Computational Statistics and Data Analysis 53 (2008) 546–553

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Confidence intervals for a common mean with missing data with applications in an AIDS study

Hua Liang [a,*], Haiyan Su [a], Guohua Zou [b]

[a] Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, NY 14642, USA

[b] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

ARTICLE INFO

Article history:
Received 11 January 2008
Received in revised form 27 August 2008
Accepted 16 September 2008
Available online 26 September 2008

ABSTRACT

In practical data analysis, nonresponse phenomenon frequently occurs. In this paper, we propose an empirical likelihood based confidence interval for a common mean by combining the imputed data, assuming that data are missing completely at random. Simulation studies show that such confidence intervals perform well, even when the missing proportion is high. Our method is applied to an analysis of a real data set from an AIDS clinic trial study.

1. Introduction

In biomedical and epidemiologic researches, data are often missing because subjects fail to report at clinical centers or refuse to answer some questions, or technicians may lose data. Simply excluding the missing data, known as complete case analysis, may waste useful information, because other observed variables associated with the missing variables are also excluded. More seriously, this simple exclusion may result in an inefficient estimation (see, for example, Liang et al. (2004); Wu (2004)) and may even lead to a false conclusion although the implementation of the complete case method is much simpler and it is a default method in most statistical software. In the literature on missing data, common approaches have been described: maximum likelihood (Ibrahim et al., 1999, 2001), weighting adjustment (c.f., Cochran (1977)), single imputation (c.f., Rao and Sitter (1995)), and multiple imputation (Rubin, 1987; Little and Rubin, 2002). This paper will focus on single imputation.

Single imputation is meant to fill a single value for the missing data. It includes mean imputation, ratio imputation, regression imputation, and hot deck imputation etc. When such an imputation is utilized to construct a confidence interval, a normal or t approximation is usually used. This may not be a very good approximation in practice. In this paper, we will propose an empirical likelihood based confidence interval by combining single imputation approach for missing data. This work was motivated by the analysis of an AIDS clinical trial data set (see Section 5). CD4+ cell count is an important biomarkers in AIDS research (Liang et al., 2003; Wu, 2004), and have commonly been used to investigate the treatment effects, which may help clinicians more deeply understand AIDS pathogenesis and improve therapy. Although antiretroviral therapy for HIV-1 infected patients has been greatly improved in recent years, and administration of drug cocktails consisting of three or more drugs can reduce and maintain the viral load below the detection limit for many patients, it is unlikely that combination therapy can eradicate HIV in infected patients because of the existence of long-lived infected cells and sites within the body where drugs may not be effective. With the success of highly active antiretroviral therapy (HAART) against HIV infection, CD4+ cell counts can come back, and the infection is considered chronic. Clinicians and patients are therefore interested in monitoring the immunologic system (measured by CD4+ cell counts). However, there is a common challenge

* Corresponding author. Tel.: +1 585 241 0704; fax: +1 585 256 2541.
E-mail address: hliang@bst.rochester.edu (H. Liang).

0167-9473/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.09.021

that CD4+ cell count is often missing because CD4+ cell counts and the viral load are measured at different time points. As discussed above, simply excluding the missing data is not wise. For the imputation of the missing CD4+ cell counts, the use of auxiliary information such as treatment time is helpful. We will utilize the methods discussed in this paper to impute the missing CD4+ cell counts and then give the confidence intervals of the mean of CD4+ cell counts (Wu et al., 2006). Careful investigation of this quantity is biologically and clinically important because it is a good biomarker for anti-HIV treatment and may be used to evaluate antiretroviral therapies.

The article is organized as follows. In Section 2, we briefly introduce two existing methods in the literature on the imputation for missing data using auxiliary information. In Section 3, we propose to use an empirical likelihood based confidence interval incorporating the imputed data. We illustrate the methods with intensive simulation experiments in Section 4, and analyze a data set from an AIDS study in Section 5. A discussion is provided in Section 6. The proof of the theoretical result is put in the Appendix.

2. Jackknife-based confidence intervals

Assume that a group of subjects with characteristic values (y, x) are independently observed n times, where y is the variable of interest with the mean θ, and x is an auxiliary variable. Let (y_i, x_i) be available for r_1 subjects, whose set is denoted by A_1, $\bar{y}_1$ and $\bar{x}_1$ be their sample means; only x_i be available for r_2 subjects, whose set is denoted by A_2, and $\bar{x}$ be the sample mean of the auxiliary variable over the sample set $s = A_1 + A_2$.

2.1. RS Method

For the missing data, Rao and Sitter (1995) used a ratio imputation approach to impute their values in the finite population inference: For $i \in A_2$, define $\hat{y}_i = \frac{\bar{y}_1}{\bar{x}_1} x_i$. Correspondingly, an estimator of θ can be given by

$$\hat{\theta}_{\mathrm{RS}} = \frac{\bar{y}_1}{\bar{x}_1}\bar{x}.$$

Under the following ratio model

$$y_i = \beta x_i + \varepsilon_i, \qquad E(\varepsilon_i) = 0, \qquad E(\varepsilon_i\varepsilon_j) = 0, \quad i \neq j, \qquad V(\varepsilon_i) = \sigma^2 x_i,$$

$\hat{\theta}_{\mathrm{RS}}$ is an unbiased estimator of θ regardless of missing mechanism. By using a Jackknife approach, a variance estimator of $\hat{\theta}_{\mathrm{RS}}$ can be obtained as

$$\mathrm{var}(\hat{\theta}_{\mathrm{RS}}) = \left(\frac{\bar{x}}{\bar{x}_1}\right)^2 \cdot \frac{1}{r_1} A + 2\left(\frac{\bar{x}}{\bar{x}_1}\right) \cdot \frac{1}{n} B + \frac{1}{n} C,$$

where

$$A = \frac{1}{r_1 - 1} \sum_{j \in A_1} \left(y_j - \frac{\bar{y}_1}{\bar{x}_1} x_j\right)^2,$$

$$B = \frac{\bar{y}_1}{\bar{x}_1} \cdot \frac{1}{r_1 - 1} \sum_{j \in A_1} \left(y_j - \frac{\bar{y}_1}{\bar{x}_1} x_j\right) x_j,$$

and

$$C = \left(\frac{\bar{y}_1}{\bar{x}_1}\right)^2 \cdot \frac{1}{n - 1} \sum_{j \in s} (x_j - \bar{x})^2.$$

The reader is referred to Haziza and Picard (in press) for the good properties of the Jackknife variance estimators in the presence of imputed data. The confidence interval with the level of $1 - \alpha$ is given by

$$\left(\hat{\theta}_{\mathrm{RS}} - z_{\alpha/2}\sqrt{\mathrm{var}(\hat{\theta}_{\mathrm{RS}})}, \hat{\theta}_{\mathrm{RS}} + z_{\alpha/2}\sqrt{\mathrm{var}(\hat{\theta}_{\mathrm{RS}})}\right), \tag{1}$$

where $z_{\alpha/2}$ is the $(1 - \alpha/2)$-quantile of the standard normal distribution.