Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 102 (2016) 106 – 112

Procedia Computer Science

12th International Conference on Application of Fuzzy Systems and Soft Computing, ICAFS 2016, 29-30 August 2016, Vienna, Austria

Development of an intelligent model to estimate the probability of having metabolic syndrome

Nuriye Sancar [a*], Mehtap Tinazli[b]

[a,*]*Department of Mathematics, Near East University, P.O.Box:99138, Nicosia, North Cyprus, Mersin 10 Turkey*
[b]*Faculty of Medicine, Near East University, P.O.Box:99138, Nicosia, North Cyprus, Mersin 10 Turkey*

Abstract

Logistic regression has now become an essential part of medical data analysis that uses a binary-response model. The model is frequently used by epidemiologists as a model for the probability (interpreted as the risk) that an individual will acquire a disease during a specified time period, during which he or she is exposed to a condition (called a risk factor) known to be or suspected of being associated with the disease. The objective is to establish a model using a minimum number of variables, and is also able to identify the relationship between the dependent variable and independent variable. Additionally, the study will determine the risk factors that can lead to the development of metabolic syndrome (MetSyn) and will establish an intelligent and biologically acceptable model for estimating the probability of having the condition, based on the NCEP ATP III criteria. In this study, binary logistic regression analysis has been employed in order to specify the risk factors that affect metabolic syndrome. Metabolic syndrome (MetSyn) is a common metabolic disorder that is increasingly caused by the pervasiveness of obesity in society and diagnosed according to the National Cholesterol Education Program (NCEP) Adult Treatment Panel III (ATP III) Identification[1]. The data has been obtained from the laboratory test results of 321 adult individuals who had consecutively been treated by the Near East University Internal Medicine Department. For this intelligent model, binary logistic regression analysis has been used. The sensitivity, specificity and accuracy rates have been detected as 94.7%, 96.0% and 95.5%, respectively. As a result, homeostatic model assessment (HOMA-IR), uric acid, body mass index (BMI), low-density lipoprotein (LDL) cholesterol, age, smoking, education level (EL) are defined as metabolic syndrome risk factors, the model has been estimated by using those variables in the acquired intelligent model. As a consequence of the research, it has been determined that the key elements that can have an impact are the changeable risk factors, meaning that the illness could be destroyed before it actually occurs, and lifestyle change, that can also prevent the illness.

Peer-review under responsibility of the Organizing Committee of ICAFS 2016

Keywords: Metabolic syndrome; risk factors; logistic regression; odds ratio.

*Corresponding Author:
Nuriye Sancar. E-mail address: nuriye.sancar@neu.edu.tr

Peer-review under responsibility of the Organizing Committee of ICAFS 2016
doi:10.1016/j.procs.2016.09.376

1. Introduction

Regression analysis is a widely used process for obtaining a predictor function for estimating the value of a dependent variable using the independent or predictor variables. In some cases, the outcome (or dependent) variable can be discrete, with two or more possible values. Logistic regression has now become an essential part of medical data analysis that uses a binary-response model. This method of statistical analysis is an extension of multiple regression methods and can be applied where the dependent variable is categorical, meaning that the values of the variable can be assigned to a countable number of categories. In the medical field, a particular outcome could be caused by the presence or absence of a specific disease. In such cases, logistic regression has become an increasingly common method used to estimate the probability that the outcome will occur as a linear function of one or more continuous and/or dichotomous independent variables[2]. The objective is to establish a model using a minimum number of variables, which is acceptable biologically and is also able to identify the relationship between the dependent variable and independent variable. The model is frequently used by epidemiologists as a model for the probability (interpreted as the risk) that an individual will acquire a disease during a specified time period, during which he or she is exposed to a condition (called a risk factor) known to be or suspected of being associated with the disease[3].

Metabolic syndrome (MetSyn) is a common metabolic disorder that is increasingly caused by the pervasiveness of obesity in society[4]. Patients with metabolic syndrome are increasingly likely to develop diabetes mellitus and cardiovascular disease and the risk of mortality from cardiovascular diseases is greater[4]. Metabolic syndrome is diagnosed according to the National Cholesterol Education Program (NCEP) Adult Treatment Panel III (ATP III) Identification[1]. According to the this definition, metabolic syndrome is present if three or more of the following five criteria are met: waist circumference over 102 cm (men) or 88 cm (women), blood pressure over 130/85 mmHg, fasting triglyceride (TG) level over 150 mg/dl, fasting high-density lipoprotein (HDL) cholesterol level less than 40 mg/dl (men) or 50 mg/dl (women) and fasting blood sugar over 100 mg/dl. The aim of this study is to establish an intelligent and biologically acceptable model for estimating the probability of having metabolic syndrome by finding risk factors that can lead to metabolic syndrome based on those criteria. For this intelligent model, binary logistic regression analysis has been applied.

The formula for a logistic regression model[5] is given by

$$\pi(x_i) = P\,(y_i=1/x_i) = \frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_{k-1}x_{k-1}}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_{k-1}x_{k-1}}} \quad (1)$$

where $y_i = \begin{cases} 1, & Having\ MetSyn \\ 0, & Not\ having\ MetSyn \end{cases}$ i=1,2,…,n

e is the base of the natural logarithm, $x_1, x_2, \ldots, x_{k-1}$ are the independent variables (or risk factors), β_0 is the constant term, β_1, β_2, …, β_{k-1} are the coefficients of the independent variables and $P\,(y_i=1/x_i)$ or $\pi(x_i)$ is the probability that the i^{th} individual will have MetSyn.

As is the case with any regression model, valuable information about the relationship between the independent variables to the binary dependent variable can be provided by the regression coefficients β_j in the logistic model by using equation (1). Logistic regression quantifies the relationship between the dichotomous (or binary) dependent variable and the predictors using odds ratios. Since the logit model provides an estimate of the odds ratio, the binary logistic regression is discussed under the logit link function in this study[6]. To describe the concept of odds ratio, odds should be defined as the probability that an event will occur divided by the probability that the event will not happen[3]. An odds ratio is a ratio of two odds and a measure of how much greater (or less) the odds are for subjects possessing the risk factor to experience a particular outcome. Therefore, we can define the odds with the following ratio:

$$\text{odds} = \frac{\pi(x)}{1-\pi(x)} \quad (2)$$

Rearranging by using equation (1), we get

$$\text{odds} = \frac{\pi(x)}{1-\pi(x)} = e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_{k-1}x_{k-1}} \quad (3)$$

$\text{logit}(\pi(x))$ is the natural logarithm of the odds of outcome, so