Comparison Logistic Regression and Discriminant Analysis in classification groups for Breast Cancer

Krieng Kitbumrungrat

Faculty of Information Technology, Rangsit University, Thailand

Summary
This research is a model group, the probability that a patient is detected at any breast cancer or not breast cancer. Assessment of characteristics of abnormal growth of breast cancer cells such as Clump Thickness (X_1), Uniformity of Cell Size (X_2), Uniformity of Cell Shape (X_3), Marginal Adhesion (X_4), Single Epithelial Cell Size (X_5), Bare Nuclei (X_6), Bland Chromatin (X_7), Normal Nucleoli (X_8), and Mitoses (X_9) are independent variable. The dependent variable is the probability that the patient is detected at any breast cancer or not breast cancer by using Logistic Regression Model and Discriminant Model. Conclude that Logistic Regression Model has 96.90% classification higher than Discriminant Model has 96.10% classification. Logistic Regression Model can used predicted variables 4 variables are Clump Thickness (X_1), Marginal Adhesion (X_4), Bare Nuclei (X_6) and Bland Chromatin (X_7). So the model in predicting the probability of collection for the classification of patients breast cancer is

$$\ln\left(\frac{P_i}{1-P_i}\right) = 10.104 - 0.535X_1 - 0.331X_4 - 0.383X_6 - 0.447X_7.$$

The study results reveal that the discriminant analysis can used predicted variables 9 variables for classifying groups of breast cancer and non- breast cancer. So the factor model for the classification of Fisher's linear discriminant functions affecting the breast cancer model. The classifying groups of breast cancer is $D_1=-23.536+1.63X_1+0.312X_2+0.495X_3+0.339X_4+0.995X_5+1.457X_6+1.198X_7+0.547X_8$ and the classifying groups of not breast cancer is $D_2=-3.288+0.742X_1+0.301X_2+0.057X_3+0.106X_4+0.708X_5+0.188X_6+0.664X_7+0.026X_8$

Keywords:
Breast cancer, Logistic Regression Model, Discriminant Model

1.Introduction

Cancer is one of the ten causes of the death of the world population. According to the World Health Organization, there were 58 million dead people worldwide in 2008, and 7.6 million of them died of cancer, which is 13% of the total dead people. At present, cancer has caused a great loss of lives of people, specifically liver cancer and breast cancer. Cancer is abnormal tumor cells growth which interfere normal cells, and divide themselves much more quickly than normal cells many times, going beyond the control of a human body. Tumor cells can spread to other parts of a human body, especially to lymph and blood without infecting former tumor cells. The property of tumor cells is they can grow very quickly. The central part of a tumor always lacks of nutrients, resulting in the dead cells of cancer. Consequently, an infection occurs easily because the dead cells and the lymph are good sources of food for diseases, and this can lead to blood infection, which finally causes cancer patients to die. Cancer causes the lack of oxygen in a human body because it consumes a lot of oxygen to help divide its cell, causing white blood corpuscle to work hard so as to eradicate cancer cell, which resulting in the low capability of the human body immune system. Thus, the human body organ in which cancer cells exist will lose its working property, and the spreading of cancer cells can also devastate working property of the nearby organs. The cancer cell will create angiogenesis substance, which causes more blood veins to feed cancerous tumors at a sufficient extent for the growth of cancer cell division. Breast cancer is a disease found mostly in females and it is a genetic transmission disease, causing by abnormal hormone, obesity, food with a high fat, and a spreading of cancer from other part of the organs to the breast [1].

The 4 ways of treatments of a patient with a breast cancer according to National Cancer Institute, Department of Medical Service, Ministry of Public Health (2008) are operation, radiation therapy, systemic therapy for the first stage cancer, and systemic therapy for the spreading stage cancer. For a medical treatment of a breast cancer, there are many ways of treatment used together to prevent a breast cancer from occurring again, and the consequences of this combination of treatments usually are unexpected side effects, such as nausea, hair falls, diarrhea, and anemia, causing a cancer patient an anxiety and a lot of mental and physical sufferings. A breast cancer needs a vast expense for a treatment, and taking leaves from a job to take care a breast cancer patient causes a cancer patient's relatives to lose a lot of income too. Notwithstanding the complete recovering of a cancer patient in stage 1 from a medical treatment, a cancer patient in a final stage has found unable to completely recover from the condition because cancer is a chronic disease.

Accordingly, the present study paid attention on the study of a prediction of tumor cell with abnormal growth of a breast cancer to detect of breast cancer or not breast cancer

Manuscript received May 5, 2012
Manuscript revised May 20, 2012

using 2 statistics methods: Logistic Regression Analysis and Discriminant Analysis. However, the potential of the Discriminant Analysis rests on ways the data are analyzed, and if an analysis is operated based on the assumption of the model or not. The Discriminant Analysis has a lot of assumptions about the model, that is to say, the interrelationship of groups and predictive variables must be a linear, predictive variables should be distributed together in a dichotomous dimension with not too much interrelationship so as to prevent a dichotomous linear. The Logistic Regression Analysis has less limitation than Discriminant Analysis.

2. Objectives

The present study aims at studying predictive group discriminant using Logistic Regression Analysis and Discriminant Analysis to predict probability of a cancer patient who already had a medical check to determine probability of having a breast cancer or not breast cancer. The breast cancer or non-breast cancer will be predicted based on tumor cells with abnormal growth used as predictive variables to see how predictive results are different and correct employing the two statistics analyses mentioned above.

3. Methodology

The present study is an experimental research conducted with the purpose to study a prediction of tumor cells with abnormal growth and the probability of breast cancer patient who has medically checked and showed not breast cancer, or who showed breast cancer using 680 sample database of Wisconsin Breast Cancer Database (WBCD). The dependent variable (y) was the probability of cancer of a breast cancer patient who has medically checked and showed not breast cancer, or who showed identified breast cancer and independent variables were Clump Thickness (X_1), Uniformity of Cell Size (X_2), Uniformity of Cell Shape (X_3), Marginal Adhesion (X_4), Single Epithelial Cell Size (X_5), Bare Nuclei (X_6), Bland Chromatin (X_7), Normal Nucleoli (X_8) and Mitoses (X_9). The techniques Logistic regression Model and Discriminant Analysis were implemented in classification groups of a breast cancer or not breast cancer in the main study.

3.1. Logistic Regression

Logistic regression is a well known statistic method to analyze problems with dichotomous (binary) dependent variables. Assuming the two outcomes of the dependent variable are 1 and 0, respectively [2].
Logistic Regression Models the posterior probability that a case is in one of two or more classes conditional on x [3].

Let $P(x)=E(Y|x)$, the expected value of Y conditional on x. The Logistic Regression Models is specified:

$$P(x)=\frac{\exp^{\alpha+\beta_1x_1+\beta_2x_2+...+\beta_ix_i}}{1+\exp^{\alpha+\beta_1x_1+\beta_2x_2+...+\beta_ix_i}} \quad (1)$$

which is nonlinear in x. The logit transformation

$$y(x)=\ln\left(\frac{p_i}{1-p_i}\right)=\alpha+\beta_1x_1+\beta_2x_2+...+\beta_ix_i \quad (2)$$

where P_i is the probability of any breast cancer, α and βi are unknown parameters that determine the shape of the logistic curve. The unknown parameters α and βi were estimated using a maximum-likelihood approach [4]. The first estimated from the training data; with theformulated model, the probability $P(Y=1)$ of a test case can be calculated using (3) and compared with the predefined threshold such that the class label of the testcase can be determined [5].

$$P(Y=1)=\frac{1}{1+\exp^{-(\alpha+\beta_1x_1+\beta_2x_2+...+\beta_ix_i)}} \quad (3)$$

3.2 **Discriminant Analysis**

The discriminant function score [7] for the i^{th} function is: $D_i=d_{i1}Z_1+d_{i2}Z_2+...d_{ip}Z_p$, (4)
where Z= the score on each predictor, and
d_i = discriminant function coefficient. The discriminant function score for a case can be produced with raw scores and unstandardized discriminant function scores. The discriminant function coefficients [8] are, by definition, chosen to maximize differences between groups. The mean discriminant function coefficient can be calculated for each group these group means are called Centroids, which are created in the reduced space created by the discriminant function reduced from the initial predictor variables. Differences in the location of these centroids show the dimension along which the groups differ. Once the discriminant function are determined groups are differentiated, the utility of these function can be examined via their ability to correctly classify each data point to their a priori groups. Classification functions are derived from the linear discriminant functions to achieve this purpose. Different classification functions are used and equations exist that are best suited for equal or unequal samples in each group. For cases with an equal sample size for each group classification function coefficient (C_j) is equal to the sum of : $C_j=c_{j0}+c_{j1}x_1+c_{j2}x_2+...+c_{jp}x_p$ (5)
for the j^{th} group, $j=1,...,k$, x = raw scores of each preditor, c_{j0} = a constant [9]. If W= within group variance

– covariance matrix, and M = column matrix of means for group j, then the constant $c_{j0} = (1/2)C_j M_j$.

For unequal sample size in each group:

$$C_j = c_{j0} + \sum_{i=1}^{p} c_{ij} x_i + \ln\left[n_j / N \right] \quad (6)$$

where n_j = size in group j, N = total sample size.

4. Result

4.1 logistic regression model

For experiments, the nine characteristics of breast cancer (Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epitheliai Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli and Mitoses) for input values (x_1-x_9) which each characteristic contains number from 1-10 and breast cancer or not breast cancer.

The statistical test in Logistic Regression model:

Step1 : An analysis of the probability of cancer of a breast cancer patients who has medically checked and showed not breast cancer, or who showed identified breast cancer. For the entire 683 breast cancer patients who had a medical check, it was found that 444 breast cancer patients were not detected a cancer, 239 breast cancer patients were detected a cancer. When testing of the coefficient of the two sets of independent variables indicated as tumor cells with abnormal growth and dependent variables indicated as breast cancer using Logit as a Link Function for the data analysis as shown in Table 1.

Table 1. : Test of a coefficient of Tumor Cells with Abnormal Growth of Breast Cancer

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	p-value
Intercept Only	884.350			
Final	102.888	781.462	9	0.000

According to Table 1, it was found that if the model comprised on constant values, the -2 Log Likelihood would equal 884.350, and if the model comprised constant values and independent variables indicated as tumor cells with abnormal growth, -2 Log Likelihood would equal 102.888. That was to say, the model with independent variables indicated as tumor cells with abnormal growth was more suitable than the model with mere constant values. To exemplify, at least one independent variable indicated as a tumor cell with abnormal growth had a correlation with a dependent variable indicated as a breast cancer, or a numbers of events of independent variables indicated as tumor cells with abnormal growth had a correlation with situations of not finding a breast cancer, or of finding a breast cancer. The equation here was Chi-Square = 781.462 with P-value = 0.000 < 0.05 rejected H_0 independent variables indicated as tumor cell with abnormal growth correlated with dependent variables indicated a breast cancer or not breast cancer.

Step2: There was a test of coefficient of Logistic Regression model. The results presented in Table 2.

Table 2: Result of coefficient of Logistic Regression model.

dependent variables	independent variables	B	Wald Test	p-value
Breast cancer	Intercept	10.104	73.958	(0.000)*
	X_1(Clump Thickness)	-0.535	14.192	(0.000)*
	X_2(Uniformity of cell size)	0.006	0.001	0.976
	X_3 (Uniformity of cell shape)	-0.323	1.958	0.162
	X_4 (Marginal Adhesion)	-0.331	7.173	(0.007)*
	X_5 (Single Epithelial cell size)	-0.097	0.381	0.537
	X_6 (Bare nuclei)	-0.383	16.659	(0.000)*
	X_7 (Bland chromatin)	-0.447	6.808	(0.009)*
	X_8 (Normal nucleoli)	-0.213	3.562	0.059
	X_9 (Mitoses)	-0.535	2.646	0.104

*: P-value < 0.05 has a statistical significance at 0.05

According to Table 2, it was found that the coefficient of Logistic Regression Model of tumor cells with abnormal growth had effects on breast cancer at a P-value of < 0.05, which showed Clump Thickness (X_1), Marginal Adhesion (X_4), Bare Nuclei (X_6) and Bland Chromatin (X_7), Therefore, there was an emergence of a model of accumulated probability prediction for classifying groups of patients based on the health conditions of tumor cells with abnormal growth yielding effects on breast cancer. The equation derived was

$$\ln\left(\frac{P_1}{1-P_1}\right) = 10.104 - 0.535X_1 - 0.331X_4 - 0.383X_6 - 0.447X_7 \quad (7)$$

where P_1 is the probability of any breast cancer.

4.2 Discriminant Analysis

The discriminant analysis to study tumor cells with abnormal growth of breast cancer patients with a medical check and found breast cancer or not breast cancer., with the test to determine if each factor helping classify groups of patients differentiates among the groups using Wilks'

Lambda revealed that the diverse factors yielded a statistical significance at a level of 0.05. These factors are 9 tumor cells presented in Table 3.

Table 3 Statistical Values of Average Equality of Each Factor

Effect	Wilks' Lambda	F-test	p-value
X_1(Clump Thickness)	0.489	711.423	(0.000)*
X_2 (Uniformity of cell size)	0.326	1406.123	(0.000)*
X_3 (Uniformity of cell shape)	0.324	1417.644	(0.000)*
X_4(Marginal Adhesion)	0.501	677.878	(0.000)*
X_5 (Single Epithelial cell size)	0.523	622.158	(0.000)*
X_6 (Bare nuclei)	0.323	1426.240	(0.000)*
X_7 (Bland chromatin)	0.425	921.010	(0.000)*
X_8 (Normal nucleoli)	0.484	727.471	(0.000)*
X_9 (Mitoses)	0.821	148.788	(0.000)*

*: P-value < 0.05

From Table 3, after classifying the factors to identify groups, it was found that the factors for group classification were Clump Thickness (X_1), Uniformity of Cell Size (X_2), Uniformity of cell shape (X_3), Marginal Adhesion (X_4), Single Epithelial Cell Size (X_5), Bare Nuclei (X_6), Bland Chromatin(X_7), Normal Nucleoli (X_8), Mitoses (X_9), which created a model for an analysis of a group classification of breast cancer patients found not having a breast cancer, and found having breast cancer using Fihser's Linear discriminant function (FLDF)

$$D=-3.442+0.183X_1+0.126X_2+0.90X_3+0.047X_4+0.058X_5 +0.261X_6+0.110X_7 +0.107X_8+0.006X_9 \quad (8)$$

which Eigenvalue = 5.383,
Wilks'Lamda = 0.157,
Chi-square = 1253.944,
P-value = 0.000 < 0.05

The discriminant analysis of breast cancer patients with a medical check and found breast cancer or not breast cancer., with the test to determine if each factor helping classify groups of patients differentiates among the groups. A model of a group of breast cancer patients with a breast cancer:

$$D_1=23.536+1.63X_1+0.312X_2+0.495X_3+0.339X_4 +0.995X_5+1.457X_6+1.198X_7+0.547X_8 \quad (9)$$

A model of a group of breast cancer patients with a not breast cancer:

$$D_2=3.288+0.742X_1+0.301X_2+0.057X_3+0.106X_4+0.708X_5 +0.188X_6+0.664X_7+0.026X_8 \quad (10)$$

4.3 Results of the Efficiency Comparison

The efficiency of Discriminant analysis and Logistic Regression Analysis considered based on the correct rations of classification was presented in Table 4.

Table 4 The number of predictive variables and percentile classify groups of breast cancer patients in correct groups of classification based on Discriminant analysis and Logistic Regression Analysis employed with 683 samples of breast cancer patients

Table 4 Comparison between the Discriminant analysis and the Multinomial Logistic Regression Analysis

Model	Discriminant	Logistic Regression
Predictive variables	X_1,X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9	X_1,X_4, X_6 ,X_7
classification (%)		
- Breast cancer	224 (92.90)	228 (95.40)
- Not breast cancer	448 (97.80)	434 (97.70)
- Total	672 (96.10)	662 (96.90)

As presented in Table 4, it was found that the discriminant analysis used all variables to predict of having a breast cancer, or not having a breast cancer , which its capability of 96.10 percent correctly predicting in overall, 97.80 percent correctly predicting breast cancer patients having not breast cancer patients, 92.90 percent correctly predicting breast cancer patients. , In contrast, Logistic Regression used some variables to predict of having a breast cancer or not having a breast cancer, which its capability of correctly predicting was higher than the Discriminant Model. The Logistic Regression was capable of 96.90 percent correctly classifying groups in overall, 97.70 percent correctly predicting breast cancer patients having not breast cancer, 95.40 percent correctly predicting breast cancer patients.

5. Conclusions and Discussion

The probability for the classification of breast cancer patients with having not breast cancer or having breast cancer was considered an essential tool to create health conditions of tumor cells with abnormal growth. In the present study, there were 9 types of tumor cells with abnormal growth indicated by 2 methods of classification : Logistic Regression Analysis and Discriminant Analysis. The results of an analysis revealed that the Logistic Regression Model has 96.90% classification higher than Discriminant Model has 96.10% classification. Logistic Regression Model can used predicted variables 4 variables are Clump Thickness (X_1), Marginal Adhesion (X_4), Bare Nuclei (X_6) and Bland Chromatin (X_7). So the model in predicting the probability of collection for the classification of patients breast cancer is

$$\ln\left(\frac{P_i}{1-P_i}\right) = 10.104 - 0.535X_1 - 0.331X_4 - 0.383X_6 - 0.447X_7.$$

The study results reveal that the discriminant analysis can used predicted variables 9 variables for classifying groups of breast cancer and non- breast cancer. So the factor model for the classification of Fisher's linear discriminant functions affecting the breast cancer model. The classifying groups of breast cancer is

$D_1=23.536+1.63X_1+0.312X_2+0.495X_3+0.339X_4+0.995X_5 + 1.457X_6+1.198X_7+0.547X_8$

and the classifying groups of not breast cancer is

$D_2=-3.288+0.742X_1+0.301X_2+0.057X_3+0.106X_4+0.708X_5 +0.188X_6 +0.664X_7 +0.026X_8$

Therefore, the models for statistical classification would be effective in case the data used for an analysis had distributions in accordance with the assumptions of statistical paramatrix of groups classification used for predictions in order to enhance health conditions of the breast cancer patients. These 2 statistical predictions were Logistic Regression Analysis and Discriminant Analysis, which later should be operated with a horizontal paramatrix for health conditions of patients, for example neutral networks, decision trees, and trait recognition approach.

References

[1] Wingo P.A., Tong, T., Bolden, S. (1995). Cancer statistics, CA Cancer J Clin, 46(1), 5-27. doi : 10.3322/canjclin.46.1.5

[2] Agreti, A. (2002). Categorical Data Analysis, (2nd ed.). New York : John Wiley&Sons.

[3] Hosmer, D.W.,&Lemeshow, S.(2000). Application logistic regression.New York : John Wiley & Sons.

[4] Stephenson, B. (2008). Binary response and logistic regression analysis. Retrieved from www.public.iastate.edu /~stat415/stephenson/stat415_chapter3.pdf.

[5] Bandhita P., Noparat T.(2006, June). Ordinal Regression Analysis in factors related to Sensorial Hearing Loss of the Employee in Industrial factory in Lampang Thailand. Mathematic, Statistics and Their Application, Penang.

[6] Klecka, William R. (1980). Discriminant Analysis. Quantitative Applications in the Social Sciences Series, No.19. Thousand Oaks, CA:Sage Publications.

[7] Lachenbruch, P. A. (1975). Discriminant Analysis. NY:Hafner. For detailed notes on computations.

[8] Overall, J.E. and C.J.Klett. (1972). Applied Multivariate Analysis. McGraw-Hill:New York.

[9] Jobson , J.D. (1992). Applied multivariate data analysis, . (2 nd ed.). Springer, New York : Berlin Heidelberg.

[10] Hayatshahi, S.H.S., Abdolmaleki, A., Safarian, S. & Khajeh, K.(2005,October). Non-linear quantitative structure-activity relationship for adenine derivatives as competitive inhibitors of adenosine deaminase. Biochem Biophys Research Comunication., 338, 1137-1142. Retrieved from http://elsevier.com/locate/ybbrc

[11] Swets, J. (1988). Measuring the accuracy of diagnostic systems. Science, 240, 1285-1293. doi : 101126/science.3287615

[12] VECCHIO, T (1966). Predictive value of a single diagnostic test in unselected populations. New England Journal of Medicine, 274, 1171-1173.

Krieng Kitbumrungrat was born in Bangkok, Thailand, in 1964. He received the B.S. degree in Statistics from Nation Institute of Development Administration (NIDA) University. Thailand in 1991. He received Doctor of Philosophy in Mathematics from Mahidol University.Thailand in 2004. His research interests include mathematics and the application of the statistical.