Available online at www.sciencedirect.com

SCIENCE DIRECT®

Computational Statistics & Data Analysis 50 (2006) 1905 – 1924

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Using principal components for estimating logistic regression with high-dimensional multicollinear data

Ana M. Aguilera*, Manuel Escabias, Mariano J. Valderrama

Department of Statistics and O.R., University of Granada, Spain

Received 3 December 2004; received in revised form 31 March 2005; accepted 31 March 2005
Available online 25 April 2005

Abstract

The logistic regression model is used to predict a binary response variable in terms of a set of explicative ones. The estimation of the model parameters is not too accurate and their interpretation in terms of odds ratios may be erroneous, when there is multicollinearity (high dependence) among the predictors. Other important problem is the great number of explicative variables usually needed to explain the response. In order to improve the estimation of the logistic model parameters under multicollinearity and to reduce the dimension of the problem with continuous covariates, it is proposed to use as covariates of the logistic model a reduced set of optimum principal components of the original predictors. Finally, the performance of the proposed principal component logistic regression model is analyzed by developing a simulation study where different methods for selecting the optimum principal components are compared.

Keywords: Logistic regression; Multicollinearity; Principal components

1. Introduction

There are many fields of study such as medicine and epidemiology, where it is very important to predict a binary response variable, or equivalently the probability of occurrence of an event (success), in terms of the values of a set of explicative variables related to it. That

* Corresponding author. Tel.: +34 958243270; fax: +34 958243267.
E-mail address: aaguiler@ugr.es (A.M. Aguilera).

0167-9473/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2005.03.011

is the case of predicting, for example, the probability of suffering a heart attack in terms of the levels of a set of risk factors such as cholesterol and blood pressure. The logistic regression model serves admirably this purpose and is the most used for these cases as we can see, for example, in Prentice and Pyke (1979).

As many authors have stated (Hosmer and Lemeshow (1989) and Ryan (1997), among others), the logistic model becomes unstable when there exists strong dependence among predictors so that it seems that no one variable is important when all the others are in the model (multicollinearity). In this case the estimation of the model parameters given by most statistical packages becomes too inaccurate because of the need to invert near-singular and ill-conditioned information matrices. As a consequence, the interpretation of the relationship between the response and each explicative variable in terms of odds ratios may be erroneous. In spite of this the usual goodness-of-fit measures show that in these cases the estimated probabilities of success are good enough. In the general context of generalized linear models, Marx and Smith (1990) and Marx (1992) solve this problem by introducing a class of estimators based on the spectral decomposition of the information matrix defined by a scaling parameter.

As in many other regression methods, in logistic regression it is usual to have a very high number of predictor variables so that a reduction dimension method is needed. Principal component analysis (PCA) is a multivariate technique introduced by Hötelling that explains the variability of a set of variables in terms of a reduced set of uncorrelated linear spans of such variables with maximum variance, known as principal components (pc's). The purpose of this paper is to reduce the dimension of a logistic regression model with continuous covariates and to provide an accurate estimation of the parameters of the model avoiding multicollinearity. In order to solve these problems we propose to use as covariates of the logistic model a reduced number of pc's of the predictor variables.

The paper is divided into four sections. Section 1 is an introduction. Section 2 gives an overview of logistic regression. Section 3 introduces the principal component logistic regression (PCLR) model as an extension of the principal component regression (PCR) model introduced by Massy (1965) in the linear case. It also proposes two different methods to solve the problem of choosing the optimum pc's to be included in the logit model. One is based on including pc's in the natural order given by their explained variances, and in the other pc's are entered in the model by a stepwise method based on conditional likelihood-ratio-tests that take into account their ability to explain the response variable. The optimum number of pc's needed in each method (stopping rule) is also boarded in Section 3 where we propose and discuss several criteria based on minimizing the error with respect to the estimated parameters. Finally, accuracy of estimations provided by the proposed PCLR models and performance of different methods for choosing the optimum models will be tested on a simulation study in Section 4. The results will also be compared with those provided by the partial least-squares logit regression (PLS-LR) algorithm proposed by Bastien et al. (2005) for estimating the logistic regression model.

2. Basic theory on logistic regression

In order to establish the theoretical framework about logistic regression we will begin by formulating the model, estimating its parameters and testing its goodness of fit.