Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 32 (2007) 987–994
Expert Systems with Applications
www.elsevier.com/locate/eswa

Comparison of logistic regression model and classification tree: An application to postpartum depression data

Handan Ankarali Camdeviren [a], Ayse Canan Yazici [b,*], Zeki Akkus [c], Resul Bugdayci [d], Mehmet Ali Sungur [a]

[a] *Biostatistics Department, Faculty of Medicine, Mersin University, Mersin, Turkey*
[b] *Biostatistics Department, Faculty of Medicine, Baskent University, Baglica Campus, 06530 Ankara, Turkey*
[c] *Biostatistics Department, Faculty of Medicine, Dicle University, Diyarbakir, Turkey*
[d] *Public Health Department, Faculty of Medicine, Mersin University, Mersin, Turkey*

Abstract

In this study, it is aimed that comparing logistic regression model with classification tree method in determining social-demographic risk factors which have effected depression status of 1447 women in separate postpartum periods. In determination of risk factors, data obtained from prevalence study of postpartum depression were used. Cut-off value of postpartum depression scores that calculated was taken as 13. Social and demographic risk factors were brought up by helping of the classification tree and logistic regression model. According to optimal classification tree total of six risk factors were determined, but in logistic regression model 3 of their effect were found significantly. In addition, during the relations among risk factors in tree structure were being evaluated, in logistic regression model corrected main effects belong to risk factors were calculated. In spite of, classification success of maximal tree was found better than both optimal tree and logistic regression model, it is seen that using this tree structure in practice is very difficult. But we say that the logistic regression model and optimal tree had the lower sensitivity, possibly due to the fact that numbers of the individuals in both two groups were not equal and clinical risk factors were not considered in this study. Classification tree method gives more information with detail on diagnosis by evaluating a lot of risk factors together than logistic regression model. But making correct selection through constructed tree structures is very important to increase the success of results and to reach information which can provide appropriate explanations.

Keywords: Classification and regression trees; Logistic regression model; Cross-Validation; Postpartum depression; Diagnostic models

1. Introduction

Classification methods are commonly used in medicine particularly with the purpose of diagnosing (Harper et al., 2003). Usability of these methods increases parallel with developments in statistical packet programs. These methods usually evaluate more than one variable together and are examined in multivariate analyses group. If dependent variable consists of two (binary) or more (multinomial) categories, taking more than one risk factor or predictor variables together into the model with the purpose of estimating the values of dependent variable or correct classifying that will be increased the success in classification. Classification models are being used commonly with this purpose in discriminant analysis, logistic regression analysis, cluster analysis and neural network (Breiman, Friedman, Olshen, & Stone, 1984; Cappelli, Mola, & Siciliano, 1998; Hosmer & Lemeshow, 1989).

Logistic regression and Classification Trees (CT) are the models being used for estimating class membership of categorical dependent variable without getting any assumption

* Corresponding author. Tel.: +90 312 2341010/1637 (O).
E-mail addresses: acyazici@baskent.edu.tr, aysecanan@yahoo.com (A.C. Yazici).

0957-4174/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2006.02.022

on independent variables (Breiman et al., 1984; Buntine, 1992; Cappelli, Mola, & Siciliano, 2002; Hosmer & Lemeshow, 1989; Kerby, 2003; Olaru & Wehenkel, 2003; Siciliano & Mola, 2000; Terin, Schmid, Griffith, D'Agostino, & Sekler, 2003). These methods are very popular in machine learning applications, computer science (data structures), botany (classification), and psychology (decision theory) and are also used as prognostic models in medicine. Nowadays, logistic regression models are used commonly with the purpose of determining risk factors in medical researches and diagnose. In last a few years CTs are attractive because they provide a symbolic representation that lends itself to easy interpretation by humans (Abu-Hanna & de Keizer, 2003; Breiman et al., 1984; Fu, 2004; Kline et al., 2003; Robnik-Sikonja, Cukjati, & Kononenko, 2003).

The aim of this study is to examine logistic regression and CT methods comparatively in term of results obtained. In direction of this purpose, summarized theoretical explanations belong to both two methods were made and results obtained by examining effects of some social-demographic features on postpartum depression with these methods were compared controversially.

2. Material and methods

2.1. Sampling procedure

This cross-sectional study was conducted in 2001, in the province of Mersin in southern Turkey on the coast of the Mediterranean. In this region, there were 58,094 women aged between 15 and 44. A multi-step, stratified (for age groups) cluster sampling method was used. *In the first step*, seven of the 20 primary health centers in Mersin Provincial Center were randomly selected. Single women and pregnant women were excluded. *In the second step*, women were separated into groups according to postpartum periods. As there is no consistently identified grouping method for postpartum periods, the time periods arbitrarily selected were: (i) 0–2 months, (ii) 3–6 months, (iii) 7–12 months, (iv) 13 months and more. *In the third step*, women were selected systematically from each group, depending on weight and age groups.

Estimating PPD prevalence as 15%, a sample size of 1477 would represent a population of 58,094 people with a reliability of 95%. We planned to reach 1550 women for four groups. The 68 women who could not be found at home after two visits and the 35 women who didn't want to participate were excluded, leaving 1447 (93.4%) women (Buğdaycı, Şaşmaz, Tezcan, Kurt, & Öner, 2004; Engindeniz, Küey, & Kültür, 1997).

2.2. Statistical analysis

2.2.1. Classification trees

The CT has a tree structure in which an internal node denotes a variable, the branches of a node denote value (or value ranges) of the corresponding risk factor and a leaf denotes a (dominant) class. The CT construction is achieved by recursively partitioning sets beginning with the whole dataset. Each partitioning of a set is based on a corresponding value partitioning of some risk factors. In each of the recursive iterations, the aim is to find the risk factor, along with its value-partitioning, that can result in subsets which are maximally homogeneous (pure) in their class value. The first node where division starts is called family node, the nodes which continue division are called child node and the nodes where division finishes or homogeneity occurs are called terminal node (Abu-Hanna & de Keizer, 2003; Fu, 2004; Lewis, 2004).

2.2.2. Logistic regression models

In logistic regression models, dependent variable is always in categorical form and has two or more levels. Independent variables may be in numerical or categorical form. The binary multiple logistic regression model is defined as below:

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \ln\left[\frac{P(y=1|x)}{P(y=0|x)}\right] = \beta_0 + \sum_{i=1}^{p}\beta_i x_i$$

The log-likelihood function is used for estimating regression coefficients (β_i) in model. Coefficients are obtained by iterative methods. Exponential value of regression coefficients (e^{β}) gives odds ratio and this value reflects the effect of risk factor in the disease and the interpreted values are odds ratios. The Wald test is used commonly, in hypothesis test of model coefficients. In addition, after model obtained a classification table is obtained as in other classification methods.

In CT and logistic regression model, seven risk factors are used. Total 1447 women were being included into the study were called Learning Sample.

In calculations, EPI-INFO 6.0 (Dean, Dean, Burton, & Dicker, 1990) and Statistica® 6.0 (STATISTICA AFA) statistical packet programs were used.

3. Results

Information about the characteristics and descriptive statistics belong to social-demographic risk factors are being included into the study were given as frequencies, percent, Mean ± SD in Table 1.

3.1. Results of CT

When Table 2 is examined, there are 30 different tree structures for this data set. Complexity of tree structures decreases from Tree 1 to Tree 30. The number of terminal nodes is used as complexity measurements. In selection of optimal tree structure, it is considered that cost-complexity measures are balanced and minimum. In condition that it is balanced, predictive accuracy of tree increases. Through the tree structures given in Table 2, the tree numbered 27 balancing the cost of misclassification (Cross-Validation