Computational Statistics and Data Analysis 57 (2013) 309–319

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Testing the fit of the logistic model for matched case-control studies

Li-Ching Chen [a], Jiun-Yi Wang [b,*]

[a] Department of Statistics, Tamkang University, New Taipei City 25137, Taiwan, ROC

[b] Department of Healthcare Administration, Asia University, Taichung 41354, Taiwan, ROC

ARTICLE INFO

Article history:
Received 23 December 2011
Received in revised form 4 June 2012
Accepted 1 July 2012
Available online 5 July 2012

Keywords:
General random effects model
Goodness-of-fit
Matched case-control data
Moment estimation
Logistic model

ABSTRACT

With numerous statistical packages being easily available to conduct the logistic regression analysis, assessment for the goodness-of-fit in the logistic case-control studies becomes more important in practice. While various methods for model checking in conventional case-control studies have been proposed in the literature, methods for checking model adequacy with matched case-control data get relatively less attention. In this study, we propose an omnibus goodness-of-fit test to assess adequacy of the conditional logistic model for matched case-control data. The proposed test can be either constructed based on the discrepancy between two moment estimations or derived to be a score-type test under a general random-effects model. Computation of the proposed test is quite simple in which it does not need to partition the covariate space or to estimate *p*-value of the test via simulations. The asymptotic null distribution and power calculation of the test are derived under a sequence of alternatives. Empirical type I error rates and powers of the test are performed by simulation studies. An example has been used to illustrate the proposed method as well.

1. Introduction

The case-control design has been widely applied in clinical and epidemiological studies to investigate the association between risk factors and a given disease. The retrospective design can be easily implemented and is more economical over prospective studies, yet it suffers from the risk of spurious association due to confounding factors. To adjust effects for confounding factors, methods such as stratification at the design stage and/or multiple regression methods at the analysis stage may be adopted. When some major confounding factors are difficult to be quantified, air pollution or sunlight exposure for example, a matching design provides an opportunity for researchers to control the confounding effects.

Usually, matched case-control data are formed by recruitment of diseased patients from a specific population and matching each case with one or more unrelated control individuals from the source population. It results in highly stratified data, and the stratum-specific effects or matching effects are parameterized by the intercepts of logistic models corresponding to all strata. The conditional approach is then adopted to eliminate the stratum-specific intercepts (Breslow and Day, 1980, Chapter 5).

With numerous statistical packages being easily available to conduct the logistic regression analysis, assessment for the adequacy of the logistic model becomes more important in practice. Several authors have proposed to develop conditional maximum likelihood-based diagnostics (Hosmer and Lemeshow, 1980; Pregibon, 1984; Moolgavkar et al., 1984, 1985) and exact conditional methods (Bedrick and Hill, 1996) for assessing goodness-of-fit of the logistic model for matched case-control studies. These methods are developed by applying diagnostic techniques in regression analysis and are useful in the

* Correspondence to: 500, Lioufeng Rd. Wufeng, Taichung 41354, Taiwan, ROC. Tel.: +886 4 23323456x1861; fax: +886 4 23321206.
E-mail addresses: wangjy.gm@gmail.com, jjwang@asia.edu.tw (J.-Y. Wang).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.07.001

detection of outliers or influential points on individual match sets. More recently, Arbogast and Lin (2004) have developed graphical and numerical methods for assessing the overall model adequacy, as well as the functional form of a covariate or the link function. Although Arbogast and Lin (2004) were the first to propose the omnibus goodness-of-fit test for matched case-control studies, the asymptotic distribution of their test is barely to be derived. Instead, numerically intensive simulations are required to perform the test. The testing procedure needs a great number of calculations based on all possible configurations of individual covariates to obtain the supremum test statistic. The number of calculations is up to the total number of individuals to the power of the number of covariates. Furthermore, Monte-Carlo-based simulations are required to generate the null asymptotic distribution of the supremum test and to estimate its critical value or p-value. This may lead to a remarkable challenge for a moderate to large sample with several covariates included in the model, especially if one would like to estimate empirical powers of the test.

In the present study, we propose to develop an omnibus goodness-of-fit test to assess the adequacy of the logistic model for matched case-control data. Computation of the proposed test is quite simple in which it does not require partitioning of the covariate space or to estimate p-values of the test via simulations. In Section 2, we first introduce the conditional probability density function of covariates given case-control status in each stratum and two moment estimators of the conditional distribution. Subsequently, we construct a moment-type test for assessing the fit of the logistic model and show that the proposed test can be also developed as a score-type test under a general random effects model. We also show the estimated type I error rates and powers of the test via some simulation studies in Section 3 and use an example to illustrate the method in Section 4. Finally, we give some concluding remarks in Section 5 and the proof of theorems in the Appendix.

2. Methods

2.1. Preliminary results

Suppose that in a matched case-control study concerning a confounding variable z, n matched sets are sampled and the ith set consists of 1 case and m_i controls. Denote y as an indicator of disease status, taking values 1 for cases and 0 for controls, and x a $p \times 1$ vector of covariates. Fitting the relationship between the disease status and the covariates in the ith set by the logistic regression model

$$P(y = 1|x, z = i) = \frac{\exp(\alpha_i + x^{\mathrm{T}}\beta)}{1 + \exp(\alpha_i + x^{\mathrm{T}}\beta)}, \quad i = 1, \ldots, n, \tag{1}$$

the parameter β can be estimated by the conditional likelihood method (Breslow and Day, 1980, Chapter 6), where α_i is the intercept or the stratum-specific effect and β is a $p \times 1$ vector of slopes.

Without loss of generality, we let x_{i0} represent the covariate vector for the case and $\{x_{ij}; 1 \le j \le m_i\}$ for the controls in the ith matched set. Suppose that the unordered $m_i + 1$ covariate vectors were observed in the ith set, but we did not know which one of these covariates corresponds to the case. Under the logistic model (1), the conditional likelihood can be written in the form (Breslow and Day, 1980, Chapter 7)

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n} \frac{P(x_{i0}|y = 1, z = i) \cdot \prod_{k=1}^{m_i} P(x_{ik}|y = 0, z = i)}{\sum_{\ell=0}^{m_i} P(x_{i\ell}|y = 1, z = i) \cdot \prod_{k=0, k\neq\ell}^{m_i} P(x_{ik}|y = 0, z = i)} \\ &= \prod_{i=1}^{n} \frac{P(y = 1|z = i, x_{i0}) \cdot \prod_{k=1}^{m_i} P(y = 0|z = i, x_{ik})}{\sum_{\ell=0}^{m_i} P(y = 1|z = i, x_{i\ell}) \cdot \prod_{k=0, k\neq\ell}^{m_i} P(y = 0|z = i, x_{ik})} \\ &= \prod_{i=1}^{n} \frac{\exp\left(x_{i0}^{\mathrm{T}}\beta\right)}{\sum_{\ell=0}^{m_i} \exp\left(x_{i\ell}^{\mathrm{T}}\beta\right)} = \prod_{i=1}^{n}\prod_{j=0}^{m_i} \mu_{ij}^{y_{ij}}, \end{aligned}$$

where

$$\mu_{ij} = \frac{\exp\left(x_{ij}^{\mathrm{T}}\beta\right)}{\sum_{\ell=0}^{m_i} \exp\left(x_{i\ell}^{\mathrm{T}}\beta\right)}$$

is the conditional probability of $y_{ij} = 1$ given that there is only one case among the $m_i + 1$ subjects in the ith set. Thus the estimating equation for β is given by

$$\sum_{i=1}^{n}\sum_{j=0}^{m_i}(y_{ij} - \mu_{ij})x_{ij} = 0. \tag{2}$$