Journal of Multivariate Analysis 131 (2014) 240–253

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Second-order asymptotic theory for calibration estimators in sampling and missing-data problems

Zhiqiang Tan

Department of Statistics, Rutgers University, 110 Frelinghuysen Road, Piscataway, NJ 08854, United States

ARTICLE INFO

Article history:
Received 23 April 2012
Available online 24 July 2014

AMS subject classifications:
62D05
62F10
62F12
65C05

Keywords:
Calibration
Control variates
Empirical likelihood
Higher-order theory
Missing data
Nonparametric likelihood
Poisson sampling
Rejective sampling
Regression estimator

ABSTRACT

Consider three different but related problems with auxiliary information: infinite population sampling or Monte Carlo with control variates, missing response with explanatory variables, and Poisson and rejective sampling with auxiliary variables. We demonstrate unified regression and likelihood estimators and study their second-order properties. The likelihood estimators are second-order unbiased but the regression estimators are not. For the missing-data problem and survey sampling, no estimator studied always has the smallest second-order variance even after bias correction. However, the calibrated likelihood estimator and bias-corrected, calibrated regression estimator are second-order more efficient than other bias-corrected estimators if a linear model holds for the conditional expectation of the response or study variable given explanatory or auxiliary variables.

1. Introduction

Consider three different but related problems with auxiliary information. The first problem is infinite population sampling or Monte Carlo with control variates (e.g., [14]). The data consist of independent and identically distributed (IID) copies of (η, ξ). The mean $E(\xi)$ is known, for example, to be 0, and $E(\eta)$ is to be estimated. The second problem is to estimate the mean of a response variable Y subject to missingness with explanatory variables X (e.g., [29]). The full data consist of IID copies of (Y, X), but some copies of Y are missing whereas all copies of X are observed. The third problem is to estimate the finite population mean of a study variable Y in survey sampling with auxiliary variables X (e.g., [7,46]). The values of Y are measured only on a sample, but those of X are obtained on the entire finite population. To highlight connections, we refer to the exercise of using auxiliary information to improve estimation as calibration.

Connections between the three problems have been previously discussed. See, for example, Meng [19] for connections between Monte Carlo computation and survey sampling, and Kang and Schafer [16] and Lumley et al. [17] for connections between missing-data problems (e.g., [29]) and survey calibration (e.g., [35]). However, subtle differences between these problems have been largely ignored. Monte Carlo and missing-data problems typically involve IID data. But survey sampling is unique in that the finite population values are fixed and the sampling indicators are random but can be neither independent nor identically distributed, depending on the sampling design.

Recently, Tan [44] built and exploited formal connections in the order from infinite population sampling to the missing-data problem and then to Poisson and rejective sampling, which are two specific sampling schemes for surveys. In fact, the

E-mail address: ztan@stat.rutgers.edu.

http://dx.doi.org/10.1016/j.jmva.2014.07.003
0047-259X/

assumption of Poisson and rejective sampling is crucial to rigorously relating survey sampling to the setup of IID data, by the independence of sampling indicators for Poisson sampling and by the conditional structure of rejective sampling. This assumption is not restrictive because rejective sampling is as broadly applicable as other sampling schemes. Moreover, fast algorithms have been developed for implementing rejective sampling [2,45].

Various estimators, called regression and likelihood estimators [37,38,41], can be transferred between the three problems. For the missing-data problem and survey sampling, there are two types of such estimators. The non-calibrated estimators are adopted from infinite population sampling, whereas the calibrated estimators are derived as modifications of the non-calibrated ones to achieve calibration on X [38,41]. Calibration on X implies double robustness against misspecification of the propensity score [31] or that of the linear model of Y given X in the missing-data problem, and is a basic requirement for calibration estimation [7] in survey sampling.

The first-order asymptotic properties of regression and likelihood estimators are closely connected from one problem to another. For infinite population sampling and the missing-data problem, the regression and likelihood estimators are known to be first-order efficient among estimators using fixed auxiliary information [37,38,41]. For Poisson and rejective sampling, Tan [44] showed that these estimators are first-order efficient in a similar sense: they are first-order asymptotically as efficient as an optimal regression estimator with fixed auxiliary variables [10,20,26].

As seen from the preceding discussion, for each of the three problems, the regression and likelihood estimators are first-order as efficient as each other. Therefore, it is interesting to compare these estimators in higher-order asymptotic properties. For infinite population sampling, the general results of Newey and Smith [21] can be used to show that the likelihood estimator is second-order unbiased but the regression estimator is not, and the likelihood estimator is second-order more efficient than the bias-corrected, regression estimator. The advantage of the likelihood estimator might be understood by the fact that it is literally a maximum nonparametric likelihood estimator [11,37].

The missing-data problem and survey sampling are more complicated than infinite population sampling. There are two likelihood estimators, calibrated and non-calibrated, based on modifications of the usual, nonparametric likelihood. Neither of them is strictly a maximum nonparametric likelihood estimator. In fact, likelihood inference is associated with fundamental difficulties for missing-data problems [28] and survey sampling [27]. There seems to be no clear intuition about how the likelihood estimators are compared with each other and with the regression estimators in higher-order asymptotic properties.

We study second-order asymptotic theory to address the foregoing questions. For the missing-data problem and Poisson and rejective sampling, a summary of our findings is as follows. The likelihood estimators are second-order unbiased but the regression estimators are not. No estimator studied always has the smallest second-order variance even after bias correction. However, the calibrated likelihood estimator and the bias-corrected, calibrated regression estimator are second-order more efficient than other bias-corrected estimators if the linear model of Y given X holds. In this sense, these two estimators can be said to be locally second-order efficient.

For the rest of the article, Sections 2–4 treat infinite population sampling, missing-data problems, and survey sampling and Section 5 provides concluding remarks. All proofs are collected as Appendix in the Supplementary Material, which can be found online at http://dx.doi.org/10.1016/j.jmva.2014.07.003.

2. Infinite population sampling

2.1. Estimators

Consider infinite population sampling, that is, the setting where independent observations are obtained from a common distribution. Suppose that $(\eta_1, \xi_1), \ldots, (\eta_N, \xi_N)$ are IID copies of (η, ξ), where η is a random variable and ξ is a random vector with $E(\xi) = 0$. The objective is to estimate $\alpha = E(\eta)$, taking advantage of the fact that $E(\xi) = 0$ to achieve variance reduction over the basic estimator $\tilde{E}(\eta) = N^{-1}\sum_{i=1}^{N}\eta_i$. Throughout Sections 2–3, $\tilde{E}(\cdot)$ denotes the sample average.

An important example is known as the method of control variates in Monte Carlo computation (e.g., [14]). Each component of ξ is called a control variate. Such variates are often constructed by exploiting specific features of individual Monte Carlo problems.

For regularity conditions, assume that

$$E(\eta^2) < \infty, \qquad E(\|\xi\|^2) < \infty, \quad V \text{ is nonsingular}, \tag{1}$$

where $\|\xi\|$ denotes the usual norm $(\xi^{\mathrm{T}}\xi)^{1/2}$ and $V = E(\xi\xi^{\mathrm{T}})$.

There are various estimators of α which admit a first-order asymptotic expansion

$$\tilde{E}(\eta) - \beta^{\mathrm{T}}\tilde{E}(\xi) + o_p(N^{-1/2}), \tag{2}$$

where $\beta = V^{-1}E(\xi\eta)$. The first-order variance $N^{-1}\mathrm{var}(\eta - \beta^{\mathrm{T}}\xi)$ is the minimum variance of unbiased estimators $\tilde{E}(\eta) - b^{\mathrm{T}}\tilde{E}(\xi)$ for b a vector of arbitrary constants and in fact the semiparametric variance bound in the model characterized by $E(\xi) = 0$ [1, Section 6.2]. The classical method of estimation is to replace β by a consistent estimator, for example,

$$\hat{\beta} = \tilde{E}^{-1}(\xi\xi^{\mathrm{T}})\tilde{E}(\xi\eta), \qquad \hat{\beta}_c = \tilde{E}^{-1}\left[\xi\left\{\xi - \tilde{E}(\xi)\right\}^{\mathrm{T}}\right]\tilde{E}\left[\xi\left\{\eta - \tilde{E}(\eta)\right\}\right].$$