IOSR Journal of Business and Management (IOSR-JBM)
e-ISSN: 2278-487X, p-ISSN: 2319-7668. Volume 18, Issue 3 .Ver. II (Mar. 2016), PP 87-92
www.iosrjournals.org

Evaluation Classification Performances of Human Development Index with Multiple Discriminant Analysis and Multinomial Regression Analysis

Özge Eren[1], LatifeSinem Sarul[2]
[1](Vocational School/ Istanbul Aydin University, Turkey)
[2](School of Business/ Istanbul University, Turkey)

Abstract: *Human Development Index is a core scale used to measure development in the world. In this study, it is performed classification by using Discriminant Analysis and Multinominal Logistic Regression on the Human Development Index data of 2013 of the United Nations Development Programme. The present study aims to evaluate and compare classification success of these methods. In the literature, Countries are classified in two or three groups. However, it is analyzed the accuracy of the assignment by four sets (much developed, developed, less developed and undeveloped) as stipulated by the UN in this study. Classification success of Discriminant Analysis has been found as 94% and also for MultinominalLogistic Regression Analysis as 97%. Analyses were performed using SPSS17.*
Keywords:*Discriminant Analysis, Human Development Index, Multinominal Logistic Regression*

I. Introduction

Economicgrowth has been described as the increase in national income per individual for the countries[1] or as an activity(programme) researches regularly the ways to improve the economic prosperity and the quality of life for a community[2]. Globally, the seeking for criterion, which consists of several welfare measures, has begun to evaluate the development after 1970s[3]. Human development index (HDI) formed at the beginning of 1990s has been the mostly preferred measure among several different criteria. This index, which measures development by regularly taking into account the specific data every year by UN, is an average rate calculates the measurement of the performance of three basic indicators which are expressed as "long and healthy life factor", "access to information factor" and "the necessary source for a life in basic standards".

This measurement rate is calculated by getting the geometric mean of normalized indexes in every three dimensions. HDI is obtained by converting to a rate between 0 and 1 with the indicator normalization process using formula given below. In the formula, $D_{healt\,h}$ refers to the rate obtained in health category, $D_{education}$ refers to the rate obtained in education category, D_{income} refers to the rate obtained in income category[4].

$$D(0,1) = \frac{ActualValue - Minimumvalue}{Maksimumvalue - Minimumvalue} \quad (1)$$

$$HDI = \sqrt[3]{D_{healt\,h} * D_{education} * D_{income}} \quad (2)$$

In this study, 187 countries are taken into account as dependent factor for; the development rate coded respectively as 1,2,3,4 and; "Life Expectancy", "Education" and "Gross Domestic Product" rates which form Human Development Index, have been taken into account as independent factor.

There is a wide literature about Human Development Index. However, in this study, we would like to take attention the classification success of the mostly used statistical methods in the analysis. In this regard, BolatB.A, Çilan C.A (2007)[5], have been classified Europe, Middle Asia-Middle East and Africa in terms of Human Development Index indicators with the discriminant analysis. In the result of analysis they demonstrated that education and gross domestic product indexes are the factors causes todifferentiate among these three groups. Burmaoğlu S, Oktay E, Özen Ü. (2009)[6] have evaluated the success of classification with Discriminant Analysis and Logistic Regression Analysis by separating the countries subject to research to two basic clusters. According to this analyze, while being 92% obtained classification with Discriminant Analysis, 100% classification success has been obtained in Logistic Regression Analysis. Yakut E., Gündüz M., Demirci A.(2015)[7], separated 81 countries according to the Human Development Index values, to classes of very high, high and average human development. They have been used in comparison of classification successwith Ordered Logistic Regression Analysis and Artificial Neural Network methods and they have 88% classification success in Ordered Logistic Regression Analysis and 97% classification successobtained with the Artificial Neural Network method.

II. Research Methods

Discriminant analysis is basically a multivariate statistical analysis, measures whether there is a significant relation between dependent and independent variables statistically. This analysis is used to predict to which population most likely belongs to the very different numbers of observations coming from different populations[8]. If the dependent factors have two categories, this analysis is called Discriminant analysis; if it has more than two categories, this analysis is called multiple discrimination analysis.This analysis contains linear combinations of the used independent factors. These functions termed also discriminant functions exposes which estimate factors affect the difference among the groups. These functions as follows[9];

$f_{km} = v_1 X_{1.km} + v_2 X_{2.km} + v_3 X_{3.km} \dots\dots\dots\dots v_p X_{1.kp}$ .... (3)

f_{km} = mth value in kth groupof the discriminant function,

$X_{1.kp}$ = observation value for mth value in kth group for variable i

v_1 = Coefficients of the discriminant function

The basis point is finding $v_i$coefficientswhich will make the rate maximum of between group variance to within group variance while the functions are formed.

$$F = \frac{Varianceofbetweengroups}{Varianceofwithingroups} = \frac{SS_b/(p-1)}{SS_w/(n-p)} \quad (4)$$

The analysis is finalized by finding $v_1, v_{2,} \dots\dots, v_p$values, which are the eigenvector of W^{-1}.B, after that the eigenvalues ofW^{-1}Bmatrix is foundto be able to estimate $v_i$coefficients.

Logistic Regression Analysis altered methodology in the present study, is used as an alternative to the classic linear regression analysis in case dependent factor is categorical while the binary logistic regression is being used when the dependent factor has two categories.In the event of that dependent variable is ordinal, the multinominal logistic regression analysis is used.Multinominal Logistic Regression Analysis is the extended form of binominal logistic regression analysis. In this study, the multinominal logistic regression analysis has been used in the analysis for the countries have defined as 1,2,3,4 values by arranging according to the level of development.

Logistic Regression Analysis is an alternative and easy in interpreting especially when the deviations from hypothesis likenormality and the homogeneity ofcovariance[10].This analysis isbasically converted to a transformation which is nonlinear with the below transformation process in case the dependent variable is to be estimated by getting value in limited edition.

This analysis isbasically converted to a transformation which is not linear with the below transformation process in case the dependent variable is to be estimated by getting value in limited ("1" or "0")values[11].

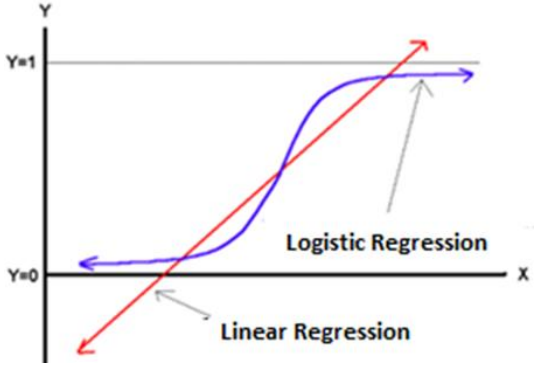
$$\log\left(\frac{p(\bar{x})}{1-p(\bar{x})}\right) = \beta_0 + \beta_1 x_1 + \dots\dots + \beta_p x_p \quad (5)$$

or

$$\frac{p(x)}{1-p(x)} = exp(\beta_0 + \beta_1 x_1 + \dots\dots + \beta_p x_p) \quad (6)$$

In these formula, transformation of $\log\left(\frac{p(\bar{x})}{1-p(\bar{x})}\right)$is called as logit.

Figure1: Graphical view of LogisticRegression Model



Logistic Regression derives values between 0 and 1 through the nature of equation as it is seen in Figure1 above. Logistic Regression Analysis which is very similar to linear regression uses natural logarithmic structures which are called as odds ratio. In the Multinominal Logistic Regression Analysis used in this study, logistic regression equation is formed to a missing of the number of category which dependent variablecontains. Any of them accepted as a reference in the scale[12].

$$g_1(x) = ln\left[\frac{P(y=1\backslash x\}}{P(y=0\backslash x)}\right] = \beta_0 + \beta_{11}x_1 + .\beta_{12}x_2 \ldots\ldots\ldots\ldots \beta_{1p}x_p$$
$$g_2(x) = ln\left[\frac{P(y=1\backslash x\}}{P(y=0\backslash x)}\right] = \beta_0 + \beta_{21}x_1 + .\beta_{22}x_2 \ldots\ldots\ldots\ldots \beta_{2p}x_p \qquad (7)$$
$$g_{n-1}(x) = ln\left[\frac{P(y=1\backslash x\}}{P(y=0\backslash x)}\right] = \beta_0 + \beta_{n-11}x_1 + .\beta_{n-12}x_2 \ldots\ldots\ldots\ldots \beta_{n-1p}x_p$$

III. Analysis Results

The analysis results must ensure that three important hypothesis, which are known as multicollinearity of the variables, equality of variance-covariance within group and multivariate normality for being able to implement. If it is possible, alternative solutions should be produced or the analysis results should not be accepted in case there is a deflection from the hypothesis[13].

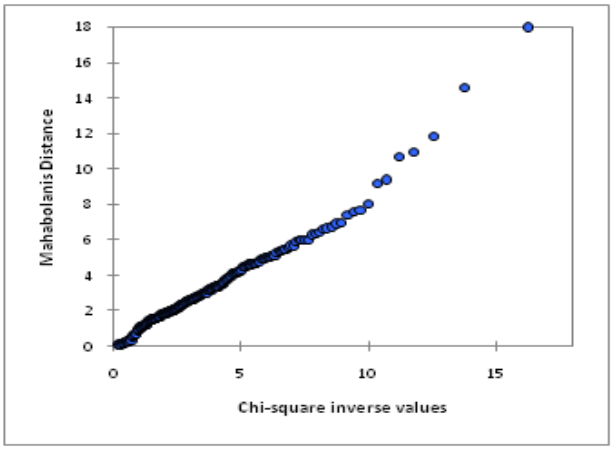
The equality of variance-covariance matrices has been checked with Box-M Tests. According to this test, the equality of variance-covariance matrices could not be achieved at the level of (p value=0.02)1% significance level (Table1). In this case, quadratic discriminant analysis can be used to resolve revealed violations but to resolve this problem;this hypothesis has been resolved as it can be obtained between the groups for it could have not obtained within the groups[14] (Table1).

Table1: Box-M Test Results

Box's M		288,379
F	Approx.	9,226
	df1	30
	df2	84523,678
	Sig.	**0,02**

Multivariate normality test has been implemented by using MahalanobisDistances[15]. It has been identified that there is a high correlation (0.889) between the adverse cumulative Chi-Square values and Mahalanobisdistances[6]. If correlation coefficient is high enough,multivariate normality test is provided.When $\bar{X}$ approximately normally distributed and sample size is large enough $(n \geq 25)$ sampling distribution of Mahalanobis Distances is approximately χ^2distributed[16].Figure2 refers to the data set normally distributed for this study graphically.

Figure 2: Mahalanobis Distances



Multicollinearity exist any model when two or more independent or input variables in the model are related to each other[17]. There are several different numerical methods for exploring multicollinearity connections. VIF and Tolerance values of these methods are statistics which the researches usually prefer. The solution was reached by using Microsoft Excel Xlstat Application.If VIF values higher than 10, it is accepted that there is multicollinearity[18].Because, all of the calculated VIF values are reasonable (Table2), there is no multicollinearity problem for the data set we analyzed.

Table2: Multicollinearity Statistics Results

	x_1	x_2	x_3	x_4
Tolerancevalues	0,334	0,258	0,317	0,58
VIF values	2,997	3,883	3,151	1,723

Three discriminant functions were created as seen at the Table2 when the discriminant analysis was applied by using SPSS17 programme to the data. The discriminant function has been obtained as much as a missing of the category number of dependent factor. Calculated eigenvalues of each discriminant function points out the explanatory level. It is seen that; the first on be of larger to smaller eigenvalues used has the highest explanatory with a 99.2% value and there is no explanatory of the third function.

Table3: Discriminant Analysis SPSS Results

Function	Eigenvalues	Variance (%)	Cumulative %	Canonical Correlation
1	10,114[a]	99,2	99,2	0,954
2	,079[a]	0,8	100	0,271
3	,003[a]	0	100	0,055

It is measured that which variable contributes to the function significantly with Wilk's Lambda Test. If this value is close to "0", it is pointed that the contribution of the variable is well. Chi-Square statistictests the significance of Wilk's Lambda. If p value is smaller than 5%, the group membership of the function is reasonably significant. Table3, points out Wilk's Lambda Test's results. According to the Table3, it is seen that, 1 and 2 variables impactfunction significantly. (p value=0.0000 for Variable1 and p value = 0.026 for Variable2).

Table4: Wilk's Lambda Test SPSS Results

Test of Function(s)	Wilks' Lambda	Chi-Square	df	Sig.
1 through 3	0,083	450,21	12	0
2 through 3	0,924	14,326	6	0,026
3	0,997	0,547	2	0,761

At the following Table 4, classification results are seen. According to the Table4, classification success of original groups is about94 %.

Table5: SPSS Classification Results

Classification Results			Predicted Group Membership				Total
			1	**2**	**3**	**4**	
	Count	1	42	1	0	0	43
		2	2	36	3	0	41
		3	0	0	53	0	53
		4	0	0	5	44	49
		Ungrouped cases	0	0	1	0	1
	%	1	97,7	2,3	0	0	100
		2	4,9	87,8	7,3	0	100
		3	0	0	100	0	100
		4	0	0	10,2	89,8	100
Original		Ungrouped cases	0	0	100	0	100

94,1% of original grouped cases correctly classified

-2 Log-likelihood Ratio Tests is similar to the Sum of Squares method in regression analysis. The relation between independent and dependent variables is analyzed with this measure, which can be also expressed as model Consistency measure. According to this, this measure value is to decrease after several iterations if there is a relation to a certain extent between dependent and independent factors. On the other hand, this significance is tested with Chi-Square test statistic[19]. Chi-Square test statistic has resulted significantly(p value=0.0000).Generally interpreting the table, independent variable is subject to dependent factor to a certain extent. The first log-likelihood value(526,209)has been calculated by not adding any independent factor. The last value indicates the calculated value when all independent factors are added to logistic regression analysis.

Table6: Consistency Table

Model	Model Consistency Cr.			
	-2 Log-Likelihood	Chi-Square	Stddev.	Significance
First Step	526,209			
Last Step	24,969	501,241	16	0.00000

Pseudo $R^2$is a value that measures the power of the relation between dependent and independent variable, similar to R^2value in multiple regression analysis.There is a strong relation between dependent and independent variable if Pseudo R^2value is close to "1", conversely there is a poor relation if Pseudo R^2value is close to "0". Pseudo $R^2$is the common name of three different measures as knownCox and Snell,Nagelkerke,Mc Fadden in the literature.In the present study, calculated Pseudo R^2 values are quite high values and the relation between dependent and independent variables is quite strong and significant.

Table7: Pseudo R^2values SPSS Results

Pseudo R^2	
Cox and Snell	0,931
Nagelkerke	0,991
McFadden	0,953

Consistency tests measure significance of each independent factor analyzed statistically. In the present study, it is seen apparently that these values, which are calculated for each value at the level of 5%significance level, are quite significant(p value= 0.0000).

Table8: Model Consistency

	Model Fitting Criteria			
Effect	-2 Log Likelihood	Chi-Square	df	Significance
Intercept	341,525	316,556	4	0,00000
loggdp	101,563	76,594	4	0,00000
leabirth	142,506	117,537	4	0,00000
Meanyearsschh	104,219	79,251	4	0,00000
expyearsofschool	57,962	32,994	4	0,00000

Finally, the classification table below points out the classification rate of all observation values handled

in the established model. This value is a quite high value with 97.3%.

Table9: Classification Table

Observed Value	Prediction Value				
	1	2	3	4	Accuracy Rate
1	41	2	0	0	%95.30
2	3	38	0	0	%92.70
3	0	0	53	0	%100.00
4	0	0	0	49	%100.00
Total Percentage	%23.50	%21.40	%28.30	%26.20	**%97.30**

IV. Conclusion

Human Development Index is a core scale used to measure development in the world. This index, which measures development by regularly taking into account the specific data every year by UN, is an average rate calculates the measurement of the performance of three basic indicators which are expressed as “long and healthy life factor”, “access to information factor” and “the necessary source for a life in basic standards”.In this study, it is performed classification by using Discriminant Analysis and Multinominal Logistic Regression on the Human Development Index data of 2013 of the United Nations Development Programme.There is a wide literature about Human Development Index. However, in this study, we would like to take attention the classification success of the mostly used statistical methods in the analysis.

The primary purpose of the present study is to measure performance of Multinomial Logistic Regression Analysis and Multiple Discriminant Analysis in the classification for a selected year. As a result of analyses, it is seen that, multiple discriminant analysis has realized 94% classification performanceand Multinominal Logistic Regression Analysis has 97%.In this study, Multinominal Logistic Regression Analysis performed better than Discriminant Analysis. Additionally, it is easier to use because it has no assumption.Discriminant analysis’s assumptions are verified for the multicollinearity and multivariate normal test but variance–covariance matrix equality could have not obtained within the groups although this problem has been resolved between the groups. Consequently, It is fair to say that Multinominal Logistic Regression Analysis is more preferable model to use because of the violation of the variance-covariance equality assumption for theDiscriminat Analysis.

In the present study, classification has been realized again by using the row metrics data of 187 world countries. The study is planned to develop by using different methods like Neural Network withmore year and additional classification for the future studies.

References

[1]. E.Kongar, “EkonomikBüyümeveKalkınma”, (Online)http://www.kongar.org/makaleler/mak_mi.php(15 Mayıs 2011)
[2]. J.D.Gatrell, R.R.Jensen, “Planning and Socioeconomic Applications”, Geotechnologies and the Environment, Springer (2009) p.6.
[3]. E.Baday, U.Sivri, M.Berber “Türkiye’deİllerinSosyo-EkonomikGelişmişlikSıralaması” BozokveÇankayaÜniversitesiUluslararasıBölgeselKalkınmaSempozyumu**,** 7-9 Ekim 2010, s.2
[4]. A.Akçiçek, “İnsaniGelişmeEndeksiveTürkiye’ninİnsaniGelişmePerformansı”, 2015, printed report
[5]. http://www.sde.org.tr/userfiles/file/2014%20insani%20geli%C5%9Fme%20endeksi%20ve%20T%C3%BCrkiye'nin%20insan%20geli%C5%9Fme%20performans%C4%B1.pdf
[6]. B.A.Bolat, Ç.A.Çilan , İnsanîGelişmeİndeksiBileşenleriAçısındanGelişmekte Olan ÜlkelerinDiskriminantAnalizi İle Karşılaştırılması, 38.Uluslararası AsyaveKuzeyAfrikaÇalışmalarıKongresi ICANAS , TÜRKİYE, 10-15 Eylül 2007, pp.1-10
[7]. S.Burmaoğlu, E.Oktay, Ü. Özen, “ BirleşmişMilletlerKalkınmaProgramıBeşeriKalkınmaEndeksiVerileriniKullanarakDiskriminantAnaliziveLojistikRegresyonAnalizininSınıflandırmaPerformanslarınınKarşılaştırılması, Kara Harp OkuluSavunmaBilimleriDergisi, Sayı 82, 2009, 23-49.
[8]. E.Yakut, M.Gündüz, A.Demirci, İnsaniKalkınmışlıkDüzeyininSınıflandırmaBaşarılarınınKarşılaştırılmasındaSıralıLojistikRegresyonAnaliziveYapaySinirAğlarıYöntemlerininKullanılması, Journal of Business Research Turk, (7)4, 2015, 172-199.
[9]. R.A.Johnson, D. W. Wichern, Applied Multivariate Statistical Analysis Pearson Prentice Hill 1998 p.217
[10]. Z.Çakmak,”EniyiAyırmaModelininBelirlenmesindeKullanılanDeğişkenSeçmeYöntemleri”, AnadoluÜniversitesiKütahyaİktisadiveİdaribilimlerFakültesi , 15. YılArmağanı ,Kütahya, Turkey, 1989, 289-299.
[11]. B.G.Tabachnick, L.S.Fidell, S.J. Osterlind, “Using Multivariate Statistics” US, Pearson/ Allyn&Bacon, BOSTON, 4th ed., 2001
[12]. Appalachian State University (2012), “An Introduction to Logistic Regression Analysis”http://www.appstate.edu/~whiteheadjc/service/logit/intro.htm
[13]. N. Bayram, “MultinominalLojistikRegresyonAnalizininİstihdamdakiİşgücüneUygulanması” İktisatFakültesiMecmuası”, 2004, (54), 61-76.
[14]. J.H.Friedman, "Regularized discriminant analysis." Journal of the American Statistical Association, 84, 1989, 165-175.
[15]. IBM (2009) “IBM support Portal”
[16]. http://www-01.ibm.com/support/docview.wss?uid=swg21479748Accessed: 20.01.2016
[17]. S.Sharma, Applied Multivariate Techniques (Wiley, USA, 1996)
[18]. R. A. Johnson, , D. W. Wichern , AppliedMultivariate Statistical Analysis (PrenticeHall, USA, 1992)
[19]. R. Matignon, “Neural Network ModelingUsing SAS Enterprise Mining”,2005, 2.
[20]. M.Kutner, C.Nachtsheim, J.Neter, Applied Linear Regression Models (McGraw-Hill Irwin Series, USA, 2004)
[21]. S.Selvin, Statistical Tools for Epidomolojik Research (Oxford University of Press, USA, 2011)