Computational Statistics and Data Analysis 53 (2009) 2673–2692

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Semiparametric analysis of randomized response data with missing covariates in logistic regression

S.H. Hsieh [a], S.M. Lee [b,*], P.S. Shen [a]

[a] Department of Statistics, Tunghai University, Taiwan

[b] Graduate Institute of Management of Technology and Department of Statistics, Feng Chia University, Taiwan

ARTICLE INFO

Article history:
Received 26 March 2008
Received in revised form 25 December 2008
Accepted 11 January 2009
Available online 30 January 2009

ABSTRACT

In this article, two semiparametric approaches are developed for analyzing randomized response data with missing covariates in logistic regression model. One of the two proposed estimators is an extension of the validation likelihood estimator of Breslow and Cain [Breslow, N.E., and Cain, K.C. 1988. Logistic regression for two-stage case-control data. Biometrika. 75, 11–20]. The other is a joint conditional likelihood estimator based on both validation and non-validation data sets. We present a large sample theory for the proposed estimators. Simulation results show that the joint conditional likelihood estimator is more efficient than the validation likelihood estimator, weighted estimator, complete-case estimator and partial likelihood estimator. We also illustrate the methods using data from a cable TV study.

1. Introduction

The randomized response techniques (RRT) have been developed (see Warner (1965), Horvitz et al. (1967) and Greenberg et al. (1969)) to obtain more valid estimates when socially sensitive topics are studied. Topics are thought to be socially sensitive when they are threatening to respondents, like for instance questions about illegal or deviant behavior, or questions concerning subjects that are very personal or stressful to respondents. Due to these threats, respondents are less willing to co-operate and when they co-operate, they tend to give more socially desirable answers. These tendencies will unavoidably result in less valid data.

Under RRT, a respondent's privacy can be protected, the tendency to refuse co-operation or to give non-incriminating or socially acceptable answers will decrease and thus the validity of the data will increase. Warner (1965) developed a related-question RRT by introducing two related questions as follows: (A) I am in favor of capital punishment; (B) I am against capital punishment. Further, Horvitz et al. (1967) and Greenberg et al. (1969) developed an unrelated-question RRT model by introducing the following two questions: (A) I am in favor of capital punishment; (C) I was born in January, February or March, where question C is not related to (A). A chance game (for instance with dice, playing cards or coins) now decides which of the two statements is answered with "true" or "untrue". Since such techniques does not reveal to the interviewer the group to which a respondent belongs, this will allow us to get an accurate estimate of the true prevalence in the population of the attitudes towards capital punishment. Several papers provide thorough reviews on RRT (e.g., Kuk (1990), Chaudhuri (2002), Chaudhuri and Mukerjee (1985), Kim and Warde (2004, 2005) Saha (2004), Kim et al. (2006) and Cruyff et al. (2008)). Recently, there have been some researches regarding non-randomized response technique models. For

[*] Corresponding author.
E-mail address: smleestat@yahoo.com.tw (S.M. Lee).

0167-9473/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2009.01.011

example, Yu et al. (2008) proposed two new models for survey sampling with sensitive characteristics and Tian et al. (2007) presented a new survey technique for assessing the association of two binary sensitive variates.

For RRT model with completely observable covariates, Scheers and Dayton (1988) presented a theory for a covariate randomized response model that is an extension of the Warner (1965) procedure and for a covariate extension of the unrelated-question RRT (Greenberg et al., 1969). Corstange (2004) proposed a method to estimate the parameters in a hidden logistic regression. As far as we know, there have been no researches regarding the analysis of data from unrelated-question RRT with missing covariates. Therefore, we consider logistic regression analysis of data from unrelated-question RRT with missing covariates. In Section 2, under the unrelated-question RRT and logistic regression model, two semiparametric estimators are proposed. One of them is an extension of the validation likelihood estimator of Breslow and Cain (1988) and the other is a joint conditional likelihood estimator based on both validation and non-validation data set. In Section 3, we derive the asymptotic properties of the proposed estimators. In Section 4, we review some existing estimates. In Section 5, a simulation study is conducted to investigate the performances of the proposed estimators. In Section 6, the proposed estimators are applied to a cable TV data set. In Section 7, we provide some concluding remarks. Technical details for the asymptotic normal theory are provided in the Appendix.

2. The proposed estimators

Let Y be a binary outcome of a sensitive question, Z be a covariate vector which is always observed, X be a covariate vector that may be missing on some subjects and W be a surrogate variable for X and independent of Y given (X, Z). We consider the following logistic regression model:

$$P(Y = 1|X, Z, W) = H(\boldsymbol{\beta}^{\mathrm{T}}\mathcal{X}), \tag{1}$$

where $H(u) = \{1 + \exp(-u)\}^{-1}$, $\mathcal{X} = (1, X^{\mathrm{T}}, Z^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\beta} = (\beta_0, \beta_1^{\mathrm{T}}, \beta_2^{\mathrm{T}})^{\mathrm{T}}$ is a vector of parameters. The (X, Z, W) is assumed to be discrete. In the presence of missing covariates, many approaches have been developed for the case when Y is observable, such as weighted methods (Zhao and Lipsitz, 1992; Robins et al., 1994; Zhao et al., 1996), imputation methods (Reilly and Pepe, 1995; Paik, 1997) and conditional methods (Breslow and Cain, 1988; Wang et al., 2002).

Under the unrelated-question RRT, Y is not observable and instead, we can only observe a binary variable Y^0, which is the response of a sensitive question or that of an innocuous question (denoted by T). Let S be a binary variable based on a chance game. Hence, we have $Y^0 = Y$ with probability $P(S = 1) = p$ and $Y^0 = T$ with probability $P(S = 0) = 1 - p$. Let $P(T = 1|X, Z, W) = c$. For the unrelated-question design, by model (1), we have

$$\begin{aligned} P(Y^0 = 1|X, Z, W) &= P(Y^0 = 1|S = 1, X, Z, W)P(S = 1) + P(Y^0 = 1|S = 0, X, Z, W)P(S = 0) \\ &= P(Y = 1|X, Z, W) \times p + P(T = 1|X, Z, W) \times (1 - p) \\ &= H(\boldsymbol{\beta}^{\mathrm{T}}\mathcal{X})p + k, \end{aligned} \tag{2}$$

where $k = c(1 - p)$. In the presence of missing covariates, we observe (Y^0, X, Z, W) or (Y^0, Z, W). Let δ indicate whether X is observed $(\delta = 1)$ or not $(\delta = 0)$. We assume that the missing mechanism is missing at random (MAR) (Rubin, 1976), i.e. the probability of X being observed (selection probability) $P(\delta = 1|Y^0, X, Z, W) = \pi(Y^0, Z, W)$, depends on (Y^0, Z, W) but not on X.

Let n be the sample size. For $i = 1, 2, \dots, n$, the validation data set $(\delta_i = 1)$ consists of (Y_i^0, X_i, Z_i, W_i), and the non-validation data set $(\delta_i = 0)$ consists of (Y_i^0, Z_i, W_i). Under MAR, selection probability is $P(\delta_i = 1|Y_i^0, X_i, Z_i, W_i) = \pi(Y_i^0, V_i)$, where $V_i = (Z_i, W_i)$. In our study, $\pi(Y_i^0, V_i)$ is nuisance parameter and unknown, although it may be prespecified at design stage in some other applications. Let $v_1, \dots, v_g$ denote the distinct values of the V_i's. For $v \in (v_1, v_2, \dots, v_g)$ and $y^0 = 0, 1$, we define a non-parametric estimator of $\pi(y^0, v)$ as follows:

$$\widehat{\pi}(y^0, v) = \frac{\sum_{i=1}^{n} \delta_i I(Y_i^0 = y^0, V_i = v)}{\sum_{i=1}^{n} I(Y_i^0 = y^0, V_i = v)}$$

where $I(\cdot)$ is an indicate function.

In the following subsections, based on $\widehat{\pi}(y^0, v)$, we will propose two estimators, namely, validation likelihood estimator and joint conditional likelihood estimator.

2.1. Validation likelihood estimator

Assume that both $P(S_i = 1) = p$ and $P(T_i = 1|X_i, Z_i) = c$ are known. For the unrelated-question design, this is a reasonable assumption since we can obtain the true p and c from the applicable design. Note that when $c = 1$ and there are no missing covariates, model (2) is reduced to the model considered by Corstange (2004).