Sankhyā : The Indian Journal of Statistics
1938, Volume 50, Series B, Pt. 2, pp. 206-214

OPTIMAL ESTIMATORS UNDER REGRESSION MODEL

By V. R. PADMAWAR
Indian Statistical Institute

SUMMARY. In view of the nonexistence of an optimal design unbiased strategy under regression model, certain optimal strategies are obtained under different restricted set-ups.

1. Introduction

Consider a finite population $U = \{1, 2, ..., N\}$ of size N. Let y be a study variate taking values y_i on units i, $1 \leqslant i \leqslant N$. Let x be an auxiliary variate, closely related to y, taking values x_i on units i, $1 \leqslant i \leqslant N$. We assume that $y_1, y_2, ..., y_N$ are a realisation of the variables $Y_1, Y_2, ..., Y_N$; the joint distribution of which is unknown but specified by the first two moments as follows. If E_ξ and V_ξ denote expectation and variance, respectively, w.r.t. the model ξ that defines a class of distributions for $Y_1, Y_2, ..., Y_N$; then

$$\begin{aligned} E_\xi(Y_i) &= \beta x_i && i = 1, 2, ..., N \\ V_\xi(Y_i) &= \sigma^2 x_i^g && i = 1, 2, ..., N \qquad \ldots (1.1) \\ E_\xi(Y_i Y_j) &= \beta^2 x_i x_j && i \neq j = 1, 2, ..., N \end{aligned}$$

where β and $\sigma^2 > 0$ are the model parameters and $g \in [0, 2]$. Model (1.1) is called regression model.

Our aim is to estimate the population mean $\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} y_i$. Given a strategy (p, t) that consists of a design p and an estimator $t(s, y)$, $y = (y_1, y_2, ..., y_N)$, we use the following measure of uncertainty to measure its performance.

$$M(p, t) = E_\xi E_p (t - \bar{Y})^2 \qquad \ldots (1.2)$$

where E_p denotes expectation w.r.t. the design p.

A strategy (p, t) is said to be p-unbiased or design unbiased for estimating the population mean $\bar{Y}$ if

$$E_p(t(s, y)) = E_p(t) = \sum_{s \in S} p(s)\, t(s, y) = \bar{Y} \ \forall\ y$$

AMS (1980) *subject classification :* 62D05.

Key words : Regression model, Population mean, Measure of uncertainty, Optimal estimators.

where S is the collection of all samples.

A strategy (p, t) is said to be ξ-unbiased or model unbiased for the population mean $\bar{Y}$ if

$$E_\xi(t(s, y)) = E_\xi(\bar{Y}) \ \forall\ s \text{ with } p(s) > 0$$

Now for a p-unbiased strategy (p, t)

$$M(p, t) = E_p V_\xi(t) + E_p(E_\xi(t - \bar{Y}))^2 - V_\xi(\bar{Y}) \qquad \ldots \quad (1.3)$$

and for a p as well as ξ-unbiased strategy (p, t)

$$M(p, t) = E_p V_\xi(t) - V_\xi(\bar{Y}) \qquad \ldots \quad (1.4)$$

Ramachandran (1982) obtained a strategy for estimating the population mean $\bar{Y}$ that, if exists, is optimal in the class of all p-unbiased strategies of given average size in the sense of minimum $M(p, t)$. Her strategy consisted of Horvitz—Thompson estimator and a sampling design that satisfies the following conditions

$$\text{(i)} \quad \sum_{s \ni i} p(s) = \frac{n\, x_i^{g/2}}{\sum_{j=1}^{N} x_j^{g/2}}, \quad 1 \leqslant i \leqslant N$$

and

$$\text{(ii)} \quad \sum_{i \in s} x_i^{1-g/2} = n \sum_{i=1}^{N} x_i \Big/ \sum_{i=1}^{N} x_i^{g/2} \ \forall\ s \text{ with } p(s) > 0.$$

Condition (i) requires the design to be a $\pi px^{g/2}$ design (inclusion probabilities proportional to $x^{g/2}$) of average size n.

As such it is difficult to construct a design satisfying the first condition. Further conditon (*ii*) puts far too many constraints on the fixed set of constants $x_1, x_2, \ldots, x_N$ and g to be able to satisfy them. Therefore the existence of the optimal strategy due to Ramachandran is ruled out in most practical situations.

In view of this it is in order to either compare different known strategies or consider the problem of obtaining optimal strategies in different restricted set-ups. The first aspect is studied in Padmawar (1981). Here we deal with the second aspect. Padmawar (1984) proved the existenc of certain optimal estimators under the regression model in the continuous survey sampling set up. We take this opportuhity to see their implications in the finite set up.

2. Optimal estimators

Given a design p, in this section, we first establish the existence and uniqueness of best linear p as well as ξ-unbiased estimator in the sense of minimum $M(p, t)$. Next, assuming the ratio $\sigma^2/\beta^2 = r$(say), in the model (1.1), to be known we prove the esistense of a best linear p-unbiased estimator again in the sense of minimum $M(p, t)$.

A linear estimator is of the type $t(s, y) = \sum_{i \in s} b(s, i) y_i$. The condition of p-unbiasedness is equivalent to

$$\sum_{s \ni i} b(s, i)\, p(s) = \frac{1}{N} \quad \forall\ i = 1, 2, \ldots, N \qquad \ldots (2.1)$$

and the condition of ξ-unbiasedness is equivalent to

$$\sum_{i \in s} b(s, i) x_i = \bar{X} \quad \forall\ s \text{ with } p(s) > 0 \qquad \ldots (2.2)$$

where $$X = N\bar{X} = \sum_{i=1}^{N} x_i.$$

For a given design p, to tackle the first problem, we minimise $M(p, t)$ subject to the conditions (2.1) and (2.2). In view of (1.4) the problem reudces to the following minimisation problem.

Minimise $$\sum_{i=1}^{N} q_i \sum_{s \ni i} a^2(s, i)\, p(s)$$

subject to $$\sum_{s \ni i} a(s, i)\, p(s) = 1 \quad \forall\ i = 1, 2, \ldots, N \qquad \ldots (2.3)$$

and $$\sum_{i \in s} a(s, i) p_i = 1 \qquad \forall s \text{ with } p(s) > 0$$

where $a(s, i) = Nb(s, i)$; $x_i = Xp_i$ and $N^2 q_i = x_i^2$.

We employ Lagrangian multipliers' technique to solve (2.3). Let

$$Q = \sum_{i=1}^{N} q_i \sum_{s \ni i} a^2(s, i)\, p(s) - 2 \sum_{i=1}^{N} \lambda_i \left[\sum_{s \ni i} a(s, i)\, p(s) - 1 \right]$$

$$- 2 \sum_{s \in S} p(s) \alpha_s \left[\sum_{i \in s} a(s, i) p_i - 1 \right]$$

where $\lambda_i, 1 \leqslant i \leqslant N$ and α_s ; $s \in S$ and $p(s) > 0$ are the Lagrangian multipliers. Solving (2.3) is equivalent to minimising Q unconditionally. Differ-

entiating and simplifying the problem boils down to solving the following system of equations

$$q_i a(s, i) = \lambda_i + \alpha_s \qquad i \in s,\ p(s) > 0$$

$$\sum_{s \ni i} a(s, i) p(s) = 1 \qquad 1 \leqslant i \leqslant N \qquad \ldots (2.4)$$

and $$\sum_{i \in s} a(s, i) p_i = 1 \qquad s \in S\ ;\ p(s) > 0.$$

After a little algebra one gets

$$a(s, i) = z_i + \eta_i \left(1 - \sum_{j \in s} z_j p_j\right) m(s) \quad i \in s\ ;\ s \in S,\ p(s) > 0 \quad \ldots (2.5)$$

where $\eta_i q_i = p_i$, $\dfrac{1}{m(s)} = \sum_{j \in s} \eta_j p_j$ and $z = (z_1, z_2, \ldots, z_N)$ is a solution to the the system of equations

$$A\, z = d \qquad \ldots (2.6)$$

with $$a_{ii} = \pi_i - \eta_i p_i \sum_{s \ni i} p(s) m(s) \quad i = 1, 2, \ldots, N$$

$$a_{ij} = -\eta_i p_j \sum_{s \ni i, j} p(s) m(s) \quad i \neq j = 1, 2, \ldots, N \qquad \ldots (2.7)$$

and $$d_i = 1 - \eta_i \sum_{s \ni i} p(s) m(s) \qquad i = 1, 2, \ldots, N$$

where $$\pi_i = \sum_{s \ni i} p(s),\ 1 \leqslant i \leqslant N.$$

Following Patel and Dharmadhikari (1977, 1978) it can be shown that the system (2.6) is consistent and the solutions (2.5) do not depend on a choice of z satisfying (2.6). We thus have the following theorem

Theorem 2.1 : *Given a design p there exists a unique best linear p as well as ξ-unbiased estimater in the sense of minimum $M(p, t)$.*

Remark 2.1 : Theorem 2.1 is true even when we have a more general variance function $\sigma^2 v(x)$, instead of $\sigma^2 x^\theta$, in the model (1.1). In that case we simply set $N^2 q_i = v(x_i)$ in (2.3).

Remark 2.2 : In the continuous set up (vide Padmawar (1984)) the existence of an analogus optimal estimator depends on the solvability of an integral equation whereas in the finite set up existence as well as uniqueness are always guaranteed. However, the construction of such an optimal estimator depends on the tractability of solutions to the system of equations (2.6). In what follows we consider two examples so as to get some idea of the above optimality result.

Example 2.1 : Considor a fixed size dosign of size n for which $\pi_i = nx_i/X$, $1 \leqslant i \leqslant N$. $z_i = X/nx_i$, $1 \leqslant i \leqslant N$, is a solution to (2.6) for the ith entry in Az is then given by

$$\frac{X}{nx_i}\left[\pi_i - \eta_i p_i \sum_{s \ni i} p(s)m(s)\right] - \sum_{j \neq i}^{N} \frac{X}{nx_j}\eta_j p_j \sum_{s \ni i,j} p(s)m(s)$$

$$= 1 - \frac{1}{n}\eta_i \sum_{s \ni i} p(s)m(s) - \frac{n-1}{n}\eta_i \sum_{s \ni i} p(s)\, m(s)$$

$$= 1 - \eta_i \sum_{s \ni i} p(s)\, m(s)$$

$$= d_i.$$

Therefore the optimal estimator, as expected, is given by $\frac{X}{nN} \sum_{i \in s} \frac{y_i}{x_i}$.

Example 2.2 : Consider the Midzuno-Sen sampling design of fixed size n for which $p(s)$ is given by $w_s/M_1 W$ where $w_s = \sum_{i \in s} w_i$, $W = \sum_{i=1}^{N} w_i$ and $M_1 = \binom{N-1}{n-1}$ with $w_i = x_i^{1-\theta}$, $1 \leqslant i \leqslant N$. For this design the coefficients (2.7) simplify to

$$a_{ii} = \pi_i - w_i/W \qquad i = 1, 2, \ldots, N$$

$$a_{ij} = -\frac{n-1}{N-1}\, x_i^{1-\theta}\, x_j/W \qquad i \neq j = 1, 2, \ldots, N$$

$$d_i = 1 - x_i^{1-\theta} X/W \qquad i = 1, 2, \ldots, N$$

where $\pi_i = \frac{n-1}{N-1} + \frac{N-n}{N-1}\, w_i/W$, $1 \leqslant i \leqslant N$, and $X = \sum_{i=1}^{N} x_i$

Now $$\sum_{j=1}^{N} a_{ij} = \pi_i - w_i/W - \frac{n-1}{N-1}\, x_i^{1-\theta}(X - x_i)/W$$

$$= \frac{n-1}{N-1}\, \{1 - x_i^{1-\theta} X/W\}$$

$$= \frac{n-1}{N-1}\, d_i.$$

Thus $z_i = \frac{N-1}{n-1}$; $1 \leqslant i \leqslant N$; is a solution to (2.6) and the optimal estimator is given by

$$\frac{1}{N}\, \frac{N-1}{n-1} \sum_{i \in s} y_i + \frac{1}{N}\left\{1 - \frac{N-1}{n-1}\, \frac{1}{X} \sum_{i \in s} x_i\right\} \sum_{i \in s} x_i^{1-\theta} y_i \frac{X}{\sum_{i \in s} x_i^{1-\theta}}$$

We have the following interesting cases.

Case 1 : For $g = 2$ in Example 2.2 the sampling design reduces to simple random sampling and the optimal estimator simplifies to

$$\frac{1}{N}\frac{N-1}{n-1}\sum_{i\in s} y_i+\frac{1}{N}\left\{1-\frac{N-1}{n-1}\frac{1}{X}\sum_{i\in s} x_i\right\}\frac{X}{n}\sum_{i\in s} y_i/x_i$$

which can be rewritten as

$$\bar{r}\bar{X}+\frac{n(N-1)}{N(n-1)}\{\bar{y}-\bar{r}\bar{x}\}$$

where $n\bar{x} = \sum_{i\in s} x_i$, $n\bar{y} = \sum_{i\in s} y_i$, $n\bar{r} = \sum_{i\in s} y_i/x_i$, $N\bar{X} = X$. Note that this optimal estimator is same as the Hartley-Ross (1954) ratio-type estimator for the population mean $\bar{Y}$.

Case 2 : For $g = 1$ in Example 2.2 the sampling design reduces to the usual Midzuno-Sen sampling scheme with $p(s) = \frac{x_s}{M_1 X}$, $x_s = \sum_{i\in s} x_i$. and the optimal estimator reduces to

$$\frac{1}{N}\frac{N-1}{n-1}\sum_{i\in s} y_i+\frac{1}{N}\left\{1-\frac{N-1}{n-1}\frac{1}{X}\sum_{i\in s} x_i\right\}\frac{X\sum_{i\in s} y_i}{\sum_{i\in s} x_i}$$

or

$$\frac{\sum_{i\in s} y_i}{\sum_{i\in s} x_i}\bar{X}$$

which is the standard ratio estimator for estimating the population mean.

We now assume that the ratio $\sigma^2/\beta^2 = r$ in (1.1) is known. Given a design p we prove the existence of an optimal p-unbiased linear estimator in the sense of minimum $M(p, t)$.

In view of (1.3) our problem is to

$$\text{Minimise} \qquad r\sum_{i=1}^{N} x_i^2\sum_{s\ni i} b^2(s, i)p(s)+\sum_{s\in S} p(s)\left[\sum_{i\in s} b(s, i)x_i\right]^2 \qquad \ldots (2.8)$$

$$\text{subject to} \qquad \sum_{s\ni i} b(s, i)\, p(s) = 1/N \;\forall\; i = 1, 2, \ldots, N$$

We again employ the Lagrangian multipliers' technique to solve the problem (2.8). Let

$$Q^* = r\sum_{i=1}^{N} x_i^2\sum_{s\ni i} b^2(s, i)p(s)+\sum_{s\in S} p(s)\left[\sum_{i\in s} b(s, i)x_i\right]^2$$

$$-2\sum_{i=1}^{N}\alpha_i\left[\sum_{s\ni i} b(s, i)\, p(s) - \frac{1}{N}\right]$$

where α_i, $1 \leqslant i \leqslant N$, are the Lagrangian multipliers.

Solving (2.8) is equivalent to minimizing Q^* unconditionally. Differentiating and simplifying we get

$$b(s, i) = \frac{1}{rx_i^\theta}\left[\alpha_i - x_i e(s) \sum_{j \in s} \alpha_j x_j^{-\theta}\right] \qquad \ldots (2.9)$$

where $r + \sum_{i \in s} x_i^{2-\theta} = \frac{1}{e(s)}$ and $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_N)$ is a solution to the following system of equations

$$B\alpha = h \qquad \ldots (2.10)$$

where

$$b_{ii} = \pi_i - x_i^{2-\theta} \sum_{s \ni i} p(s)e(s) \qquad i = 1, 2, \ldots, N$$

$$b_{ij} = -x_i x_j^{-1\theta} \sum_{s \ni i,j} p(s)e(s) \qquad i \neq j = 1, 2, \ldots, N$$

and $\qquad \ldots (2.11)$

$$Nh_i = rx_i^\theta \qquad i = 1, 2, \ldots, N$$

$$\pi_i = \sum_{s \ni i} p(s) \qquad i = 1, 2, \ldots, N$$

Note that the left hand side of the equation $\det B = 0$ can be thought of as a polynomial in r of degree k (say). Therefore $\det B = 0$ has atmost k solutions $r_1, r_2, \ldots, r_k$ (say). For all other values of r $\det B$ is nonzero and hence $B\alpha = h$ will have a unique solution. We therefore have the following theorem.

Theorem 2.2 : *If the ratio r in (1.1) is known then for a given design p there exists a unique best linear p-unbiased estimator for estimating the population mean $\bar{Y}$ for all but a finite number of values of r in the sense of minimum $M(p, t)$.*

Remark 2.3 : Theorem 2.2 is true for a more general variance function $\sigma^2 v(x)$, instead of $\sigma^2 x^\theta$, in the model (1.1).

Remark 2.4 : In the continuous set up (vide Padmawar (1984)) the existence of an analogus optimal estimator depends on the solvability of an integral equation whereas in the finite set up existence as well as uniqueness are guaranteed for all but a few values of r. However, the construction of such an optimal estimator depends on the tractability of a solution to the system of equations (2.10). We consider an example so as to get an idea of the above optimality result.

Example 2.3 : Considor a Midzuno-Son sampling design of fixed size n for which $p(s)$ is given by $\frac{w_s}{M_1 V}$ where $w_s = \sum_{i \in s} w_i$, $V = \sum_{i=1}^{N} w_i$ and $M_1 = \binom{N-1}{n-1}$ with $w_i = x_i^{2-\theta} + r/n$, $1 \leqslant i \leqslant N$. For this dosign the quantities in (2.11) are given by

$$b_{ii} = \pi_i - x_i^{2-\theta}/V \qquad i = 1, 2, \ldots, N$$

$$b_{ij} = -(n-1)x_i x_j^{1-\theta}/(N-1)V \qquad i \neq j = 1, 2, \ldots, N$$

where $\pi_i = (n-1)/(N-1) + (N-n)w_i/(N-1)V$, $1 \leqslant i \leqslant N$.

It is easy to see that $\alpha_i = ax_i + bx_i^{\theta}$, $1 \leqslant i \leqslant N$, is a solution to (2.10).

where $$a = \frac{(n-1)bX}{r(N-1)} \text{ and } b = \frac{N-1}{N} \Big/ \left(\frac{n-1}{r} + \frac{N-n}{nV}\right) \quad \ldots (2.12)$$

Note that $\sum_{j=1}^{N} b_{ij}\alpha_j = \pi_i\alpha_i - \alpha_i x_i^{2-\theta}/V - \frac{(n-1)x_i}{(N-1)V} \sum_{j \neq i}^{N} \alpha_j x_j^{1-\theta}$

$$= \left[\frac{n-1}{N-1} + \frac{(N-n)r}{n(N-1)V}\right](ax_i + bx_i^{\theta}) - \frac{(n-1)x_i}{(N-1)V}\left[a\sum_{j=1}^{N} x_j^{2-\theta} + b\sum_{j=1}^{N} x_j\right]$$

$$= \left[\frac{n-1}{N-1} + \frac{(N-n)r}{n(N-1)V}\right] bx_i^{\theta} + \frac{x_i}{(N-1)V}\left[r(N-1)a - (n-1)b\sum_{j=1}^{N} x_j\right]$$

$$= \frac{r}{N} x_i^{\theta}.$$

Therefore (2.10) is consistent and the optimal estimator is of the form

$$\frac{b}{r}\sum_{i \in s} y_i + \left(a - \frac{b}{r}\sum_{i \in s} x_i\right)\frac{\sum_{i \in s} x_i^{1-\theta} y_i}{r + \sum_{j \in s} x_j^{2-\theta}}$$

where a and b are given by (2.12).

Remark 2.5 : Note that for the sampling design in the above example the system (2.10) is consistent for all values of r. Therefore the optimal estimator exists for all values of r.

Remark 2.6 : Instead of assuming the ratio r to be known if we assume a prior $\rho(\sigma^2, \beta)$ for σ^2 and β jointly with known $E_\rho(\sigma^2)$ and $E_\rho(\beta^2)$, where E_ρ denotes expectation under the prior ρ, then $E_\rho M(p, t)$ is a natural choice for an optimality criterion. In this set up we can compare any two stratogies for estimating the population mean w.r.t. the optimality criterion $E_\rho M(p, t)$.

The strategy that consists of a design for which $\pi_i \propto x_i^{\rho/2}$, $1 \leq i \leq N$, and the corresponding Horvitz-Thompson estimator is relatively more and more efficient as the ratio $r_\rho = E_\rho(\sigma^2)/E_\rho(\beta^2)$ increases and for large values of r_ρ it is likely to be the best strategy in the class of all p-unbiased strategies of given average size in the sense of minimum $E_\rho M(p, t)$. For a linear p-unbiased strategy $E_\rho M(p, t)$ is given by

$$E_\rho(\sigma^2) \sum_{i=1}^{N} x_i^\rho \left\{ \sum_{s \ni i} a^2(s, i)\, p(s) - \frac{1}{N^2} \right\} + E_\rho(\beta^2)\, V(p, t)\big|_x$$

where $V(p, t)\big|_x$ denotes the variance of the strategy (p, t) at $x = (x_1, \ldots, x_N)$.

The aforesaid strategy minimises the first term in the class of all p-unbiased linear strategies of given average size n.

Acknowledgements. The author is thankful to the referee for his useful suggestions.

References

HARTLEY, H. C. and ROSS, A. (1954): Unbiased ratio estimates. *Nature*, 174, 270-271.

PADMAWAR, V. R. (1981): A note on the comparison of certain sampling strategies. *J. R. Statist. Soc.* B, 43, 321-326.

——— (1984): Two existence theorems in survey sampling of continuous populations. *Sankhyā*, B, 46, 217-227.

PATEL, N. C. and DHARMADHIKARI, S. W. (1977): On linear invariant unbiased estimators in survey sampling. *Sankhyā*, C, 39, 21-27.

——— (1978): Admissibility of Murthy's and Midzuno's estimators within the class of linear unbiased estimators of finite population totals. *Sankhyā*, C, 40, 21-28.

RAMACHANDRAN, G. (1982): Horvitz-Thompson estimator and generalised IIps designs. J.S.P.I., 7, 151-153.

Paper received: *May, 1985.*

Revised: *July, 1987.*