FINDING OPTIMAL ESTIMATORS IN SURVEY SAMPLING USING UNBIASED ESTIMATORS OF ZERO

Tapan K. Nayak

George Washington University, Department of Statistics

Washington, DC 20052

e-mail: tapan@gwu.edu

Abstract. Minimum mean squared error linear unbiased estimation of the total of a finite population is investigated utilizing a characterization of such estimators involving unbiased estimators of zero. Several criteria emerging from taking either or both of the design and a model into account are discussed. For each criterion, necessary and sufficient conditions for existence of a best estimator and the optimal estimator, when it exists, are derived using our unified approach. The results are presented for an equi-correlated regression superpopulation model, which includes many commonly used models as special cases.

1. INTRODUCTION

We consider a finite population of \( N \) units, and a variable \( y \) whose value for unit \( i \) is \( y_i \). It is assumed that \( y_i \) is the realized values of \( Y_i \) with

\[
Y_i = x_i' \Theta + \Omega_i, \quad i = 1, \ldots, N,
\]

where \( x_1, \ldots, x_N, \Theta \) are given vectors, \( \Omega \), and \( \Theta \) are unknown parameters, and \( a_1, \ldots, a_N \) are known nonzero numbers. Many special cases of our model have been discussed by Godambe (1955), Cassel et al. (1976, 1977), Isaki and Fuller (1982) and others. We consider estimation (or prediction) of the population total \( t = \sum_{i=1}^{N} y_i \) based on \( d = \{(k, y_k); k \in s\} \), where \( s \) is a subset of the population units selected according a non-informative sampling design \( p \).

In this paper, we compare linear unbiased estimators of \( t \) based on their mean squared errors (MSE). An estimator \( e \) is called linear if it is of the form \( e = w_{s0} + \sum_{i \in s} w_i y_i \), and if \( w_{s0} = 0 \), it is called homogeneous linear. The bias and MSE of an estimator \( e \) are the expected values of \( (e - t)^2 \) and \( (e - t)^2 \), respectively. Three measures of bias and MSE arise from taking the expectations, respectively, under (i) the sampling design \( p \) keeping \( y \) fixed, (ii) the model (1) keeping \( s \) fixed, and (iii) both the sampling design \( p \) and the model (1).

The prefixes \( \text{pm}_- \), \( \text{m}_- \), and \( \text{p}_- \) will be used to distinguish the biases and MSEs defined by these three approaches. We investigated minimizing the MSE of linear estimators subject to unbiasedness for several choices of bias and MSE measures. Our technique, based on the following generalization of a well known result in estimation theory, is new and quite general.

**Theorem 1.** Let \( D \) be a random vector and \( T \) be a random variable with joint distribution \( f(d, t|D), D \in \Omega \). Suppose \( C \) is a class of predictors of \( T \) based on \( D \), and \( C^0 \) is a class of functions of \( D \) with finite variances such that (i) for all \( e_1, e_2 \in C \), \( e_1 - e_2 \) is in \( C^0 \) and (ii) \( C \) contains \( e + ku \) for all \( e \in C, u \in C^0 \) and \( \|u\| < k < \infty \). Then, \( e_0 \in C \) minimizes \( E_{\Omega}[e^2 \Omega T] \) for all \( \Omega \in \Omega \) and among all predictors in \( C \) if and only if \( E_{\Omega}[(e_0 \Omega T)u] = 0 \) for all \( \Omega \in \Omega \) and \( u \in C^0 \).
2. OPTIMALITY RESULTS

In the following we state some of the optimality results we obtained using the new approach.

**Result 1.** Under the criterion of minimizing the $p$-MSE subject to $p$-unbiasedness, (a) a best estimator within all linear estimators does not exist, and (b) a best estimator within all homogeneous linear estimators exists if and only if the design is unicluster and $\mathbb{P}_i > 0, i = 1, \ldots, N$ (where $\mathbb{P}_i$ is the inclusion probability of unit $i$), in which case the best estimator is $\hat{e}_{HT} = \sum_{i \in s} y_i / \mathbb{P}_i$, the Horvitz-Thompson estimator.

**Result 2.** Under the criteria of minimizing the $m$-MSE or the $pm$-MSE under $m$-unbiasedness, the best estimator of $t$ exists if and only if at least one of the following three conditions hold: (i) $\mathbb{P} = 0$, (ii) for each $s, a_i = x_i' k_s, i \in s$, for some vector $k_s$, (iii) for each $s, \sum_{i \in s} w_{si} a_i = \sum_{i=1}^N a_i$. Further, if the the best estimator exists, it is given by

$$
\hat{e}_0 = \sum_{i \in s} y_i + (\sum_{i \notin s} x_i) / \mathbb{P},
$$

where $\mathbb{P} = (\sum_{i \in s} x_i' x_i / a_i^2)^{-1} (\sum_{i \in s} x_i y_i / a_i^2)$ is a weighted least squares estimator of $\mathbb{P}$.

**Result 3.** Under the criterion of minimizing the $pm$-MSE subject to $p$-unbiasedness, a best estimator exists if and only if $\sum_{i \in s} x_i / \mathbb{P}_i = \sum_{i=1}^N x_i$ for all $s$, and either $\mathbb{P} = 0$ or $\sum_{i \in s} a_i / \mathbb{P}_i = \sum_{i=1}^N a_i$ for all $s$. Further, the best estimator, when it exists, is the Horvitz-Thompson estimator.

We also considered minimizing the $pm$-MSE under $pm$-unbiasedness, and found that when a best estimator exists, it is the estimator in (2). However, some conditions need to be satisfied for it to be optimal. Also, if (2) is optimal according to this criterion, then it is also optimal according to the criterion in Result 2, but the converse is not true. We also discussed many special cases of our results leading to the regression, the ratio, the mean of ratios, and other standard estimators. Thus, conditions for optimality of those estimators are also obtained.

We would like to emphasize that our approach is new and it has two main advantages: it is general and applies to various criteria, and it directly leads to necessary and sufficient conditions for existence of a best estimator, and the best estimator, when it exists.

REFERENCES


