Computational Statistics and Data Analysis 53 (2008) 394–404

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Minimum MSE regression estimator with estimated population quantities of auxiliary variables☆

Mingue Park*, HyungJun Cho

Korea University, South Korea

ARTICLE INFO

Article history:
Received 7 June 2007
Received in revised form 26 July 2008
Accepted 1 August 2008
Available online 8 August 2008

ABSTRACT

Construction of a regression estimator in which the population means of auxiliary variables are estimated with a larger sample is considered. Using the variances of the estimated population means, and the correlation between auxiliary variables and the variable of interest, a design consistent regression estimator that has minimum model mean squared error under a working model is derived. A limited simulation study shows that the minimum model mean squared error regression estimator performs well compared to the generalized least squares regression estimator, even when the working model is inappropriate.

1. Introduction

Use of information about the population in constructing design and estimation procedures is common in survey sampling. The information, sometimes called, *auxiliary information*, often comes from official sources such as a national census. In a survey of land use, the total surface area and the area in permanent water bodies may be available from national data sources. Based on the type of auxiliary information, we can use the information in designing a survey, in constructing an estimator or in both of these phases. If we have detailed information, for example $\mathbf{x}_i$ for every element in the population, we can use the information to select a sample through a stratification or to order the population for the selection of a systematic sample. Regression estimation is one of the important procedures that use auxiliary information in the estimation stage to construct efficient estimators. The information that is necessary to define a regression estimator is only the population means of auxiliary variables. For discussion of the efficiency of a regression estimator, see Cochran (1977) and Särndal et al. (1992). A review of regression estimation for sample surveys is given by Fuller (2002).

One regression estimator of the population mean is a linear estimator

$$\bar{y}_{\text{reg}} = \sum_{i \in A} w_i y_i \tag{1}$$

where the w_i minimize

$$\sum_{i \in A} (w_i - \alpha_i)^2 \phi_{ii} \,, \tag{2}$$

subject to

$$\sum_{i \in A} w_i (1, \mathbf{x}_i) = (1, \boldsymbol{\mu}_x) \,, \tag{3}$$

☆ This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD, Basic Research Promotion Fund) (KRF-2007-331-C00063).

* Corresponding address: Department of Statistics, Korea University, Anam-Dong, Seongbuk-Gu, Seoul, 136-701, South Korea. Tel.: +82 2 3290 2243; fax: +82 2 924 9895.
E-mail address: mpark2@korea.ac.kr (M. Park).

0167-9473/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.08.003

ϕ_{ii} is the weight used to define a regression coefficient estimator, α_i is the sampling weight, A is a set of indices selected in the sample and $\boldsymbol{\mu}_x$ is the population mean vector of auxiliary variables $\mathbf{x} = (x_1, \ldots, x_p)$. One possible weight α_i is

$$\alpha_i = \left(\sum_{j\in A} \pi_j^{-1}\right)^{-1} \pi_i^{-1}, \tag{4}$$

where the π_i are the selection probabilities. The choice of weight ϕ_{ii} used to define the distance measure between α_i and the final weight w_i is arbitrary. Usually ϕ_{ii} is assumed to be known up to a constant. The commonly used ϕ_{ii} is the function of the model variance of the error in the linear regression superpopulation model. The choice of ϕ_{ii} is discussed well in Wright (1983). The linear restrictions (3), called the *calibration equation*, improve the efficiency of the regression estimator when the study variable and auxiliary variables are linearly related and $\boldsymbol{\mu}_x$ is known.

It is often the case that the population means of auxiliary variables that are necessary to define a regression estimator are not available, or are not the fixed true values. Auxiliary information from a large scale survey may have a significant uncertainty. For example, the US census values, counted every ten years, are known to have an overall undercount on the order of about 1.2%. Using regression estimation, coverage studies have been conducted to estimate the undercount as a part of the US census. See Isaki et al. (2000). When the population means of auxiliary variables are not available, two phase sampling can be used to collect the necessary information by selecting a relatively larger sample in the first phase. The estimation for two phase samples is discussed in Särndal and Swensson (1987) and Hidiroglou and Särndal (1998). Recently, a chain regression type estimator for the population mean and total with two phase sampling designs has been introduced by Singh et al. (2006). For the variance estimation for two phase samples, see Sitter (1997), Fuller (1998) and Kim and Sitter (2003).

In our study, we consider the regression estimation of the population mean of a study variable when the population means of auxiliary variables were estimated using another large sample. We consider two possible cases depending on the structure of the two samples. In one of the cases, the estimator of the population mean vector of auxiliary variables from a larger sample is uncorrelated with the one from a smaller sample. In the other case, we consider a two phase sampling design where the estimators from the first phase sample and second phase sample are correlated. Hidiroglou (2001) and Judkins and Hidiroglou (2004) discussed the regression estimation with these two cases. With an estimator of the population mean vector of auxiliary variables, $\hat{\boldsymbol{\mu}}_x$, the restriction $\sum_{i\in A} w_i \mathbf{x}_i = \hat{\boldsymbol{\mu}}_x$ may result in a loss of efficiency and a bias of the regression estimator when $\hat{\boldsymbol{\mu}}_x \neq \boldsymbol{\mu}_x$.

For the appropriate use of incomplete information on the population means of auxiliary variables, we consider a procedure that replaces the linear constraints $\sum_{i\in A} w_i \mathbf{x}_i = \hat{\boldsymbol{\mu}}_x$ with a component in the objective function (2), and compare the procedure to the generalized least square predictor. The estimator obtained by relaxing the linear constraints has the form of a ridge regression estimator. To provide an optimal property for the estimator, we derive the optimal ridge coefficient matrix under a working linear model and use the ridge coefficient matrix to define a ridge type regression estimator. We also show the defined optimal estimator is design consistent, thus it is robust to model failure in a large sample framework. Note that the property of design consistency does not depend on the working model. A detailed outline of the paper is as below.

In Section 2, we define a generalized least squares regression estimator that is appropriate under a general linear model for the estimators of the means of auxiliary variables and a study variable. The result of Hidiroglou (2001) is reconfigured in the generalized least squares prediction framework in this section. In Section 3, we construct a new regression estimator that has approximately a minimum model mean squared error (MSE) under the working model, and is design consistent by using the idea that is also useful to derive a nonnegative or a non-extreme regression weight. In Section 4, through a simulation study, we demonstrate that the minimum MSE regression estimator performs well, even when the working model fails to explain the relationship between study variables and auxiliary variables. In Section 5, we describe our application of the minimum MSE regression estimator to a soil carbon study for estimating the mean soil carbon near Mead, Nebraska.

2. Generalized least squares prediction

To define a generalized least squares regression estimator when the population means of auxiliary variables are not available, consider the linear model for $\bar{\mathbf{z}} = (\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \bar{y}_2)'$

$$\bar{\mathbf{z}} = \mathbf{D}_z \boldsymbol{\mu}_z + \mathbf{e}_z, \tag{5}$$

where $\mathbf{D}_z = \text{Blockdiag}[(\mathbf{I}, \mathbf{I})', 1]$, $\boldsymbol{\mu}_z = (\boldsymbol{\mu}_x, \mu_y)'$, $\mathbf{e}_z = (\bar{\mathbf{u}}_{x_1}, \bar{\mathbf{u}}_{x_2}, \bar{u}_{y_2})'$, $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ are the estimators of the population mean of $\mathbf{x}$ based on samples A_1 and A_2, respectively, $\bar{y}_2$ is an estimator of the population mean of y based on sample A_2, $\boldsymbol{\mu}_x$ and μ_y are the population means of $\mathbf{x}$ and y and $\bar{\mathbf{e}}_z$ has a mean vector of zeros and a variance-covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{x_1x_1} & \boldsymbol{\Sigma}_{x_1x_2} & \boldsymbol{\Sigma}_{x_1y_2} \\ \boldsymbol{\Sigma}_{x_2x_1} & \boldsymbol{\Sigma}_{x_2x_2} & \boldsymbol{\Sigma}_{x_2y_2} \\ \boldsymbol{\Sigma}_{x_2y_1} & \boldsymbol{\Sigma}_{x_2y_2} & \sigma_{y_2y_2} \end{pmatrix}.$$

Estimators of the $\boldsymbol{\mu}_x$ and μ_y, $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$ and $\bar{y}_2$, could be weighted means of sampled measurements. Let n_1 and n_2 denote the size of two samples A_1 and A_2, respectively and assume $n_1 > n_2$. The values of $\mathbf{x}_i \in A_1$ may not be available. For example, if the estimator $\bar{\mathbf{x}}_1$ is obtained from a large national survey or from a census then only $\bar{\mathbf{x}}_1$ and its variance estimator are usually available. However, if $\bar{\mathbf{x}}_1$ is obtained from the first phase sample in a two phase sampling design, we can observe the value of $\mathbf{x}_i$ for all units in the first sample.