Computational Statistics and Data Analysis 56 (2012) 4450–4461

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Bootstrap variance estimation with survey data when estimating model parameters

Jean-François Beaumont [a,*], Anne-Sophie Charest [b]

[a] Statistics Canada, Statistical Research and Innovation Division, 100 Tunney's Pasture Driveway, R.H. Coats Building, 16th floor Ottawa, Ontario, Canada, K1A 0T6

[b] Department of Statistics, Carnegie Mellon University, 5000 Forbes Avenue, Baker Hall 132, Pittsburgh, PA 15213, United States

ARTICLE INFO

Article history:
Received 7 December 2009
Received in revised form 9 February 2012
Accepted 12 March 2012
Available online 27 March 2012

Keywords:
Bootstrap weights
Estimating equations
Generalized bootstrap
Heteroscedasticity
Taylor linearization

ABSTRACT

When estimating model parameters from survey data, two sources of variability should normally be taken into account for inference purposes: the model that is assumed to have generated data of the finite population, and the sampling design. If the overall sampling fraction is negligible, the model variability can in principle be ignored and bootstrap techniques that track only the sampling design variability can be used. They are typically implemented by producing design bootstrap weights, often assuming that primary sampling units are selected with replacement. The model variability is often neglected in practice, but this simplification is not always appropriate. Indeed, we provide simulation results for stratified simple random sampling showing that the use of design bootstrap weights may lead to substantial underestimation of the total variance, even when finite population corrections are ignored. We propose a generalized bootstrap method that corrects this deficiency through a simple adjustment of design bootstrap weights that accounts for the model variability. We focus on models in which the observations are assumed to be mutually independent but we do not require the validity of any assumption about their model variance. The improved performance of our proposed generalized bootstrap weights over design bootstrap weights is illustrated by means of a simulation study. Our methodology is also applied to data from the Aboriginal Children Survey conducted by Statistics Canada.

1. Introduction

It has long been the tradition for national statistical agencies to design surveys mostly for the estimation of descriptive parameters of a finite population of interest. In this context, inference is typically made with respect to the sampling design; i.e., the variability of estimators is due to the selection of a sample from the finite population. However, there is now a growing interest from the user community in producing more sophisticated analyses of the collected survey data. Survey analysts are often interested in studying complex relationships between survey variables by postulating models. The parameters to be estimated are no longer descriptive parameters, but model parameters, sometimes called analytical parameters.

If sampling is informative, it is well known that classical model-based procedures that ignore sampling design features may lead to inconsistent estimators of model parameters and invalid inferences. One possible approach to dealing with informative sampling, and by far the most commonly used in practice, is to weight the sample estimating equations by the inverse of selection probabilities, or by some other survey weights that account for the sampling design such as calibration

* Corresponding author.
E-mail addresses: Jean-Francois.Beaumont@statcan.gc.ca (J.-F. Beaumont), acharest@stat.cmu.edu (A.-S. Charest).

0167-9473/$ – see front matter Crown Copyright © 2012 Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.03.011

weights. With this approach, two sources of variability should normally be taken into account for inference purposes: the model that is assumed to have generated the finite population data and the sampling design (e.g. Pfeffermann, 1993; Binder and Roberts, 2003; Rubin-Bleuer and Schiopu-Kratina, 2005; Demnati and Rao, 2010). These authors showed that, if the overall sampling fraction is negligible, the model variability can be ignored, which simplifies variance estimation. However, there are many practical cases in which this simplification is not appropriate (see Graubard and Korn, 2002) and both sources of variability must be taken into account. Note that when we refer to model variability we mean the variability in the observed data which is caused by the underlying statistical model, not a change in the model used to describe the data.

The bootstrap is an attractive variance estimation method for survey analysts, especially when a proper set of bootstrap weights accompanies the survey data file. It is often used at Statistics Canada for social surveys. There exists a number of methods for generating design bootstrap weights that track the sampling design variability (e.g., Rao and Wu, 1988; Rao et al., 1992; Sitter, 1992; or Beaumont and Patak, 2012). If the overall sampling fraction is negligible, so that the model variability can be ignored, the design bootstrap weights can be used to obtain valid variance estimates, confidence intervals or hypotheses tests (Beaumont and Bocci, 2009). If the model variability cannot be ignored, these methods need refinements. Such refinements have not yet appeared in the literature. We have thus developed a generalized bootstrap methodology that consists of adjusting design bootstrap weights so as to account for the model variability.

We give some background on survey sampling in Section 2. In Section 3, we present our bootstrap variance estimation method for survey data. We also briefly discuss the special case of a census. Our proposed method is evaluated in a simulation study for stratified simple random sampling. Results are presented in Section 4. We compare our method to the strategy of generating design bootstrap weights under the assumption that sampling units have been selected with replacement. This ad hoc strategy is often suggested and used in practice in the absence of a suitable alternative. We have also applied our technique to data from the Aboriginal Children Survey conducted by Statistics Canada. Results are presented in Section 5. We briefly conclude in the last section.

2. Background and notation

2.1. Model parameters and finite population parameters

When making inferences from survey data, survey analysts often assume that data of the finite population U of size N are generated according to a model and they are interested in drawing conclusions about a vector $\boldsymbol{\beta}$ of unknown model parameters. A typical model describes the conditional distribution $F(\mathbf{y}_U \mid \mathbf{X}_U; \boldsymbol{\beta}, \boldsymbol{\theta})$, where the N-vector $\mathbf{y}_U$ contains the population values of a dependent variable y, $\mathbf{X}_U$ is an N-row matrix that contains the population values of a vector of independent variables $\mathbf{x}$ and $\boldsymbol{\theta}$ is a potential vector of additional unknown model parameters that are not of interest. We assume that $y_i, i \in U$, are mutually independent conditional on $\mathbf{X}_U$. This set-up covers linear and logistic regression as special cases, two models often used for survey data.

We denote by $\boldsymbol{\beta}_U$ the population parameter that would be used to estimate $\boldsymbol{\beta}$ if a census were conducted and we assume that it is implicitly defined by some estimating equation

$$\mathbf{S}_U(\boldsymbol{\beta}_U) = \sum_{i \in U} \mathbf{S}_i(\boldsymbol{\beta}_U; y_i, \mathbf{x}_i) = \mathbf{0}, \tag{2.1}$$

where the function $\mathbf{S}_i(\boldsymbol{\beta}_U; y_i, \mathbf{x}_i)$ is such that $\mathbf{S}_U(\boldsymbol{\beta}_U) = \mathbf{0}$ is an unbiased estimating equation for $\boldsymbol{\beta}$; that is, $E_m\{\mathbf{S}_i(\boldsymbol{\beta}; y_i, \mathbf{x}_i)\} = \mathbf{0}$ and thus $E_m\{\mathbf{S}_U(\boldsymbol{\beta})\} = \mathbf{0}$. The subscript m indicates that the expectation is evaluated with respect to the model; i.e., with respect to the distribution $F(\mathbf{y}_U \mid \mathbf{X}_U; \boldsymbol{\beta}, \boldsymbol{\theta})$. For instance, if the linear model $E_m(y_i) = \mathbf{x}_i'\boldsymbol{\beta}$ is considered then a possible unbiased estimating function is $\mathbf{S}_i(\boldsymbol{\beta}; y_i, \mathbf{x}_i) = (y_i - \mathbf{x}'_i\boldsymbol{\beta})\mathbf{x}_i$. For simplicity, we will use from now on the notation $\mathbf{S}_i(\cdot)$ instead of $\mathbf{S}_i(\cdot; y_i, \mathbf{x}_i)$.

2.2. Survey-weighted estimators

A sample s of size n is selected from the finite population U according to a probability sampling design $p(s)$; thus, $\boldsymbol{\beta}_U$ cannot be computed directly. An estimator $\boldsymbol{\beta}_s$ of $\boldsymbol{\beta}$ is usually obtained by considering the following weighted estimating equation (e.g. Binder, 1983; Rao et al., 2002; Demnati and Rao, 2010):

$$\mathbf{S}(\boldsymbol{\beta}_s) = \sum_{i \in s} w_i \mathbf{S}_i(\boldsymbol{\beta}_s) = \mathbf{0}, \tag{2.2}$$

where $w_i = 1/\pi_i$ is the survey weight of unit i and π_i is the probability that unit i is selected in the sample. This choice of survey weight ensures that $E_p\left(\sum_{i \in s} w_i \mathbf{S}_i(\tilde{\boldsymbol{\beta}})\right) = \sum_{i \in U} \mathbf{S}_i(\tilde{\boldsymbol{\beta}})$, for any fixed vector $\tilde{\boldsymbol{\beta}}$; the subscript p indicates that the expectation is evaluated with respect to the sampling design. As a result, $E_{mp}\{\mathbf{S}(\boldsymbol{\beta})\} = \mathbf{0}$, so that (2.2) is an unbiased estimating equation for $\boldsymbol{\beta}$. Note that the expectation is now taken with respect to the joint distribution induced by the model and the sampling design.