Predictive Estimation In Finite Population Sampling

M. C. Agrawal [1]

Abstract

Recent developments in Survey Sampling have presented the estimation of the population total (or the mean) in a new perspective called "predictive estimation". In this expository article, the highlights of the various contributions in this area have been reviewed and knit together to facilitate appreciation of a number of interesting results under the "fixed population set-up" and the "superpopulation" set-up.

1. Introduction

We consider a population of N units arbitrarily labelled 1, 2, ..., N. Let the characteristic of interest say, y take the y_i on the ith unit. We wish to estimate the population total $Y = \sum_{i=1}^{N} y_i$ from observations made on the units of a sample s (obtained by any sampling scheme whatsoever). The criteria of optimality such as unbiasedness and minimum variance are usually invoked to derive or assess an estimator without consideration to the fact that the sampling procedure divides the population into a completely known part and a completely unknown part in respect of the characteristic y. It is precisely in pursuance of such a consideration that the predictive theory of estimation (also called "prediction approach") has been brought about to play a pivotal role in shedding light on the estimation of a population total (or the population mean). Writing the total as

$$Y = \sum_{i \in s} y_i + \sum_{i \in s} y_i = \sum_{i \in s} y_i + Y_s, \qquad (1.1)$$

we note that, in order to estimate Y, the problem essentially is to predict the second component on the right side, viz., $\sum_{i \in s} y_i$ or Y_s (s being the complement of s) since the first one on the right side is exactly known. Rather, we have to predict Y_s on the basis of $\sum_{i \in s} y_i$ and, in order that the latter should provide information on the former, there must be some link between the two components. Such a link is available to us in survey sampling in the form of a sampling design or a superpopulation model or simply a model. We may designate Y_s as "unknown component" or "unobserved residuum". In what follows, we shall present and scan the different approaches with regard to the prediction of this unknown component.

2. Predictive Estimation

In the context of the superpopulation set-up, Royall (1970) and, in the context of the fixed population set-up, Basu (1971) were concerned with the prediction of the unknown component, Y_s, each in his own way as would be clear hereinafter. Basu (1971) has suggested what has come to be known as "predictive approach" [vide Smith (1976)] for examining the plausibility and face-validity of an estimator. The approach presents the estimation of

Keywords: Predictive estimation, predictor, double sampling procedures, Superpopulation.

[1] U.P. Visiting Propesor

the population total (or the mean) in an interesting perspective which was described by Basu (1971) as the "heart of the matter. In his approach designated "prediction theory or approach", Royall (1970) worked out an optimal predictor of Y_s (and hence that of Y) under a specified superpopulation model.

We would now discuss the problem of predicting the unknown component under the following set-ups:

A. The Fixed Population Set-up (F-P Set-up): The value of the characteristic of interest, under this set-up, for each unit of the population is a fixed but unknown real number.

An estimator Y of Y given in (1.1) can be expressed as

$$Y = \sum_{i \in s} y_i + \sum_{i \in s} y_i$$

or (2.1)

$$Y = \sum_{i \in s} y_i + Y_s$$

where y_i and Y_s are the implied predictors of y_i and Y_s, respectively.

Model-free Prediction

Basu (1971) examines the Desraj ordered estimator from the "predictive" standpoint and finds it non-conforming. The Desraj ordered estimator is defined by

$$Y_D = y_1 + y_2 + \ldots + y_{n-1} + y_n/p_n\ (1-p_1-p_2-\ldots-p_{n-1})$$
$$= y_1 + y_2 + \ldots + y_n + y_n/p_n\ (1-p_1-p_2-\ldots-p_n)$$

where $y_1, y_2, \ldots, y_n$ are the observations on the n units in order of their draw in the sample, and $p_1, p_2, \ldots, p_n$ are the initial probabilities of the respective units in the sample. To fix the idea, let p_i be proportional to some measure of size, say, x_i and then, on comparing with (2.1), it can be observed from the Desraj estimator that the predictor of Y_s is $y_n/x_n\ (X_s)$ where X_s is the sum of X-values on the non-sampled units. Here, the predictor [i.e. $y_n/x_n\ (X_s)$] exploits the information on the nth unit only, and not on all the n units, for the purpose of predicting Y_s and as such, it lacks in plausibility. However, if we consider a predictor based on suitable pooling of the sample observations, i.e., $(\sum_{i \in s} y_i / \sum_{i \in s} x_i)\ X_s$ for predicting Y_s; then we get the conventional ratio estimator of Y.

It would be interesting to scan certain widely used conventional estimators from the viewpoint as to whether they conform to the predictive form or not. Starting with simple random sampling without replacement, we note that the simple expansion estimator of the population total is expressible as

$$Y_{ran} = Ny = \sum_{i \in s} y_i + \sum_{i \in s} y,$$

implying that the predictor of y_i in unobserved residuum is y (the sample mean of y-values), which is quite sensible under the circumstances when no extra or ancillary information is available.

In regard to simple random sampling without replacement, we may treat the problem of predicting Y_s as being essentially one of propounding a linear function of sample observations, say $\sum_{i \in s} d_i y_i$ where d_i's are to be determined optimally so that the function has minimum variance and is unbiased for the expectation of Y_s, i.e.

$$E\left(\sum_{i \in s} d_i y_i\right) = E(Y_s)$$

$$\sum_{i \in s} d_i = N - n$$

Employing an optimisation method to achieve the minimum variance subject to unbiasedness, we obtain $d_i = (N - n)/n$ and hence the usual simple expansion estimator Ny of Y.

Given an auxiliary variable x (x_i being the value for unit i) related to y, the well-known ratio and regression

estimators for simple random sampling can, respectively, be split as

$$Y_r = N\frac{y}{x}X = \sum_{i\varepsilon s} y_i + \sum_{i\varepsilon s}\frac{y}{x}x_i$$

and

$$Y_l = N[y + b(X - x)] = \sum_{i\varepsilon s} y_i + \sum_{i\varepsilon s}[y + b(x_i - x)],$$

where x and X are the sample and population means of x-values, respectively, and b is the sample regression coefficient. The predictors of y_i in these two cases are intuitively appealing and compatible with the situations visualised for the use of these estimators, and thus these estimators are reasonably endowed with a predictive form.

We would now allude to some estimators which lack in a predictive character such as the product and the Horvitz-Thompson estimators which could, respectively, be decomposed as

$$Y_{Prod} = N\frac{yx}{X} = \sum_{i\varepsilon s} y_i + \sum_{i\varepsilon s}\frac{yx}{x_s}\cdot\frac{N-n}{(N-n)}\cdot\frac{X_s}{X}$$

and

$$Y_{HT} = \sum_{i\varepsilon s}\frac{y_i}{\pi_i} = \sum_{i\varepsilon s} y_i + \sum_{i\varepsilon s} x_i - \sum_{i\varepsilon s}\left[\frac{y_i}{n x_i}(X - nx_i)\right],$$

where π_i is the probability of inclusion of unit i in the sample, and is taken proportional to x_i, i.e., $\pi_i = nx_i/X$ and X and X_s are, respectively, the population total of x-values and the mean of x-values for the non-surveyed units. Neither of these estimators apparently conforms to a predictive form, for there is no appealing explanation in favour of the respective preditors for Y_s. However, in the case of the Horvits-Thompson estimator, we find, on a closer scrunity, that this is the only unbiased estimator under the proposed sampling design if we employ a linear predictor (i.e., a linear function of sample observations) for Y_s with a view to minimising its variance subject to unbiasedness, i.e.,

$$E\left(\sum_{i\varepsilon s} d_i y_i\right) = E(Y_s) = E\left(Y - \sum_{i\varepsilon s} y_i\right)$$

$$\sum_{i=1}^{n} \pi_i d_i y_i = Y - \sum_{i=1}^{n} \pi_i y_i$$

$$d_i = \frac{1-\pi_i}{\pi_i}$$

and this leads to the Horvitz-Thompson estimator.

In the above context, we wish to suggest that, wherever possible, we should try to predict Y_s by employing a linear predictor whose variance is sought to be minimised subject to unbiasedness.

Some other known estimators, e.g., the ratio-type estimator and the usual estimator in simple random sampling with replacement also lack in predictive character. It is, therefore, not surprising that some of the estimators not conforming to the predictive form have generated a lot of discussion and controversy.

An interesting aspect of the predictive theory of estimation is that, in respect of a conventional estimator as is not endowed with a predictive form, it would be a worthwhile idea if one works out a "predictive" estimator of Y by planting an "apt" predictor of y_i (or Y_s) in (2.1). To design such as "apt" predictor, we can exploit the background information that is available to us in the form of the conventional estimator. Subsequently, a comparison of the conventional and the corresponding predictive estimators in respect of bias and mean square error may be undertaken. As a matter of fact, some work in this direction has already been initiated [see Srivastava (1983), Agrawal and Jain (1987), Agrawal & Kulldorff (1987)].

Prediction for Incomplete Data

It is well known that the Hansen-Hurwitz technique is applied to tackle incomplete data due to non-response in mail surveys. This technique consists in collecting information from a subsample of non-respondents, say, m out of n_2 non-respondents, in the second attempt, while n1 in a sample of size n_1 have responded in the first attempt. To clarify the matter, we introduce the following notations:

s_1 : set of those sampled and responding in the first attempt, its size being n_1.

s_2 : set of those sampled and non-responding in the first attempt, its size being n_2.

s_3 : set of the non-sampled units.

Here, the predictive estimator of Y can be viewed as

$$Y = \sum_{i \in s_1} y_i + \sum_{i \in s_2} y_i + \sum_{i \in s_3} y_i.$$

We can now split the customary Hansen-Hurwitz estimator as

$$Y_{HH} = Ny^* = \sum_{i \in s_1} y_i + \sum_{i \in s_2} y_m + \sum_{i \in s_3} y^*.$$

where $y^* = (n_1 y_{n1} + n_2 y_m)/(n_1 + n_2)$, and y_{n1} and y_m are the means based on n_1 and m units, respectively. In this case, it would be quite natural to predict y_i in s_2 by y_m and y_i in s_3 by y^*. Thus, Y_{HH} is predictive in character.

Predictive Estimation in Double Sampling Procedures

Agrawal and Jain (1987) have examined the ratio, ratio-type and regression estimators in double sampling procedures from the predictive standpoint under the following two approaches:

(A) Having drawn the first-phase sample S_1, the second-phase sample S_2 is drawn as a sub-sample from the first-phase sample.

(B) Having drawn the first-phase sample S_1, the second-phase sample S_2^* is drawn independently from the whole population.

The authors have split the population total under the respective approaches (A) and (B) as

$$Y_A = \sum_{i \in s_2} y_i + \sum_{i \in s_1 s_2} y_i + \sum_{i \in s_1} y_i$$

and,

$$Y_B = \sum_{i \in s_2^*} y_i + \sum_{i \in s_3} y_i + \sum_{i \in s_4} y_i$$

where S_3 and S_4 are the sets containing (v-n) and (N-v) units (v being the number of distinct units in the two samples S_1 and S_2^*), and S_1 and S_2 are the complements of S_1 and S_2, respectively. Since the first component in both Y_A and Y_B is known, the predictors of Y_A and Y_B can be written as

$$Y_A = \sum_{i \in s_2} y_i + \sum_{i \in s_1 s_2} y_i + \sum_{i \in s_1} y_i$$

and,

$$Y_B = \sum_{i \in s_2^*} y_i + \sum_{i \in s_3} y_i + \sum_{i \in s_4} y_i.$$

The ratio and regression estimators under approach (A) are expressible as

$$Y_{rd} = N\frac{y}{x}x' = \sum_{i \in s_2} y_i + \sum_{i \in s_1 s_2} \frac{y}{x} x_i + \sum_{i \in s_1} \frac{y}{x} x'$$

and

$$Y_{1d} = N(y + b(x' - x)) = \sum_{i \in s_2} y_i + \sum_{i \in s_1 s_2} (y + b(x_i - x)) + \sum_{i \in s_1} (y + b(x' - x)),$$

where x', is the mean of the auxiliary variable based on S_1 and b is the sample regression coefficient of y on x. Here, the predictors of y_i in the second and third components of both the estimators are quite intuitive and sensible, and as such, these estimators are endowed with

predictive form. Similarly, these two estimators unfold themselves in predictive form under approach (B).

The customary ratio-type estimator (biased) in double sampling is found to be lacking in predictive form under both the approaches (A) and (B). Agrawal and Jain (1987) work out "predictive" ratio-type estimators under the approaches (A) and (B), and compare their biases with those of the conventional ratio-type estimators.

Model-based Prediction

The foregoing discussion under the F-P set-up was model-free. However, one could use models under the F-P set-up, and consider the problem of prediction in the presence of a model. Models have been employed in the context of ratio and regression methods of estimation [see Sukhatme and Sukhatme (1970, p. 146, p.197)]. In such cases, we choose, as usual, a linear predictor for Y_s and seek to achieve minimum variance coupled with unbiasedness under the specified models. For models just referred to, we would arrive at the conventional ratio and regression estimators.

B. The Superpopulation Set-up (S-P Set-up):

Under this set-up, a random variable with a specified stochastic structure is associated with each unit of the population. The actual value observed on a population unit is regarded as the realization of this random variable.

In the context of the S-P set-up, it would be quite apt to quote from Cassel et al. (1977) regarding the distinction between superpopulation and Bayesian models. The Bayesian models express personal subjective belief in the form of a prior distribution $y_1, ..., Y_N$, while the superpopulation models need not be Bayesian in this sense and, at times, the latter could be as objective as some of the models in classical statistical theory.

We would now consider the prediction of the unknown component when we specify, in respect of the superpopulation, some of its moments or its complete distributional form.

Specification of moments

Royall (1970) invokes a superpopulation model to predict Y_s, and designates the inferential process as "prediction approach" under the superpopulation models are allowed a primary and dominent role, while the sampling designs are relegated to the background. He obtains an optimal predictor of Y_s for the following superpopulation model ξ

$$E\,\xi(y_i) = \beta x_i \quad (i = 1, 2, \ldots N)$$

$$V\xi(y_i) = \sigma^2 v(x_i), \qquad (2.2)$$

where the function v is known, and β and σ^2 are unknown constants. A result of special interest discussed by Royal (1970) is when $v(x_i) = x_i$, for it leads to the conventional ratio estimator of Y.

Under a superpopulation model with certain specified moments, a general approach that we suggest is to predict Y_s by optimizing a linear function of sample observations with a view to achieving minimum model variance subject to model unbiasedness. In notations, we are proposing

$$Y = \sum_{i \in s} y_i + \sum_{i \in s} d_i y_i$$

as an estimator of

$$Y = \sum_{i \in s} y_i + Y_s,$$

wherein we seek minimisation of the model (ξ) variance $E\xi(\sum_{i \in s} d_i y_i - Y_s)^2$ subject to the model (x) unbiasedness

defined by

$$E\xi\left(\sum_{i \in s} d_i y_i\right) = E\xi\left(Y_s\right),$$

where ξ denotes any superpopulation model. It would not be out of place to point out the Cochran (1977, pp. 158-59) has discussed the optimal estimation under the model (2.2) with $v(x_i) = x_i$. We, however, notice a discrepancy in his proof when he invokes the concept of model variance in two senses at two points, i.e., he minimises the model variance $E\xi\,[Y - E(Y)]^2$ but, finally, determines the model variance $E\xi\,(Y - Y)^2$.

Specification of distributional form

Hajek (1981) has considered the prediction of Y_s by speculating on $y_1, y_2, \ldots, y_n$ as random variables having the same normal distribution $N(\mu, \sigma^2)$ where μ and σ^2 are unknown parameters. For the purpose of predicting Y_s, he suggest the statistics $t(y_1, y_2, \ldots, y_n)$ and then minimises the quantity

$$E[y_{n+1} + y_{n+2} + \ldots + y_N - t(y_1, y_2, \ldots, y_n)]^2$$
$$= E[(N-n)\mu - t(y_1, y_2, \ldots, y_n)]^2 + (N-n)\sigma^2$$

which amounts to estimation of m in a normal sample, thus yielding

$$t(y_1, y_2, \ldots, y_n) = (N-n)\,y.$$

Hence, Y_s is predicted by $(N-n)y$ and Y by Ny.

In another example considered by Hajek (1981), the Y_i $(i = 1, 2, \ldots, N)$ terms are assumed to be independent random variables that have the normal distributions $N(ax_i, b^2x_i)$ where $a, b > 0$ are unknown parameters. Using the same technique as in the preceding case, the predictor of Y_s is obtained as

$$Y_s = \frac{\sum_{i \in s} y_i}{\sum_{i \in s} x_i} \sum_{i \in s} x_i$$

and hence, Y will be predicted by $(y/x)\,X$ which is the conventional ratio estimator.

Thus, viewed in totality, we can look upon the predictive estimation as an instrument to judge the face-validity of an existing estimator and, in case, the estimator does not conform to a predictive form, the prediction approach may be used to construct a "predictive" estimator with a view to comparing the latter estimator with the existing one. Further, such an approach can be invoked to provide an optimal estimator through recourse to prediction, under a given set of conditions, of the unknown component, irrespective of whether we have a fixed population or a superpopulation set-up.

3. A case of convergence under F-P and S-P set-ups

The customary ratio-type (biased) estimator is given by

$$y_{rt} = \frac{1}{n} \sum_{i \in s} \frac{y_i}{x_i} X = rX,$$

where $r = 1/n \sum_{i \in s} r_i$ and $r_i = y_i/x_i$. Agrawal and Kulldorff (1987) have indicated that, in the light of the estimator y_{rt}, an apt predictor of y_i for $i \in s$ under the F-P set-up would be rx_i, or alternatively (and equivalently), the predictor of Y_s would be rX_s. Thus (1.2) would lead to

$$Y = \sum_{i \in s} y_i + rX_s = rX + n(y - rx)$$

or

$$Y = rX + n/N(y - rx),$$

where the different notations have the same meaning as in the preceding sections. Agrawal & Kulldorff (1987) designate Y as predictive ratio-type estimator of the population mean and point out that this estimator is optimal (in the sense of model unbiasedness and minimum

model mean square error) under the following super-population model ξ [see Royall (1970)]:

$$E\xi(y_i) = \gamma x_i,\ V\xi(y_i) = \lambda x_i^2,\ \text{cov}(y_i, y_j) = 0$$

for $i, j = 1, 2, ..., N\ (i \neq j)$

where γ and λ are parameters and x_i $(i = 1, 2, ..., N)$ are known. Thus the predictive ratio-type estimator y under the F-P set-up is optimal under the S-P set-up.

References

Agrawal, M.C. and Jain, Nirmal (1987). Predictive estimation in double sampling procedures, to appear in American Statistician.

Agrawal, M.C. and Kulldorff, G. (1987). Precision of ratio-type estimators and a predictive counterpart, Statistical Research Report No. 1987-3, Institute of Mathematical Statistics, University of Umea (Sweden).

Basu, D. (1971). An essay on the logical foundations of Survey Sampling, Part one, in V.P. Godambe and D.A. Sprott, eds., Foundations of Statistical Inference, Toronto: Holt, Rinehart and Winston of Canada, Ltd., 203-233.

Cassel, C., Sarndal, C. and Wretman, J.H. (1977). Foundations of Inference in Survey Sampling, John Wiley & Sons, New York.

Cochran, W.G. (1977). Sampling Techniques, John Wiley & Sons, New York.

Hajek, J. (1981). Sampling from a Finite Population, Marcel Dekker, Inc., New York.

Royall, R.M. (1970). On finite population sampling under certain linear regression models, Biometrika, 57, 377-387.

Smith, T.M.F. (1976). The foundations of Survey Sampling: a review, Journal of the Royal Statistical Society, A 139, 183-204.

Srivastava, S.K. (1983). Predictive estimation of finite population mean using product estimator, Metrika. 30, 93-99.

Sukhatme, P.V. and Sukhatme, B.V. (1970). Sampling Theory of Surveys with Applications, Asia Publishing House, London.