ESTIMATING FUNCTIONS : AN OVERVIEW

By

V. P. Godambe
Department of Statistics
University of Waterloo
Waterloo (Ont) N2L 3G1
Canada

B.K.Kale
Department of Statistics
University of Poona
Pune - 411007 India

1. HISTORICAL BACKGROUND

For nearly two centuries now, the two most prevalent methods of estimation are the method of 'least squares (LS) and the method of 'maximum likelihood' (ML), using the present day terminology. Legendre put forward the LS method in 1805, on more or less intuitive grounds. He also coined the term 'LS method'. Gauss (1809, 1823) provided two statistical justifications to the LS method. (I) He related it to the ML estimation and the normal model by showing that within the location family of distributions, the two methods LS and ML coincide uniquely for the normal family. This indicated that the acceptance of both LS and ML methods was in line with the acceptance of the normal model. (II) He showed that with conditions only on the first two moments of the distribution but otherwise irrespective of its distributional form, the LS estimate has minimum variance within the class of 'linear unbiased estimates'. This is the wellknown Gauss-Markov (GM) theorem. For our subsequent discussion, it is important to note that Gauss did not invoke the present day concept of ' unbiased estimates' . Instead he used the concept of ' consistent estimates ', that is, ' estimates which equal the true vlaue when the observations are without error ' (Bertrand, 1889, Sprott (1983). This ' consistency ' which is a ' finite sample ' property should be distinguished from the 'asymptotic

consistency ' often mentioned in the literature. It is of course true that for linear models, the linear estimates which are consistent in the Gauss sense are also unbiased estimates and conversely. (Heyde and Seneta, 1977, Chapter 4).

As seen above, Gauss provided justifications for the LS method by relating it to the ML and unbiased minimum variance (UMV) methods of estimation. These 'relationships' among the three methods LS, ML and UMV, when exploited fully, as discussed below, provides a 'theory' of estimation which combines the strengths of the three methods and at the same time eliminates their weakness. This theory, called the theory of Estimating Functions will be reviewed in the subsequent sections.

2. ESTIMATING FUNCTIONS AND GAUSS-MARKOV (GM) THEOREM

To emphasize the basic concepts, we consider the simplest case of the LS estimation and the GM theorem, the extensions to the general linear models being straightforward. Let $y_1, y_2, \ldots, y_n$ be independent real random variates with common expectation and variance given by

$$E(y_i) = \theta \,, \quad \mathrm{Var}(y_i) = \sigma^2 \,, \quad i = 1, 2, \ldots, n \,. \qquad (1)$$

The unknown parameter θ is known to belong to a specified real interval. To estimate θ, on the basis of observations $y_1, y_2, \ldots, y_n$ by the LS method we minimize $\Sigma_1^n (y_i - \theta)^2$, for the variations of θ. The minimum is attained at the LS estimate of θ, namely the sample mean $\bar{y} = (\Sigma_1^n y_i)/n$. Again as a very special case we have here

GM THEOREM. Let $(y_1, y_2, \ldots, y_n)$ be independent real

random variates satisfying (1), and ℓ be any linear unbiased estimate of θ. That is,

$$\ell = \Sigma\ a_i y_i \qquad (2)$$

where a_i , $i = 1, \ldots, n$ are any ' constants ' such that the expectation $E(\ell) = \theta$, implying

$$\Sigma\ a_i = 1 \quad . \qquad (3)$$

Then the variance of ℓ is minimized for $a_i = 1/n$, $i = 1,\ldots,n$. That is the LS estimate=sample mean $= \bar{y}$ is the linear UMV estimate for θ.

Now instead of restricting our attention to the functions of observations only, that is estimates, we also consider real functions g of observations and the parameter θ, $g = g(y_1, y_2, \ldots, y_n, \theta)$. Such functions g are called <u>estimating functions</u> in view of their central role in estimation, which will be evident from what follows. Consider an estimating function g of the form

$$g = \Sigma\ (y_i - \theta)\ b_i \qquad (4)$$

where b_i, $i = 1, \ldots, n$ are any constants. It is easy to see that any linear unbiased estimate in (2) can be obtained as a solution in θ of the equation

$$g = 0 \qquad (5)$$

for suitably chosen constants b_i, $i = 1, \ldots, n$ in (4). Noting that because of (1) in (4) the expectation

$$E(g) = 0 \ , \qquad (6)$$

we have the following alternative version of the GM theorem.

GM THEOREM (A). If $(y_1, y_2, \ldots, y_n)$ are independent real random variates satisfying condition (1), the variance of the estimating function g in (4), is minimized for the variations of b_i, $i = 1, \ldots, n$ subject to the condition

$$\Sigma\ b_i = C \quad , \text{a constant} \tag{7}$$

for $b_i = C/n$, $i = 1, \ldots, n$. With this substitution, $b_i = C/n$ in (4), the equation (5) provides the sample mean $\bar{y}$ as the estimate of θ.

In the above theorem we first consider shifting the emphasis from the criteria of UMVness within linear unbiased estimates to an optimality criteria applicable to estimating function $g = \Sigma\, b_i(y_i - \theta)$. In order that $g = 0$ defines an estimator $\hat{\theta}_b = \Sigma\, b_i\, y_i / \Sigma\, b_i$, we must have $\Sigma\, b_i \neq 0$. Thus we consider the class G_o of estimating functions given by

$$g(y, \theta) = \Sigma\, b_i(y_i - \theta) \tag{8}$$

for any constants b_i satisfying $\Sigma\, b_i \neq 0$. Note that g is Gauss consistent in the sense that if all $y_i = \theta$ then $g = 0$. Further we call an estimating function g, <u>unbiased</u> if it satisfies $E(g) = 0$. Thus $g \in G_o$ is Gauss consistent as well as unbiased and $\mathrm{Var}(g) = \sigma^2 \Sigma\, b_i^2$. The equations $g = 0$ and $kg = g' = 0$ where $k \neq 0$ are equivalent and define the same estimate $\hat{\theta}_b$ which is unbiased and has variance $\sigma^2 \Sigma\, b_i^2 / (\Sigma\, b_i)^2$. But $\mathrm{Var}(g') = k^2\, \mathrm{Var}(g)$ can however be made arbitrarily small and thus the comparison of two estimating functions on the basis of variance alone is not meaningful unless some standardization is introduced. The standardized version of g is defined by

$$g_s = \Sigma b_i(y_i - \theta) / \{-\Sigma b_i\} \qquad (9)$$

Note that $g = 0$ and $g_s = 0$ determine the same estimate $\hat{\theta}_b$ and $Var(g_s) = \sigma^2 \Sigma b_i^2/(\Sigma b_i)^2 = Var(g_s')$. An estimating function $g^* \in G_o$ is said to be optimal in G_o if

$$Var(g_s^*) \leq Var(g_s) \text{ for any } g \in G_o \qquad (10)$$

We observe that $Var(g_s) = Var(\hat{\theta}_b)$ and the optimal estimating function g^* is unique up to a constant multiple.

Motivation behind the standardization (9) and the optimality criteria in (10) is provided by the following argument. In order to be used as an estimating equation the estimating function g should be as near to zero as possible when θ is the true value. This requires that Var(g) be made as small as possible. Further $g(y, \theta + \delta\theta)$ should be as far away from zero as possible when θ is the true value "$\delta\theta$" being any departure. This requires

$$[E\ (\frac{\partial g}{\partial \theta})]^2 = (\Sigma\ b_i)^2$$

be as far away from zero as possible. Both of these requirements can be combined in to one requiring that $E[g/E\ (\frac{\partial g}{\partial \theta})]^2 = Var(g_s)$ be made as small as possible. In the above setup this is equivalent to minimizing $Var(\hat{\theta}_b)$ as every linear unbiased estimate can be written as a solution of $\Sigma b_i(y_i - \theta) = 0$ and conversely. GM Theorem can therefore be reformulated as

GM THEOREM (A) : If $(y_1, y_2, \ldots, y_n)$ are independent random variates with $E(y_i) = \theta$ and $Var(y_i) = \sigma^2$ then $g^* = \Sigma(y_i - \theta)$ is an optimum estimating function in G_o. The equation $g^* = 0$ provides the sample mean $\bar{y}$ as an estimate of θ.

We emphasize that the above GM theorem (A) is logically equivalent to the GM theorem and as such it lends as much justification to the LS estimate $\bar{y}$ as the GM theorem does. However, unlike the GM theorem its alternative version GM theorem (A) admits the following extension.

3. AN EXTENSION OF GAUSS - MARKOV THEOREM

As said before, we can get all the linear unbiased estimates of the form (2) by sloving for θ, the equations (5), for different estimating functions g in G_o. Conversely every member of G_o corresponds to some unbiased linear estimate in (2). But it is easy to see that, unlike GM Theorem, the GM theorem (A) remains valid if 'b_i' in (4) are allowed to be functions of the parameter θ, $b_i = b_i(\theta)$, $i = 1,2,\ldots,n$. However in this case the solutions in θ, of the equations (5) provide a much <u>wider</u> class of estimates than that of linear unbiased estimates implied by G_o. In this class of estimating functions every implied estimate need not be unbiased but it will necesarily satisfy the property of Gauss consistency discussed in Sec. 1. i.e. when all y_i are observed without any error, that is $y_i = \theta$, $i = 1,\ldots,n$, $g = 0$. Thus the <u>extension</u> of the GM theorem (A), namely GM theorem (B) , given subsequently, where in coefficients 'b_i' are allowed to depend on θ provide a greater justification, for the LS estimate $\bar{y}$, than what GM theorem does.

Thus we now consider an extended class of estimating functions $G_1 = \{g\}$ where $g(y, \theta) = \Sigma\, b_i(\theta)\,(y_i - \theta)$ and where $b_i(\theta)$ are differentiable functions of θ. Note that $g(y, \theta)$ is Gauss consistent with $\mathrm{Var}(g) = \sigma^2\, \Sigma\, b_i^2(\theta)$.

Further $E\left(\frac{\partial g}{\partial \theta}\right) = -\Sigma b_i(\theta)$ and the standardized version of g is given by

$$g_s = \Sigma b_i(\theta)\,(y_i - \theta)/\{-\Sigma b_i(\theta)\} \qquad (11)$$

The only difference between (9) and (11) is that b_i's are now allowed to depend on θ. We observe that $Var(g_s) = \sigma^2 \Sigma b_i^2(\theta)/[\Sigma b_i(\theta)]^2$ and is minimized for every $\theta \in \Omega$ when $b_1(\theta) = \ldots = b_n(\theta) = k(\theta) \neq 0$. Thus $k(\theta)\,\Sigma(y_i - \theta)$ is optimal estimating function in the extended class G_1 satisfying the optimality criteria (10) with G_o replaced by G_1. Hence $(\bar{y} - \theta)$ can be taken as optimal estimating function upto a multiplicative constant depending only on θ. We thus have a more general version of GM Theorem (A) given by

GM THEOREM (B) : If $(y_1, y_2, \ldots, y_n)$ are independent random variates with $E(y_i) = \theta$ and $Var(y_i) = \sigma^2$ then $g^* = \Sigma(y_i - \theta)$ is an optimal estimating function in the class G_1 and the equation $g^* = 0$ provides $\bar{y}$ as an estimate of θ.

As said before the GM theorem (B) provides a greater justification to the LS estimate $\bar{y}$ than what the classical Gauss Markov theorem does. This added versatality is clearly the result of shifting emphasis from estimates to estimating functions. It is pointed out that not all estimates covered by the GM Theorem (B) need be unbiased yet the corresponding estimating function is unbiased. We can thus in general define any function $g(y_1, y_2, \ldots, y_n, \theta)$ such that $E(g) = 0$ as an unbiased estimating function. The argument for standardization of g can be generalized to

define the standardized version of g as

$$g_s = g/E\left(\frac{\partial g}{\partial \theta}\right) \qquad (12)$$

The optimality criteria based on minimization of Var (g_s) leads us to define g^* to be optimal in the class G if $g^* \in G$ and

$$\text{Var}(g_s^*) \leq \text{Var}(g_s) \quad \text{for any} \quad g \in G \ . \qquad (13)$$

In the next section we show how the flexibility provided by the standardization and the above optimality criteria allows us to obtain optimal estimating function in a situation in which the classical Gauss-Markov approach does not correspond to LS approach.

4. A FAILURE OF GAUSS-MARKOV (GM) APPROACH

Now we discuss a situation where the GM theorem fails completely to relate itself to the LS estimate, yet the extended GM theorem (B) of Section 3, shows a way out. Let $(y_1, y_2, \ldots, y_n)$ be independent real random variates.

$$E(y_i) = \alpha_i(\theta) \ , \quad \text{Var}(y_i) = \sigma^2 \ , \quad i = 1, 2, \ldots, n \qquad (14)$$

where α_i are some differentiable functions of θ with unique inverse functions α_i^{-1}. The LS estimate is obtained by minimizing $\Sigma (y_i - \alpha_i(\theta))^2$ which is as intuitive in this case as it is when $\alpha_i(\theta) = \theta$. Yet it is justified by the GM theorem only for those functions $\alpha_i(\theta)$ which are linear in θ. The failure of GM approach is mainly due to the fact that even though y_i is unbiased for $\alpha_i(\theta)$, $\alpha_i^{-1}(y_i)$ (except for a linear function) is not unbiased for θ. However, if y_i is

Gauss consistent for $\alpha_i(\theta)$ then $\alpha_i^{-1}(y_i)$ is Gauss consistent for θ. The LS approach gives immediately the estimating equation

$$\Sigma(y_i - \alpha_i(\theta)) \frac{\partial \alpha_i}{\partial \theta} = 0 \qquad (15)$$

Then following the same argument as used in deriving GM Theorem (B) at the end of section 3, we consider the class G_2 of estimating functions $\tilde{g}$ of the type

$$\tilde{g} = \Sigma\ b_i(\theta)\ (y_i - \alpha_i(\theta))\ . \qquad (16)$$

Using (12) we have,

$$\tilde{g}_s = \Sigma\ b_i(\theta)\ (y_i - \alpha_i(\theta))/(\ -\ \Sigma\ b_i \frac{\partial \alpha_i}{\partial \theta}\ )\ . \qquad (17)$$

Now $Var(\tilde{g}_s) = \sigma^2\ \Sigma\ b_i^2/(\Sigma\ b_i \frac{\partial \alpha_i}{\partial \theta})^2$ is minimized for $b_i(\theta) = k(\theta) \frac{\partial \alpha_i}{\partial \theta}$ showing there by that an optimal estimating function within G_2 is given by $\Sigma(y_i - \alpha_i(\theta)) \frac{\partial \alpha_i}{\partial \theta}$ leading to LS estmating equation, in (15).

It is important to note that in case when $(y_1, y_2, \ldots, y_n)$ are normally distributed all the three methods LS, ML and optimal estimating function lead to the same estimating equation. On the other hand the approach based on UMVness of the estimate of θ fails in case where $\alpha_i(\theta)$ are non-linear even if we assume the normality of $(y_1, y_2, .., y_n)$ Yet by transferring the UMVness of an estimate to that of standardized estimating function we could establish the optimality of LS estimating equation. However, the estimate θ obtained from (15) would be biased for θ. But this, if at all, is a small price to pay as in many instances unbiased

estimates of θ, even under normality, would not exist. This brings out the flexibility of the approach to estimation by estimating functions. In the next section we show how this approach suggests a modification to the LS approach in some situations where the classical LS approach fails.

5. A FAILURE OF THE LEAST SQUARE (LS) APPROACH

In all the previous discussion we have assumed tacitly that $\mathrm{Var}(y_i) = \sigma^2$ is independent of θ. Now consider $(y_1, y_2, \ldots, y_n)$ independent random variates such that

$$E(y_i) = \alpha_i(\theta) \quad \text{and} \quad \mathrm{Var}(y_i) = c\,\sigma_i^2(\theta) \tag{18}$$

where $\sigma_i^2(\theta)$ are specified differentiable functions of θ and c is an unknown positive constant not depending on θ. To estimate θ, the LS approach suggests minimizing $\Sigma(y_i - \alpha_i(\theta))^2/\sigma_i^2(\theta)$ leading to the LS estimating equation

$$\tilde{g}_1^* + B = 0 \tag{19}$$

where $\tilde{g}_1^* = \Sigma(y_i - \alpha_i(\theta))\,\frac{\partial \alpha_i}{\partial \theta}\,/\sigma_i^2$ and

$B = \Sigma(y_i - \alpha_i(\theta))^2\,\frac{\partial \sigma_i}{\partial \theta}\,/\sigma_i^3(\theta)$. We first note that whereas $E(\tilde{g}_i^*) = 0$, $E(B) = \Sigma\,\frac{\partial \sigma_i}{\partial \theta}/\sigma_i = \frac{\Sigma\,\partial \log\,\sigma_i}{\partial \theta}$ is in general non-zero and thus $\tilde{g}_1^* + B$ is not an unbiased estimating function. Even for large samples although $\frac{1}{n}\,\tilde{g}_1^*$ will converge in probability to zero this may not be the case with B/n and in fact B/n could diverge to $\pm\,\infty$ depending on the nature of $\frac{1}{n}\,\Sigma\,\frac{\partial\,\log\,\sigma_i}{\partial \theta}$. As such for large n, the

solution of the equation $\tilde{g}_1^* = 0$, under some regularity conditions, will converge in probability to the true value whereas the solution of the LS estimating equation may not. In fact, it could converge to a value far away from the true value. Many a time it is suggested that we take $\tilde{g}_1^* = 0$ as an estimating equation, a sort of modified LS estimating equation. We now show that $\tilde{g}_1^*$ is in fact an optimal estimating function in the class G_2 defined by (16). let

$$\tilde{g}_1 = \Sigma \ (y_i - \alpha_i(\theta)) \ b_i(\theta)$$

Now using the standardization (12) we have,

$\text{Var}(\tilde{g}_{1s}) = c \ \Sigma \ b_i^2(\theta) \ \sigma_i^2(\theta) / \{\Sigma \ b_i \frac{\partial \alpha_i}{\partial \theta}\}^2$. Further $\text{Var}(\tilde{g}_{1s})$ is minimized for $b_i(\theta) = k(\theta) \ . \ \frac{\partial \alpha_i}{\partial \theta} \ / \ \sigma_i^2(\theta)$, $i = 1,2,\ldots, n$. This leads to the optimum estimating function in G_2 given by

$$\tilde{g}_1^* = \Sigma (y_i - \alpha_i(\theta)) \ \frac{\partial \alpha_i}{\partial \theta} \ / \ \sigma_i^2(\theta) \ . \qquad (20)$$

Hence the modified LS estimating equation $\tilde{g}_1^* = 0$ corresponds to the optimum estimating function in G_2.

It is interesting to compare the optimum estimating equation $\tilde{g}_1^* = 0$, with the ML equation for a specified value $c = c_o$ say, in (18) when $(y_1, y_2, \ldots, y_n)$ are assumed to be normally distributed. The likelihood equation then is the same as LS equation corrected for bias and is given by

$$g = \tilde{g}_1^*/c_o \ + \{ B/c_o - \Sigma \frac{\partial \log \sigma_i}{\partial \theta} \} = 0 \qquad (21)$$

We note that $E(g) = 0$ only when $c = c_o$ in (18). Otherwise

$$E(g) = \{ \frac{c}{c_o} - 1 \} \Sigma \frac{\partial \log \sigma_i}{\partial \theta} \quad .$$

Thus if the true value of c in (18) is such that c/c_o is far away from one then the likelihood equation is again biased and leads to the same problems as in the case of the LS estimating equation even for large n. On the other hand the optimum estimating equation $\tilde{g}_1^* = 0$ is unaffected by the value of c. Of course if c is not specified, the ML method is undefined in principle even if we assume normality and it is also in general undefined for the model specifying only the first two moments of random variables. (See Section 8). Yet for the parametric sub-model obtained from (18) by assuming normality and a specified value of c, the ML equation is reasonably well approximated by the optimal estimating equation $\tilde{g}_1^* = 0$. This is in line with the connection, established by Gauss, between the LS and ML estimating equations. Further the above optimality property of g_1^* is mathematically and statistically analogous to the optimality property of the score function $\frac{\partial \log p}{\partial \theta}$ first established by Godambe (1960) in case of a (full) parametric model given by $\{ p(x, \theta), \theta \in \Omega \subseteq R_1 \}$.

In the following section we will consider the estimating functions and equations in the context of parametric models and indicate how this approach provides a logical frame for estimation of parameter(s) of interest in the presence of nuisance parameter(s). This also shows that the estimating functions provide a common connecting link for estimation in parametric and semi-parametric models. Our restriction to the classes G_o, G_1 and G_2 of estimating functions in the context of LS method (which assumes only first two moments) is justified by the nature of the corresponding semi-parametric models. (Godambe & Thompson, 1985, 1989,

Godambe & Heyde, 1987). The classes G_o, G_1, G_2 are of the type $\Sigma\, a_i(\theta)y_i + b_i(\theta)$, i.e. linear in $(y_1, y_2, \ldots, y_n)$ but possibly non-linear in θ. Durbin (1960) considered linear optimum estimating functions with many applications to time series. In the parametric models we will define a more general class G of estimating functions and search for an optimal estimating function in G.

6. ESTIMATING FUNCTIONS - OPTIMALITY

Let x be an abstract random variable on a sample space $(\mathcal{X}, \mathcal{B})$ with a probability density function $p(x, \theta)$, w.r.t. a σ-finite measure μ on $(\mathcal{X}, \mathcal{B})$ and where θ is a real or vector valued parameter which is assumed to be a labelling or indexing parameter. Thus if we know θ, then the probability distribution of x becomes completely known and hence our object is to estimate θ on the basis of the observed value of x. Conventionally the problem of estimation is tackled by proposing an estimator T(x) and then studying the properties of the estimator T(x) which depend on the sampling distribution of x. These optimal properties include sufficiency, unbiasedness, minimum variance or minimum mean squared error etc. The estimator T(x) is obtained by some standard methods such as the least squares, maximum likelihood, minimum chi-square, method of moments among others. Most often these methods though intutive are adhoc and do not directly follow from the optimality properties demanded or expected from the "best" estimator, notable exception being the construction of minimum variance unbiased estimator following Rao-Blackwell, Lehmann-Scheffe approach.

A common feature of the methods such as least squares, maximum likelihood, moments, minimum chi-square is that these methods lead to an estimating equation $g(x, \theta) = 0$, in case θ is real or a set of estimating equations $g_i(x, \theta) = 0$

$i = 1, 2, \ldots, m$ in case of vector valued parameter $\theta = (\theta_1, \theta_2, \ldots, \theta_m)'$. Indeed the phrase " equation for estimation " occurs as early as Fisher (1935).

Following the ideas of the earlier sections we now consider an estimating function $g(x, \theta)$ defined on $\mathcal{X} \times \Omega$ rather than a statistic $T(x)$. We will first consider θ real valued and we assume that $g(x, \theta)$, for each $\theta \in \Omega$ is such that $E_\theta(g) = 0$ and $Var_\theta(g)$ is finite. The idea of using an estimating function, a function of observations as well as the parameter has been around for quite some time. Pivotal quantities or 'pivots' used by Fisher (1935) are prime examplesof these. The distribution of a pivot does not depend on θ and this property is exploited for making inference on θ. On the other hand we require that only the first moment of g be independent of θ which we can take to be zero without loss of generality. Our assumption that $Var_\theta(g)$ is finite can be rephrased by requiring that $Var_\theta(g') = 1$ by redefining $g' = g/\sqrt{var(g)}$, equivalently assuming that the first two moments of g are independent of θ. However, we will not pursue this line of thought considered by Barnard (1973) but instead follow Godambe's (1960) optimality criteria informally discussed and illustrated in previous sections.

Since the estimating function g is initially used to obtain an estimator by solving the equation $g = 0$, the unbiasedness condition

$$E_\theta(g) = 0 \ , \quad \forall \quad \theta \in \Omega \tag{22}$$

becomes a natural one and is not very restrictive since as mentioned above we are any way assuming the existence of first two moments of g. In order that the estimating

DR. B K. KALE
PROFESSO STATISTICS
DEPT. OF STATISTICS
UNIV. OF POONA
PUNE-411 007

equation $g(x, \theta) = 0$ should determine an estimator $\hat{\theta}_g$ as a function of x, it is necessary that the implicit function theorem holds. For this a sufficient condition is $\frac{\partial g}{\partial \theta} \neq 0$. As $g(x, \theta)$, for each θ fixed, is a random variable, we must have for each θ, $P_\theta\{ \frac{\partial g}{\partial \theta} \neq 0 \} = 1$. A weaker assumption is that

$$E_\theta\left(\frac{\partial g}{\partial \theta} \right) \neq 0 \qquad \forall\ \theta \in \Omega \qquad (23)$$

Let G be the class of all estimating functions which satisfy (22), (23) and such that $Var(g) = E(g^2)$ is finite. Godambe (1960) defined $g^* \in G$ as <u>optimal</u> estimating function if for any $g \in G$

$$\frac{E\{(g^*)^2\}}{\{E(\frac{\partial g^*}{\partial \theta})\}^2} \leq \frac{E\{(g^2)\}}{\{E(\frac{\partial g}{\partial \theta})\}^2} \qquad \forall\ \theta \in \Omega \qquad (24)$$

Again the motivation behind this optimality criteria is similar to the one discussed earlier: We require that $Var_\theta(g) = E_\theta(g^2)$ is as small as possible and $E_\theta\{(g(x, \theta + \delta\theta))^2\}$ is as large as possible, a kind of measure of sensitivity of the estimating function for the departure from the true value θ. Both these objectives are achieved if we minimize $E_\theta(g^2)/\{E(\frac{\partial g}{\partial \theta})\}^2$ for all $g \in G$ uniformly in $\theta \in \Omega$. In terms of the standardization given by (12) the criterion (24) is equivalent to

$$Var(g_s^*) \leq Var(g_s) \ , \quad \forall\ \theta \in \Omega \ . \qquad (25)$$

The interpretation of the standardization (12) in terms of the variances of the estimating functions g and 'constant'xg,

discussed earlier is due to Barnard. Godambe's (1960) main result was that under mild regularity conditionson the class of estimating functions G and the class of density funcitions $\{p(x, \theta), \theta \in \Omega\}$ the score function $\frac{\partial \log p}{\partial \theta}$ is optimal. In this sense the likelihood equation is an optimal estimating equation. It is easy to demonstrate that optimal estimating function is essentially unique in the sense that if g_1^* and g_2^* are both optimal then $g_{1s}^* = g_{2s}^*$.

The optimality criterion of Godambe given above and the standardization (12) it introduces can be viewed in many different ways. Kale (1962a) following the general theory of estimating functions as developed by Kimball (1946) and Wilks (1948) derived an extension of Cramer-Rao inequality for estimating functions. Under mild regularity conditions, Kale (1962a) proved that for any $g \in G$

$$\mathrm{Var}_\theta(g) \geq \{E(\frac{\partial g}{\partial \theta})\}^2/I(\theta) \qquad (26)$$

where $I(\theta)$ is the Fisher information. He also pointed out that $\frac{\partial \log p}{\partial \theta}$, attains the extended C-R lower bound to Var(g) in inequality (26), for $g \in G$ and is therefore optimal. The extended C-R inequality given by (26) can be written as

$$\mathrm{Var}(g_s) \geq \frac{1}{I(\theta)} \qquad (27)$$

where $g_s = g/E(\frac{\partial g}{\partial \theta})$ the standardized version of g. The advantage of the standardized version (27) over (26), is that the lower bound $1/I(\theta)$ is independent of $g \in G$.

Another justification of Godambe's optimality criteria

(24) is provided by the fact that under mild regularity conditions the estimator $\hat{\theta}_g^*$ provided by solving the optimal estimating equation $g^*(x, \theta) = 0$ minimizes, atleast asymptotically, the mean square error $E(\hat{\theta}g - \theta)^2$ where $\hat{\theta}g$ is the estimator provided by $g(x, \theta) = 0$ for $g \in G$. This follows from the argument that g_s and $\hat{\theta}_g - \theta$ are stochastically equivalent as $n \to \infty$ and minimizing $Var(g_s)$ is same as minimizing mean squared error of $\hat{\theta}_g$. For details we refer to Kale (1985) as well as Small & McLeish (1988). A further justification in terms of asymptotically shortest confidence intervals - in a very general setting of stochastic processes - is due to Godambe & Heyde (1989). An early work in this direction is due to Wilks (1939).

A different kind of justification of Godambe's optimality criteria and the standardization of the estimating function is provided by its connection with the Newton-Raphson process for solving the corresponding estimating equation $g(x, \theta) = 0$. In many cases $\hat{\theta}g$, the estimator defined by the solution of the estimating equation, can not be obtained explicitly and has to be determined by an iterative procedure. A commonly employed procedure is the Newton- Raphson procedure given by

$$\theta_{r+1} = \theta_r - \{g(x, \theta)/ \frac{\partial g}{\partial \theta} \}_{\theta=\theta_r} \qquad (28)$$

with a trial value θ_1 as any consistent estimator of θ. A modification of the Newton-Raphson process suggested by Fisher (1925) in the context of the likelihood equation, leads to the well known method of scoring for parameters. Thus the Fisher-Newton-Raphson iterative procedure is given by

$$\theta_{r+1} = \theta_r - \{ g(x, \theta)/E(\frac{\partial g}{\partial \theta}) \}_{\theta=\theta_r} \qquad (29)$$

Thus the correction term to the successive iterates is $g_s(x, \theta)$ the standardized version of g.. Minimizing $Var(g_s)$ thus corresponds to choosing the estimating function g with smallest correction term at least on average. The optimality of the likelihood equation now translates into the fact that the method of scoring for parameters converges very fast even though it is only a first order process, a phenomenon observed by Kale (1962b).

7. MULTIPARAMETRIC CASE

Durbin (1960) considered the estimating functions for the vector valued parameter. When the indexing parameter θ is vector valued say $(\theta_1, \theta_2, \ldots, \theta_m)' \in \Omega_m \subset R_m$ then we consider a vector valued estimating function $(g_1(x, \theta), \ldots, g_m(x, \theta))' = g(x, \theta)$ such that $E(g) = 0$, the variance-covariance matrix $M_g = E(g \, .g')$ exists and is positive definite. In the single parameter case we imposed the condition that $E(\frac{\partial g}{\partial \theta}) \neq 0$. Analogously in the vector valued situation we assume that Dg is non-singular where Dg denotes the mxm matrix with elements $E(\frac{\partial g_i}{\partial \theta_j})$.

Kale (1962a) proved that for a vector valued $g \in G^{(m)}$

$$M_g - D_g \, J^{-1} \, D_g' \text{ is non-negative definite} \qquad (30)$$

where J is the Fisher Information Matrix and $G^{(m)}$ is the class of m- dimensional vector valued estimating functions such that M_g is positive definite and D_g is non-singular. The standardization (12) in the present case immediately

leads to the standardized version of g, given by $D_g^{-1}g = g_s$ and the extended Cramer-Rao inequality for a regular standardized estimating function now is given by

$$M_{g_s} - J^{-1} \text{ is non-negative definite} \qquad (31)$$

The above standardization was first proposed by Ferreira (1982) and also independently by Chandrasekar (1983).

For the vector valued g, several different optimality criteria can be proposed. The most common among these are

(i) Matrix Optimality : $M_{g_s} - M_{g_s^*}$ is non-negative define.

(ii) Trace Optimality : $Tr(M_{g_s}) \geq Tr(M_{g_s^*})$

(iii) Determinant Optimality : $|M_{g_s}| \geq |M_{g_s^*}|$

Chandrasekar and Kale (1984) proved that the three criteria are equivalent in the sense that if g* is optimal with respect any one of these criteria then it is also optimal with respect to the remaining two.

From the extended Cramer-Rao inequality (30) or (31) it immediately follows that, in the class of all regular unbiased estimating functions the vector score function

$$\frac{\partial \log p}{\partial \theta} = \left(\frac{\partial \log p}{\partial \theta_1}, \ldots, \frac{\partial \log p}{\partial \theta_m} \right)'$$

is optimal. The essential uniqueness of the optimal estimating function follows from the fact that if g_1 and g_2 are both optimal then their standardized versions g_{1s} and g_{2s} are identical. In particular this implies that the likelihood equation is an optimal estimating equation as the estimating

equations $g = 0$ and $g_s = 0$ are identical. Bhapkar (1972) defined the concept of the efficiency of an estimating equation corresponding to the Trace and Determinant optimality criteria and he also introduced the Rao- Blackwellization of an estimating function with respect to minimal sufficient statistics. Another interesting reference in this connection is Morton (1981).

8. INVARIANCE AND NUISANCE PARAMETERS

Another important property of the estimating functions is the invariance under one to one transformation of the parameter θ. Thus if $g(x,\theta)$ is an estimating function then under any one to one differentiable transformation $\Phi = \alpha(\theta)$ where $\frac{\partial \alpha}{\partial \theta}$ is non-singular, then $g(x, \alpha(\Phi)) = g_1(x, \Phi)$ is an estimating function for Φ. If $\hat{\theta}_g$ and $\hat{\Phi}_{g_1}$ are the estimates obtained from the equations $g = 0$ and $g_1 = 0$ respectively, then $\Phi_{g_1} = \hat{\theta}_g$. It is well known that this invariance property is enjoyed by the maximum likelihood estimation but does not hold for the unbiased minimum variance estimation. On the other hand unlike unbiased estimation, the maximum likelihood estimation faces difficulties when nuisance parameters are involved i.e. when we are interested in $\theta^{(1)} = (\theta_1, \theta_2, \ldots, \theta_r)'$ only while $\theta^{(2)} = (\theta_{r+1}, \ldots, \theta_m)'$ acts as a nuisance parameter. In fact the maximum likelihood estimator of only $\theta^{(1)}$, the parameter of interest, is technically undefined. A naive approach, based on obtaining maximum likelihood estimator of the entire parameter $\hat{\theta} = (\hat{\theta}^{(1)}; \hat{\theta}^{(2)})$ and then using $\hat{\theta}^{(1)}$, in general leads to anamolies as is well known from the Neyman -Scott problem. We will now show how the estimating functions approach alleviates the difficulties

arising in the nuisance parameter case. This will also show that the approach based on estimating functions unifies the maximum likelihood estimation as well as unbiased minimum variance estimation by eliminating their respective weaknesses namely the nuisance parameter case and non-invariance under a one to one transformation of the parameter space.

8. OPTIMALITY : NUISANCE PARAMETER CASE

Consider now an abstract random variable x with pdf belonging to $\{p(x, \theta), \theta \in \Omega_m\}$ where $\theta = (\theta^{(1)}, \theta^{(2)})'$ with $\theta^{(1)} = (\theta_1, \ldots, \theta_r)'$ as the parameter of interest and $\theta^{(2)} = (\theta_{r+1}, \ldots, \theta_m)'$ as a nuisance parameter. A regular estimating function for $\theta^{(1)}$ is an r-dimensional function $g(x, \theta^{(1)})$ such that $E(g) = 0$, $\forall\ \theta \in \Omega_m$ and $M_g = E(gg')$ is positive definite and the rxr matrix $D_g = E(\frac{\partial g}{\partial \theta^{(1)}})$ is non-singular. Let $g_s = D_g^{-1} g$ denote the standardized version of g. Note that g_s may depend on $\theta^{(2)}$, the nuisance parameter, although the corresponding estimating equations $g_s = 0$ and $g = 0$ are same. An estimating function $g^*(x, \theta^{(1)})$ is optimal in $G(\theta^{(1)})$, the class of all regular estimating functions for estimating $\theta^{(1)}$ in the presence of the nuisance parameter $\theta^{(2)}$ if $M_{g_s} - M_{g_s^*}$ is non-negative definite $\forall\ \theta \in \Omega_m$, and $\forall\ g \in G(\theta^{(1)})$. As seen earlier this is equivalent to Trace and Determinant optimality.

Godambe & Thompson (1974) considered the case r = 1 and m = 2, a single parameter of interest and a real valued nuisance parameter and showed that in the case of $N(\theta_1, \theta_2)$

the optimal estimating function for θ_1 ignoring θ_2 is $(\bar{x} - \theta_1)$ while $s^2 - (n-1)\theta_2$ is optimal estimating function for θ_2 ignoring θ_1. An interesting reference in this connection is Barnard (1973). Godambe (1976) considering the case $r = 1$ and $m > 1$ with $\theta = (\theta_1, \theta^{(2)}) \in \Omega_1 \times \Omega_2$ introduced an interesting structure for estimating θ_1 ignoring $\theta^{(2)}$. Godambe (1976) assumed that there exists a statistic $T(x)$ such that

$$p(x, \theta) = f_t(x; \theta_1) \cdot h(t, \theta_1, \theta^{(2)})$$

where h is the pdf of T and $f_t(x, \theta_1)$ is the conditional pdf of x given t which depends only on θ_1, the parameter of interest. Further he assumed that the class $\{h(t, \theta_1, \theta^{(2)}), \theta_1$ fixed $\theta^{(2)} \in \Omega_2\}$ is complete. Under mild regularity conditions on p, f_t, h and $G(\theta^{(1)})$ Godambe (1976) showed that the conditional score function $\frac{\partial \log f_t}{\partial \theta_1}$ is optimal estimating function for θ_1 ignoring $\theta^{(2)}$. Using this theory Godambe (1976) showed how the Neyman-Scott problem can be resolved and how the optimum estimating function for estimating the error variance ignoring block means leads to the minimum variance unbiased estimator which is consistent when the number of blocks goes to infinity. Ferreira (1982) and Chandrasekar (1983) generalized these results for $r > 1$ and showed that the conditional vector score function $\frac{\partial \log f_t}{\partial \theta^{(1)}}$ is an optimal estimating function. Kale (1987a) pointed out the analogy of Godambe's structure with the Neyman-structure used

DR. B. K. KALE
PROFESSOR OF STATISTICS
DEPT. OF STATISTICS
UNIV OF POONA
PUNE-411 007

in construction of UMP tests in the presence of nuisance parameters.

Chandrasekar & Kale (1984) proved a Cramer-Rao type inequality namely $M_{g_s} - J^{11}$ is non-negative definite $\forall\ \theta \in \Omega$, $\forall\ g \in G(\theta^{(1)})$, where J^{11} is the rxr matrix when J^{-1} is partitioned corresponding to $(\theta^{(1)}, \theta^{(2)})'$. Kale (1987b) proved the essential uniqueness of the optimal estimating function. Following the extension of Cramer-Rao inequality approach Subramanyam & Naik-Nimbalkar (1988) obtained a generalization of (31) for a Hilbert space valued parameter and proved that Aalen's (1978) estimator of cumulative intensity function emerges as a solution of optimal estimating equation.

Godambe (1976), (1980) uses the optimality of conditional score function to define partial sufficiency and ancillarity of a statistic T(x) for $\theta^{(1)}$ the parameter of interest when $\theta^{(2)}$ acts as a nuisance parameter. Kale (1987a) has shown that for multiparameter exponential family, Godambe's approach for defining partial sufficiency and ancillarity succeeds whereas many other approaches fail. Recently Bhapkar (1988) has extensively studied this problem along with the problem of defining Fisher Information about $\theta^{(1)}$ ignoring $\theta^{(2)}$.

In the above set-up we assumed that the statistic T(x) exists uniquely for all values of $\theta^{(2)} \in \Omega_2$. Lindsay (1982) deals with the case when T depends on $\theta^{(2)}$.

10. EXTENSIONS

It is thus clear that the approach based on estimating

functions unifies the method of maximum likelihood and the method of minimum variance unbaised estimation in case of the parametric models. It is no wonder that this theory has been applied successfully for estimation problems in such diverse fields from survey-sampling to time-series and stochastic processes as examplified by the papers of Godambe & Thompson (1986), Thavaneswaran & Abraham (1987) and Godambe (1985).

We have also seen how the estimating functions theory successfully tackles the situation where the usual Gauss-Markov or least square theory fails to give a reasonable solution. Now the Gauss Markov or LS approach as said before is for semi-parametric models where we do not assume the exact form of the density $p(x, \theta)$ but assume only the knowledge of the first few moments. Since the form of p is not known estimating functions based on $\frac{\partial \log p}{\partial \theta}$ are not available here. However, as Halmos (1946) showed that $\bar{x}$ is minimum variance unbiased estimator of $E(x) = \theta$ in the class $\mathcal{F}_0$ of all continuous distribution functions with mean θ, Godambe and Thompson (1978) proved that $(\bar{x} - \theta)$ is optimal estimating function within the sub-class $\mathcal{F}_1 \subset \mathcal{F}_0$ with location parameter θ.

In general a parameter of a distribution in a semi-parametric model is a well defined functional of the underlying population distribution function and therefore the definition of such a parameter is closely connected with the method of estimation of this parameter. Suppose this parameter $\theta(F)$ for $F \in \mathcal{F}$ is a parameter of interest and $\varphi(F)$ is a nuisance parameter such that $(\theta(F), \varphi(F))'$ is a labelling parameter for $\mathcal{F}$. Godambe & Thompson (1984)

obtained an optimal estimating function for estimating $\theta(F)$ which in turn could also be used to define the parameter $\theta(F)$. This line of work has not been followed very vigorously and deserves more attention.

The theory of optimum estimating functions has provided a new and fruitful perspective on 'quasi-likelihood' (Wedderburn, 1974) by identifying 'quasi-score function' with the 'optimal estimating function (Godambe & Heyde, 1987; Godambe & Thompson, 1989). This is also true in connection with 'partial likelihood' (Cox, 1975; Godambe, 1985).

To indicate the varied applications of estimating function theory to areas of Biostatistics, we just refer to Liang & Zeger (1987) and Prentice (1988) .

It is now clear that, among researchers in different areas of statistics there is an increasing trend to utilize estimating function theory for statistical model building, inference and the like. A purpose of this 'Overview' is to accelarate this already existing trend.

ACKNOWLEDGEMENTS

Both authors are grateful to National Science and Engineering Council of Canada and the University Grants Commission, India for providing financial support during their leave periods. They are also grateful to B. Abraham for some comments.

REFERENCES

Aalen, O. (1978). Nonparametric Inference for a family of counting processes, Ann. Stats 6, 701-706.

Bernard, G. A. (1973). Maximum likleihood and nuisance parameters. Sankhya A, 35, 133-138.

Bhapkar, V. P. (1972). On a measure of efficiency of an estimating equation, Sankhya A 34, 467-472.

Bhapkar, V. P. (1988). On generalized principles for inference in the presence of nuisance parameter, Tech. Report No. 266, Department of Statistics, Univ. of Kentucky.

Chandrasekar, B. (1983). Contributions to the theory of unbiased statistical estimation functions. Ph.D. thesis submitted to Univ. of Poona, Pune - 7, India.

Chandrasekar, B. and Kale, B. K. (1984). Unbiased statistical estimation functions in presence of nuisance parameter. Jour. Stat. Plan.Inf. 9, 45-54.

Cox, D. R. (1975). Partial Likelihood. Biometrika 62, 269-276.

Durbin, J. (1960). Estimation of parameters in time series regression models. Jour. Roy. Stat.Soc. Ser B 22 138-153.

Ferreira, P. E. (1982). Multiparametric estimating equations. Ann. Inst. Stat. Math. 34A , 423-431.

Fisher, R. A. (1925). Theory of Statistical Estimation. Proc. Cambridge Phil. Soc. 22, 700-725.

Fisher, R. A. (1935). The fiducial argument in statistical Inference. Ann. Eugenics, 6, 391-396.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. Ann. Math. Stat. 31, 1208-1212.

Godambe, V. P. and Thompson, M. E. (1974). Estimating equations in the presence of nuisance parameters. Ann. Stat. 2, 568-571.

Godambe, V. P. (1976). Conditional likelihood and unconditional optimum estimating equations. Biometrika, 63, 277-284.

Godambe, V. P. and Thompson, M. E. (1976). Some aspects of the theory of estimating equations. Jour. Stat. Plan. Inf. 2, 95-104.

Godambe, V. P. (1980). On the sufficiency and ancillarity in the presence of nuisance parameter, Biometrika, 67, 269-276.

Godambe, V. P. and Thompson, M. E. (1984). Robust estimation through estimating equations. Biometrika, 71 , 115-125.

Godambe, V.P. (1985). The foundations of finite sample estimation in stochastic processes. Biometrika, 72, 419-428.

Godambe, V. P. and Thompson, M. E. (1986). Parameters of super population and survey population, their relationship and estimation. Int. Statist. Rev. 54, 127-138.

Godambe, V. P. and Thompson, M. E. (1985). Logic of least squares revisited. Pre-print.

Godambe, V. P. and Thompson, M.E. (1989). An extension of quasilikelihood. To appear in Jour. Statist. Plan and Inference.

Godambe, V. P. and Heyde, C. C. (1987). Quasi-likelihood and optimal estimation. Int. Stat. Rev. 55, 231-244.

Halmos, P. (1946). The theory of unbiased estimation. Ann. Math. Stat. 17, 43-54.

Kale, B. K. (1962a) An extension of Cramer-Rao inequality for statistical estimation functions. Skand. Aktur. 45, 60-89.

Kale, B. K. (1962b). On the solution of likelihood equations by iteration processes. Biometrika, 49, 479-486.

Kale, B.K. (1985). Theory of unbiased statistical estimation functions. Lecture Notes, Dept. of Statistics, Iowa State Univ. Ames, Iowa, USA 50011, and Tech. Report 81, Dept. of Statistics, Univ. of Poona, Pune 411007, India.

Kale, B. K. (1987a). Optimal estimating function in multiparameter exponential family. Tech. Report 87-02 Dept. of Stat. and Acturial Sc. Univ. of Waterloo, Waterloo, Canada.

Kale, B. K. (1987b). Essential uniqueness of optimal estimating functions. Jour. Stat. Plan. Inf. 17, 405-407.

Kimball, B. F. (1946). Sufficient statistical estimation functions for the parameters of the distribution of maximum values. Ann. Math. Stat. 17, 299-309.

Liang, K. Y. and Ziger, S.L. (1987). Longitudional data analysis using generalized Linear models. Biometrika, 73, 13-22.

Lindsay, B. (1982). Conditional score functions : some optimality results. Biometrika, 69, 503-512.

Mortan, R. (1981). Efficiency of estimating equations and the use of pivots. Biometrika 68 227-233.

Prentice,R. L. (1988). Correlated Binary regression with covariates specific to each binary observation. Pre-print.

Small C. and McLeish D. L. (1988). The theory and applications of statistical inference functions. Lecture Notes in Statistics No. 44, Springer Verlag, Heidelberg, New York London.

Subramanyam, A., and Naik-Nimbalkar, U. V. (1989). Optimal unbiased statistical estimating functions for Hilbert space valued parameters. To appear in J. Stat. Plan. Inference.

Thavaneshwaran, A. and Abraham, B. (1988). Estimation for non-linear time series models using estimating equations Jour. Time Series Analysis, 9, 99-108.

Wilks, S. S. (1939). Shortest average confidence intervals from large samples. Ann. Math. Stat. 9, 166-175.

Wilks, S. S. (1948). Mathematical Statistics. John Wiley and Sons, New York.