The Use of Logistic Regression Model in Estimating the Probability of Being Affected By Breast Cancer Based On the Levels of Interleukins and Cancer Marker CA15-3

Fadhaa Ali[1], Sara Ayyed Najm[2]

[1,2] Department of Statistics, College of Administration and Economics-University of Baghdad, Baghdad, Iraq

---

Abstract

Breast cancer has got much attention in the recent years as it is a one of the complex diseases that can threaten people lives. It can be determined from the levels of secreted proteins in the blood. In this project, we developed a method of finding a threshold to classify the probability of being affected by it in a population based on the levels of the related proteins in relatively small case-control samples. We applied our method to simulated and real data. The results showed that the method we used was accurate in estimating the probability of being diseased in both simulation and real data. Moreover, we were able to calculate the sensitivity and specificity under the null hypothesis of our research question of being diseased or not.

Keywords: Logistic Regression, K-Means, Classification, Breast Cancer

Corresponding Author:
Fadhaa Ali- Dept. Of Statistics-College of Administration and Economics-
University of Iraq-Baghdad-Iraq
Ali_Fadhaa@coadec.uobaghdad.edu.iq

استخدام نموذج الانحدار اللوجستي لتقدير احتمالية الإصابة بسرطان الثدي بناءً على مستويات الإنترلوكينات وعلامة السرطان CA15-3

فضاء علي [1] **سارة عايد نجم[2]**
[1],[2] كلية الإدارة والاقتصاد / جامعة بغداد

المستخلص :
لقد حظي سرطان الثدي باهتمام كبير في السنوات الأخيرة لأنه أحد الأمراض المعقدة التي يمكن أن تهدد حياة الناس. يمكن تحديده من مستويات البروتينات المفرزة في الدم. في هذا المشروع ، قمنا بتطوير طريقة لإيجاد عتبة لتصنيف احتمالية التأثر بها في مجموعة سكانية بناءً على مستويات البروتينات ذات الصلة في عينات صغيرة نسبيًا من حالات التحكم. طبقنا طريقتنا على بيانات محاكاة وحقيقية. أظهرت النتائج أن الطريقة التي استخدمناها كانت دقيقة في تقدير احتمالية الإصابة بالمرض في كل من بيانات المحاكاة والبيانات الحقيقية. علاوة على ذلك ، تمكنا من حساب الحساسية والخصوصية في ظل فرضية العدم لمسألة بحثنا المتمثلة في الاصابة من عدمها.

الكلمات المفتاحية: الانحدار اللوجستي ، K - متوسطات ، التصنيف

ISSN (1681- 6870)

1. Introduction

Breast cancer is the malignant tumor that results from the unusual growth of abnormal breast cells. It can affect tissues which are responsible for milk production (Ductal and lobular tissues). It became the most common malignancy in women and one of the greatest health issues in the recent years[1]. Many reports show that there are about one million new cases diagnosed worldwide. Such cases are representing 18% of the total number of cancer in women. It has been reported by many healthcare centers that one out of eight women in the USA [2] and one out of 10 women in the UK [3] will develop breast cancer at some point in their lives.

Several studies have been conducted worldwide to look into the causes of this disease. Preventive Services Task Force suggested screening of breast cancer for women age 50 to 74 years old for the past ten years[5]. Switzerland also conducted a program-based national strategy in 2013, recommending screening of breast cancer twice a year for women aged over 50[6,7], as many studies have been concluded that age is a risk factor for entering a population screening program[8-10]. When it comes to considering the women who are aged below 50 in any screening program, it has been found that only about 25% of women identified as infected by breast cancer[11,12].

Many studies started looking into factors underlying breast cancer. Some of which have been concluded that the disease could be in a causal relationship with the levels of interleukins in women. Interleukins are small proteins secreted mainly by CD3+ and CD4+ T lymphocytes that mediate the interactions between cells which is essential for cancer progression. Interleukins are necessary to develop and differentiate different cells (NK, B, and T leukocytes) . Generally, they have been identified as a causal factors in many diseases, including breast cancer, as they have a unique participation in systemic inflammation and immune system modulation [4]. Some researchers have also studies the effect of cancer marker CA15-3 on breast cancer in sample of women taken from Iraqi population[22].

One of the most techniques that have been used for several years is machine learning (ML) forecasting. This procedure is an alternative tool to standard one that can provide accurate prediction of being affected by breast cancer[13]. Accuracy and reliability estimates have been achieved by the use of ML in models related to cancer prognosis and survival [16–18].

One of the machine learning techniques is the logistic regression which can play an important role in classifying the case-control samples in terms of their similarities and differences [14]. An approach has been proposed to classify disease through DNA microarray data. They suggested a penalized LR to reduce the number of genes and select specific variables. The results were accurate enough to conclude that their proposed approach perform better than classical LR[14].

Vector Machine (SVM) and K Nearest Neighbour (KNN) have also been used as a classifier technique of machine learning in which the machine is learned from the past data to predict the category of new input as far as breast cancer is concerned. Their idea was applied to datasets taken from the UCL repository[15]. ML methods is also used in few studies for personalized breast cancer risk prediction or for comparing the predictive accuracy with models commonly used in clinical practice [19]. In this study, we proposed a threshold-based simulation to be used as a cut-off point in detecting individuals with high probability of developing breast cancer.

2. Methodology

2.1 -Fitting the GLM

Let $x_1, x_2, x_3, \dots, x_p$ be explanatory variables and y be a disease status variable taking 1 for individuals with disease and 0 for individuals with no disease. As far as LR model is concerned, we are interested in calculating the probability of being diseased, given the explanatory variables, $p_r(Y = 1|X)$. Then, the logistic regression model can be fitted in the following form[23]:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \qquad (1.1)$$

where $i = 1,2,\dots,n \text{ and } j = 1,2,\dots,p$, β_j is the unknown model coefficients,

and $\pi_i = p_r(Y_i = 1|X_i;\boldsymbol{\beta}) = \frac{e^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}}}{1+e^{\beta_0+\sum_{j=1}^{k}\beta_j x_{ij}}}.$ (1.2)

The model coefficients can be estimated through maximizing the likelihood function as follow:

$$L(\boldsymbol{\beta};\mathbf{X}) = \prod_{i=1}^{n} \pi_i^{y_i}[1-\pi_i]^{1-y_i} \qquad (1.3)$$

$$l(\boldsymbol{\beta}) = ln[L(\boldsymbol{\beta})] = \sum_{i=1}^{n}\{y_i ln[\pi_i] + (1-y_i) ln[1-\pi_i]\} \qquad (1.4)$$

Hence, eq(1.4) is difficult to be solved analytically. Therefore, many computational algorithms are available to estimate the model coefficients ($\widehat{\boldsymbol{\beta}}$). The most popular one is R function (glm).

As we get the estimated coefficients, we shall use their values to estimate $\hat{\pi}_i$ from eq.(1.2). We then cluster individuals into two groups based on their $\hat{\pi}_i$'s . The two groups are defined as follow:

$$S_1 = \{ind_i \ ; \hat{\pi}_i \geq \tau_m\}$$

$$S_0 = \{ind_i \ ; \hat{\pi}_i < \tau_m\},$$

where τ_m is a threshold (cuttoff point) which is specified by using bootstrap sampling. Clearly, it can be said that individuals in group $S_1$is having a higher probability of being affected by the disease, whereas the group S_2 includes the individuals with low probability. We then define the estimated disease status as follow:

ISSN (1681- 6870)

$$\hat{y}_i = \begin{cases} 1 & if\ ind_i \quad belong\ to\ S_1 \\ 0 & if\ ind_i \quad belong\ to\ S_0 \end{cases}$$

What is important to be mentioned that according to this procedure, it would be easy to estimate the disease status of a new individual (ind_{i+1}) based on the value of X_{i+1} and the estimated coefficients ($\boldsymbol{\beta}$).

2.2 Specifying the threshold τ_m by non-parametric bootstrap

We propose the following method to specify the threshold we used as a cut-off point $\boldsymbol{\tau_m}$. The method is based upon using the case- control data under study and it can be summarised in the following steps:

1-We divide each independent variable into two variables based on the corresponding disease status $\{0,1\}$. Therefore, the design matrix becomes $X = (X_0, X_1)$, where

$X_0 = (X_{01}, X_{02}, X_{03}, \dots\dots, X_{0p})$ that is corresponding to $Y = 0$, and

$X_1 = (X_{11}, X_{12}, X_{13}, \dots\dots, X_{1p})$ that is corresponding to $Y = 1$.

2- For $k = 1, 2, 3, \dots\dots, m$, we repeat the below [a- f] steps:

a- Generate bootstrap sample from $X = (X_0, X_1)$ and denote it by $X^* = (X_0{}^*, X_1{}^*)$
b- We then use X^* and $\widehat{\boldsymbol{\beta}}$ to calculate $\pi_i{}^*$ by eq. (1.2) and then generate Y^* from Bernoulli trials Ber($\pi_i{}^*$)
c- Fit the GLM for the new data (Y^*, X^*) to find the estimated $\widehat{\boldsymbol{\beta}}^*$
d- Calculate $\hat{\pi}_i{}^*$ from eq.(1.2) using X^* and $\widehat{\boldsymbol{\beta}}^*$
a- Use K-means algorithms to classify $\hat{\pi}_i{}^*$ into two groups and denote the group with the lower mean by G
b- Let $\tau_k = \inf G$

3- We finally consider $\boldsymbol{\tau_m} = \max\{\tau_k, k = 1,2,3, \dots . m\}$

As we mentioned above, the calculated threshold can be vital in estimating the disease status of a new individual

2.3 Assessing the accuracy of the proposed threshold

To assess the accuracy of estimated disease status, we use the sensitivity and specificity to examine the classifier performance[21]. It can be calculated in the same manner of calculating type I and type II error in measuring the accuracy of a test statistics. In doing so, we set the two hypotheses as below:

$H_0: ind_i$ has infected by disease (belong to case sample)

$H_1: ind_i$ has not infected by disease (belong to control sample).

The confusing matrix can be designed as in Table 1 ,

Table 1: confusing matrix

	Predicted: NO	Predicted: YES
Actual: NO	TN	FP
Actual: YES	FN	TP

Note that.

True Negative(TN): Number of healthy individuals ($y_i = 0$) that have correctly detected as healthy ($\hat{y}_i = 0$)

True Positive (TP): Number of diseased individuals ($y_i = 1$) that have correctly detected as diseased ($\hat{y}_i = 1$)

False Positive (FP): Number of healthy individuals ($y_i = 0$) that have incorrectly detected as diseased ($\hat{y}_i = 1$)

False Negative (FN): Number of diseased individuals ($y_i = 1$) that have incorrectly detected as healthy ($\hat{y}_i = 0$).

After constructing the confusing matrix, the sensitivity and specificity can be calculated as:

Sensitivity= $\frac{TP}{TP+FN}$ and Specificity = $\frac{TN}{TN+FP}$

3. Application to real data

We collect data from Iraqi population which represents case-control sample. The total size is 136 of which 106 represent case sample size (diseased individuals) and 30 represents control sample size. The data is base on six independent variables which can be defined as

X_1: Age, X_2: Interleukin_6(IL6), X_3: Interleukin_8(IL8), X_4: Interleukin_10(IL10),

X_5: Interleukin_18(IL18), X_6: Cancer marker (CA15_3).

The disease status variable is Y and taking 0 if an individual is not infected by a disease of our study and taking 1 if an individual is infected by disease.

ISSN (1681- 6870)

We divide the data into two sample. The first one is training data and consist from 76 cases and 20 controls. The remaining 30 cases and 10 controls are to be considered as a new data to which we apply the proposed method. The initial fit of the logistic regression to training data have been done by the use of R function(glm). The resulted coefficients were

$\hat{\beta}_j$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$
value	−38.116	0.014	0.99	0.0334	1.3	0.000062	0.0108

Afterward, we applying the three steps of our proposed threshold that have been described in section (2). It has been found that the threshold value was ($\tau_m = 0.52$). The confusing matrix has been calculated as below

The sensitivity and specificity calculated

and $Specificity = \frac{TN}{FP+TN} = \frac{20}{0+20} = 1$ $Sensitivity = \frac{TP}{TP+FN} = \frac{76}{76+0} = 1$

Table 2: The calculated confusing matrix of the training data

	Predicted: NO	Predicted: YES	Total
Actual: NO	TN=20	FP=0	20
Actual: YES	FN=0	TP=76	76
Total	20	76	96

It is obvic . In much the same spirit, the healthy individuals have been also identified as healthy. This means that the proposed threshold performed well enough to consider it as a good cutoff in classifying individuals into healthy and unhealthy in terms of their estimated probability of being infected by breast cancer. The same manner applied to the remaining data that was consist of 30 cases and controls. The remaining data can be considered as new entries to detect whether they are affected by the disease or not. Here we use the same parameters that we estimated by the training data and the threshold value we found previously. The confusing matrix of the remaining data can be represented as below

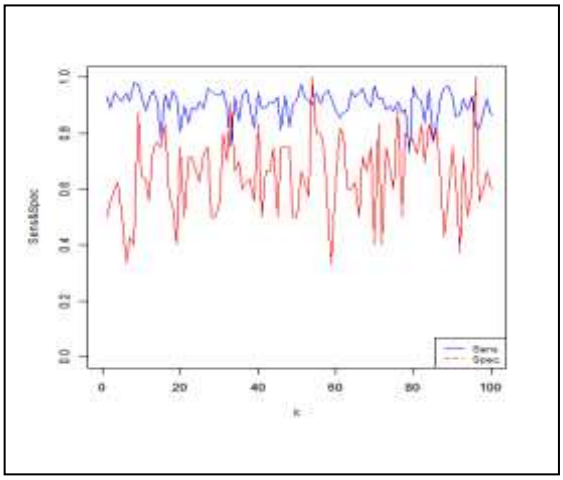
Table 3: The confusing matrix of the remaining data.

	Predicted: NO	Predicted: YES	Total
Actual: NO	TN=10	FP=0	10
Actual: YES	FN=0	TP=30	30
Total	10	30	40

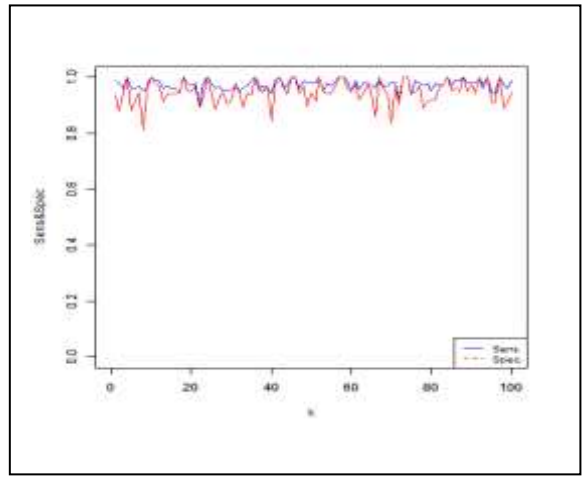
The results above clearly showed that all healthy individuals are detected as healthy as well as the infected ones in both the training and remaining data. This can justify the accuracy of our threshold. In the next section, we carry on a simulation analysis to assess the accuracy in several sample sizes.

4. Simulation Study

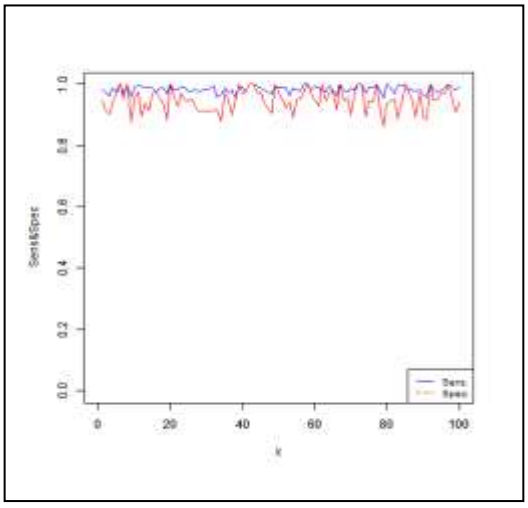
In this section, we conducted a further analysis of the proposed threshold. We first simulated X_{0j} from $N(2,1)$ and X_{1j} from $N(6,1)$. The parameters are chosen to be $\beta_0 = -2, \beta_1 = 0.2, \beta_2 = 0.4, \beta_3 = 0.5, \beta_4 = 0.7, \beta_5 = 0.8, \beta_6 = 0.2$. The disease status was generated by Bernoulli trials with respect to π_i, where π_i is calculated according to eq(1.2). We replicated this simulation k=100 times. At any of which, we calculated the sensitivity and specificity after applying the proposed method to simulated data. The below figures showed the results of sens. and spec.



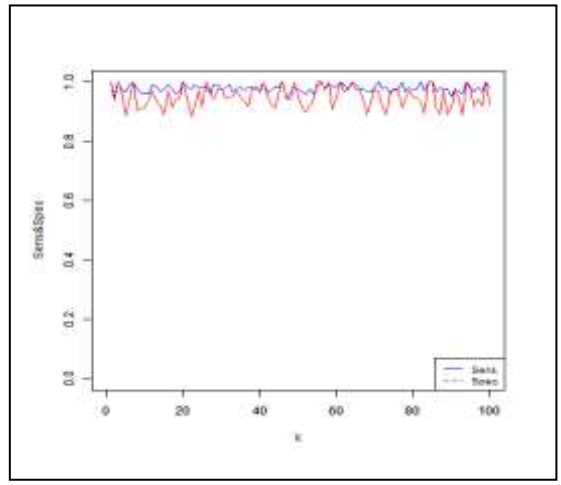
N=90



N=136



N=160



N=200

Figure 1: The calculated sensitivity and sensitivity of different sample sizes

ISSN (1681- 6870)

5. Conclusions

We propose a threshold that showed a good performance of detecting individuals whether they are diseased or not after calculating the probability of being diseased. In this work we ignore the problems that can reduce the accuracy of fitting the generalized linear model as they are not having much influence in performance of our classifier. The proposed method has been applied to real data taken from Iraqi population and representing 106 women who are infected by breast cancer and 30 women who are not having the disease of interest. The aim was to classify the whole sample into two groups by the use of threshold after successfully estimating the probability of having the disease. The results showed that all infected women have correctly been identified as infected by breast cancer. The healthy women have also been identified as healthy women. The accuracy of the proposed classifier has been assessed by sensitivity and specificity after constructing the confusing matrix. The proposed has also been applied to simulated data. The graph of sensitivity and specificity showed the performance of the proposed method in several samples' sizes. As it can be seen from the figures the sensitivity and specificity is close to 1 which means that the proposed threshold is well performed. This method can be programmed in specific medical equipment and used in the laboratory as a test for having breast cancer or not. The calculated probability may be considered as a result of the test and compared to the normal level which we found by the simulation to be less than 0.52. Of course we have not consider all available interleukins in the study as we do not have their data, but it can be considered in a further study by other researchers in future which may lead to much accuracy in estimating the probability

Acknowledgments

We thank the unknown reviewers of this article for their comments to improve the quality of it. We also thank Dr. Mohammed A. Najm at Alfaluja General Hospital for providing us with the data of this study. We thank the Department of Statistics at College of Administration and Economics-University of Baghdad and Iraqi Statistical Association for their efforts in publishing this work.

References

1. Madhavan, M.; Priya, S.; Elizabeth, A.; Iqbal, A.; Vijayalekshmi, N. R. and Prabha, B. (2002). Down regulation of endothelial adhesion molecules in node positive breast cancer: possible failure of host defence mechanism. Patho. Onco. Res. 8 .125-128.
2. Wolff, M. S.; Gwen, W.; Collman, J.; Carl, B. and James, H.(1996). Breast cancer and environmental risk factors: epidemiological and experimental findings. Annu. Rev. Pharmacol.and Toxicol. 36: 573-596.
3. Evans, D. G. R. and Lalloo, F. (2002). Risk assessment and management of high risk familial breast cancer. J. Med. Genet. 39 .865–871.
4. Fasoulakis Z, Kolios G, Papamanolis V, et al. (November 05, 2018) Interleukins Associated with Breast Cancer. Cureus 10(11): e3549. doi:10.7759/cureus.3549
5. Nelson HD, Tyne K, Naik A, Bougatsos C, Chan BK, Humphrey L. Screening for breast cancer: an update for the U.S. Preventive Services Task Force. Ann Intern Med. 2009;151(10):727–37 w237–42.
6. Arie S. Switzerland debates dismantling its breast cancer screening programme. BMJ. 2014;348.

7. Christine Bouchardy PP, Lorez M, Clough-Gorr K, Bordoni A, the NICER Working Group. Trends in Breast Cancer Survival in Switzerland. NICER. Zurich: Schweizer Krebsbulletin(Nr. 4/2011); 2011.
8. Mainiero MB, Moy L, Baron P, Didwania AD, diFlorio RM, Green ED, et al. ACR Appropriateness Criteria((R)) breast cancer screening. J Am Coll Radiol. 2017;14(11s):S383–s90.
9. Qin X, Tangka FK, Guy GP Jr, Howard DH. Mammography rates after the 2009 revision to the United States Preventive Services Task Force breast cancer screening recommendation. Cancer Causes Control. 2017;28(1):41–8.
10. Sardanelli F, Aase HS, Alvarez M, Azavedo E, Baarslag HJ, Balleyguier C, et al. Position paper on screening for breast cancer by the European Society of Breast Imaging (EUSOBI) and 30 national breast radiology bodies from Austria, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Israel, Lithuania, Moldova, The Netherlands, Norway, Poland, Portugal, Romania, Serbia, Slovakia, Spain, Sweden, Switzerland and Turkey. Eur Radiol. 2017;27(7):2737–43.
11. King MC, Levy-Lahad E, Lahad A. Population-based screening for BRCA1 and BRCA2: 2014 Lasker Award. Jama. 2014;312(11):1091–2.
12. Azim HA Jr, Partridge AH. Biology of breast cancer in young women. Breast Cancer Res. 2014;16(4):427.
13. Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. N Engl J Med. 2016;375(13):1216–9.
14. González S., Robles V., Peña J.M., Cubo O. (2009) EDA-Based Logistic Regression Applied to Biomarkers Selection in Breast Cancer. In: Omatu S. et al. (eds) Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living. IWANN 2009. Lecture Notes in Computer Science, vol 5518. Springer, Berlin, Heidelberg.
15. Ch. Shravya, K. Pravalika, Shaik Subhani. 2019. Prediction of Breast Cancer Using Supervised Machine Learning Techniques. JITEE vol(8), Issue (6), pp 1106-1110.
16. Team RC. R: a language and environment for statistical computing. Vienna: Team RC. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2017.
17. Zhang F. Breast cancer risk assessment. 2.0 ed; 2018.
18. Dinov ID. Data science and predictive analytics: biomedical and health applications using R. Cham: Springer; 2018.
19. Heidari M, Khuzani AZ, Hollingsworth AB, Danala G, Mirniaharikandehei S, Qiu Y, et al. Prediction of breast cancer risk using a machine learning approach embedded with a locality preserving projection algorithm. Phys Med Biol. 2018;63(3):035020.
20. J. A. Hartigan and M. A. Wong, 1979. A K-Means Clustering Algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 28, No.1pp. 100-108.
21. Tom Fawcett ,(2005). An introduction to ROC analysis. Pattern Recognition Letters, Volume 27, Issue 8, June 2006, Pages 861-874
22. Amina N. Althwani, Mohammed A. Najm (2011). The importance of CA15-3 in the follow up of Metastatic Invasive Ductal Carcinoma Iraqi Women. Iraqi Journal of Cancer and Medical Genetics, volume 4, Issue 1,Pages 7-10.
23. David W. Hosmer, Stanley Lemeshow. Applied Logistic Regression(2000). Wiley Series in Probability and Statistics.