I.I.S. - 47e Session
Paris, 29 août - 6 Septembre 1989

I.S.I. - 47th Session
Paris, August 29 - September 6, 1989

C1-1.2

FISHER LECTURE

QUASI–LIKELIHOOD AND OPTIMALITY FOR ESTIMATING FUNCTIONS: SOME CURRENT UNIFYING THEMES

C.C. Heyde

Department of Statistics, IAS
The Australian National University
GPO Box 4, Canberra, ACT 2601
Australia

Key words: Optimality in estimation, quasi–likelihood, efficiency, information, minimal confidence zones, estimating functions, least squares, maximum likelihood

1. INTRODUCTION

This paper is concerned with recent progress towards a general unification of two of the principal themes in statistical estimation theory, those of least squares, which is founded on finite sample considerations, and maximum likelihood, whose justification is primarily asymptotic. The gestation period for this progress has been long and very many authors have contributed. Many key references are given in Godambe and Heyde (1987).

The methods underlying the abovementioned unification are based on extensions of the concept of Fisher information and it is especially appropriate to remember Fisher at the 47th Session of the ISI, the one which is closest to the centenary of his birth (February 17, 1890 in London). His introduction of maximum likelihood (although not yet under that name) in Fisher (1912) and his seminal papers Fisher (1920), (1922), (1925) introducing information, consistency, sufficiency, ancillarity, etc., provided a basis on which all current discussions are built.

R. Maller et al. (eds.), *Selected Works of C.C. Heyde*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5823-5_50, © Springer Science+Business Media, LLC 2010

It is through the use of estimating functions (functions of both the data and the parameter) rather than the estimators themselves that it is possible to incorporate the principal advantages of the methods of least squares and of maximum likelihood. The virtue of this indirect approach of estimating an unknown function which has the true value of the parameter as a root rather than estimating the parameter directly may be less than transparent at first sight. This is despite the long history of the subject, which dates back at least to K. Pearson's (1894) introduction of the method of moments, and the fact that the standard methods of estimation: maximum likelihood, least-squares, conditional least-squares, minimum chi-squared, M-estimation, etc., are included under minor regularity conditions. Also, Fisher's information is an estimating function property, namely of the score function (the derivative of the log-likelihood with respect to the parameter) rather than of the maximum likelihood estimator. Indeed, the rationale lies in the more fundamental character of the estimating function than that of an estimator derived therefrom. For example, under minor regularity conditions, the score function provides a minimal sufficient partitioning of the sample space. However, there is often no single sufficient statistic. Also, the asymptotic properties of an estimator are almost invariably obtained, as in the case of maximum likelihood, via the asymptotics of the estimating function and then transferred to the parameter space via local linearity. For details concerning the derivation of the asymptotic properties of the MLE from those of the score function together with a further justification for use of estimating functions in terms of marginalization for incomplete data problems, see Chapter 1 of McLeish and Small (1988). To these considerations we shall add another, namely the ready capacity to combine separate estimating function, each with information to offer about the unknown parameter (Heyde (1987), (1989)).

Within broad classes of unbiased estimating functions it is possible to develop a theory of optimality, founded on information based ideas, which encapsulates the virtues of least squares and maximum likelihood. This carries with it, as simple corollaries, such diverse results as the Gauss-Markov Theorem, the Cramér-Rao Inequality and the

minimum size asymptotic confidence zones property possessed by the maximum likelihood estimator (subject to the usual regularity conditions). The estimating functions so produced can be interpreted as quasi-score estimating functions and the estimators derived therefrom as quasi-likelihood estimators.

2. THE SETTING

Let $\{X_t, 0 \le t \le T\}$ be a sample in discrete or continuous time drawn from a process taking values in r-dimensional Euclidean space whose distribution involves a "parameter" θ taking values in an open subset Θ of p-dimensional Euclidean space. The setting may be parametric or nonparametric; θ could be, for example, the mean of stationary process. The true value of the "parameter" is θ_0 and this is to be estimated.

Suppose that the possible probability measures for $\{X_t\}$ are $\{P_\theta\}$ and that each $(\Omega,\mathcal{F},P_\theta\}$ is a complete probability space.

We shall confine attention to the class $\mathcal{G}$ of zero mean square integrable estimating functions $G_T = G_T(\{X_t, 0 \le t \le T\}, \theta)$ for which $EG_T(\theta) = 0$ for each P_θ. Here G_T is a vector of dimension p. Estimators $\hat{\theta}$ are found by solving the estimating equation $G_T(\hat{\theta}) = 0$.

We consider a class of $\mathcal{H} \in \mathcal{G}$ of estimating functions G_T which are a.s. differentiable with respect to the components of θ and such that

$$E\dot{G}_T = (E\partial G_{T,i}/\partial\theta_j)$$

and EG_TG_T' are nonsingular, the prime denoting transpose.

Recent studies on the performance of estimating functions G_T involve focusing on the quantity

$$\mathcal{E}(G_T) = (E\dot{G}_T)' (EG_TG_T')^{-1} (E\dot{G}_T)$$

which may be thought of as an information matrix (and is indeed the Fisher information matrix when $G_T = U_T$, the score function). Our object is to maximize this estimating

function information within specified classes of useful estimating functions and to this end we adopt the following definition whose origins date back to Godambe (1960) and Durbin (1960).

Definition 1. *Suppose that* $G_T^* \in \mathcal{H}_1 \subseteq \mathcal{H}$. *If*

$$\mathcal{E}(G_T^*) - \mathcal{E}(G_T)$$

is nonnegative definite for all $G_T \in \mathcal{H}_1$, *we say that* G_T^* *is* O_F*–optimal within* $\mathcal{H}_1$.

The terminology O_F–optimal refers to finite sample optimality. There is also a closely related concept of O_A–optimality (asymptotic optimality) (see Godambe and Heyde (1987), Heyde (1988)) and both hold under broad conditions. We shall confine attention to O_F–optimality in this paper for clarity of exposition.

Note that the definition of O_F–optimality does not require the existence of the score function, and indeed the setting may be nonparametric. Nevertheless, to motivate the definition it is useful to suppose that the score function U_T exists and is a.s. differentiable with respect to the components of θ, and that differentiation and integration can be interchanged in EG_TU_T' and EU_TG_T' for all G_T under consideration.

Now, as noted above, the score function typically provides a minimal sufficient partitioning of the sample space so we should use this estimating function as a basis for inference if it is known. However, if the true underlying parametric family and hence the score function is unknown we can seek an estimating function G_T which has minimum dispersion distance from U_T. This is precisely what an O_F–optimal solution provides. That is, for some fixed matrix α depending on θ and T,

$$E(\alpha_T U_T - G_T)(\alpha_T U_T\ G_T)' - E(\alpha_T U_T - G_T^*)(\alpha_T U_T - G_T^*)'$$

is nonnegative definite for all $G_T \in \mathcal{H}_1$. (Of course an optimal estimating function is defined only up to a constant (matrix) multiplier.) Furthermore, if the dimension of

θ, $p = 1$, another equivalent formulation is that G_T^* has maximum squared correlation with U_T. These are, of course, least squares principles built into the formulation. For details of the equivalences see Godambe and Heyde (1987).

Additionally, under a fairly broad range of conditions the covariance matrix of the estimator $\bar{\theta}_T$ given by $G_T(\bar{\theta}_T) = 0$ is asymptotically $(\mathcal{E}(G_T(\theta_0)))^{-1}$ and, indeed,

$$(EG_T(\theta_0)\, G_T'(\theta_0))^{-1/2}\, (E\dot{G}_T(\theta_0))\, (\bar{\theta}_T - \theta_0) \xrightarrow{d} N_p(0,I_p)\,. \tag{1}$$

where "d" denotes convergence in distribution to $N_p(0,I_p)$, the standard p-variate normal. Hence, asymptotic confidence zones of minimum size are associated with an O_F-optimal estimator for which $(\mathcal{E}(G_T(\theta_0)))^{-1}$ is minimized in the partial order of nonnegative definite matrices. This is of course, the characteristic property of maximum likelihood but here there is a restricted class of competitors. Nevertheless, the prime advantage of maximum likelihood is built into the formulation.

The heuristics are as follows. By Taylor expansion

$$0 = G_T(\bar{\theta}_T) = G_T(\theta_0) + \dot{G}_T(\theta_{1,T})\, (\bar{\theta}_T - \theta_0)$$

where $\|\theta_0 - \theta_{1,T}\| \le \|\theta_0 - \bar{\theta}_T\|$ and under certain conditions,

$$\dot{G}_T(\theta_{1,T})\, (E\dot{G}_T(\theta_0))^{-1} \xrightarrow{p} I_p\,,$$

$$(EG_T(\theta_0)\, G_T'(\theta_0))^{-1/2}\, G_T(\theta_0) \xrightarrow{d} N_p(0,I_p)$$

as $T \to \infty$ from which (1) readily follows.

The criterion of Definition 1 is often not of direct practical value but an equivalent from is given in the following theorem of Heyde (1988) which is easier to use and is preferabe to other equivalent terms such as those mentioned above.

Theorem 1. *Suppose that $\mathcal{H}_1 \subseteq \mathcal{H}$ is a convex set. Then, $G_T^* \in \mathcal{H}_1$ is O_F-optimal within $\mathcal{H}_1$ if and only if*

$$(E\dot{G}_T)^{-1} EG_T G_T^{*\prime} = (E\dot{G}_T^*)^{-1} EG_T^* G_T^{*\prime}$$

for all $G_T \in \mathcal{X}_1$.

3. SOME RAMIFICATIONS

An estimating function which is O_F–optimal can be regarded as a *quasi–score* estimating function and an estimator derived therefrom as a *quasi–likelihood estimator*. We shall henceforth employ this terminology which was introduced by Wedderburn (1974) and has been widely used in the context of the general linear model (e.g. McCullagh and Nelder (1983, Ch.8)) as it is a helpful description and underlines its useful properties. A quasi–score estimating function will indeed be a true score in a wide variety of situations, such as in an exponential family environment; for some general discussion of this issue see Sorensen (1989).

The classical quasi–likelihood setting concerns estimation of θ where the sample is of independent random vectors Y_t , $1 \le t \le T$, with means $\mu_t(\theta)$ and dispersions

$$E(Y_t - \mu_t(\theta))\,(Y_t - \mu_t(\theta))' = v_t(\theta) .$$

The quasi–score estimating function is defined as

$$Q_T = \sum_{t=1}^{T} \dot{\mu}_t' v_t^+ (Y_t - \mu_t) = \dot{\mu}' V^+ (Y - \mu)$$

where

$$\mu' = (\mu_1' \vdots \ldots \vdots \mu_T') , \quad Y' = (Y_1' \vdots \ldots \vdots Y_T') , \quad V^+ = \mathrm{diag}(v_1^+ \vdots \ldots \vdots v_T^+)$$

and for a matrix A , A^+ denotes its Moore–Penrose pseudoinverse, the unique matrix possessing the properties $AA^+A = A$, $A^+AA^+ = A^+$, $A^+A = AA^+$. That Q_T is O_F–optimal within the relevant class of weighted sums of $(Y_t - \mu_t)$'s follows immediately from Theorem 1.

As another application of Theorem 1 we note that under the regularity conditions which are imposed,

$$E\dot{G}_T = -EG_T U_T'$$

and if $U_T \in \mathcal{Y}_1$, then

$$(E\dot{G}_T)^{-1} EG_T U_T' = -I_p .$$

Thus, U_T is O_F–optimal within $\mathcal{Y}_1$ via Theorem 1. This means that for $G_T \in \mathcal{Y}_1$,

$$(E\dot{G}_T)^{-1} (EG_T G_T') ((E\dot{G}_T)^{-1})' - (EU_T U_T')^{-1} \tag{2}$$

is nonnegative definite which encapsulates the multivariate *Cramer–Rao Inequality*. This is usually presented in the case where G_T is of the form $G_T = S_T - \theta$ when S_T is a statistic with $ES_T = \theta$, and then (2) gives the nonnegative definiteness of

$$E(S_T - \theta)(S_T - \theta)' - (EU_T U_T')^{-1} .$$

For random sampling $EU_T U_T' = n\Sigma$, say, where Σ is the Fisher information contained in a single observation.

A final application of Theorem 1 treats the *Gauss–Markov Theorem*. This has traditionally been set in the context of a multiparameter linear regression model, for example

$$Y = X\theta + \varepsilon$$

where $Y = (y_1,...,y_n)'$, X is an $n \times p$ design matrix of rank p, $\theta = (\theta_1,...,\theta_p)'$ is a parameter vector and ε is a zero mean error vector with independent components and covariance matrix V.

Unbiased linear estimates of θ are of the form AY where A is a $p \times n$ matrix such that $AX = I_p$, and we may think of these as coming from a corresponding linear unbiased estimating function $A(Y - X\theta)$ which, when set to zero, yields the estimator AY.

The standard approach via least squares is to minimize the covariance matrix

$$E(AY - \theta)(AY - \theta)' = AVA' ,$$

subject to $AX = I_p$, in the partial order of nonnegative definite matrices. This leads to

the solution

$$A^* = (X'V^{-1}X)^{-1} X'V^{-1}, \tag{3}$$

and estimator A^*Y for θ, provided V is nonsingular.

Now an alternative approach is to seek the O_F-optimal estimating function within the family of estimating functions

$$\mathcal{H} = \{A(Y - X\theta), \ AX = I_p\}.$$

Then, writing

$$G = A(Y - X\theta), \ G^* = A^*(Y - X\theta)$$

with A^* as in (3), we have, via Theorem 1, that G^* provides the desired solution if

$$(E\dot{G})^{-1} EGG^{*\prime}$$

is a constant matrix. This is, furthermore, easily checked since

$$E\dot{G} = -I$$

and

$$EGG^{*\prime} = AE(\varepsilon\varepsilon')A^{*\prime} = AVA^{*\prime} = (X'V^{-1}X)^{-1}.$$

This theory can be put in a more general setting of best linear unbiased estimators (BLUE's) for stochastic processes, as in Grenander (1981, Ch.4) with minimal change to the mathematical details. Here it is assumed that $Y(t)$ has mean $m(t)$ where $m(t)$ is expressible in the form

$$m(t) = \sum_{\nu=1}^{p} a_\nu \psi_\nu(t) = \alpha'\Psi$$

where the coefficients $(\alpha_1,...,\alpha_p)' = \alpha$ are real or complex unknown constants and the $(\psi_1,...,\psi_p)' = \Psi$ are given functions. Similar considerations to the above lead to Theorem 2, Chapter 4 of Grenander (1981) and we can interpret the BLUE as coming from the O_F-optimal estimating function within the class

$$\{cY, \ c'\Psi = I_p\}.$$

4. FURTHER CONSIDERATIONS AND EXTENSIONS

In this concluding section we mention a number of related issues of importance.

Many estimation problems are based in a natural way on a semimartingale model which thereby leads to consideration of martingale estimating functions. For such models, O_A-optimality, which is briefly mentioned above and is very closely related to O_F-optimality, is generally a more pertinent concept. For the beginnings of a comparison of the different criteria see Heyde (1988).

O_F and O_A-optimality are exact properties and in some applications optimality holds only in an asymptotic sense. For a treatment of asymptotic quasi-likelihood see Heyde and Gay (1989).

Quasi-score estimating functions based on different classes of estimating functions can be derived in many contexts and it is important to be able to compare them and to combine them if this should be advantageous. Furthermore, various problems present a collection of conditional or marginal processes leading to separate quasi-score or score functions which can potentially be combined. This may occur, for example, when likelihood methods are very difficult to use or when it is only part of the model that can be treated satisfactorily. In either case, one has to consider a set of quasi-score functions $Q_{i,T}$, $1 \le i \le k$, and the maximizing of information over the class of estimating functions $\{\Sigma_{i=1}^{k} w_i Q_{i,T}\}$. New results on this problem are given in Heyde (1989) building upon discussions of the case $k = 2$ in Heyde (1987) and of the combinations of true score functions in Lindsay (1988).

Much more remains to be said on the subject of optimal inference in problems where there are nuisance parameters. For some discussion and references see McLeish and small (1988). Their approach to optimal inference is via concepts of E-ancillarity and E-sufficiency which permit inferential reduction analogous to the usual sufficiency and ancillarity reductions but within the class of unbiased estimating functions.

Finally, rather little theory has yet been developed for extensions of optimal

estimation beyond the case of finite dimensional parameters. Contributions include the thesis of Thavaneswaran (1986) and treatment of the special case of linear inference (Grenander (1981)). Many important applications involving estimating of functions await this theory and *ad hoc* methods abound.

REFERENCES

Durbin, J. (1960). Estimation of parameters in time-series regression models. *J.R. Statist. Soc.* B **22**, 139–153.

Fisher, R.A. (1912). On an absolute criterion for fitting frequency curves. *Messeng. Math.* **41**, 155–160.

Fisher, R.A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error. *Mon. Not. Roy. Astron. Soc.* **80**, 758–770.

Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans.* A **222**, 309–368.

Fisher, R.A. (1925). Theory of statistical estimation. *Proc. Camb. Phil. Soc.* **22**, 700–725.

Godambe, V.P. (1960). An optimal property of regular maximum likelihood estimation. *Ann. Math. Statist.* **31**, 1208–1212.

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

Grenander, U. (1981). *Abstract Inference*. Wiley, New York.

Heyde, C.C. (1987). On combining quasi-likelihood estimating functions. *Stoch. Processes Applic.* **25**, 281–287.

Heyde, C.C. (1988). Fixed sample and asymptotic optimality for classes of estimating functions. *Contemporary Mathematics* **80**, 241–247.

Heyde, C.C. (1989). On efficiency for quasi-likelihood and composite quasi-likelihood methods. In Proceedings of International Conference on Recent Developments in Statistical Data Analysis and Inference, Neuchâtel, August 1989. Elsevier, Amsterdam, to appear.

Heyde, C.C. and Gay, R. (1989). On asymptotic quasi-likelihood estimation. *Stoch. Processes. Applic.* 31, in press.

Lindsay, B.G. (1988). Composite likelihood methods. *Contemporary Mathematics* 80, 221-239.

McCullagh, P. and Nelder, J.A. (1983). *Generalized Linear Models*. Chapman and Hall, London.

McLeish, D.L. and Small, C.G. (1988). *The Theory and Application of Statistical Inference Functions*. Springer Lecture Notes in Statistics No. 44, Springer, New York.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Phil. Trans. Roy. Soc.* A 185, 71-110.

Sorensen, M. (1989). On quasi-likelihood for semimartingales. *Stoch. Processes Applic.*, to appear.

Thavaneswaran, A. (1986). *Estimations of Semimartingales*. PhD thesis, University of Waterloo, Canada.

Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* 61, 439-447.

SUMMARY

This paper outlines the recent development of a general theory of quasi-likelihood which embraces the principal features of the methods of least squares and maximum likelihood.

RESUMÉ

Cet article expose le developpement nouveau de une théorie générale de quasi-vraisemblance. par où on embrasse les méthodes de moindre carrés et de maximum de vraisemblance.