Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Statistical Planning and Inference 136 (2006) 108–124
journal of statistical planning and inference
www.elsevier.com/locate/jspi

Bias-corrected maximum semiparametric likelihood estimation under logistic regression models based on case–control data

Biao Zhang*

Department of Mathematics, The University of Toledo, Toledo, OH 43606, USA

Received 20 June 2004; accepted 17 June 2004
Available online 14 August 2004

Abstract

We consider the corrective approach (Theoretical Statistics, Chapman & Hall, London, 1974, p. 310) and preventive approach (Biometrica 80 (1993) 27) to bias reduction of maximum likelihood estimators under the logistic regression model based on case–control data. The proposed bias-corrected maximum likelihood estimators are based on the semiparametric profile log likelihood function under a two-sample semiparametric model, which is equivalent to the assumed logistic regression model. We show that the prospective and retrospective analyses on the basis of the corrective approach to bias reduction produce identical bias-corrected maximum likelihood estimators of the odds ratio parameter, but this does not hold when using the preventive approach unless the case and control sample sizes are identical. We present some results on simulation and on the analysis of two real data sets.

MSC: primary 62G05; 62G20

Keywords: Biased sampling problem; Case–control data; Fisher information matrix; Mixture sampling; Profile likelihood; Score derivative matrix; Score function

1. Introduction

Logistic regression models are commonly used for modeling binary data and in the analysis of case–control studies (Breslow and Day, 1980). Let Y be a binary response

* Tel.: +1-419-530-2568; fax: +1-419-530-4720.
E-mail address: bzhang@utnet.toledo.edu (B. Zhang).

0378-3758/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2004.06.026

variable and let X be the associated $p \times 1$ vector of explanatory variables. Then the standard logistic regression model assumes that

$$P(Y = 1|X = x) = \frac{\exp(\alpha^* + \beta^\tau x)}{1 + \exp(\alpha^* + \beta^\tau x)} \equiv \phi(x; \alpha^*, \beta), \tag{1.1}$$

where α^* is a scale parameter and $\beta=(\beta_1, \ldots, \beta_p)^\tau$ is a $p\times 1$ vector of odds ratio parameters. Under case–control sampling as described by Prentice and Pyke (1979), data are collected retrospectively in the sense that the value of X is observed for samples of subjects having $Y = 1$ (cases) and having $Y = 0$ (controls). Specifically, let $X_1, \ldots, X_{n_0}$ be a random sample from $P(x|Y = 0)$ and, independent of X_i, let $Z_1, \ldots, Z_{n_1}$ be a random sample from $P(x|Y = 1)$. Prentice and Pyke (1979) studied maximum likelihood estimation under model (1.1) based on case–control data and showed that the maximum likelihood estimators of the odds ratio parameters and their asymptotic covariance matrices with case–control sampling may be obtained by applying the prospective logistic regression model (1.1) to the case–control study as if the data had been obtained in a prospective study.

Let $g(x) = f(x|Y = 0)$ and $h(x) = f(x|Y = 1)$ be, respectively, the conditional density or frequency functions of X given $Y = 0$ and $Y = 1$. Qin and Zhang (1997) showed that model (1.1) is equivalent to the following two-sample semiparametric model:

$$\begin{aligned} &X_1, \ldots, X_{n_0} \text{ are independent with density } g(x), \\ &Z_1, \ldots, Z_{n_1} \text{ are independent with density } h(x) = \exp(\alpha + \beta^\tau x)g(x), \end{aligned} \tag{1.2}$$

where $\alpha = \alpha^* + \log\{(1 - \pi)/\pi\}$ with $\pi = P(Y = 1) = 1 - P(Y = 0)$. Throughout this paper, let $G(x)$ be the cumulative distribution function corresponding to $g(x)$ and let $(x_1, \ldots, x_{n_0}, z_1, \ldots, z_{n_1})$ be the observed value of $(X_1, \ldots, X_{n_0}, Z_1, \ldots, Z_{n_1})$. Note that (1.2) is a biased sampling model with weight function $\exp(\alpha + \beta^\tau x)$ depending on the unknown parameters α and β. For a complete survey of developments in biased sampling problems, see Vardi (1982, 1985), Gill et al. (1988), Qin (1993), and Gilbert et al. (1999) among others. Qin and Zhang (1997) considered maximum semiparametric likelihood estimation of (α, β) under model (1.2) based on a semiparametric profile likelihood function of (α, β). Their Lemma 1 matches Prentice and Pyke's (1979) results.

In the parametric likelihood setting for a single-sample problem, it is well known that maximum likelihood estimators may be biased when the sample size or the total Fisher information is small. Under a parametric model subject to the appropriate regularity conditions, Cox and Hinkley (1974, p. 310) showed that the asymptotic bias of the maximum likelihood estimator $\hat\theta$ of a q-dimensional vector parameter θ may be written as $E(\hat\theta-\theta)=b(\theta)/n+o(n^{-1})$, where n is the sample size. Substituting $\hat\theta$ for the unknown θ in $b(\theta)/n$, the bias-corrected maximum likelihood estimator of θ is calculated as $\hat\theta_{BC} = \hat\theta - b(\hat\theta)/n$, which removes the first-order term $b(\theta)/n$ from the asymptotic bias of $\hat\theta$. According to Firth (1993), this approach to bias reduction is 'corrective' rather than 'preventive' in the sense that the maximum likelihood estimator $\hat\theta$ is first calculated, then corrected. Because the application of the bias-corrected estimator $\hat\theta_{BC}$ in practice requires the existence of $\hat\theta$ for a given finite sample, Firth (1993) proposed a 'preventive' approach to bias reduction on the basis of a modified score function and showed in regular parametric problems that the $O(n^{-1})$ bias may be removed from the maximum likelihood estimator $\hat\theta$ by introduc-