استخدام طرق التصنيف في تحديد فئات قوة العمل في فلسطين

Using Classification Methods in Identifying the Labor Force Categories in Palestine

Abdalla M. EL-HABIL , Husam Salama

Faculty of Economics and Administrative Sciences;
Al-Azhar University, Gaza - Palestine.

E-mail: abdalla20022002@yahoo.com

Received 11/12/2014 *Accepted* 3/3/2015

Abstract:

Multinomial logistic regression(MLR) and Discriminant Analysis (DA) are two techniques that commonly used for data classification. Both of them are applied at Labor Force in Palestine data 2012 in order to predict the probability of a specific categorical of Labor Force (LF) based upon several explanatory variables. we used real data on LF, from a survey of LF 2012 which was conducted by Palestinian Central Bureau of Statistics(PCBS). The data sample size had been 25353 observations from West Bank and Gaza Strip. The target group was the age group (15- 65) years for both sexes. Labor Force data has 12 variables; the dependent variable is nominal with three categories and 11 independent variables. So, we have two models for each techniques. Correct classification is 83.5% for MLR model compared with 81.1% for DA. In addition that the area under the ROC curve is 91.89% for MLR and 52.8% for DA These results demonstrate that MLR can be more powerful analytical technique.

Key Words: Confusion Matrix – Roc curve – Multinomial Logistic Regression – Discriminant Analysis - Odds ratio

ملخص:

يعتبر أسلوبا الانحدار اللوجستي المتعدد والتحليل التمييزي الخطي من أكثر الأساليب استخداما لتصنيف البيانات، وقد تم تطبيق كلا الأسلوبين على بيانات القوى العاملة في فلسطين 2012 لتحديد فئات قوة العمل على أساس العديد من المتغيرات المفسرة وتم الحصول عليها من مسح القوى العاملة 2012 وذلك بالتواصل مع الجهاز المركزي

للإحصاء الفلسطيني حيث بلغ حجم العينة 25353 مشاهدة موزعة على الضفة الغربية وقطاع غزة ومستهدفة الفئة العمرية (15− 65) لكلا الجنسين. وتحتوي البيانات على 12 متغير منها 11 متغير مستقل بالإضافة إلى المتغير التابع وهو متغير وصفي بثلاث فئات ولذلك فان لدينا نموذجين لكل أسلوب . ولقياس دقة التصنيف تم حساب المساحة تحت المنحنى (AUC). حيث بلغت نسبة التصنيف الصحيح 83.5% لنموذج الانحدار اللوجستي المتعدد مقارنة ب 81.1 % لنموذج التحليل التمييزي، إضافة إلى أن المساحة المحصورة تحت منحنى ROC هي 91.89 % لأسلوب الانحدار اللوجستي المتعدد مقارنة ب 52.8% لأسلوب التحليل التمييزي وتبين النتائج أن أسلوب الانحدار اللوجستي المتعدد يمكن أن يكون تقنية تحليلية أكثر قوة.

1. Introduction

Labor Force in Palestine is all persons aged 15 years and over who are either employed or unemployed (PCBS, 2012). Unemployment rates are rapidly growing due to the recessive economic status in Palestinian territories. It is useful to investigate this category in order to propose policy actions toward reducing the prevalence rates. Using ILO (International Labor Organization) standards, the number of unemployed was about 260 thousand in the 4th quarter 2012: about 139 thousand in the West Bank and about 121 thousand in Gaza Strip. The unemployment rate in Gaza Strip was 32.2% compared with 18.3% in the West Bank, and the unemployment rate for males in Palestine was 20.7% compared with 31.7% for females (PCBS, 2012). The deterioration of the Palestinian economy continued in 2014, particularly in Gaza where the situation was dire even before the recent conflict. The average yearly economic growth exceeded 8% between 2007 and 2011 but declined to 1.9% in 2013, and reached minus 1% for the first quarter of 2014. A quarter of the Palestinian population lives in poverty, with the rate in Gaza twice as high as that in the West Bank. According to the Palestinian Central Bureau of Statistics, in mid-2013, prior to the closure of the tunnels to Egypt, 24,200 individuals worked in the construction sector. Today, it employs only 6,800 people. Between the second half of last year and the end of the first quarter of 2014, about 17,400 individuals who had made their living in the construction sector lost their jobs.

We analyzed Labor Force survey data 2012 which are provided by the Palestinian Central Bureau of Statistics (PCBS) and built a statistical model that can identify the important risk factors on Labor Force in Palestine. The classification methods are used in order to categorize certain data of statistical community on different groups based on one or more of the basic properties of these data. The nature of data help or restrict us to choose the best classification method. It is meaningful to address how the analyst can deal with data representing multiple independent variables and a categorical dependent variable. The main objective of this paper is to explore the importance and assumptions of a statistical model that can be used when the response variable is categorical and that can identify the most important risk factors associated with various forms of Labor Force in Palestine . The goal is to find the best model according to model selection criteria and ROC curve will validate the model. The most popular methods are the Multinomial Logistic Regression (MLR) and Linear Discriminant Analysis (LDA), many studies had been used them to analyze categorical data. These studies had varied on different subjects including those related to social and medical issues, behavioral, and some scientific experiments. See (Chao and Rebecca, 2002), (Nichols et al., 2005), (Raymo and Sweeney, 2006), (Slingerland et al., 2007), (Antonogeorgos et al., 2009), (Al-khatib and Al-Horani, 2012).

We will discuss and evaluate MLR and LDA methods as classification models. We will see how these models can classify LF cases into one of three categories.

The paper is organized as follows : Section 2 Theoretical Aspect, Section 3 LF data analysis and Section 4 Conclusion.

2. Theoretical Aspects

2.1 Multinomial Logistic Regression (MLR) Model

There is an important difference between logistic regression model and the linear regression model concerning the nature of the relationship between the outcome and independent variable (Hosmer & Lemeshow, 2000). Logistic Regression can be binomial or multinomial. Binomial or binary logistic regression deals with situations in which the observed outcome for a dependent variable can have only two possible types. In binary logistic regression, the outcome is usually coded as "0" or "1" . MLR does not make any assumptions of normality, linearity, and homogeneity of variance for the independent variable. It deals with situations where the outcome can have three or more possible types. MLR was chosen to answer the research question for two reasons; first, MLR provides an effective and reliable way to obtain the estimated probability of belonging to a specific population and the estimate of odds ratio; second, MLR is a procedure by which estimates of the net effects of a set of explanatory variables on the response variable can be obtained. MLR can be used to predict a response variable on the basis of continuous and/or categorical explanatory variables to determine the percent of variance in the response variable explained by the explanatory variables; to rank the relative importance of independents to assess interaction effects; and to understand the impact of covariate control variables. MLR allows the simultaneous comparison of more than one contrast ,that is the log odds of three or more contrasts are estimated simultaneously (Garson, 2009). If the response variable has more than two values, and there is no natural ordering of the categories, it called Multinomial Logistic Regression. Suppose we have n independent variable has k categories, to construct the logits in the multinomial case, one of the categories must be considered the base level and all the logits are constructed relative to it. Any category can be taken as the base level, so we will choose any category k as the base level.

Let πj denote the multinomial probability of an observation falling in the jth category, to find the relationship between this probability and the p explanatory variables, $x_{1,}$ x_2, … x_p, the MLR model is

$$\log[\tfrac{\pi_j(x_i)}{\pi_k(x_i)}] = \alpha_{0i} + \beta_{1j}x_{1i} + \ldots + \beta_{pj}x_{pi} \qquad (1)$$

where $j = 1 , 2 , . . . , (k - l)$, i = 1 , 2 , …n. This equation reduces to

$$\pi_j(X_i) = \frac{\exp(\alpha_{0i} + \beta_{1j}x_{1i} + \ldots\beta_{pj}x_{pi})}{1 + \sum_{j=1}^{k-1}\exp(\alpha_{0i} + \beta_{1j}x_{1i} + \beta_{2j}x_{2i}\ldots + \beta_{pj}x_{pi})} \qquad (2)$$

for j = 1,2, . . ,(k-1), the model parameters are estimated by the method of MLE. Practically, we use statistical software to do this fitting (Chatterjee and Hadi, 2006).

An adjusted odds ratio is an odds ratio comparing two categories of the variable after controlling for the other variables in the model. For example, an adjusted odds ratio comparing two categories of the variable (Moorman and Carr, 2008)

$$\overset{\wedge}{OR}_{X_1=1 vs\ X_1=0} = \frac{\overset{\wedge}{Odds}(Y = 1 | X_1 = 1, X_2, \ldots, X_k)}{\overset{\wedge}{Odds}(Y = 1 | X_1 = 0, X_2, \ldots, X_k)} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 X_2 + \ldots \hat{\beta}_{,} X_k}}{e^{\hat{\beta}_0 + \hat{\beta}_2 X_2 + \ldots \hat{\beta}_{,} X_k}} = e^{\hat{\beta}_1} \qquad (3)$$

2.2 Discriminant Analysis (DA)

Discriminant Analysis is a classic method of classification that has stood the test of time. It is the first multivariate statistical classification method used for decades by researchers and practitioners in developing classification models (Hamid & Hashibah, 2010). It often produces models whose accuracy approaches (and occasionally exceeds) more complex modern methods. It can be used only for classification (i.e., with a categorical target variable), not for regression. The target variable may have two or more categories. The number of functions required to maintain maximum separation for a subset of the original variables is called the rank or dimensionality of the separation. The goals of (DA) are to construct a set of discriminants that may be used to describe or characterize group separation based upon a reduced set of variables, to analyze the contribution of the original variables to the separation, and to evaluate the degree of separation (Timm, 2002).

Suppose the training set consists of a random sample size n_i from population π_i, $i = 1, 2, \ldots, k$. Denote the $n_i \times p$ data set, from population π_i, by X_i and its j^{th} row by x'i.j . Sample mean vectors is $\bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$. and $\bar{x}$ is the total sample mean vector

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i \bar{x}_i}{\sum_{i=1}^{k} n_i} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}}{\sum_{i=1}^{k} n_i} \qquad (4)$$

Which is the $p \times 1$ vector average taken over all of the sample observations in the training set . We define the sample between groups matrix B which includes the sample sizes

$$B = \sum_{i=1}^{k} n_i (\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})' \quad (5)$$

An estimate of $\sum$ is based on the sample within groups matrix

$$W = \sum_{i=1}^{k} (n_i - 1) s_i = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (\bar{X}_{ij} - \bar{X}_i)(\bar{X}_{ij} - \bar{X}_i)' \qquad (6)$$

Where n_i is the number of samples in the kth class and $\bar{x}_i$ is the mean vector of class i, k is the number of classes , $\bar{x}_{ij}$ is the jth sample of class *i* .

So, the goal of DA is to maximize the between-class measure while minimizing the within-class measure. This objective function can be described by :

$$F = \max_{\ell} \frac{\ell' \mathrm{B} \ell}{\ell' w \, \ell} = \max_{\ell} \frac{tr(\ell' \mathrm{B} \ell)}{tr(\ell' w \, \ell)} \qquad (7)$$

A linear combination of variables ' ℓ ' maximizing the ratio of the between-groups sums of squares B and the within-groups sums of squares W .

where tr(.) denotes the matrix trace. It is defined to be the sum of the elements on the main diagonal (Andrew, 2002).

2.3 Validation and Evaluation of Classification Models

There are several criteria available to evaluate a set of classification rules, The simplest and the most frequently used criterion for comparison between two methods is error rate or misclassification rate. In practice population parameters are unknown. Therefore, most research on the error rate estimation focused on the actual hit rate (Hussain et al., 2002). Probably the simplest and most widely used method for estimating prediction error is cross-validation. This method directly estimates well the expected prediction error (Hastie et al, 2009). The classification table, also called a confusion table, is a table in which the rows are the observed categories of the dependent variable and the columns are the predicted categories. When prediction is perfect, all cases will lie on the diagonal. The percentage of cases on the diagonal is the percentage of correct classifications.

Cross-Validation Method

Cross-validation, sometimes called rotation estimation, is a technique for assessing

how the results of a statistical analysis will generalize to an independent data set. In most real applications, only a limited amount of data is available, which leads to the idea of splitting the data: Part of data (the training sample) is used for training the algorithm, and the remaining data (the validation sample) are used for evaluating the performance of the algorithm. The validation sample can play the role of new data (Arlot and Celisse, 2010).

Leave-One-Out Cross-Validation

Leave-one-out cross-validation (LOOCV) is a special case of k-fold cross-validation where k equals the number of instances in the data.

The fitting process optimizes the model parameters to make the model fit the training data as good as possible This is called over fitting , and is particularly likely to happen when the size of the training data set is small, or when A confusion matrix shows the number of correct and incorrect predictions made by the classification model compared to the actual outcomes (target value) in the data.

ROC Curve

it's a technique used for visualizing the performance of a classifier, it has been extended to visualize, and rank, the performance of a competing set of classifiers for selecting the best of them; by plotting them together in the same graph (Fawcett, 2004). The ROC curve plots the sensitivity or (true positives) of a model on the vertical axis against 1-specificity or (false positives) on the horizontal axis. The result is a bowed curve rising from the 45 degree line to the upper left corner, the sharper bend and the closer to the upper left corner, the greater accuracy of the model.

Area under the curve (AUC)

ROC curves of better models that are closer to the left and top edges of the unit square . The different area under a ROC curve for a good model should be close to 1 (the area of the unit square). This suggests that the area under the ROC curve (AUC) might be a reasonable single number summary to use to compare the ROC curves of different models. Although their ROC curves two models may cross each other, the ROC curve of the better model will on average enclose a greater area.

3. Labor Force Data Analysis

Real data of LF Survey, 2012 where conducting by Palestinian Center Bureau of

Statistics (PCBS) used for application of the MLR and LDA. The data were used for the purposes of scientific research according to a special agreement between PCBS and Al –Azhar University- Gaza. The sample size had been 25353 observations. 62.7% residing in the West Bank 37.3% residing in Gaza Strip. Dataset contains 12 variables. We are interested on LF status(1) variable. This variable has been used as a dependent variable in this analysis. It involves three categories (30% Employment, 8.7% Unemployment, and Outside of LF is 61.3% by using SPSS and R software statistical package programs. The goal is to find the best model which can describe the relationship between different types of LF and other factors.

Concepts and Definitions

Employed: Persons aged 15 years and over who were at work at least one hour during the reference period, or who were not at work during the reference period, but held a job or owned business from which they were temporarily absent (because of illness, vacation, temporarily stoppage, or any other reason) he\ she was employed, unpaid family member or other. The employed person is normally classified in one of two categories according to the number of weekly work hours.

Unemployed :Unemployed persons are those individuals 15 years and over who did not work at all during the reference period, who were not absent from a job and were available for work and actively seeking a job during the reference period by one of the following methods news paper, registered at employment office, ask friends or relatives or any other method.

Outside Labor Force: The population not economically active comprises all persons 15 years and over, who were neither employed nor unemployed accordingly to the definitions above. Classifies persons outside labor force by reason in the following categories: Student, Housekeeping, Abstinent from work, Guest , Old, Illness, Retired

Table (3.1) : The explanatory variables

NO.	Var.	Description	Categories	Marginal Percentage
1	LF(1)	Labor Force Status (1)	1-Employment	30.0%
			2-Unemployment	8.7%
			3-Outside of LF	61.3%
2	Sex	Sex	1-Male	51.0%
			2-Female	49.0%
3	Age	Age at last birth day	15-65	100.0%
4	Attend	Does…currently attending school	1-Currently Attending	30.5%
			2-Attended and left	32.7%
			3-Attended and graduated	32.5%
			4- Never attended	4.2%
5	PR4	Educational Attainment(higher Qualification)	1-Illiterate	4.3%
			2-Can Read and Write	6.0%
			3-Elementary	22.4%
			4-Preparatory	33.1%

			5-Secondary	19.3%
			6-Associatte Diploma	4.6%
			7-BA\ BSc	9.3%
			8-Higher Diploma	.1%
			9-Master Degree	.6%
			10-Ph.D	.2%
6	PR6	Training course attendance (such as training course that managed by ministry of labour, Qalandia institute) must present certificate at the end of the training course	1-Currently Attending	.5%
			2-Attended and graduated	8.3%
			3-Attended and left	.3%
			4- Never attended	90.8%
7	HR5	Refugee Status	1-Registered	43.8%
			2-not Registered	.4%
			3-Not Refugee	55.8%
8	ID7	Locality Type	1-Urban	65.8%
			2-Rural	22.0%
			3-Camp	12.3%
9	WBGS	Region	1-West Bank	62.7%
			2-Gaza Strip	37.3%
10	Marital	Marital Status	1-Never Married	45.2%
			2-Ever Married	50.5%
			3-Other	4.3%
11	Industry	Industry group	1-Agriculture	6.4%
			2-Manufacturing	4.2%
			3-Construction	5.9%
			4-Commerce-Hotels	6.4%
			5-Transport-storage	2.1%
			6-services	74.9%
12	HR4	Relationship to the Head of Household	1-head	26.2%
			2-spouse	23.1%
			3-son\daughter	46.8%
			4-father\mother	1.0%
			5-brother\ sister	.9%
			6-grand father\mother	*
			7-grand child	.3%
			8-Son Wife\ Daughter Husband	1.3%
			9-Other relative	.4%

			10-Others	*

Note:"" less than 0. 1%*

In table (3.1) describe all variables in the study. There are 11 independent variables. All of them are nominal except age variable is quantitative.

3.1 Statistical Analysis of LF using MLR

In this section we will perform multinomial logistic regression analysis on the Labor force data set. One of the main assumptions of the MLR is the independence among the dependent variable choices (i.e Employment, Unemployment, Outside if LF). We will use Hausman-McFadden Test on a subset of alternatives. If IIA (Independence of Irrelevant Alternatives) holds, the two sets of estimates should not be statistically different, chisq = = 6.1196, df = 20, p-value = = 0.9987, Alternative hypothesis: IIA is rejected. So the dependent variable categories are uncorrelated.

Table (3.2) : Likelihood Ratio Tests

	Model Fitting Criteria	**Likelihood Ratio Tests**		
Effect	**-2 Log Likelihood of Reduced Model**	**Chi-Square**	**Df**	**Sig.**
Intercept	18565.678	0.000	0	
Age	19194.472	628.794	2	0.000
Sex	19146.736	581.058	2	0.000
Attend	21187.523	2621.845	6	0.000
PR4	21076.054	2510.377	18	0.000
PR6	18830.996	265.318	6	0.000
HR5	18573.680	8.002	4	0.091
ID7	18573.449	7.771	4	0.100
Wbgs	18674.448	108.771	2	0.000
Marital	18617.642	51.964	4	0.000
Industry	23917.124	5351.446	10	0.000
HR4	18826.274	260.596	18	0.000

Table(3.2) demonstrates the likelihood ratio test evaluates the overall relationship between an independent variable and the dependent variable. we checked the same point with all explanatory variables used to build model separately. The result was referred that the existence of a relationship between each of the explanatory variables and the response variable was supported. It is seen that there is a statistically significant relationship between all the independent variables and the dependent variable except two variables HR5 and ID7.

Since the dependent variable has three categories, so we have two models.

We know that the "Exp(B)" is predicted change in odds for a unit increase in the corresponding explanatory variable. Odds ratios less than 1 correspond to decreases and odds ratio more than 1.0 correspond to increases. Odds ratios close to 1.0 indicates that unit changes in that explanatory variable does not affect the response variable.

Interpretation of the odds ratio of the parameter Estimation

1- **Sex variable Categories**

Likelihood ratio test information (chi – square = 628.794)

Table(3.3) : The parameter Estimates of response variable versus Sex

Response variable	Sex variable	B	Std. Error	Df	Sig.	Exp(B)
Employment	Intercept	-11.18		1		
	[sex=1]	1.847	0.088	1	0	6.342
	[sex=2]	0^b		0		
Unemployment	Intercept	-29.86	3246.573	1		
	[sex=1]	1.794	0.101	1	0	6.015
	[sex=2]	0^b		0		

"b "means the variable is base category

Sex has two categories, 1-male, 2-female. For response *LF(1)* by *Employment*, the odds ratio of Sex as "male" is 6.342. we can therefore say that the odds of being *Employment* rather than *Outside of LF* is increased by a factor of 6.342 by being the *"sex"* of the person is male rather than the person is female, controlling for other variables in the model. In the same way, we can interpret the parameter estimates of the response *LF(1)* by *Unemployment* the odds ratio for *"sex"* as male is 6.015. We can say that the odds of being

Unemployment rather than Outside of LF is increased by factor of 6.015 by being the person is male rather than the person is female, controlling for other variables in the model.

2- Region (Wbgs) variable Categories

Likelihood ratio test information (chi – square = 108.771)

Table (3.4) : The parameter Estimates of response variable versus Region

Response variable	Region variable	B	Std. Error	Df	Sig.	Exp(B)
Employment	Intercept	-11.18	3747.045	1		
	[wbgs=1]	0.298	0.059	1	0	1.347
	[wbgs=2]	0^b		0		
Unemployment	Intercept	-29.86	3246.573	1	0.993	
	[wbgs=1]	-0.335	0.071	1	0	0.715
	[wbgs=2]	0^b		0		

Region variable has two categories, 1-West Bank, 2-Gaza Strip For response *LF(1)* by *Employment*, the odds ratio of *Region* as " West Bank " is 1.347. we can therefore say that the odds of being Employment rather than Outside of LF is increased by a factor of 1.347 by being the " *Region* " of the person is in West Bank rather than the person is in Gaza Strip, controlling for other variables in the model. In the same way, we can interpret the parameter estimates of the response *LF(1)* by *Unemployment* the odds ratio for " *Region* " as West Bank is 0.715. We can say that the odds of being *Unemployment* rather than Outside of LF is decreased by factor of 0.715 by being the person is in West Bank rather than the person is in Gaza Strip, controlling for other variables in the model. P- value for Region variable categories is 0.0 which is less than 0.05 , so region variable is significance and is contained in the model.

3- Industry variable

Likelihood ratio test information (chi – square = 5351.446)

Table (3.5) : The parameter Estimates of response variable versus Industry

Response variable	Age variable	B	Std. Error	Df	Sig.	Exp(B)
Employment	Intercept	-11.18		1		
	[Industry=1]	4.538	0.1	1	0	93.528
	[Industry=2]	4.163	0.152	1	0	64.269
	[Industry=3]	2.919	0.126	1	0	18.53
	[Industry=4]	3.913	0.12	1	0	50.034
	[Industry=5]	2.768	0.197	1	0	15.923
	[Industry=6]	0^b		0		
Unemployment	Intercept	-29.86		1		
	[Industry=1]	3.409	0.133	1	0	30.246
	[Industry=2]	2.968	0.18	1	0	19.443
	[Industry=3]	3.029	0.139	1	0	20.675
	[Industry=4]	2.499	0.149	1	0	12.168
	[Industry=5]	1.812	0.225	1	0	6.124
	[Industry=6]	0^b		0		

Table (3.5) demonstrates the *Industry* variable has the maximum value of chi – square among all response variables (chi-square = 5351.446, df =10)

Industry has six categories, 1- Agriculture, 2- Manufacturing, 3- Construction, 4 -Commerce, Hotels and Restaurant, 5- Transport, Storage, and 6- Communication (services). For response *LF (1)* by *Employment*, the odds ratio of *Industry* as *Agriculture* is 93.528. So the odds of being *Employment* rather than *Outside of LF* is increased by a factor of 93.528 by being the *Industry* is *Agriculture* rather than *Services,* controlling for other variables in the model. The odds of being *Unemployment* rather than *Outside of LF* is increased by factor of 30.246 by being *Agriculture* rather than *Services*, controlling for other variables in the model.

The odds of being *Employment* rather than *Outside of LF* is increased by a factor of 18.530 by being *Industry* as *Construction* rather than *Services*, controlling for other variables in the model. By comparison with the odds of *Construction* in *Unemployment* is 20.675. All Refugee Status variable aren`t significant so, HR5 isn`t contained in the model of Employment but *HR5=1(Registered)* is significant only in Unemployment model. Other risk factors have approximately similar interpretation.

Table (3.6): The parameter Estimation

		B	**S. E**	**Df**	**Sig.**		**B**	**S.E**	**Df**	**Sig.**
Employment	Intercept	-11.18				**Unemployment**	-29.86			
	Age	-0.052	0.003	1	0		-0.078	0.004	1	0
	[attend=1]	-1.987	0.663	1	0.003		-2.369	1.189	1	0.046
	[attend=2]	1.318	0.657	1	0.045		2.638	1.179	1	0.025
	[attend=3]	0.946	0.661	1	0.152		2.286	1.182	1	0.053
	[attend=4]	0^b		0			0^b		0	
	[PR4=1]	-4.279	1.012	1	0		14.83	3246.57	1	0.996
	[PR4=2]	-5.462	0.788	1	0		13.55	3246.57	1	0.997
	[PR4=3]	-5.112	0.782	1	0		13.68	3246.57	1	0.997
	[PR4=4]	-4.831	0.782	1	0		13.79	3246.57	1	0.997
	[PR4=5]	-4.241	0.781	1	0		14.29	3246.57	1	0.996
	[PR4=6]	-2.245	0.783	1	0.004		16.45	3246.57	1	0.996
	[PR4=7]	-1.296	0.782	1	0.097		17.34	3246.57	1	0.996
	[PR4=8]	0.282	1.026	1	0.783		18.46	3246.57	1	0.995
	[PR4=9]	-0.582	0.827	1	0.481		16.79	3246.57	1	0.996
	[PR4=10]	0^b		0			0^b		0	
	[PR6=1]	-1.646	0.299	1	0		-2.618	0.426	1	0
	[PR6=2]	0.99	0.084	1	0		1.199	0.097	1	0
	[PR6=3]	-0.312	0.427	1	0.465		0.616	0.446	1	0.167
	[PR6=4]	0^b		0			0^b		0	
	[HR5=1]	0.021	0.059	1	0.719		0.173	0.071	1	0.015
	[HR5=2]	-0.277	0.367	1	0.45		-0.068	0.463	1	0.884
	[HR5=3]	0^b		0			0^b		0	
	[ID7=1]	-0.122	0.08	1	0.129		-0.041	0.096	1	0.668
	[ID7=2]	-0.025	0.094	1	0.789		-0.102	0.115	1	0.376
	[ID7=3]	0^b		0			0^b		0	
	[Maritals=1]	-0.263	0.176	1	0.135		-0.022	0.235	1	0.924

	[Maritals=2]	0.314	0.159	1	0.049		-0.135	0.231	1	0.559
	[Maritals=3]	0^b		0			0^b		0	
	[HR4=1]	14.536	3747.045	1	0.997		13.522	0.502	1	0
	[HR4=2]	13	3747.045	1	0.997		12.375	0.5	1	0
	[HR4=3]	13.517	3747.045	1	0.997		12.978	0.493	1	0
	[HR4=4]	13.967	3747.045	1	0.997		-1.201	942.452	1	0.999
	[HR4=5]	13.632	3747.045	1	0.997		13.25	0.569	1	0
	[HR4=6]	0.373	6242.142	1	1		0.119	8389.08	1	1
	[HR4=7]	13.243	3747.045	1	0.997		13.318	0.751	1	0
	[HR4=8]	11.598	3747.045	1	0.998		12.38	0.538	1	0
	[HR4=9]	12.987	3747.045	1	0.997		13.069	0	1	
	[HR4=10]	0^b		0			0^b		0	

Estimates MLR Models

MLR Model (1) Employment

Logit(π_1) = -11.18 -0.052[Age] +1.847[Sex=1] -1.987[Attend=1]+ 1.318[Attend=2] -4.279[PR4=1] -5.462[PR4=2] -5.112[PR4=3] -4.831[PR4=4] -4.241[PR4=5] -2.245[PR4=6] -1.646[PR6=1] +0.99[PR6=2] +0.298[WBGS=1] +0.314[Marital=2] +4.538[Industry=1] +4.163[Industry=2] +2.919[Industry=3] +3.913[Industry=4] +2.768[Industry=5]

MLR Model (2) Unemployment

Logit(π_2) = -29.86 -0.078[Age] +1.794[Sex=1] -2.369[Attend=1] +2.638[Attend=2] -2.618[PR6=1] +1.199[PR6=2] +0.173[HR5=1] -0.335[WBGS=1] +3.409[Industry=1] +2.968[Industry=2] +3.029[Industry=3] +2.499[Industry=4] +1.812[Industry=5] +13.522[HR4=1] +12.375[HR4=2] +12.978[HR4=3] +13.25[HR4=5] +13.318[HR4=7] + 12.38[HR4=8]

Any of the categories can be chosen to be the baseline. The model will fit equally well, achieving the same likelihood and producing the same fitted values. Only the values and interpretation of the coefficients will change. Outside of LF category is chosen to be base category.

3.2 Statistical Analysis of LF using DA

Wilks' lambda is a test statistic used in multivariate analysis of variance (MANOVA) to test whether there are differences between the means of identified groups of subjects on a combination of dependent variables. It results chi-square tests of significance for the function. The associated chi-square statistic tests the hypothesis that the means of the functions listed are equal across groups. The small significance value means that the discriminant function is good because it does at separating the groups well.

Table (3.7) :Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	Df	Sig.
1 through 2	.406	22842.511	22	.000
2	.949	1327.807	10	.000

It is clear that the second function is significant and the combination of the two functions are significant too. Wilks' lambda combines both discriminant functions allows you to predict all but 0.406 of the variation in level of LF Status(1). we can see that the Wilks' Lambda is big (.949) and has a probability of 0.0 which was less than the level of significance of 0.05 .This means about 94.9% of the variance unexplained. But when we add the first function to the predictive equation, we reduce the unexplained variance to only about 40.6% .

The standardized canonical discriminant function coefficients

The standardized coefficients can compare variables measured on different scales. Coefficients with large absolute values correspond to variables with greater discriminating ability, so it measure the relative importance of the selected variables, the larger absolute value of the coefficient corresponds to greater discriminating ability, and mean that the groups differ a lot on that variables. The coefficients will see how the original variables combine to make a new one that maximally "separates" the LF Status(1) categories. You can interpret the standardized discriminant function coefficients as a measure of the relative importance of each of the original predictors.

Table (3.8) Discriminant Functions Coeff.

Independent variables	Standardized Canonical Discriminant Function		Canonical Discriminant Function Coeff.	
	1	2	1	2
Sex	.516	-.114	1.159	-.256
Age	.282	-.538	.017	-.033
Attend	-.514	.892	-.614	1.067

PR4	-.478	.037	-.331	.025
PR6	.143	-.098	.247	-.170
HR5	.017	-.085	.018	-.086
ID7	-.008	-.022	-.011	-.031
WBGS	.051	.264	.106	.549
Marital	-.135	-.404	-.239	-.717
Industry	.734	.225	.575	.176
HR4	.158	.307	.138	.267
(Constant)			-3.651	-1.211

Table (3.8) provides two functions. At function one, the sign indicates the direction of the relationship. Industry group (.734) was the strongest predictor while sex group (0.516) was next in importance as a predictor. These two variables with large coefficients stand out as those that strongly predict allocation to the LF categories.

At function two, the strongest predictor is *Attend* variable with value .892. The second one is Age variable with value -0.538. The discriminant function coefficients *b* or standardized form *beta* both indicate the partial contribution of each variable to the discriminate function controlling for all other variables in the equation. These unstandardized coefficients (b) are used to create the discriminant function (equation). At function 1 discriminant score is -3.651. New cases would be classified into groups depending on the group whose centroid their own vector of scores was closest to it.

DA Models

D_1= -3.651 + 1.159 Sex + 0.017 Age -0.614Attend -0.331 PR4 + 0.247PR6 +0.018 HR5 - 0.011ID7

+0.106WBGS -0.239Maritals + 0.575Industry +0.138 HR4

D_2 = -1.211 -0.256 Sex -0.033 Age +1.067Attend +0.025 PR4 -0.170 PR6 -0.086HR5 -0.031ID7

+0.549WBGS -0.717Marital +0.176 Industry +0.267HR4

3.3-Comparison between MLR &DA

A- Classification Methods

The classification table, also called a confusion matrix, is simply a matrix in which the rows are the observed categories of the dependent and the columns are the predicted categories. When prediction is perfect all cases will lie on the main diagonal. The percentage of cases on the diagonal is the percentage of correct classifications.

Table (3.9) Correct Classification of two statistical methods

L F Status (1)	MLR	LDA
Employment	82.3%	75.6%
Unemployment	18.3%	12.2%
Outside of LF	93.3%	93.50%
Overall Classification of model	83.5%	81.10%

Table (3.9) shows the comparison between the two models in terms of their accuracy rate using LOOCV method . MLR model can correctly classify the first category Employment with accuracy (82.3%) compared to (75.6%) for DA model . At the second category Unemployment, Correct classification accuracy of MLR model (18.3%) is higher than DA model (12.2%) . Both of them have nearly the same percentage for Outside of LF which equal 93.5%. The correct classification rates for all categories by the MLR model is better than DA model. The classification accuracy of DA is estimated at 81.1% , and the misclassification rate is 18.9 % . The classification accuracy of MLR using the cross-validation is estimated to be equal 83.5 % , and the misclassification rate to be equal 16.5 % .

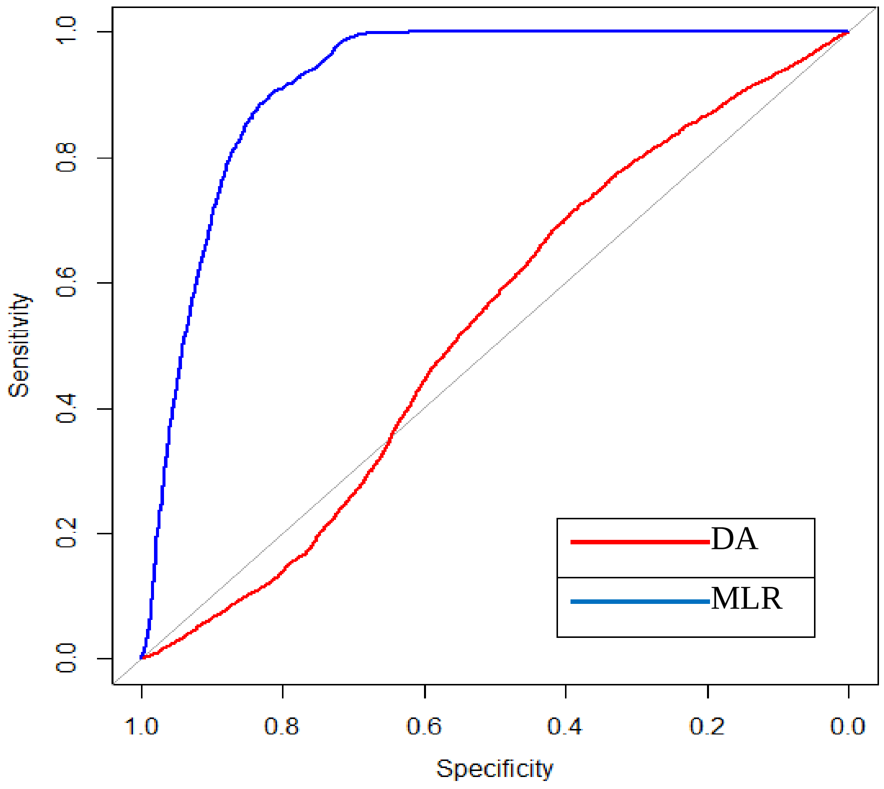
(b) ROC Curves

For evaluation the two models, ROC curves show the area. Moreover the two ROC curves are displayed in one figure to make the comparison easier.

Table (3.10): Area under the curve for two models

Model	Area (MLR)	Area(DA)
Area under the ROC curve	91.89%	52.8%

Figure (1):ROC Curves of MLR & DA Models



ROC curve of all the two statistical models have been drawn in the same graph. Figure (1) can clearly show that, the ROC curve of the MLR is always higher than the other model. This indicates that the MLR model provides classification for all categories with much higher accuracy than DA model when the two methods are applied at LF (2012) data. The areas under the ROC curves are computed and presented in table (3.10). The area under curve for MLR model is 91.89 % compared to that of DA is (52.8%) . Results of the MLR proved to be much better than DA. Finally, we can conclude that MLR model is the best model since its curve is the closest to the upper right hand corner in the graph.

The result of classification are implemented by R statistical package with some functions (DA) and libraries: library (Mass), library(foreign), library(mlogit), library (ROCR), library (fmsb), and library (pROC).

4- Conclusion and Recommendations

In this study, we have used two different classification methods, MLR and DA. Using different assessment techniques in order to achieve to the best model that represents the dataset of Labor Force. We compared the performance of DA and MLR on LF data. The sample size has the most obvious impact on the difference between and the errors it makes in prediction two techniques. DA assumes normality but MLR assumes nothing about it. Both methods are different in results. Correct classification is 83.5% for MLR model compared with 81.1% for DA. In addition that the area under the ROC curve is 91.89 % for MLR and 52.8% for DA. The model means that any one (observation) in Palestinian region (West bank and Gaza Strip) can answer 11 questions (independent variables) and the age is between (15- 65). MLR and DA models can classify it into one of three groups (Employment, Unemployment and Outside LF) with misclassification 16.5 % and 18.9% respectively. These results demonstrate that MLR gain popularity. According to the results, we may recommend that a researcher should use Multinomial logistic regression method because of its efficiency in predicting and classifying, and use it in other fields, such as medical and physical researches. The obtained results are consistent with the findings of Pohar et al. (2004), where the MLR performs better than DA when the number of categories lower than 5. We recommend to use Multinomial logistic regression methods in classifying and prediction technique in other fields, such as medical research, genetics research and physics research. It`s important to give priority to unemployment problem especially in Gaza Strip in any program that aims to limit the unemployment rate. More attention and focus should be given to governorates with high level of unemployment. In regard to unemployment rate, it increased from 38.5% in the 4th quarter 2013 to 40.8% in the 1st quarter 2014 while it remained steady in the West Bank at 18.2% in the same period. highest unemployment rate in Khan Younis 46.4%. Using MLR for analysis Labor Force data with additional independent variables to achieve a model with higher qualification and less error rate.

References:

Al-khatib, H and Al-Horani , A (2012), " *Using Logistic Regression and Discriminant Analysis to predict financial distress of publicly listed companies in Jordan* ", European Scientific Journal, Vol.8, No. 15, PP. 1 -17.

Andrew R. Webb. (2002), *"Statistical Pattern Recognition"*, 2nd Edition, John Wiley & Sons, Ltd.

Antonogeorgos G , Demosthenes B , Kostas N , Anastasia T. (2009), "*Logistic Regression and Discriminant Analyses in Evaluating Factors Associated with Asthma Prevalence*: *Divergence and Similarity of the Two Statistical Methods"*, International Journal of Pediatrics, Vol. 26, No. 3, PP. 211- 229.

Arlot Sylvain, Celisse Alain. (2010), "*A Survey Of Cross-Validation Procedures For Model Selection"*, Statistics Surveys: Vol. 4, PP. 40 –79.

Chao.Y.P and Rebecca.N.N (2002), "*Using multinomial logit models to predict adolescent behavioral risk"*, Journal of Modern Applied Statistical Methods, Spring 2003, Vol. 2, No.1.

Chatterjee S, and Hadi A (2006) "*Regression Analysis by Example*", John Wiley & Sons.

Fawcett, Tom (2004), *"ROC Graphs: Notes and Practical Considerations for Researchers"*. Technical report HPL No. 4, PP. 1- 38.

Garson D. (2009), *"Logistic Regression With SPSS"*, North Carolina State University, Public administration Program.

Hamid & Hashibah (2010). *A new approach for classifying large number of mixed Variables*. World academy of science, Engineering and Technology.

Hastie T., Friedman J., Tibshirani R.(2009), "*The Elements of Statistical Learning"*, 2nd Edition, Springer-Verlag, New York.

Hosmer, David & W. Lemeshow, Stanley (2000). *Applied logistic regression*. 2nd edition, New York: Wiley.

Hussain M, Wright S, Petersen L.(2002), "*A Comparing performance of multinomial logistic regression and discriminant analysis for monitoring access to care for acute myocardial infarction"*, Journal of Clinical Epidemiology, Vol. 55, No. 4, PP. 400-406.

Moorman, S.M., and Carr, D. (2008), " *Spouses effectiveness as end-of-life health care surrogates: Accuracy, uncertainty, and errors* *of overtreatment or under treatmen"*, The Gerontologist, 48, 811.

Nichols J.J, Ziegler C, Mitchell G.L, and Nichols K.K (2005), "*Self-Reported Dry Eye Disease across Refractive Modalities"*, Association for Research in Vision and Ophthalmology, Vol. 46, No. 6, PP. 1911 -1914.

PCBS, (2012), "*User guide, Labor Force Survey 2012*",Ramallah, Palestine: Palestinian Central Bureau of Statistics.

Pohar M., Blas M., and Turk S. (2004), "Comparison of logistic regression and linear discriminant analysis: a simulation study", Metodoloski Zvezki, Vol. 1, No. 1, pp. 143–161.

Raymo J.M and Sweeney M.M, (2006), "*Work-Family Conflict and Retirement*

Preferences", The Journals of Gerontology series B : Psychological Sciences and Social Sciences, Vol.61, No. 3 PP. S161-S169.

Slingerland A.S, Van Lenthe F.J, Jukema J.W, Kamphuis C.B.M, Looman C, Giskes.K, Huisman M, Narayan K.M.V, Mackenbach J.P, and Brug J. (2007), "*Aging Retirement, and Changes in Physical Activity : Prospective Cohort Findings from GLOBE Study*", American Journal of Epidemiology, Volume 165, Number 12, pp. 1356-1363.

Timm Neil H. (2002), *"Applied Multivariate Analysis"*, Springer, Verlag, New York, USA.