Computational Statistics and Data Analysis 95 (2016) 109–121

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

On quadratic logistic regression models when predictor variables are subject to measurement error

Jakub Stoklosa [a], Yih-Huei Huang [b], Elise Furlan [c], Wen-Han Hwang [d,*]

[a] School of Mathematics and Statistics, and Evolution & Ecology Research Centre, The University of New South Wales, Australia
[b] Department of Mathematics, Tamkang University, Taiwan
[c] Institute for Applied Ecology, University of Canberra, Australia
[d] Institute of Statistics, National Chung Hsing University, Taiwan

ARTICLE INFO

Article history:
Received 20 April 2015
Received in revised form 23 September 2015
Accepted 26 September 2015
Available online 8 October 2015

Keywords:
Functional measurement error
Quadratic logistic regression
Regression calibration
Weighted corrected score

ABSTRACT

Owing to its good properties and a simple model fitting procedure, logistic regression is one of the most commonly used methods applied to data consisting of binary outcomes and one or more predictor variables. However, if the predictor variables are measured with error and the functional relationship between the response and predictor variables is non-linear (e.g., quadratic) then consistent estimation of model parameters is more challenging to develop. To address the effects of measurement error in predictor variables when using quadratic logistic regression models, two novel approaches are developed: (1) an approximated refined regression calibration; and (2) a weighted corrected score method. Both proposed approaches offer several advantages over existing methods in that they are computationally efficient and are straightforward to implement. A simulation study was conducted to evaluate the estimators' finite sample performance. The proposed methods are also applied on real data from a medical study and an ecological application.

1. Introduction

Predictor variables (or covariates) are readily used for logistic regression models in a wide variety of applications, such as: biostatistics, ecological and environmental sciences, engineering, epidemiology, finance, genomics, medical research studies, public health, social sciences, etc. A typical example is seen in medical studies, where the response variable is some binary outcome (e.g., does a patient have diabetes), and the predictor variables are usually recorded characteristics, attributes or measurements taken on patients (e.g., age or recorded body mass index values). Usually, the objective is to understand the nature of the relationship between the response and predictor variables. If the observed predictor variables are measured with error, i.e., they are measured imprecisely, then there may be a loss of statistical power, substantial bias in parameters estimates and loss of features (also referred to as the "triple whammy of measurement errors", Carroll et al., 2006), which will subsequently result in invalid statistical inference in most regression analysis.

Numerous methods have been proposed to account for the measurement error in predictor variables, early work dates back to the 1940 and 1950s—e.g., see Wald (1940) and Berkson (1950). To adjust for the bias inherent in the estimation of the model parameters, several well-known methods have been developed, these include: regression calibration (Carroll et al., 2006, Section 4), refined regression calibration (Wang, 2000), simulation extrapolation (SIMEX, Cook and Stefanski, 1994), conditional score (Stefanski and Carroll, 1987), corrected score (Stefanski, 1989; Nakamura, 1990), Bayesian methods

* Corresponding author.
E-mail address: wenhan@nchu.edu.tw (W.-H. Hwang).

http://dx.doi.org/10.1016/j.csda.2015.09.012
0167-9473/

(Gustafson, 2003), Monte Carlo Expectation Maximization (MCEM, Wei and Tanner, 1990; Stoklosa et al., 2015) and many others, see Carroll et al. (2006). Among these methods, the conditional score approach is shown to be locally efficient (Stefanski and Carroll, 1987) but it is not applicable for regression models with a quadratic term, whereas the refined regression approach is easy to implement and gives satisfactory outcomes in general and practical situations (Liang and Liu, 1991; Wang, 2000).

We are motivated by two data sets consisting of imprecisely measured covariates. The first data set was collected on diabetics in Taiwan in 2005. These survey data were obtained by the National Health Research Institute of Taiwan. Of particular interest is understanding how the incidence of diabetes relates to a patients' characteristics, attributes or some recorded body measurement. Huang et al. (2015) previously analysed these data using body mass index (kg/m^2) and age of patients as two predictor variables. The body mass index data were collected by self-report (where the measurements were collected from a questionnaire) and thus contained some information on the uncertainty associated with the measurements, see Section 5.1 for further details. By fitting several logistic regression measurement error models, Huang et al. (2015) found that age was not significant in their analysis, however a quadratic model (i.e., using both linear and squared body mass index terms) fits the data well—suggesting that some non-linearity between the response and the linear predictor was evident. We will return to this case study in Section 5.1.

In many ecological applications, it is also common to observe both: non-linearity between non-normal responses and the linear predictor (Austin, 2007; Bolker, 2008); and error in variables (Hwang and Huang, 2003; Stoklosa et al., 2011; Xu and Ma, 2014; Stoklosa et al., 2015). Thus, our second example focuses on an ecological study that uses capture–recapture data collected from sub-adult/adult platypus *Ornithorhynchus anatinus* on Kangaroo Island, Australia. These data were previously analysed in Furlan et al. (2012) where the estimation of capture probabilities and abundance of platypus was of interest.

With the exception of a semi-parametric efficient estimator proposed by Tsiatis and Ma (2004), the majority of measurement error literature has been primarily developed for parametric general linear regression models with little attention given to quadratic models; this study focuses on the latter. Importantly, methods such as regression calibration or MCEM can incorporate quadratic structures however the distribution of the predictor variables is usually assumed to be normal, which in practice can be quite restrictive. As an example, in Fig. 1 we present a *qq*-plot and histogram of the self-reported body mass index in the diabetes data described above. The actual distribution of the true covariate is of course never known as it is masked by the measurement error, but it is evident that there is some right-skewness in the observed covariate. If this measurement error were to follow a normal distribution (which to a good approximation usually holds in practice, see Carroll et al., 2007), then it is evident that the true covariate is non-normally distributed. Assuming normality when in fact the true covariate is non-normally distributed will lead to inconsistent estimates of model parameters (Carroll et al., 2006).

To overcome this restriction, Tsiatis and Ma (2004) proposed a locally semi-parametric efficient estimator that provides consistent estimates using a functional measurement error model—i.e., no distributional assumptions are made on the true covariate, and the model remains flexible enough to handle non-linearity. However, their proposed method requires solving an integral equation that is difficult to implement, especially in the multivariate predictor case. More recently, Huang et al. (2015) developed a much simpler approach through an extensively corrected score method that is also able to handle the quadratic logistic model, but this method requires error augmentations in computation and could encounter divergent estimates if the measurement error is severe. Here, we further investigate the method of Huang et al. (2015) and additionally propose two new approaches: (1) an approximated refined regression calibration; and (2) a weighted corrected score method. Both methods are flexible and easier in computation whilst handling quadratic models with only minor bias reported in the parameter estimates.

In Section 2 we give some notation and review several existing methods. We present the proposed approaches in Section 3 and discuss an extension to binomial models in 3.3. We then investigate the finite sample performances for each measurement error method (including the error-free case) using several simulation studies in Section 4, and in Section 5 we fit each method to real data using the two examples described above. We conclude with some further extensions and discussion in Section 6.

2. Notation and a review of existing methods

For $i = 1, \ldots, n$, let Y_i be a random sample of independent binary response variables. Let Z_i be a categorical covariate and X_i be a continuous covariate. In what follows, and for the sake of simplicity, we assume that both Z_i and X_i are univariate (scalar) variables, nevertheless, all the methods presented in this study can be easily extended to the multivariate predictor case. Suppose that the covariates Z_i and X_i are given, then we have

$$P(Y_i = 1 \mid Z_i, X_i) = H(\alpha_1 + \alpha_2 Z_i + \beta_1 X_i + \beta_2 X_i^2), \tag{1}$$

where $H(u) = \{1+\exp(-u)\}^{-1}$ is the logistic function. Suppose that X_i is measured with additive random error and we only have the observed surrogate variable W_i. We assume that $W_i = X_i + \epsilon_i$ for all i, where ϵ_i is the measurement error which is stochastically independent of any other covariates and the response variable. Suppose that ϵ_i is normally distributed with mean 0 and variance σ^2, denoted by $\epsilon_i \sim N(0, \sigma^2)$. We also assume that the measurement error variance σ^2 is known. In practice, the measurement error variance can be estimated from replicate surrogate measurements W_i or validation data