Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Computational Statistics & Data Analysis 49 (2005) 85–97
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

A new algorithm for solving binary discrimination in conditional logistic regression, with two choices of strata

Chong Yau Fu[a,b,*], Jeng-Hsiu Hung[b,c], Shih-Hua Liu[d], Yung-Lin Chien[a,b]

[a] *The Institute of Public Health, National Yang Ming University, Taiwan*
[b] *School of Medicine, National Yang-Ming University, Taiwan*
[c] *Department of Obstetrics and Gynecology, Taipei Veterans General Hospital, Taiwan*
[d] *Department of Humanities & Science, National Yunlin University of Science & Technology, Taiwan*

Received 25 December 2003; received in revised form 27 April 2004; accepted 27 April 2004

Abstract

When conditional logistic regression is based on the exact conditional distribution for inference, the intercept is eliminated. This becomes a problem when the predicted probability is a key issue for binary discrimination. This report details a new algorithm for risk score instead of predicted probability for stratified data in binary discrimination. From the statistical point of view, data partition will reduce the variation of data. Comparing the data-inherent strata and strata generated from the Classification and Regression Tree (CART), the strata generated from CART had greater variation reduction than did the data-inherent strata. Finally, the conditional logistic regression algorithm, used for discrimination when modeling fetal biometric data, resulted in cost savings and computer time savings benefits.

Keywords: Conditional logistic model; Binary discrimination; Strata; Risk score; ROC; CART; Fetal biometric data

1. Introduction

The technique of stratification is commonly used in data analysis. For example, matched design is often used in epidemiology research. The technique of stratification

* Corresponding author. Institute of Public Health, National Yang-Ming University, 155, Sec. 2, Li-Long Street, Shih-Pai, Taipei, 112, Taiwan. Tel.: +886-2-28267054; fax: +886-2-28210514.
E-mail address: chong@ym.edu.tw (Chong Yau Fu).

0167-9473/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2004.04.013

is used to partition data into more homogeneous subgroups before model building. Thus, the strata are formed in the post hoc stage instead of in the design stage, and the strata parameters are treated as nuisance parameters in the model. The number of nuisance parameters depends on the number of strata. The larger the number of strata used in one data set, the finer the strata size and the larger the number of nuisance parameters. Thus, the parameters estimation will yield biased and inconsistent results (Cox and Hinkley, 1974, Chapter 9).

When the outcome variable is binary, logistic regression is used for modeling. For binary outcome data with stratification, Breslow and Day (1980, Chapters 5 and 6) developed the conditional likelihood function to resolve biased and inconsistent estimation. It is based on an exact conditional distribution, and thus, the nuisance parameters (intercepts) are eliminated in the numerator and denominator. When only the estimated coefficients of risk variables are used for interpretation, for example, in epidemiology research, the intercept is not required. However, when the predicted probability is a key issue, the absence of an intercept becomes problematic.

In medical decision making, logistic regression is a commonly used classification technique (Hosmer and Lemeshow, 2000, Chapter 5; Zhang and Singer, 1999, Chapter 3; Asparoukhov and Krzanowski, 2001). The predicted probability of the fitted model is the basis of further classification. Thus, when the conditional logistic model is also used for classification, as in logistic regression, a modified measure, instead of predicted probability, is required. We develop the use of the risk score of fitted models in place of predicted probability for further binary classification.

From the statistical point of view, a stratified data set removes extra variation due to stratified variables. The variation reduction depends on the number of strata and the rules of stratification (Cochran, 1968). In our study of fetal biometric data, there were two types of strata; one was the inherent strata (gestational age), while the other was generated using the classification and regression tree, CART 4.0 (Breiman et al., 1984; Steinberg and Colla, 1995). These two types of strata are compared for reduction in variation.

2. Methods

2.1. Conditional logistic model

Logistic regression is a model dealing with binary outcome variable. Using a sampling design with special considerations, model with a modified intercept can embed with this information (see Scott and Wild, 1986). For stratified data, a logistic model for a specified stratum, k, is written as

$$\pi_k(x) = \frac{\exp(\beta_{0k} + \beta' x)}{1 + \exp(\beta_{0k} + \beta' x)}, \quad \text{where } k = 1, 2, 3, \ldots, K \tag{1}$$

and where $\pi_k(x)$ is the probability of $Y = 1$ ($Y = 0$ or 1), β_{0k} is a nuisance parameter, with constant contribution within the kth stratum, $\beta' = (\beta_1, \beta_2, \ldots, \beta_p)$ are coefficients with respect to covariates, $x = (X_1, X_2, X_3, \ldots, X_p)$.