10, 11, 12) are similar with there is the very small differences. We stopped at the 12th iteration and considered its parameters as the best result to be obtained for the parameter.

3.2.1.1 Explanation of parameters of the model

To examine the statistical significance of each variable, Table 5 shows the estimates of the parameters of the optimal model obtained at the twelfth session.

Table 5 shows all the parameters of the estimated model in Log-odd units, and the equation of the model as in formula:

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -244.67 + 0.001X_1 + 2.238X_2 + 6.083X_3 - 1.524X_4 + 0.286X_5 + 1.549X_6 - 3.094X_7 \qquad (10)$$

Such that :
$\hat{P}$: The probability of obtaining a country with very high HD (very rich). These estimates show the relationship between dependent variable and independent variables in logit units.

Table 5: Estimated model

		β_i	S E	Wald	d f	Sig	exp(β)	95% C I	
								Lower	Upper
Step 1a	Constant	-244.76	115.73	4.437	1	0.034	0.000		
	X_1	0.001	0.000	4.341	1	0.038	1.001	1.000	1.001
	X_2	2.238	1.073	4.351	1	0.037	9.376	1.145	76.792
	X_3	6.083	2.887	4.440	1	0.035	438.263	1.529	125627.64
	X_4	-1.524	0.916	2.765	1	0.096	0.218	0.036	1.313
	X_5	0.286	0.163	3.094	1	0.079	1.331	0.968	1.831
	X_6	1.549	0.929	2.781	1	0.095	4.708	0.762	29.073
	X_7	-3.094	1.669	3.473	1	0.064	0.045	0.002	1.193

The table also shows that the reference to per capita Gross national income (GNI) per capita is positive, this mean, that the higher the GNI per capita by one dollar, the higher the logit or the logarithm of the weighting factor, the dependent variable is closer to the value of one $(Y = 1)$, that is , the country has a very high HD level of 0.001 with the constancy of the effect of other variables, Also, for the life expectancy at birth, we note that the increase by one year will increase the logit or the logarithm of the weighting coefficient and will be approached to $Y = 1$; that is, the country has a very high HD level of 2.238 times, with the constancy of the effect of other variables.

The increase in the mean years of schooling by one year will increase the logit or logarithm of the weighting coefficient by 6.083 times, with the constancy of the effect of other variables, is the indexes adopted in building the HD index, and these variables have a significant effect on the classification of countries where, $p-$value < 0.05 and 95% confidence level, that is, these variables are of great importance in the classification of countries according by the HD index. While the variable under-5 mortality rate has shown a negative sign, this means that the higher of the number of under-5 mortality by one unit, the lower the logit or logarithm of the weighting coefficient by 1.524 times, with the constancy the effect of the other variables. Also the continental site, which showed a statistically significant relationship to the two continents of Europe and North America, and it is decisive factor in determining the belonging of the countries to the rich or very rich, and non-significant for the rest continents. The second column represents the standard error according to the formula:

$$S.E(\hat{\beta}_i) = Z_{ii}$$

Where: $Z_i i$ represents the Diagonal elements of the estimated co-variance matrix by the formula

$$\text{Cov}(\hat{\beta}_0) = \{X\text{Diag}[n\hat{p}_i(1-\hat{p}_i)]X\}^{-1} \tag{11}$$

While the third column represents the Wald statistic to test the significance of the model's coefficients according to the formula: $\text{Wald} = \left(\frac{\hat{\beta}_i}{S.E(\hat{\beta}_i)}\right)^2$ it distributed as a X distribution with freedom degree 1. The column exp(β) indicates the value of the exponential function of the LR coefficient, which expresses the multiplier in which the weighting ratio changes. From the above we observe the mean of years of schooling was in the first order to influence the dependent variable Y in the classification of countries, followed by influence the life expectancy variable at birth and GNI per capita variable, all of which showed a high significance on the dependent variable. For to the variables (x_4, x_5, x_6, x_7) non-significant in influencing the dependent variable.

Table 6: Efficiency of classification of the model

		Predicted			Percentage
Observed		Rich	Very rich	Total	Correct
Step 1	Rich	47	2	49	95.9
	Very rich	2	45	47	95.7
	Overall Parentage			96	95.8
Step 0	Rich	49	0	49	100
	Very rich	47	0	47	0
	Overall Parentage			96	51

Table (6) special for testing the efficiency of classification of the model, which is one of the methods of examining the quality of the model's conformity to the data, the performance standard has been used $(AER) = \frac{n_{01}-n_{10}}{n_0-n_2}$ and $(ACCR) = \frac{n_{00}-n_{11}}{n_0-n_2}$, to evaluate the efficiency of the model.

The table shows that 47 of the 49 rich countries were properly classified, also 45 of the 47 countries of the very rich group were properly classified, it also shows that the probability of correct classification of the two groups was 95.8%, that only four of them were wrongly classified and that the probability of error is 4.2%, that is acceptable percentage indicating that the model represents the data well.

The results of the analysis also indicate the improvement in the correct classification rate achieved by the model after the inclusion of independent variables 95.8%, while without including the independent variables in step (0) about 51%. And that the probability of the correct classification of very rich countries was 95.9% and the probability of the wrong classification of the group of very rich countries 4.1% while the probability of the correct classification of rich countries 95.7% and the probability of the wrong classification of rich countries group 4.3%. Thus, we conclude that the classification table is able to classify 92 countries of the 96 countries a correctly classification to the group to which it belongs, thus showing that the probability of wrong classification was too small and equal to 4.2% indicating that the model has the ability to classify.

3.2.2 Results of discriminant function

3.2.2.1 Application of discriminatory function model

Table 7: The arithmetic means of the factors affecting the classification of countries

Variables / classification	X_1	X_2	X_3	X_4	X_5	X_6	X_7
Very rich	39618.19	79.607	11.606	5.136	10.255	4.345	1.638
Rich	14937.34	74.631	9.316	14.347	36.391	12.589	2.918
All States	27020.67	77.116	10.438	9.838	23.595	8.553	2.292

By calculating the mathematical means of the variables involved in building the model, there were significant differences between the two groups of variables (x_1, x_2, x_3) for the benefit of the very rich countries, while the differences in the variables (x_4, x_5, x_6, x_7) were for the benefit of the rich countries only. The differences between the mean of each variable of the formula are calculated:

$$d_i = \overline{X}_{k(1)} - \overline{X}_{k(2)} = \begin{bmatrix} 39618.69 - 14937.34 \\ 79.7064 - 74.6306 \\ 11.6064 - 9.3163 \\ 5.1362 - 14.43469 \\ 10.2553 - 36.3906 \\ 4.3447 - 12.5898 \\ 1.6383 - 2.9184 \end{bmatrix} = \begin{bmatrix} 24680.87 \\ 5.0758 \\ 2.2901 \\ -9.2107 \\ -26.1353 \\ -8.2451 \\ -1.2801 \end{bmatrix}$$

Variance and Co-variance Matrix of the Group of Rich Countries

$$V_{(0)} = \begin{bmatrix} 109212631.6 & -175.839 & -3314.87 & -9018.439 & -54339.5 & -7922.5 & -1400.8 \\ -175.839 & 6.531 & -1.222 & -5.187 & -7.647 & -5.282 & 0.080 \\ -3314.87 & -1.222 & 2.186 & -1.287 & -5.097 & -0.694 & -0.601 \\ -9018.439 & -5.187 & -1.287 & 39.981 & 42.969 & 25.312 & 3.262 \\ -54339.54 & -7.674 & -5.097 & 42.969 & 352.268 & 35.365 & 6.348 \\ -7922.51 & -5.282 & -0.0649 & 25.312 & 35.365 & 27.685 & 2.214 \\ -1400.76 & 0.080 & -0.601 & 3.262 & 6.348 & 2.214 & 1.493 \end{bmatrix}$$

Variance and Co-variance Matrix of the Group of very Rich Countries

$$V_{(1)} = \begin{bmatrix} 397479984.8 & 14485.7 & -2782.3 & 30.26 & -2.15141 & -1352.496 & 3900.25 \\ 14485.705 & 9.812 & 0.831 & -5.708 & -13.386 & -4.830 & 0.207 \\ -2782.3 & 0.831 & 1.512 & -1.691 & -3.878 & -1.562 & -0.272 \\ 30.258 & -5.708 & -1.691 & 7.856 & 17.973 & 6.858 & 1.187 \\ -21514.072 & -13.386 & -3.878 & 17.973 & 76.412 & 15.977 & 4.225 \\ -1352.496 & -4.830 & -1.562 & 6.858 & 15.977 & 6.122 & 1.017 \\ 3900.245 & 0.207 & -0.272 & 1.187 & 4.225 & 1.017 & 1.410 \end{bmatrix}$$

Matrix of variance and co-variance of the two combined groups

$$V_{(01)} = \begin{bmatrix} 401467606.7 & 38559.8 & 11250.6 & -61947.8 & -200760.3 & -56045.19 & -6797.19 \\ 38559.822 & 14.556 & 2.720 & -17.191 & -43.858 & -15.576 & -1.500 \\ 112560.58 & 2.270 & 3.161 & -6.795 & -19.567 & -5.853 & -1.175 \\ 61947.78 & -17.191 & -6.795 & 45.429 & 91.202 & 35.287 & 5.201 \\ 200760.26 & -43.858 & -19.567 & 91.202 & 387.473 & 80.021 & 13.701 \\ -56045.19 & -15.576 & -5.853 & 35.287 & 80.021 & 34.119 & 4.276 \\ -6797.196 & -1.500 & -1.175 & 5.201 & 13.701 & 4.276 & 1.851 \end{bmatrix}$$

Finding the equation of the discriminatory function

$$\hat{L} = \hat{a}_1 x_1 + \hat{a}_2 x_2 + \hat{a}_3 x_3 + \hat{a}_4 x_4 + \hat{a}_5 x_5 + \hat{a}_6 x_6 + \hat{a}_7 x_7 \tag{12}$$

We obtain the estimated values of the parameters $\hat{a}_i$ of the formula $\hat{a}_i = v^{-1}d$ and by the substitution, we get:

$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_4$	$\hat{a}_5$	$\hat{a}_6$	$\hat{a}_7$
0.477	0.391	0.513	-0.029	-0.134	-0.211	-0.127

And the equation of the discriminatory function becomes as follows:

$$\hat{L} = 0.477X_1 + 0.391X_2 + 0.513X_3 - 0.029X_4 - 0.134X_5 - 0.211X_6 - 0.127X_7$$

After obtaining the discriminatory function, we can show the relative importance of each of the variables involved in building the model compared to the other variables by formula:

$$\hat{a}_j^* = \hat{a}_j \sqrt{v_{ii}} \tag{13}$$

$\hat{a}_1^*$	$\hat{a}_2^*$	$\hat{a}_3^*$	$\hat{a}_4^*$	$\hat{a}_5^*$	$\hat{a}_6^*$	$\hat{a}_7^*$
9557.485	1.492	0.912	-0.195	-2.638	-1.233	-0.173

By neglecting the negative sign with the ascending order, the largest value indicates that the corresponding variable is the most important variable and has the ability to distinguish between the two groups. The second largest value corresponding to the second variable in importance and has the ability to distinguish between groups. Table 8 illustrates this importance for each variable and its discriminatory capacity.

Table 8: show the relative importance of each variable indicated in descending order.

Variable symbol	Variable name	The relative importance of the variable
X_1	Gross national income (GNI) per capita	9557.48
X_5	Maternal mortality ratio	-2.638
X_2	Life expectancy at birth	1.492
X_6	Infant mortality rate	-1.233
X_3	Mean years of schooling	0.912
X_4	Under - 5 mortality rate	-0.195
X_7	Continental Location	-0.173

Table 8 shows that more of the variable important in the most in the classification of countries (very rich or rich) are the per capita GNI, followed by the maternal mortality ratio , and followed the life expectancy at birth as shown in the table.

3.2.2.2 Test of significant the discriminatory function

To test the significance of the linear discriminant function, the measures shown in Table (9) indicate that the variant between the two groups for all variables were explained by the discriminatory function, which explained 100% of the variance. And its legal correlation coefficient was 0.846. The table also showed that the model is highly accurate in countries' forecast in terms of rich or very rich.

Table 9 shows the value of the Wilks' Lambda scale according to the formulas $\Lambda = |W|/|T|$, Where, T: Matrix of variance and co-variance of the two combined groups, W:variance and covariance matrix within the two groups, and its value is approximately between 0, 1. If it is approached to or equal to one, this indicates that group means are equal,

Table 9: Test of significant the discriminatory function

Function	Eigenvalue	% of variance	Cumulative	Canonical Correlation
1	2.526	100.0	100.0	0.846
Test of function	Wilks' Lambda	Chi-Square	d f	Sig
1	0.284	114.05	7	0.000

thus, there is no distinction between the two groups, but if the value is approached to zero, it indicates the strength of discrimination. In this research, the value of the Wilks' Lambda test is near to zero and this indicates that there is a high distinction between the two groups which are statistically significant at the level of significance $p < 0.05$, the measure χ^2 confirms that , which is calculated from the formula: $\chi^2 = -\log(\Lambda)$, with degree freedom $p(k-1)$, where was its value 114.05. From the above we find that the discriminatory function is able to explain the variance between the two groups. And for to test the significance of all variables to determine the importance of each variable in the discriminatory function, and the extent of its effect, were the results according to table (10).

Table 10: Test the effect of independent variables

Variables	Wilks' Lambda	F	df1	df2	Sig
X_1	0.617	58.39	1	94	0.000
X_2	0.553	75.96	1	94	0.000
X_3	0.581	67.78	1	94	0.000
X_4	0.528	83.89	1	94	0.000
X_5	0.555	75.42	1	94	0.000
X_6	0.497	95.19	1	94	0.000
X_7	0.776	27.07	1	94	0.000

Table 10 shows that all variables have high moral significance and have a significant impact in terms of classification and distinction between rich and very rich countries.

3.3 Cutoff Point

The discriminatory function reached has the ability to distinguish states into (rich and very rich). We will calculate a value $L_{(1)}, L_{(2)}$ in the equation of the discriminatory function of both rich and very rich countries by compensating for values X_i in the equation as appendix. We then find an mean of L in each group.

$$\overline{L}_{(1)} = \frac{989614.3}{51} = 19263.71, \quad \overline{L}_{(2)} = \frac{285411.4}{45} = 6342.476$$

We note that if the discriminatory value of the country to be classified is greater than the Cutoff point $\overline{\overline{L}} = 12803.088$, it belongs to the group of very rich countries, but if the discriminatory value is less than the Cutoff point, it belongs to the group of rich countries.

3.4 The correct classification

Based on the Cutoff point , the results of the classification were obtained according to table (11)

The table 11 shows that 46 of the 49 rich countries were properly classified, also 43 of the 47 countries of the very rich group were properly classified. It also shows that the probability of correct classification of the two groups was 92.7%, and that the probability of the correct classification of very rich countries was 91.5%, and the probability of the wrong classification of the group of very rich countries 6.1% while the probability of the correct classification of rich countries 93.9% and the probability of the wrong classification of rich countries group 8.5%. Thus, we conclude that the classification table is able to classify 89 countries of the 96 countries a correctly classification to the group to which it belongs, thus showing that the probability of wrong classification was too small and equal to 7.3% indicating that the model has the ability to classify.

Table 11: Results of the classification process

		Predicted Group Membership		
Observed		Rich	Very rich	Total
Original	Rich	46	3	49
	Very rich	4	43	47
		Ratio of rich countries	Ratio of very rich countries	96
	Rich	93.9	6.1	100
	Very rich	8.5	91.5	100

Table 12: Classification of Drug data by Logistic Regression and Discriminant Function Methods

		Predicted Group Membership			
		Logistic Regression		Discriminate Analysis	
Actual Group	No. of cases	Rich	Very rich	Rich	Very rich
Rich	45	47(95.9%)	2(4.1)	46(93.9%)	3(6.1%)
Very rich	51	2(4.3%)	45(95.7%)	4(8.5%)	43(91.5%)
Overall% correctly classified		95.8		92.7%	

The overall percentage of the probability of the correct classification was 92.7% and 95.8% for the discriminatory analysis and the logistic regression method, respectively. Therefore, the results showed that the overall classification rate for both methods is higher in predicting the possibility of discovering belonging of the country to any group.

We conclude that the method of multivariate logistic regression analysis is more efficient than analyzing the discriminant function in predictive accuracy. This contradicts the findings of the research [28], which say : that the analysis of the discriminant function is more efficient than the multivariate logistic regression analysis method in prediction accuracy).

4 Conclusions

1. The application of the two methods (the binary logistic regression method, the discriminatory function) on the data of HDR , the high and the very high, gave accurate and consistent results with the real classification of countries.
2. The results of the research showed that the method of logistic regression analysis is more efficient than analyzing the discriminate function in the accuracy of the prediction.
3. The logistic regression model classified 92 countries out of 96 to the right group and by ratio correct classify 95.8%, while the discriminatory function classified 89 out of 96 countries correctly and by ratio correct classify 92.7% respectively.
4. Statistical model that has been created has a statistical significance; indicating the variables have great significance and statistically significant impact in the classification of countries to very rich and rich countries.
5. The results indicate that the variables that had a significant impact on the classification of countries in the case of logistic regression are: the mean years of schooling followed by life expectancy at birth and the rate of per capita GNI. In the case of the discriminatory function, the per capita GNI was ranked first, then the maternal mortality rate, and finally the life expectancy at birth.

Appendix:

The discriminatory values of the group of very rich countries (1-51) and the group of rich countries (52-96)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
32289.04	32289.89	20464.08	26924.8	21503.1	21273.09	37320.53	22135.05	20642.8	17718.35	20349.92	25434.4	15716.69	22099.97	18130.79	17815.39
17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
16513.2	14928.05	29835.64	18204.13	19709.94	18577.26	20839.04	13709.48	16052.33	15673.19	13461.64	11870.61	34779.42	12610.26	14088.6	14108.41
33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
62004.68	11540.53	12439.07	10369	24514.01	12802.64	12486.98	31613.84	11193.08	10808.69	10020.26	9714.291	17795.06	7385.945	11139.04	9299.387
49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64
36321.15	7489.067	16443.24	7165.284	9167.296	7791.16	10567.19	11871.2	9317.532	8589.394	13406.83	6713.626	5853.865	3590.039	7851.789	4259.389
65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
8950.29	7248.207	4925.721	6381.232	7847.48	7861.286	6778.853	5519.637	4842.603	5945.427	6498.004	3940.016	3538.259	4853.185	5420.665	6951.813
81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96
5058.959	6396.83	3963.143	5013.166	4699.169	4010.862	6114.903	4845.564	7667.283	4915.001	6115.985	4974.055	2555.626	6851.884	3545.936	5035.794

References

[1] Krieng, K. Comparison Logistic Regression and Discriminant Analysis in classification group for breast cancer. Faculty of Information Technology, Rangsit University, Thailand, International Journal of Computer Science and Network Security, 12(5);111-115 (2012).

[2] Lai, P. Koehly, L .M.(2003). Linear discriminant analysis versus Logistic Regression: a comparison of classification errors in the two-group case. Journal of Experimental Education.72: 25-49.

[3] Halperin, M. Blackwelder W.C and Verter J.I.(1971). Estimation of the Multivariate Logistic risk function: A Comparison of the discriminant function and Maximum Likelihood Approaches. J. Chronic Dis. 24:125-158

[4] Press, J. and Wilson, S. (1978). Choosing Between Logistic Regression and Discriminant Analysis: Journal of the American Statistical Association. 73(364): 699-705.

[5] Kleinbaum, .D.G, Kupper, .L.L, Muller, .K.E., and Morgens-Tern, H. (1982): Epidemiologic Research: Principles and Quantitative Methods. Van No strand Reinhold Company, New York, p: 281-417.

[6] Dattalo, P. (1995). A Comparison of Discriminant Analysis and Logistic regression. Journal of Social Services Research, Volume 19, Issue 3-4, p. 121-144.

[7] Hyunjoon, K. and Zheng, G.u (2010). Predicting Restaurant bankruptcy: A Logit model in Comparison With Discriminant Model; Tourism and Hospitality Research Journal, Vol. 10, P 171-187.

[8] Edokpayi, A.A., Agho, C., Ezomo, J.E., Edosomwan, O.S. and Ogiugo, O.G. (2013). A Comparison of the Classification Performance of Discriminant Analysis and the Logistic Regression Methods in Identification of Oil Palm fruit Forms. A Paper Presented at the Annual Conference of Nigerian Statistical Association. 11-13th, pp 20-26.

[9] Balogun, O.S., Balogun, M.A., Abdulkadir, S.S. and Jibasen, D. (2014). A Comparison of the performance of Discriminant Analysis and the Logistic Regression methods in Classification of Drug Offenders in Kwara State. International Journal of Advanced Research, Vol. 2, Issue 10, p: 280-286.

[10] Balogun, O.S . Akingbade, T. J and Oguntunde, P.E.(2015). An assessment of the performance of discriminant analysis and the logistic regression methods in classification of mode of delivery of an expectant mother. Mathematical Theory and Modeling ; 5:147-154

[11] Zahra, S.h. Naser, M. Leila, S.h and Parisa N.(2016). Prediction of Depression in Cancer Patients with Different Classification Criteria, Linear Discriminant Analysis versus Logistic Regression. Global Journal of Health Science ; 8(7): 41-46.

[12] Lea, S.(1997). Multivariate Analysis II: Manifest Variables Analysis. Topic 4: Logistic Regression and Discriminant Analysis. University of Exeter Department of Psychology. Revised 11th March,1997, available at: http://www.exeter.ac.uk/~SEGLea/multivar/diclogihtml.

[13] Gebotys , R.(2000). Examples: Binary Logistic Regression.Janury,2000.

[14] Dutta, A., Bandopadhyay, G. and Sengupta, S. (2012). Prediction of Stock Performance in the Indian Stock Market Using Logistic Regression. Vol. 7, No. 1, June.

[15] Walker, M.D.(1998). Discriminant Function Analysis" Lesson 8.

[16] Abbas , A. K.(2012). Using Logistic Regression Model to predict the Functions with Economic categorical Dependent variables. Journal of the University of Kirkuk for administrative and economic sciences, Vol 2, Issue 2 ,p 234-253 (in Arabic).

[17] Walker, J.(1996). Methodology Application Logistic Regression Using CODES Data. Developed at :http//www.nrd.nhtsa.dot.gov/pubs/96843.PDF.

[18] Ghanem, A. & Al Jaoun F. K.(2011): Using of a double-response logistic regression technique in the study of the economic and social determinants of household income adequacy. Applied Study on a Random Sample of Households in Damascus Governorate, Journal of University. Damascus for Economic and Legal Sciences, Vol 27, No. 1: 119-120 (in Arabic).

[19] Dallal, C.M.(2001). Logistic Regression. Available at : www.tufts.edu/~gdallal/Logistic.htm.

[20] Farhood , S.H . A.(2014): The Use of Logistic Regression in studying the Factors Influencing the Performance of Stocks (An Empirical Study on the Kuwait Stock Exchange). Journal of Al Azhar University-Gaza (Natural Sciences), 2014, 16 : 47-68 (in Arabic).

[21] Poston, Duley, L. (2004). Sociological Research: Quantitative Methods (Lecture notes, Lecture 7). Spring .
[22] Garson, D.(2006). Logistic Regression. Available at: http://www2.clss.ncsu.edu/garson/pa765/Logistic.htm.
[23] Menard, S.W (2002). Applied logistic regression analysis. 2nd edition, Sage Publication Series: Quantitative Application in the Social Sciences, No. 106, Thousand Oaks, CA: Sage.
[24] Raykov, T . Marcoulides, G . A. (2008). An Introduction to Applied Statistical Analysis . Second Edition, JOHN WLLEY& Sons, NEW York.
[25] Johnson, R. A. Wichern ,D.W.(2007). Applied Multivariate Statistical Analysis. 26 th ed, Pearson Education, Inc , USA
[26] Myers, E.W(1986). An O(ND): difference algorithm and its variations. Algorithmica, Springer New York 1(1): 251266.
[27] Hosmer, W. Lemeshow, S. (2000). Applied Logistic Regression. 2nd Edition, New York: Johnson WI leg & Sons, Lnc.
[28] Elgohari, H.(2017). Efficiency of Discriminant Analysis and Multivariate Regression for the Detection of Anemic Children with Logistic Chronic Kidney Disease. International Journal of Statistics and Applications, 7(2): 131-13

Mohammed M. Ahmed Almazah received his B.S. degree in Mathematics from College of Science Sana'a Yemen, 1994, the M.S. degree in Mathematics(Statistics) from Khartoum University , Sudan 2001, and the Ph.D degree in Statistics from Khartoum University, Khartoum, Sudan , in 2005. 7/2012. Now Associate Professor in Department of Mathematics and computer in Ibb University , Ibb, Yemen . Certified Trainer. He has many research articles and Books in Statistics and probability.