Computational Statistics and Data Analysis 66 (2013) 32–54

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Logistic regression with outcome and covariates missing separately or simultaneously

Shu-Hui Hsieh[a], Chin-Shang Li[b], Shen-Ming Lee[c,*]

[a] Center for Survey Research, Research Center for Humanities and Social Science, Academia Sinica, Taiwan
[b] Department of Public Health Sciences, Division of Biostatistics, University of California, Davis, CA, USA
[c] Department of Statistics, Feng Chia University, Taiwan

ARTICLE INFO

Article history:
Received 29 November 2011
Received in revised form 1 February 2013
Accepted 6 March 2013
Available online 6 April 2013

Keywords:
Outcome missing
Covariate missing
Validation likelihood
Joint conditional likelihood

ABSTRACT

Estimation methods are proposed for fitting logistic regression in which outcome and covariate variables are missing separately or simultaneously. One of the two proposed estimators is an extension of the validation likelihood estimator of Breslow and Cain (1988). The other is a joint conditional likelihood estimator that uses both validation and non-validation data. Large sample properties of the proposed estimators are studied under certain regularity conditions. Simulation results show that the joint conditional likelihood estimator is more efficient than the validation likelihood estimator, weighted estimator, and complete-case estimator. The practical use of the proposed methods is illustrated with data from a cable TV survey study in Taiwan.

1. Introduction

Logistic regression is used to describe the relationship between a dichotomous response variable and a set of covariate or explanatory variables; see, e.g., Cox (1970) and Pregibon (1981). The covariate variables may be continuous or (with dummy variables) discrete. Researchers often use logistic regression to estimate the effects of various covariates on some binary outcome of interest. It is basically assumed that in a logistic regression model the log-odds of the outcome is a linear function of the covariates. That is, the variables (Y, X, Z) are assumed to follow the model

$$P(Y = 1|X, Z) = H(\beta_0 + \beta_1^{\mathrm{T}}X + \beta_2^{\mathrm{T}}Z) = H(\boldsymbol{\beta}^T \mathcal{X}). \tag{1}$$

Here the Y is a binary outcome. (X, Z) is a vector of covariates. $H(u) = [1 + \exp(-u)]^{-1}$. $\boldsymbol{\beta} = (\beta_0, \beta_1^T, \beta_2^T)^T$ is a vector of regression parameters for $\mathcal{X} = (1, X^T, Z^T)^T$. The maximum likelihood method is usually used to estimate the $\boldsymbol{\beta}$.

It is required that data consist of precise measurements for the Y and (X, Z) while the maximum likelihood method is used. However, the data as entered are often not measured perfectly. It has been an active research area in practical problems to study logistic regression with missing covariates. For example, Breslow and Cain (1988) proposed a pseudo conditional likelihood method for a two-stage case-control study in which at the second stage some X's are observed on each stratum classified by (Y, W) for W being a categorical surrogate. When the missingness of X does not depend on both the outcome and missing values, Carroll and Wand (1991) and Pepe and Fleming (1991) proposed semiparametric estimation methods to approximate the likelihood without modeling the distribution of X given (Z, W). Little (1992) reviewed related methods in this field. A mean-score method was proposed by Reilly and Pepe (1995) for discrete covariates when X is missing at

[*] Corresponding author. Tel.: +886 424517250.
E-mail addresses: shhsieh@gate.sinica.edu.tw (S.-H. Hsieh), cssli@ucdavis.edu (C.-S. Li), smlee@fcu.edu.tw (S.-M. Lee).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.03.007

random (MAR) (Rubin, 1976). Robins et al. (1994) proposed an efficient estimation method by computing an optimal score function in semiparametric models. Wang et al. (2002) combined the validation and non-validation data to propose a joint conditional likelihood method. Additionally, Chatterjee and Li (2010) have recently developed three estimators, i.e., mean score, pseudo-likelihood, and semiparametric maximum likelihood, for the regression model under partial questionnaire design and other study settings that can generate nonmonotone missing data in covariates.

Unfortunately, there is another common problem in a logistic regression analysis when outcome data is missing. The topic has been studied by Pepe (1992) and Cheng and Hsueh (1999). They discussed bias correction in the estimation of parameters of a logistic regression model when the binary outcome is subject to missing and misclassification. Cheng and Hsueh (2003) proposed estimation methods for a logistic regression model fitting when the binary outcome and covariate values are both subject to measurement errors. Note that they assumed the validation data set consists of a primary sample plus a smaller validation subsample, which is obtained by a double sampling scheme. Lee et al. (2012) proposed a semiparametric method to estimate the parameters of a logistic regression model when both covariates and outcome data are missing simultaneously. Zhao et al. (2009) extended the semiparametric maximum likelihood method for missing covariate problems to deal with more general cases where covariates and/or responses are missing by design in which they estimated asymptotic variances and confidence intervals using the profile log likelihood and EM algorithms for each case, but there has been no study on fitting a regression model to a data set in which covariates and outcome may be missing separately or simultaneously. Therefore, we are motivated by this to propose two estimation methods to deal with the aforementioned case.

Let Y^0 and W be surrogate variables for Y and X, respectively. Note that the W is available and independent of Y given (X, Z). Moreover, the X, Z, and W are assumed to be categorical. Two semiparametric methods are proposed to estimate the logistic regression parameters $\boldsymbol{\beta}$, where the missing data possibly depends on the observed data. The first method is an extension of the validation likelihood approach of Breslow and Cain (1988). The second one is an extension of the joint conditional likelihood method of Lee et al. (2012) that uses the validation and non-validation data. We do not make any model assumptions for the probability of missingness and specification of the conditional distribution of the missing covariates given the observed covariates for both the methods.

The proposed estimators are described in detail in Section 2. In Section 3, we study the asymptotic properties and relative efficiencies of these estimators. Simulation experiments are conducted to investigate their finite-sample performance in Section 4. In Section 5, we apply the proposed methodology and other existing methodology to the cable TV survey data set from Taiwan. Finally, Section 6 provides some concluding remarks.

2. The proposed estimators

The binary outcome Y and vector of covariates X may be separately or simultaneously missing on a subject, so only one of the $(Y_i^0, Y_i, X_i, Z_i, W_i)$, (Y_i^0, Y_i, Z_i, W_i), (Y_i^0, X_i, Z_i, W_i), and (Y_i^0, Z_i, W_i), for each $i = 1, 2, \ldots, n$, can be observed. Note that the binary surrogate outcome Y_i^0, the surrogate variable W_i, and a covariate vector Z_i are always observed. The term of auxiliary data refers to data that is not in the regression model, but thought to be informative about the Y_i and X_i.

The missingness statuses are defined as follows: $\delta_{i1} = 1$ if both X_i and Y_i are observed; 0 otherwise. $\delta_{i2} = 1$ if X_i is missing and Y_i is observed; 0 otherwise. $\delta_{i3} = 1$ if X_i is observed and Y_i is missing; 0 otherwise. $\delta_{i4} = 1$ if both X_i and Y_i are missing; 0 otherwise. The selection probability of observing Y_i and X_i is then assumed as follows:

$$P(\delta_{ij} = 1 | Y_i, Y_i^0, X_i, Z, W_i) = \pi_j(Y_i^0, Z_i, W_i) = \pi_j(Y_i^0, V_i).$$

Here $V_i = (Z_i^T, W_i^T)$, $j = 1, 2, 3, 4$. $\sum_{j=1}^4 \pi_j(Y_i^0, V_i) = 1$. The missing statuses of Y_i or X_i are then MAR (Rubin, 1976) under these assumptions. It is noted that the $\pi_j(Y_i^0, V_i)$'s may be prespecified at the design stage in some other applications, but they are unknown nuisance parameters.

All covariate variables are categorical under these assumptions, so we let $v_1, v_2, \ldots, v_g$ denote the distinct values of the V_i's. The nonparametric estimator of $\pi_j(Y_i^0, V_i)$ is then given by

$$\widehat{\pi}_j(y^0, v) = \frac{\sum_{i=1}^n \delta_{ij} I(Y_i^0 = y^0, V_i = v)}{\sum_{i=1}^n I(Y_i^0 = y^0, V_i = v)}, \quad j = 1, 2, 3, 4. \tag{2}$$

Here $y^0 = 0, 1$ and $v \in \{v_1, v_2, \ldots, v_g\}$.

2.1. Validation likelihood estimator

When the Y_i's are binary and X_i's are observable, by using the conditional likelihood of Y_i given X_i, Z_i, and $\delta_{i1} = 1$, Breslow and Cain (1988) and Lee et al. (2012) proposed a semiparametric estimator of $\boldsymbol{\beta} = (\beta_0, \beta_1^T, \beta_2^T)^T$. One can then show