PARIMAL MUKHOPADHYAY (*) - KAIPILLIL VIJAYAN (**)

Estimating functions in response dependent sampling from finite populations

1. INTRODUCTION

Let $\wp$ be a finite population of labelled units, $\wp = \{1, \ldots, i, \ldots, N\}$. Associated with each i is a pair of real numbers (y_i, x_i), y_i being the value of a response variable y and x_i the value of a closely related auxiliary variable (covariate) 'x'. We assume that (y_i, x_i) is a realisation of a random vector (Y_i, X_i), the joint distribution of $\{(Y_1, X_1), \ldots, (Y_N, X_N)\}$ being given by

$$\xi_\theta = \xi(\mathbf{x}, \mathbf{y}; \theta) = \prod_{i=1}^{N} f_{1i}(y_i | x_i; \theta)\, f_{2i}(x_i) \tag{1.1}$$

where $z = (z_1, \ldots, z_n)$, f_{2i}, is the marginal density of x_i, f_{1i} the conditional density of y_i given x_i, the random vector (Y_i, X_i) being distributed independently of (Y_j, X_j) $(i \neq j = 1, \ldots, N)$. In the formulation (1.1), the conditional density of y_i given x_i involves the population parameter θ whereas the marginal density of x_i is independent of θ. We assume that $\theta \subset \Theta \subset R_1$. The family of densities $C = \{\xi_\theta : \theta \,\varepsilon\, \Theta\}$ is the superpopulation model. Our problem is to estimate θ.

A function $g(\mathbf{y}, \mathbf{x}; \theta)$ is said to be a estimating function $(E.F)$ for θ if an estimate of θ can be obtained by solving the equation

---

(*) Indian Statistical Institute, Calcutta.

(**) University of Western Australia.

The research was supported by a grant from the UWA.

$$g\,(\mathbf{y}, \mathbf{x}; \theta) = 0 \tag{1.2}$$

An $E.F.g$ may be said to be unbiased for θ if is an unbiased estimate of zero.

Following Godambe and Thompson (1986) an $E.F.g^*$ is optimal in the class G of all unbiased estimating functions iff

$$\lambda_{g^*}(\theta) \leq \lambda_g(\theta) \;\; \forall \;\; g \varepsilon G$$

where

$$\lambda_g(\theta) = \frac{E[g(\mathbf{y},\mathbf{x};\theta)]^2}{\left[E\dfrac{\partial g(\mathbf{y},\mathbf{x};\theta)}{\partial\theta}\right]^2}$$

The corresponding solution for θ will be an optimal estimate. If g^* is an optimal function for θ then $g^* + B$ is optimal for $\theta + B$ where B is independent of θ. We assume for the time being that x' s are given fixed quantities. Thus g will be an unbiased $E.F$ for θ if

$$\varepsilon\,[g\,(\mathbf{Y}, \mathbf{x}; \theta) \mid \mathbf{x}] = 0 \tag{1.3}$$

when $\varepsilon(.\mid z)$ denotes the conditional expectation of (.) given z. We shall denote

$$\varepsilon(Y_i \mid x_i; \theta) = \mu_{Y_i\mid x_i}(\theta)$$

$$\gamma(Y_i \mid x_i) = \sigma^2 \vartheta_i \tag{1.4}$$

$$\varepsilon(Y_i) = \mu_{Y_i}(\theta)$$

where $\gamma(.\mid z)$ denotes the conditional variance given z, $\in(.)$ denotes the unconditional expectation of (.) and ϑ_i are known constants, and σ^2 may not be known.

Clearly, the $E.F$

$$h = \sum_{i=1}^{n} \phi_i a_i(\theta) \tag{1.5}$$

$$\phi_i = y_i - \mu_{Y_i} \mid x_i(\theta),$$

where $a_i(\theta)$ are functions of θ, satisfies (1.3) and hence is an unbiased estimating function. We shall restrict ourselves to the class H of *E.F's* h of the form (1.5). Suppose now θ is required to be estimated on the basis of observations on units in a sample s selected from $\wp$ according to the sampling design (s.d.) p with probability $p(s)$, $s \varepsilon S = \{s\}$.

Following Kalbfleisch and Lawless (1988), we consider the situation where the response vector $\mathbf{y}$ is fully observed but the values of the covariate x are observed only for $i \varepsilon s$. Problems of this type arise in the study of reliability of industrial products. Suppose that N items are in field use and that associated with ith item is a time to failure y_i and value x_i of a regressor variable X_i. Suppose further that (y_i, x_i) $(i = 1, \ldots, N)$ arises a random sample from a distribution with joint pdf

$$f_{1i}(y \mid x; \theta) f_2(x)$$

where the conditional pdf of Y_i given x, $f_{1i}(y \mid x, \theta)$ is completely specified up to a parameter θ to be estimated and $f_2(x)$ is the pdf of X. Our main interest is in estimating θ and thus the conditional distribution of failure time given x. This is often done on the basis of failure-record data where the failure-time y_i is observed for all the units in the population but the x-values are observed only for those units whose failure-time $y_i \leq T$, an warranty period for this batch of items. Units are sampled iff $y_i \leq T$.

The *s.d.* in this case is response dependent so that

$$p(s) = p(s \mid \mathbf{y}) \quad \forall s \, \varepsilon \, S \tag{1.6}$$

Our problem, therefore, boils down to that of estimating the parameter θ of the distribution ξ, given the data $\chi_s = \{\mathbf{y}, s, x_i : i \, \varepsilon \, s, \; p(s \mid \mathbf{y})\}$. Since the s.d. does not depend on θ, the derived *E.F.* from g in (1.2) should be

$$g_1 = \int \ldots \int g(\mathbf{y}, \mathbf{x}; \theta) \prod_{i \varepsilon \bar{s}} f_{2i}(x_i) \, d x_i \tag{1.7}$$

where $\bar{s} = \wp - s$.

Considering h in (1.5), the corresponding derived *E.F.* h_1 is

$$h_1 = \sum_{i \varepsilon s} \phi_i a_i(\theta) + \sum_{i \varepsilon \bar{s}} \{(y_i - \mu_{Y_i}(\theta)\} a_i(\theta) \tag{1.8}$$

It is known from Godambe and Thompson (1986) that an optimal *E.F.* in the class H is of the form

$$\nu = \sum_{i=1}^{N} \phi_i \tau_i$$

when

$$\tau_i = \frac{a(\theta)\mu'_{Y_i|x_i}(\theta)}{\vartheta_i}, \tag{1.9}$$

$a(\theta)$ is some function of θ and $\mu'_{Y_i|x_i}(\theta) = \partial\mu_{Y_i|x_i}(\theta) / \partial\theta$.

Hence the derived optimal *E.F.* from ν is

$$\nu_1 = \sum_{s} \phi_i \tau_i + \sum_{\bar{s}} (Y_i - \mu_{Y_i})\psi_i \tag{1.10}$$

where

$$\psi_i = \frac{a(\theta)}{\upsilon_i} \frac{\partial \mu_{Y_i}(\theta)}{\partial\theta},$$

$\mu_{Y_i}(\theta)$ being the mean of the marginal distribution of Y_i.

To estimate θ we shall use the optimal *E.F.*'s ν and ν_1 and find an *E.F.* $e(\chi_s)$ based on χ_s which is optimal in a certain class in a certain sense for estimating ν and ν_1.

2. Optimal estimating functions

Following Godambe and Vijayan (1992) we define a class $F(\mathbf{y})$ of estimating functions $e\{(i, x_i) : i \,\varepsilon\, s, \mathbf{y}, \theta\}$ as follows. Let

$$F_1 = \{e : E(e) = \nu \text{ for all } \mathbf{x}, \theta \,\varepsilon\, \Theta\} \tag{2.1}$$

where E denotes expectation with respect to *s.d.p.* $(s \mid \mathbf{y})$.

Similarly, let

$$F_2 = \{e : \varepsilon_{\mathbf{y}}(e) = \nu_1 \; \forall s : p(s \mid \mathbf{y}) > 0, \; \theta \,\varepsilon\, \Theta\} \tag{2.2}$$

where $e_{\mathbf{y}} = \varepsilon(. \mid \mathbf{y})$ denotes the expectation with respect to the distribution ξ in (1.1) for a fixed value of $\mathbf{y}$. Since x_i's are observed only for $i \,\varepsilon\, s$ we shall, following Godambe and Vijayan (1992), first keep y's to vary.

Let

$$F(\mathbf{y}) = F_1(\mathbf{y}) \cap F_2(\mathbf{y}) \tag{2.3}$$

Hence any *E.F.* $e \,\varepsilon\, F$ is approximately unbiased both for ν and ν_1. An *E.F.* e^* is said to be conditionally optimal in F if $e^* \varepsilon\, F$ and if it simultaneous satisfies the following inequalities:

$$\begin{aligned}
&\{\varepsilon_y E(e^* - \nu)^2\} \,/\, \{\varepsilon_y E(\partial e^* / \partial\theta)\}^2 \\
&\leq \{\varepsilon_y E(e - \nu)^2\} \,/\, \{\varepsilon_y E(\partial e / \partial\theta)\}^2 \\
&\{\varepsilon_y E(e^* - \nu_1)^2\} \,/\, \{\varepsilon_y E(\partial e^* / \partial\theta)\}^2 \\
&\leq \{\varepsilon_y E(e - \nu_1)^2\} \,/\, \{\varepsilon_y E(\partial e / \partial\theta)\}^2
\end{aligned} \tag{2.4}$$

$\forall\, e \,\varepsilon\, F,\ \theta \,\varepsilon\, \Theta$. Since $E(\partial e / \partial\theta)$ is constant for all $e \,\varepsilon\, F$ and because of (2.1) and (2.2), the inequality in (2.4) reduces to

$$\varepsilon_y E(e^{*2}) \leq \varepsilon_y E(e^2) \ \forall\, e \,\varepsilon\, F,\ \theta \,\varepsilon\, \Theta \tag{2.5}$$

Clearly an *E.F.* which is conditionally (for fixed values of $\mathbf{y}$) optimal is also unconditionally (whatever be the values of $\mathbf{y}$) optimal.

We now prove

THEOREM 2.1

If the sampling design $p(s\,|\,\mathbf{y})$ is such that a sample s for which

$$\sum_{i \in s} \frac{(y_i - \mu_{Y_i})\psi_i}{\pi_i} = \sum_{i=1}^{N} (y_i - \mu_{Y_i})\psi_i \tag{2.6}$$

is selected with probability one (and with probability K^{-1} if there are K such samples) then the optimal estimating function (in the sense of (2.5)) is given by

$$e^*(\chi_s) = \sum_{i \in s} \frac{\phi_i \tau_i}{\pi_i} \tag{2.7}$$

where $\pi_i = \sum_{s \ni i} p(s\,|\,\mathbf{y})$, the first order inclusion-probability of the sampling design.

Proof. Obviously $e^* \in F_1$. We shall now show that e^* satisfies (2.5) and $e^* \in F_2$.

Now

$$\nu = \sum_{i=1}^{N} (y_i - \mu_{Y_i})\psi_i$$

$$+ \sum_{i=1}^{N} \left[\left(\mu_{Y_i}\psi_i - \mu_{Y_i|x_i}\tau_i \right) + (\tau_i - \psi_i) y_i \right]$$

$$= \sum_{i=1}^{N} u_i + \sum_{i=1}^{N} \upsilon_i \text{ (say)}$$

Following Godambe and Thompson (1986), for a fixed $\mathbf{y}$, the optimal *E.F.* for $\sum_{i=1}^{N} \upsilon_i$ is given by $\sum_{i \in s} \upsilon_i / \pi_i$. Hence by the invariance property of the optimal *E.F.*'s

$$\sum_{i \in s} \frac{\upsilon_i}{\pi_i} + \sum_{i=1}^{N} \upsilon_i$$

is the optimal estimating function for ν. Now, by (2.6) for samples s for which $p(s|\mathbf{y}) > 0$,

$$\sum_{i \in s} \frac{\upsilon_i}{\pi_i} + \sum_{i=1}^{N} \upsilon_i = \sum_{i \in s} \frac{\upsilon_i + u_i}{\pi_i}$$

$$= e^*.$$

To show that $e^* \in F_2$, we note that

$$\nu_1 = \sum_{i \in s} \upsilon_i + \sum_{i=1}^{N} u_i$$

Hence, from (2.6), for all samples for which $p(s|\mathbf{y}) > 0$;

$$\nu_1 - e^* = \sum_{i \in s} \left(1 - \frac{1}{\pi_i} \right) \upsilon_i,$$

so that

$$\varepsilon_y(\nu_1 - e^*)$$

$$= \sum_{i \in s} \left(1 - \frac{1}{\pi}\right) \varepsilon_y(\upsilon_i)$$

$$= 0 .$$

Hence the theorem.

ACKNOWLEDGEMENT

The authors are thankful to the referee for many suggestions which substantially improved both presentation and mathematical details of the paper.

REFERENCES

KALBFLEISH, J.D. and LAWLESS, J. (1988) Estimation of reliability in field performance studies, *Technometrics*, 30, 365-388.

GODAMBE, V.P. and THOMPSON, M.E. (1986) Parameters of a superpopulation and survey population: their relationships and estimation, *Int. Statist. Rev.*, 54, 127-128.

GODAMBE, V.P. and VIJAYAN, K. (1992) *Optimal estimation for response dependent retrospective sampling*, Univ. of Western Australia, Tech. Rep. No. 18/1992.

Estimating functions in response dependent sampling from finite populations

SUMMARY

Godambe and Vijayan (1992) considered the problem of estimating a population parameter θ involved in the joint distribution of a response variate y and a covariate x, using the likelihood as estimating functions in sampling from a finite population where the sampling design depends on y. In this note we confine ourselves to a class of estimating functions and find an optimal function in the class.

Funzioni di stima nel caso di campionamento da popolazione finita con variabile ausiliaria

RIASSUNTO

Godambe a Vijayan (1992) hanno considerato il problema di stimare il parametro θ della distribuzione congiunta di una variabile risposta y e una covariata x, usando le verosimiglianze come funzioni di stima nel campionamento da una popolazione finita, quando il disegno campionario dipende da y. In questa nota gli autori considerano il caso di una classe particolare di funzioni di stima nella quale trovano una funzione ottimale.