Statistics and Probability Letters 81 (2011) 947–956

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

An empirical likelihood approach to data analysis under two-stage sampling designs

Ming Zheng, Wen Yu*

Department of Statistics, School of Management, Fudan University, Shanghai 200433, PR China

ARTICLE INFO

Article history:
Received 3 November 2009
Received in revised form 4 October 2010
Accepted 17 January 2011
Available online 28 January 2011

MSC:
primary 62G05
secondary 62G20

Keywords:
Empirical likelihood
Missing data
Over-identified
Two-stage sampling
Validation sample

ABSTRACT

A new empirical likelihood approach is developed to analyze data from two-stage sampling designs, in which a primary sample of rough or proxy measures for the variables of interest and a validation subsample of exact information are available. The validation sample is assumed to be a simple random subsample from the primary one. The proposed empirical likelihood approach is capable of utilizing all the information from both the specific models and the two available samples flexibly. It maintains some nice features of the empirical likelihood method and improves the asymptotic efficiency of the existing inferential procedures. The asymptotic properties are derived for the new approach. Some numerical studies are carried out to assess the finite sample performance.

1. Introduction

When resource constraints prohibit collecting exact measurements on all variables of interest for the participants in a study, a two-stage or double sampling design may be used. A primary sample of cheaper (proxy) measurements are first drawn from the target population. At the second stage, a validation subsample is drawn from the primary sample. For each subject in the validation subsample, the exact information is collected in addition to the rough data. As a result, two sets of data, the proxy data on all subjects, along with the validation data on the subsample, are available for statistical inference.

Some efforts have been made to provide statistical methods for data analysis under two-stage sampling designs. For example, Tenenbein (1970) discussed estimation for binomial data. Breslow and Cain (1988) focused on logistic regression. Carroll and Wand (1991) and Pepe and Fleming (1991) proposed a nonparametric likelihood approach which can be used when the continuous part of the data is of low dimension. Since exact information for the whole sample is not observed completely, this two-stage sampling design can be put into the framework of missing data and some missing data analysis techniques can be applied. When the full parametric models for both primary and validation samples are specified, one can use the EM algorithm to get the MLE (Dempster et al., 1977). However, the validity of the MLE relies heavily on the model assumptions. Robins et al. (1994) provided a general class of estimators for missing data under the missing at random assumption. Their estimators include all possible regular asymptotic linear estimators. The semiparametric efficiency bound is attainable by choosing the optimal estimating function in that class. Nonetheless, the optimal one involves the knowledge about the underlying joint distribution of the two samples, and is therefore difficult to construct. More recently, Chen and Chen (2000) proposed a simple estimation procedure when the validation subsample is randomly chosen from the

* Corresponding author.
E-mail address: wenyu@fudan.edu.cn (W. Yu).

0167-7152/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2011.01.011

primary sample. Their estimator belongs to the general class of Robins et al. (1994). Although it is not optimal, their estimator is not difficult to implement and possesses good efficiency. Chen and Chen's (2000) method is designed for regression problems, so the number of unknown parameters is required to be equal to the number of estimating functions. This restricts its extension to more complicated situations.

In many problems, people may have more estimating functions than unknown parameters and want to make full use of all the information. Qin and Lawless (1994) introduced the empirical likelihood method to utilize all the estimating functions. Empirical likelihood was first introduced by Owen (1988, 1990) for constructing generalized likelihood ratio test statistics and corresponding confidence regions. In Qin and Lawless's (1994) work, they used the method to obtain point estimation and showed that asymptotically, the empirical likelihood function is able to make the optimal linear combination of the original estimation functions automatically. Therefore, empirical likelihood has been widely used in these so-called over-identified problems. Some recent papers extended the empirical likelihood approach to validation sample problems; c.f., Wang and Rao (2002) and Stute et al. (2007), among many others. However, all these papers focus on regression problems and require the response variables to be observable. Moreover, to the best of our knowledge, no existing literature has discussed the over-identified situation under the two-stage sampling scheme.

In this paper, we develop a new empirical likelihood approach to data analysis under the two-stage sampling design when the validation data is a random subsample of the primary data. The new method is not confined to the regression framework and is able to deal with the over-identified situation flexibly. As a semiparametric method, our approach does not require a fully parametric model nor a correct specification of the association between the primary and validation samples. All the knowledge about the underlying distribution is contained in a series of estimating functions. The new empirical likelihood approach can utilize the information from all the estimating functions and the rough data robustly and effectively. Moreover, when making inferences, our approach does not require estimating any variance–covariance matrices, which is an important feature of the empirical likelihood method. Instead, some suitable optimization techniques are necessary to implement the new approach.

The rest of the paper is organized as follows. In Section 2, we describe the new empirical likelihood approach and state its asymptotic properties. Some numerical results are presented in Section 3. Section 4 concludes. Technical details are given in the Appendix.

2. Main results

2.1. Notation and model specification

Let N denote the size of the primary sample. Assume that the rough or proxy data, $\tilde{X}_i,\ i = 1, 2, \ldots, N$, are i.i.d. copies of a d-dimensional random vector $\tilde{X}$. Let V be a random subset of $\{1, 2, \ldots, N\}$ with size of n $(n < N)$. For each subject in V, the exact information $X \in \mathbb{R}^d$ is recorded. Thus, the two available data sets are the primary sample $\{\tilde{X}_i,\ i = 1, 2, \ldots, N\}$ and the validation sample $\{X_i, i \in V\}$.

Let the distribution of X be denoted by F. The parameter of interest, θ, is a p-dimensional parameter associated with F. The information about θ is contained in a set of estimating functions $\mathbf{g}(X, \theta) = (g_1(X, \theta), g_2(X, \theta), \ldots, g_r(X, \theta))^T$, where $r \geqslant p$. Let θ_0 be the true value of θ. The model we specify takes the form of $E[\mathbf{g}(X, \theta_0)] = 0$. Correspondingly, for the proxy data, assume that we have $\mathbf{h}(\tilde{X}, \gamma) = (h_1(\tilde{X}, \gamma), h_2(\tilde{X}, \gamma), \ldots, h_s(\tilde{X}, \gamma))^T$, where γ is a q-dimensional parameter associated with the distribution of $\tilde{X}$ and $\mathbf{h}(\cdot, \cdot)$ is an s-dimensional function of $\tilde{X}$ and γ chosen by the researchers. A key requirement of the choice is that there exists some γ_0 such that $E[\mathbf{h}(\tilde{X}, \gamma_0)] = 0$. Let $\mathbf{X} = (X^T, \tilde{X}^T)^T$, $\boldsymbol{\alpha} = (\theta^T, \gamma^T)^T$ and $\mathbf{m}(\mathbf{X}, \boldsymbol{\alpha}) = (\mathbf{g}(X, \theta)^T, \mathbf{h}(\tilde{X}, \gamma)^T)^T$.

Let $\delta = 1$ if an observation belongs to the validation sample and $\delta = 0$ otherwise. Then the available data can be written as $\{(\tilde{X}_i, \delta_i, \delta_i X_i),\ i = 1, 2, \ldots, N\}$, which can be viewed as N i.i.d. copies of $(\tilde{X}, \delta, \delta X)$. Note that when $\delta_i = 0$, X_i is unobservable, or equivalently, is missing. Thus, this two-stage sampling design can be placed within the framework of missing data. That the validation data is a random subsample means that δ_i is independent of $(\tilde{X}_i, X_i)$, which corresponds to the assumption that the missing data are missing completely at random (MCAR).

Under the MCAR assumption, one may use the validation data to make valid inference about θ_0, but the proxy data is also useful in improving the efficiency because of the association between the primary and validation samples. When $p = r = q = s$, the estimating equations for θ_0 and γ_0 are well defined based on $\mathbf{g}(X, \theta)$ and $\mathbf{h}(\tilde{X}, \gamma)$. Therefore, Chen and Chen's (2000) estimation procedure can be applied here without any difficulty. However, when $r > p$ or $q \neq s$, their method cannot be extended directly. To fully utilize all the available information, a new inferential procedure is needed.

2.2. The empirical likelihood approach

When $r > p$, Qin and Lawless (1994) defined an empirical likelihood function to combine all the estimating functions $g_1(X, \theta), g_2(X, \theta), \ldots, g_r(X, \theta)$. Based on the validation sample only, Qin and Lawless's (1994) empirical likelihood function takes the form of

$$\mathcal{L}_{\mathrm{QL}}(\theta) = \sup\left\{\prod_{i\in V} p_i \mid \sum_{i\in V} p_i \mathbf{g}(X_i, \theta) = 0, \sum_{i\in V} p_i = 1, p_i \geq 0\right\},$$