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

ISSN 2307-4531
(Print & Online)

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

The Logistic Regression Models and the Accuracy Measures in Analyzing Medical Data

Taghreed Al-Said *

A lecturer of Statistics at AL AZHAR University, Faculty of Commerce, Department of Statistics, Cairo, Egypt
Email: taghreed_minna@yahoo.com / taghreed_1177@yahoo.com

Abstract

There are many statistical models for analyzing categorical outcome variable such as the classification models, the discriminant models, and the logistic regression models. This research introduces a comparison study between these regression models to analyze medical data on the basis of several measures of predictive accuracy to set proper choice between them. A reviewed of such models and the assumptions of each model will introduce in the research. Many evaluation and accuracy measures for evaluating models are presented. Two applications in the medical field are introduced using two different real data sets to make the comparisons between models. The analysis of the results and conclusions are concluded in the research.

Keywords: Ordinary regression models; logistic regression models; the R-squared measure; the Wilks' Lambda measure.

1. Introduction

The ordinary regression models, the discriminant linear models and the logistic regression models are the widely used statistical models for analyzing data sets and explain the relationship between variables. Both discriminant and logistic regression models are appropriate and widely used in analyzing categorical outcome variables especially the medical data sets. Also, these models success in classifying observations to pre-defined groups.

----

* Corresponding author.

Linear discriminant models have many assumptions to give useful results such as the normality assumption, the significance difference centroids, and the equality variance- covariance matrices. In practice these assumptions are nearly violated especially with real data, therefore researchers have to use the logistic models in analyzing the data. The logistic models have not any assumptions on the distribution of the explanatory variables or about the variance - covariance matrices. Sometimes these models named the free model assumptions. Many researchers use the ordinary regression models, or the discriminant models without insure the existence of the assumptions or although the violation of the assumptions. This research interested in studying and analyzing the logistic regression models in a comparison with the ordinary regression models and the linear discriminant models. Also, the evaluation measures that are used with these models are introduced in the research to reveal the best model and evaluate the supposed model. Two suitable applications of these models in the medical field are done to complete the comparison. A review of the regression models, the ordinary, discriminant, and logistic regression models are stated in section (2). Details of the evaluations measures used with these models are in section (3). Section (4) has two applications of the logistic models by using two medical real data sets, the breast cancer data set, and the thyroid gland diseases data set. The analysis of the results and conclusions are presented in section (5). Section (6) has the recommendations for future studies. Finally, at the end of the research the references are stated.

2. A Review of The Regression Models

This section has the details of three regression models, its assumptions and forms. The reviewed models are the ordinary regression models, the discriminant analysis models, and the logistic regression models.

2.1 The ordinary regression models

The ordinary regression models are the most frequently used in analyzing data sets. There are two ordinary regression models, the simple and multiple regression models. The ordinary models deal with continuous outcomes, and the explanatory variables. The general form of these modes can be defined as follows:

$$Y= f\left(X_1, X_2, \dots, X_p\right) + \varepsilon \qquad (1)$$

where the random error ε represents the discrepancy in the approximation. It accounts the failure of the model to fit the data exactly. The function $f\left(X_1, X_2, \dots, X_p\right)$ describes the relationship between the continuous dependent variable Y, and the predictor variables $X_1, X_2, \dots, X_p$. If there is one predictor variable, the simple regression model arises, whereas many predictor variables arise the multiple regression models [4, 8]. For categorical outcomes, the discriminant models and logistic models can be used instead of the ordinary models. It is a big mistake to use ordinary models in analyzing such cases, and results will be suspected. The details of linear discriminant and logistic regression models will be defined.

2.2 The linear discriminant models

The regular and the basic methods used to differentiate the between two or more groups are the discriminant models when the response variable is categorical or non- metric variable [11]. The observations can be classified to a suitable group on the basis of the predictor variables. All discrimination models assume categorical outcomes, normality assumption of the explanatory variables, and the equality variance – covariance matrices for groups [7]. The discriminant analysis models are appropriate when the dependent variable is a categorical, nominal or nonmetric variable and the independent variables are metric or non-metric variables [8]. The simplest linear discriminant, Fisher model for two groups is defined as a linear discriminant function passes through the centroids of the two groups. The conditional distribution of $x \setminus y$ has multivariate normal distribution with mean vector is μ_y and common covariance matrix Σ. It can be defined as follows:

$$p(y_i \setminus x_i) = \quad = \frac{1}{(1+e^{\alpha+\beta x})^{-1}} \tag{2}$$

Where the coefficient $\alpha = -\,log\frac{\pi_1}{\pi_0} + 0.5\,(\mu_1+\mu_0)'\,\Sigma^{-1}\,(\mu_1-\mu_0)$, and the parameters vector $\beta = (\mu_1 - \mu_0)^T\,\Sigma^{-1}$. The prior probabilities π_0, π_1 are the probabilities of belonging to group one, and to group two respectively. In practice all parameters are unknown, and will be replaced by the sample estimate, where $\hat{\pi}_0 = \frac{n_0}{n}$, $\hat{\pi}_1 = \frac{n_1}{n}$, $\widehat{\mu_1} = \frac{1}{n_1}\sum_{y_i=1} x_I$, $\widehat{\mu_0} = \frac{1}{n_0}\sum_{y_i=0} x_I$, and $\sum$ is defined as follows: $\sum = \left[\sum_{yi=1}(x_i - \bar{x}_1)(\,x_i - \bar{x}_1\,)^t \;+\; \sum_{yi=0}(x_i - \bar{x}_0)\,(\,x_i - \bar{x}_0\,)^t\right]/n$ [2].

The form of equation (2) is equivalent the form of logistic regression models. Hence, the two models do not differ in the functional form. The forms will differ only in the estimation methods of the coefficients. All linear forms of discriminant models can be used but on the basis of strict condition about normality, or equality variance – covariance matrices. Also, there are many fixable models of discriminant analysis such as quadratic discriminant models that based on only one condition about the equality of variance – covariance matrices, and the logistic regression models have not any conditions or restrictions [6].

2.3 The Logistic regression models

There are many objectives, types, applications of the logistic regression models. The logistic models have many objectives such as finding the best-fitting model, describe the relationship between the categorical outcome (dependent or response) variables, and a set of independent (predictor or explanatory) variables [10]. These models are robust, flexible, and easily used. It has not any assumptions regarding the distribution of the explanatory variables such as linear discriminant models. The logistic regression models are the suitable models for dichotomous, binary outcome variable Y [9]. The new cases can be classified to only one group by using the logistic models. This is the second goal of the logistic models [3].

The details of the relation between the regression models and the logistic models are defined in [6]. The two types of the logistic models are also defined, the binary and multinomial logistic models.

The logistic discriminant method was applied to differentiate malignant from benign proven breast lesions in a group of patients based on ultrasonic parameters using a database including 273 patients' ultrasonography pictures consisting of 14 quantitative variables.

The logistic regression models have many applications in many fields especially the medical fields. The logistic models can be defined as follows,

$$p(y_i \backslash x_i) == \frac{e^{\beta^t x_i}}{1 + e^{\beta^t x_i}} \qquad (3)$$

Where the response variable y_I has Bernoulli distribution. The coefficient of this model is estimated using the maximum likelihood method.

The categories have to be mutually exclusive and exhaustive; new cases have to classify to only one group [1].

3. The Evaluation Measures of the Logistic Models

There are many measures can be used with the logistic regression models and other regression models such as the Fisher models, and the ordinary regression models. This section reviewed some of these measures.

3.1 *The classification error* measure

The classification error, C.F. is the simplest and most frequently used criteria. It is the percent of incorrectly classified objects. However this measure is very simple and many researchers use it in many studies; it is a very insensitive and statistically inefficient measure [3].

3.2 The B-indexes measure

There are many different measures to determine and evaluate the predictive accuracy of models. The proposed evaluating measures have intuitive clearness, predictive the accuracy of models, and sometimes add to the classification error (C.E.). The B - index measures the average of squared difference between the estimated and the actual value. It can be defined as follows:

$$B=1-\sum_{i=1}^{n}[(P_i - Y_i)^2 /n] \qquad (4)$$

Where the probability of classification into groups are P_i, the sample size of both populations are n, and the actual group membership is Y_i (i has 1 or 0 value).

The B-indexes ranged between (0, 1) values, where 1 indicates perfect prediction. For cases of equally sized groups, B-indexes is 0.75[14]

3.3 The Q-indexes measure

This measure is similar to the B-indexes. It has the following form

$$Q = \sum_{i=1}^{n}[1 + log_2(p_i^{y_i}(1 - p_i)^{1-y_i})]/n \qquad (5)$$

The Q-indexes measure also ranged between (0,1). One indicates perfect prediction while 0 indicates random predictions. The Values less than 0 indicate worse prediction than random. When predicted probabilities equal to 0 or 1, the Q-indexes is undefined and cannot be calculated.

The two criteria B-indexes and Q-indexes consider the accuracy of prediction besides discrimination. These measures are the most used in specialized researches [10].

3.4 The Wilks's Lambda measure

The Wilks's Lambda measure is sometimes called the U statistic. It is used to test the significance of the independent variables. It has the same objective as ANOVA, F- test to test the null hypothesis that the canonical correlations are zero.

The small value of Wilks's Lambda for an independent variable, include more contributes of the variable. The test statistic takes the following form:

$$\lambda = \frac{|W|}{|W+H|} \qquad (6)$$

Where W is the residual variance, H is the variance due to the linear relationship, and W+H is the total variance. Wilks's Lambda has approximately chi-square χ^2 distribution with one degree of freedom. The Wilks's Lambda values vary from zero to one [12].

3.5 The Wald test

The Wald test is used to test the statistical significance of each coefficient (β) in the model. The Wald test can be defined as follows:

$$W = \frac{\hat{\beta}}{SE(\hat{\beta})} \qquad (7)$$

Where $\hat{\beta}$ is the maximum likelihood s of the parameter β, and SE($\hat{\beta}$) is the standard error of the maximum likelihood estimate. The squared of the statistics has chi-square distribution with one degree of freedom [12].

3.6 The likelihood-ratio test

The likelihood-ratio tests the overall significance of the p coefficients for the independent variables in the model. It uses the ratio of the maximised value of the likelihood function for the full model (L_1) and the maximum of the simpler model (L_o). This test has the following form:

$$-2\,log\left(\frac{L_o}{L_1}\right) = 2\,[log(L_o) - log(L_1)] \qquad (8)$$

The likelihood-ratio test is more replicable for small sample size than the Wald test. [2]

4. Applications of The Logistic Regression Models

This section provides two applications of the logistic regression models in the medical field by using real data sets, the breast cancer data set, and the thyroid gland diseases data set.

4.1 The application of the logistic model using the breast cancer data set

Breast cancer is a kind of cancer diseases that affects the breast cells of women or men. It usually starts in the inner lining of the milk ducts or the lobules that supply them with milk. A malignant tumour can invasive to other parts of the body starts in the lobule that is known as lobular carcinoma. If it develops in the ducts is named, ductal carcinoma [18]. The breast cancer is common and invasive in females worldwide. It is approximately %22.9 in women.

For both males and females this type is approximately %18.2 of all cancer deaths worldwide 18.2% [18]. According to the Ministry of Health in Saudi Arabia, in (2015) the breast cancer is the most common disease in the KSA, where 2,741 cases (%19.9) from women having breast cancer comparing with other types of cancer.

In the US and Arab countries, including Saudi Arabia, %50 of new cases are women for ages ranged from 65-52 [16]. After a woman is diagnosed with breast cancer, doctors have to determine whether it has spread, invasive or not. This step is named staging.

The stage helps in determining the seriousness of the cancer and the best treatment options the case need. There are three important stages of the interest and available in the data set. Stage one will be when tumour size less than 2 cm, there is no lymph node metastasis. Stage two will be when tumour size between 2-5 cm, there is no lymph node on the same side of breast, and there is or metastasis. Stage three will be when tumour size more than 5 cm, there is a lymph node on the same side of breast and there is no metastasis [13].

The data set of the application of the breast cancer stages was taken from King Abdul-Aziz Hospital from the pathology and laboratory medicine section (Histopathology Reports) reports from 2015–2016. There are 36 cases available, 12 cases belong to stage I, 12 cases belong to stage II and 12 cases belong to stage III. The details of the three stages data set are shown in Table 1 as follows.

The data is analyzed by using the SPSS package, and two models are applied, the logistic model by supposing the outcome variable has two stages (Early Cancer and Advance Cancer), also when the outcome variable have three stages I, II, and III as defined before. Three predictive variables are used which are helpful in the diagnostic, the number of lymph nodes involved, the cancer has reached nearby lymph nodes or not, and the tumor size.

The classification results of applying the logistic regression model for the two stages will be stated in Table 3 as follows:

Table 1: The breast cancer data for the three stages

Stage	Number of lymph nodes involved	The cancer has reached nearby lymph nodes	Tumor size
1	0	FALSE	1.2
1	0	FALSE	1.5
1	0	FALSE	1.2
1	0	FALSE	1.5
1	0	FALSE	1
1	0	FALSE	1.5
1	0	FALSE	1.1
1	0	TRUE	1
1	0	FALSE	1
1	0	FALSE	1
1	0	FALSE	0.5
1	0	TRUE	1
2	0	FALSE	3
2	2	TRUE	3.2
2	0	FALSE	2.2
2	3	TRUE	2.5
2	1	TRUE	3
2	0	FALSE	3
2	0	FALSE	3
2	0	FALSE	2.5
2	2	FALSE	3.5
2	0	TRUE	2.5
2	0	TRUE	2.5
2	0	FALSE	5
3	0	TRUE	5
3	0	TRUE	8
3	0	TRUE	11
3	0	TRUE	7
3	2	TRUE	9.3
3	1	TRUE	5
3	0	TRUE	5.5
3	3	TRUE	6.5
3	13	TRUE	3
3	9	TRUE	2.5
3	21	TRUE	5.5
3	0	TRUE	7

The details of the two stages data set are shown in Table 2 as follows:

Table 2: The breast cancer data for the two stages

Stage	Number of lymph nodes are involved	The cancer reaches nearby lymph nodes	Tumor size
Early	0	FALSE	1.2
Early	0	FALSE	1.5
Early	0	FALSE	1.2
Early	0	FALSE	1.5
Early	0	FALSE	1
Early	0	FALSE	1.5
Early	0	FALSE	1.1
Early	0	TRUE	1
Early	0	FALSE	1
Early	0	FALSE	1
Early	0	FALSE	0.5
Early	0	TRUE	1
Early	0	FALSE	3
Early	2	TRUE	3.2
Early	0	FALSE	2.2
Early	3	TRUE	2.5
Early	1	TRUE	3
Early	0	FALSE	3
Early	0	FALSE	3
Early	0	FALSE	2.5
Early	2	FALSE	3.5
Early	0	TRUE	2.5
Early	0	TRUE	2.5
Early	0	FALSE	5
Advance	0	TRUE	5
Advance	0	TRUE	8
Advance	0	TRUE	11
Advance	0	TRUE	7
Advance	2	TRUE	9.3
Advance	1	TRUE	5
Advance	0	TRUE	5.5
Advance	3	TRUE	6.5
Advance	13	TRUE	3
Advance	9	TRUE	2.5
Advance	21	TRUE	5.5
Advance	0	TRUE	7

Table 3: Classification result of logistic regression of breast cancer for two stages

Observed		Predicted stage		
		Early cancer	Advance cancer	
Stage	Early cancer	24	0	100.0
	Advance cancer	0	12	100.0
Overall percentage				100.0

The classification results of applying the logistic regression model for the three stages are stated in Table 4 as follows:

Table 4: Classification result of logistic regression for breast cancer for the three stages

Observed	Predicted			
	1	2	3	Percent correct
1	12	0	0	100.0%
2	0	12	0	100.0%
3	0	0	12	100.0%
Overall percentage	33.3%	33.3%	33.3%	100.0%

4.2 The application of the logistic model using the thyroid gland diseases data set

The thyroid is a gland at the base of the throat near the windpipe. It has a butterfly shaped, with a right lobe and a left lobe. A thin piece of tissue connects the two lobes. The thyroid makes hormones that help control heart rate, blood pressure, body temperature, and weight [17].

Thyroid diseases are very common in human, especially women. There are two types of thyroid diseases, hypothyroid and hyperthyroid; these cases give many effects such as weight gain, weight loss, and stress... Detecting the disease in earlier stage gives proper treatment to the patients. The TSH is thyroid stimulating hormone, and it is the best way to test the thyroid function. A high TSH level indicates that the thyroid gland is failing, and the thyroid is producing low hormone (hypothyroid). The opposite situation, when the TSH level is low, indicates that the person has an overactive thyroid that is producing too much thyroid hormone (hyperthyroidism). In most healthy individuals, a normal TSH value ranges from 0.4 to 4.0 m U/L (mill units per liter) means. Also the tests T3 and T4 are often useful to diagnosis hyperthyroidism, and often patients have hyperthyroid will have an elevated T3 level [5]. The data set is drown and loaded from the website (http: //repository. seasr.org/Datasets /UCI/ arff). There is a 3772 available case. The chosen random sample size is 100 cases divided into two groups, 47 cases have hypothyroid and the other group has 53 cases that have hyperthyroid [15]. The detail of data set is in table 5 as follows:

Table 5: The thyroid gland disease data

Group	T3	TSH	Group	T3	TSH
Hypothyroid	2.1	4.2	Hypothyroid	0.	2.7
Hyperthyroid	6.2	0.05	Hyperthyroid	2.9	0.25
Hypothyroid	1.4	0.09	Hyperthyroid	2	1.2
Hypothyroid	2.4	9.6	Hyperthyroid	1.9	0.22
Hypothyroid	1.1	4.23	Hyperthyroid	1.7	0.23
Hyperthyroid	2.3	0.6	Hyperthyroid	1.5	0.015
Hyperthyroid	2.4	0.9	Hyperthyroid	4	0.02
Hyperthyroid	2.7	0.03	Hyperthyroid	7.3	0.15
Hyperthyroid	3.4	0.4	Hyperthyroid	3.8	0.2
Hyperthyroid	3.6	0.3	Hyperthyroid	4	0.02
Hypothyroid	2.7	0.56	Hyperthyroid	3.9	0.005
Hypothyroid	2.2	7.9	Hypothyroid	1.7	6.2
Hypothyroid	2.9	5.23	Hyperthyroid	2.3	1.1
Hyperthyroid	4	0.015	Hyperthyroid	1.7	0.35
Hypothyroid	2.3	1.4	Hypothyroid	1.9	9.8
Hypothyroid	1.7	8.9	Hypothyroid	0.9	8.3
Hypothyroid	2.1	4	Hypothyroid	2.4	5.8
Hyperthyroid	1.9	0.15	Hyperthyroid	3.4	0.03
Hyperthyroid	1.8	0.1	Hypothyroid	2.2	4.0
Hyperthyroid	1.5	1.1	Hypothyroid	2.1	3.4
Hyperthyroid	5.5	0.15	Hypothyroid	2.1	1.8
Hypothyroid	2.2	1.6	Hypothyroid	0.8	9.9
Hyperthyroid	1.7	0.26	Hypothyroid	1.1	1.83
Hyperthyroid	1.9	0.1	Hypothyroid	1.7	8.2
Hyperthyroid	5.2	0.16	Hypothyroid	2.4	6.4
Hyperthyroid	2.5	0.3	Hypothyroid	1.2	8.9
Hyperthyroid	1.5	0.2	Hypothyroid	0.2	4.2
Hyperthyroid	2.5	0.035	Hypothyroid	1.4	1.2
Hypothyroid	5	0.2	Hypothyroid	2.5	20
Hypothyroid	3.9	1.9	Hypothyroid	1.4	9.4
Hyperthyroid	4.7	0.06	Hypothyroid	0.7	2.3
Hyperthyroid	3.8	0.01	Hypothyroid	0.2	2.3
Hyperthyroid	2.3	0.15	Hypothyroid	0.9	2.5
Hyperthyroid	3.7	0.02	Hypothyroid	2.1	8.3
Hyperthyroid	3.5	0.005	Hypothyroid	1.1	8
Hyperthyroid	3.2	0.24	Hypothyroid	1.3	13
Hyperthyroid	2.3	0.06	Hyperthyroid	2.4	0.01

Hyperthyroid	2.3	0.02	Hyperthyroid	2	0.13
Hyperthyroid	4	4.1	Hyperthyroid	1.7	0.12
Hyperthyroid	2	0.005	Hyperthyroid	1.9	0.04
Hyperthyroid	4.2	0.03	Hypothyroid	1.5	1.6
Hypothyroid	4.4	27	Hyperthyroid	1.9	0.25
Hypothyroid	2.7	6.3	Hypothyroid	0.3	8.6
Hyperthyroid	3.9	0.25	Hypothyroid	0.5	8.6
Hypothyroid	2.4	0.23	Hypothyroid	0.5	8.6
Hyperthyroid	3.8	0.72	Hypothyroid	1	1.78
Hypothyroid	2.6	31	Hyperthyroid	1.8	0.2
Hypothyroid	1.5	230	Hyperthyroid	1.4	0.22
Hyperthyroid	4	0.22	Hyperthyroid	2.1	0.2
Hypothyroid	2.2	1.6	Hyperthyroid	2.3	0.005

The package SPSS is used to analysis the data set and apply the logistic model. The classification results of the logistic regression for the two groups will be stated in Table 6 using the logistic model as follows:

Table 6: Classification result of logistic regression for thyroid gland disease data

Observed			Predicted group		
			Hypothyroid	Hyperthyroid	Percentage correct
Step 1	Group	Hypothyroid	40	7	85.1
		Hyperthyroid	1	52	98.1
	Overall percentage				92.0

5. The Analysis of the Results and Conclusions

This section has analysis of the results for the application of the two data sets, the breast cancer and thyroid gland diseases.

5.1 The analysis of the breast cancer data set

The classification results of the logistic model for the two stages reveal that the overall percentage is %100, and all respondents are classified correctly into early-stage cancer or advanced cancer. It is clear that the logistic model is the suitable model to deal with medical data. Also, the discriminant model is suitable model if the assumptions are present. By examining the equality means of the two groups (early stage and advanced stage), and three groups (stage I, stage II and stage III) respectively. The p-value of the tests are significance for all variables that means all variables are differ and separated at (sig. <0.05) for the two groups and the three groups.

The normality assumption will be tested by using the Shapiro-Walk’s test that is used on the breast cancer data for the two groups and three group applications. For the two stage groups, the advance stage is only normal distributed while the early stage is not normal distributed. For the three stage groups, stage I and stage III are normal distributed. Also, the homogeneity assumption can be detected by using the Box's M test. The results for the breast cancer data reveal no significance difference between the two groups case, while there is a significance difference between the three groups case.

The results of Walks’ Lambda measure that test the significance of discriminant function, show highly significance function (p-value = 0.000) and shows 16.3% unexplained for function 1, whereas there is 99.2% unexplained for function 2 for the two stage groups. For the three stage groups, results show significance functions where (p-value=0.000) at significance level 5% and 24.8% unexplained.

The test of chi-square goodness of fit and Pseudo R-square for three stages of breast cancer, reveal that there is a good classification of the observed values. The Cox and Snell’s R^2 indicate that %88.9 of the variation is explained by the logistic regression model, and the Nagelkerke and Mcfadden R^2 will be 1 indicate strong relationship of 100% between the predictors and the prediction. For the three and two stage groups, the results reveals that there are %100 overall classification by using the Fisher's model, although the severalty assumption is not satisfied for the early stage in the two stage groups, and not satisfied for the second stage for the three stages groups..

5.2 The analysis of the thyroid gland diseases data set

The classification results of the logistic regression for the two groups reveals %92 overall correct classification of all observed values. For group (Hypothyroid) the correct classification is %85.1, whereas the correct classification for the other group (Hyperthyroid) is %98.1. It is clear that the logistic model successes in classifying cases to the suitable group. Also, the discriminant model if its assumptions are present it gives suitable results. By examining the significance of the discriminant analysis function, Wilks’ Lambda test shows highly significance function (p-value <0.000) and shows 72.6% unexplained. The T3 hormone was the strongest predictor.

The results of the classification of the thyroid gland for the two groups by using the Fisher's model reveals there are %68.1 correctly classification for hypothyroid and 71.7% correctly classification for hyperthyroid although the severalty assumption is not satisfied for two groups.

The chi-square test value reveals the logistic model is significance that indicates a good fit model and the predictors do have a significant effect for explaining the relationship, while this test not significance for The Fisher model. The normality assumption tested by using the Shapiro-Wilks test. It shows that the two groups have not normal distributed. The homogeneity assumption will be tested and the results show significance difference between the two group variables.

5.3 Conclusions

This research interests in studying and applying the logistic regression models with medical data. It introduces two applications of the logistic regression models using two real data sets, the breast cancer data set and thyroid gland data set. The results of the logistic regression model for the breast cancer data set showed %100 correct classifications either for two stages or three stages application. For the gland disease application, the correct classifications are %92. The problems of classifying observations are solved without considering any conditions with increasing correct percentages. Although the discriminant analysis sometimes introduces good classification results, but the assumptions of discriminant models such as normality, or equality variances-covariances are violated in practice and applications.

6. Recommendations

The classical models such as ordinary regressions, and discriminant analysis may give good results although the violation of the assumptions. The suitable and frequently models for dialing with medical data sets are the logistic regression models either the binary or the multinomial models. These models have not any assumptions about the outcome variable and the independent variables. There are also many estimating methods can be used with the logistic regression models to estimate the parameters of the models. The available measures for evaluating the parameters, and models available in packages such as Minitab and SPSS or any else packages may help especially researcher not specialized; it is very helpful to classify cases and patients to the suitable group and lead to good decision. The researcher can also use the discriminant or any alternative statistical models if the assumptions are present or the free assumption methods such as logistic models else.

References

[1] P. Abdolmaleki, M.M. Dizagi, M. Vahead, and M. Gity. “Logistic discriminant anlysis of breast cancer using ultrasound measurements.” Iran Journal Radiat, Research, 2004, pp. 1-8.

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

[3] S. Ben Youssef and A. Rebai. “Comparison between Statistical Approaches and Linear Programming for Resolving Classification Problem.” International Mathematical Forum, 2007, pp. 3125-3141.

[4] S. Chatterjee and A. Hadi. Regression Analysis by Example. The United States of America, 2006, John Wiley & Sons, Inc.

[5] A. Christianson , and H. Bender. The Complete Idiot's Guide to Thyroid Disease (Idiot's Guides). New York, 2011, Penguin Group.

[6] C. M. Dayton. “Logistic regression analysis statistics and evaluation: linear programming for resolving classification problem”. International Mathematical Forum, 1992, pp. 3125-3141.

[7] S. Green, N.J. Salkind, and T.M. Akey. Using SPSS for Windows and Macintosh: Analyzing and understanding data. New Jersey, 2008, Prentice Hall.

[8] J.F. Hair, W.C. Black, R.E. Anderson, and B.J. Babin. Multivariate Data Analysis. (Vol. 7th), 2010, Pearson Prentice Hall.

[9] D.W. Hosmer and S. Lemeshow. Applied Logistic Regression. Canada, John Wiley & Sons Inc., 2000, Second edition.

[10] M. Pohar, M. Blas, and S. Turk. Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study. Metodoloski zvezki, 2004, pp.143-161.

[11] D.W. Stockburger. Muitivariate Statistics :Concept, Models, and Applications (Vol. III). Missouri State University, 1998

[12] S. Suleiman, I. Suleman, U. Usman, and Y.O. Salami. “Predicting an Applicant Status Using Principal Component, Discriminant and Logistic Regression Analysis”. International Journal of Mathematics and Statistics Invention, 2014, pp. 05-15.

Sites

[13] American Cancer Socity. What is thyroid cancer? Retrieved from American Cancer Socity. http: // www.cancer.org/cancer/thyroidcancer/detailedguide/thyroid-cancer-what-is-thyroid-cancer. [2016]

[14] F.E. Harrell and Jr. K. L. Lee. https:// pdfs.semanticscholar.org / 2f1a / c6401855dd86aacd47d8eb85b1c6b3f21615.pdf, [1985]

[15] M. Lichman. {UCI} Machine Learning Repository. (University of California, Irvine, School of Information and Computer Sciences) Retrieved from ". http://archive.ics.uci.edu/ml". [2013]

[16] Ministry of Health. National Campaign for Breast Cancer Awareness. Retrieved from:

http://www.moh.gov.sa/en/HealthAwareness/Campaigns/Breastcancer/Pages/ efault.aspx. [2015]

[17] National Cancer Institute. A Snapshot of Thyroid Cancer. Retrieved from National Cancer Institute: https: // www.cancer.gov / research /progress/snapshots / thyroid. [2014]

[18] C. Nordqvist. Breast Cancer: Causes, Symptoms and Treatments. Retrieved from Medical News Today. http: // www.medicalnewstoday.com/articles / 37136.php. [2016]