Computational Statistics and Data Analysis 53 (2009) 2596–2604

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Variance estimation of survey estimates calibrated on estimated control totals—An application to the extended regression estimator and the regression composite estimator

Yves G. Berger[a,*], Juan F. Muñoz[b], Eric Rancourt[c]

[a] University of Southampton, Southampton Statistical Sciences Research Institute, Highfield Campus, Southampton, SO17 1BJ, UK
[b] Department of Quantitative Methods in Economics and Business, University of Granada, Granada, 18071, Spain
[c] Business Survey Methods Division, Statistics Canada, Ottawa (Ontario), Canada, K1A 0T6

ARTICLE INFO

Article history:
Received 22 August 2007
Received in revised form 2 December 2008
Accepted 23 December 2008
Available online 8 January 2009

ABSTRACT

Calibration on control totals is commonly used for survey weighting. It is usually assumed that these totals are values known without sampling errors. However, they can be estimated from other sources. A variance estimator that takes into account the randomness of control totals is derived. Several situations such as calibration on external sources and calibration with sampling on two occasions are investigated. The methodology proposed is general and can be implemented in various situations when control totals are estimated.

1. Introduction

Beyond reducing the variance and improving the statistical qualities of estimators, weighting (or calibration) on control totals is widely used in practice for its highly desirable feature of producing estimates that are consistent with external sources. However, these control totals can themselves be estimated quantities. For example, the Canadian Labour Force Survey uses a form of calibration whereby some of the control totals are estimates of quantities obtained on the previous wave or occasion of the Canadian Labour Force Survey. Similarly, control totals in business surveys may be obtained for partially reported values in a register. In the Canadian system, some business surveys are calibrated on Gross Business Income, which is often present in Statistics Canada's Business Register, but otherwise estimated.

The estimation of control totals can be ignored at the variance estimation stage by assuming that their impact on the variance of a characteristic of interest is likely to be small. However, this impact might not be negligible. In this paper, we derive a variance estimator that takes the estimation of control totals into account.

In Section 2, we consider the widely used regression estimator which is an estimator calibrated on a (fixed) control total known without error. For simplicity, we assume that we have one control total. In Section 9, the method in this paper is generalised to the case of more than one control total. In Section 3, we introduce the extended regression estimator which is a regression estimator with an estimated control total. In Section 4, we propose an estimator for the variance of the extended regression estimator. In Section 5, we show how this estimator can be used when the control total is estimated from an independent source. In Section 6, we show how the extended regression estimator can be used for sampling on two occasions. In Section 7, we show how to estimate the variance when the control total is estimated from an earlier wave of a repeated survey. In Section 8, we define the composite estimator which is a particular case of the extended regression estimator, and we show how to estimate its variance. Sen (1973) proposed an approximation of the variance of the composite

* Corresponding author.
E-mail addresses: y.g.berger@soton.ac.uk (Y.G. Berger), jfmunoz@ugr.es (J.F. Muñoz), eric.rancourt@statcan.gc.ca (E. Rancourt).

0167-9473/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.12.011

estimator. However, this estimator is not designed to handle unequal probability sampling. The proposed method is more general, as it can be used with unequal probability sampling designs. We also propose an adjustment to take into account the imputation component of the composite estimator. Finally, in Section 9, we support our result with a simulation study based upon two waves (April and May 2003) of the Canadian Labour Force Survey.

2. The customary regression estimator

In this section, we introduce some notation used in this paper. We also define the customary regression estimator and the usual estimator for its variance.

Consider a finite population denoted by $U = \{1, \ldots, i, \ldots, N\}$ where N is the number of units in the population. Suppose we wish to estimate the population total $Y = \sum_{i\in U} y_i$, where y_i is the value of a variable of interest y, for a unit labelled i.

A sample s is a sub-set of n distinct units from U. We assume the sample is selected according to a probability sampling design $p(s)$. For simplicity, we assume complete response. A design unbiased estimator of Y is given by the π-estimator (Narain, 1951; Horvitz and Thompson, 1952),

$$\hat{Y} = \sum_{i\in s} d_i y_i, \tag{1}$$

where $d_i = 1/\pi_i$ with π_i denoting the first-order inclusion probability of unit i.

Consider the *regression estimator* (Cassel et al., 1976, 1977) which is a particular form of the calibration estimator (Huang and Fuller, 1978; Deville and Särndal, 1992). The regression estimator (e.g. Särndal et al. (1992, page 225)) is defined by

$$\hat{Y}_g = \hat{Y} + \sum_{j=1}^{J} (X_j - \hat{X}_j)\hat{B}_j,$$

with $\hat{B}_1, \ldots, \hat{B}_2$ are the components of the J vector

$$\hat{\mathbf{B}} = (\hat{B}_1, \ldots, \hat{B}_J)^\tau = \left(\sum_{j\in s} d_j \mathbf{x}_j \mathbf{x}_j^\tau\right)^{-1} \sum_{k\in s} d_k \mathbf{x}_k y_k, \tag{2}$$

where $\mathbf{x}_i = (x_{ij}, \ldots, x_{iJ})^\tau$ is column vector of auxiliary variables, $\mathbf{x}_i^\tau$ denotes the transpose of $\mathbf{x}_i$, $X_j = \sum_{i\in U} x_{ij}$ and $\hat{X}_j = \sum_{i\in s} d_i x_{ij}$.

For simplicity, we consider that the model that motivates the regression estimator has one quantitative auxiliary variable x and an intercept; that is, $\mathbf{x}_i = (1, x_i)^\tau$ where x_i denotes the values of the auxiliary variable for a unit labelled i. This implies (Särndal et al., 1992, page 229)

$$\hat{Y}_g = \tilde{Y} + (X - \tilde{X})\hat{\beta}, \tag{3}$$

with $X = \sum_{i\in U} x_i$, $\tilde{Y} = (N/\hat{N})\hat{Y}$, $\tilde{X} = (N/\hat{N}) \sum_{i\in s} d_i x_i$, $\hat{N} = \sum_{i\in s} d_i$ and

$$\hat{\beta} = \frac{\sum_{i\in s} d_i (y_i - \hat{Y}/N)(x_i - \hat{X}/N)}{\sum_{j\in s} d_j (x_j - \hat{X}/N)^2}.$$

Linearization can be used to derive an estimator for the variance of the regression estimator. This estimator for the variance is given by (Särndal et al., 1992, page 237)

$$\hat{V}(\hat{Y}_g) = \sum_{i\in s}\sum_{j\in s} \breve{\Delta}_{ij}\, d_i \hat{e}_i\, d_j \hat{e}_j, \tag{4}$$

where $\hat{e}_i = y_i - \hat{Y}/N - \hat{\beta}(x_i - \hat{X}/N)$ and $\breve{\Delta}_{ij} = (\pi_{ij} - \pi_i\pi_j)\pi_{ij}^{-1}$, with π_{ij} denoting the joint inclusion probability of units i and j. An alternative estimator of the variance uses the regression weights instead of the sampling weights d_i. This alternative estimator can also be used instead of (4).

Huang and Fuller (1978) introduced a more elaborate version of (3) (see also Deville and Särndal (1992)) often called calibrated estimator. For simplicity, we consider the simplified version (3) which is asymptotically equivalent to the calibration estimator, as far as the variance is concerned (Deville and Särndal, 1992).

3. The extended regression estimator

In this section, we define the extended regression estimator (Rancourt, 2001). It is often assumed that the control total X is known. However, in practice X might be unknown and estimated from other sources. Let $\hat{X}_0$ be an estimator of X based on another sample s_0. The estimator $\hat{X}_0$ can be either a π-estimator or a regression estimator. For simplicity, we assume that