# Estimating Functions and Equations: An Essay on Historical Developments with Applications to Econometrics[1]

**Anil K Bera**
E-mail: abera@uiuc.edu
*Department of Economics, University of Illinois*,
*1206 S. 6th Street, Champaign, IL 61820, USA*
**Yannis Bilias**
E-mail: bilias@ucy.ac.cy
*Department of Economics, University of Cyprus*,
*P. O. Box 20537, 1678 Nicosia, Cyprus*
**Pradosh Simlai**
E-mail: psimlai@uiuc.edu
*Department of Economics, University of Illinois*,
*1206 S. 6th Street, Champaign, IL 61820, USA*

[1]We are most grateful to Kerry Patterson for his constant encouragement and very helpful comments on an earlier draft. Without his many kind and gentle proddings this chapter would not have been completed. We are also most thankful to Peter Phillips and Jun Yan for many pertinent comments and suggestions. However, we retain the responsibility for any remaining errors. A part of the work was done during a visit by the first author to the Department of Economics, University of Cyprus, whose financial support is gratefully acknowledged.

**Abstract**

The idea of using estimating functions goes a long way back, at least to Karl Pearson's introduction to the method of moments in 1894. It is now a very active area of research in the statistics literature. One aim of this chapter is to provide an account of the developments relating to the theory of estimating functions. Starting from the simple case of a single parameter under independence, we cover the multi-parameter, presence of nuisance parameters and dependent data cases. Application of the estimating functions technique to econometrics is still at its infancy. However, we illustrate how this estimation approach could be used in a number of time series models, such as random coefficient, threshold, bilinear, autoregressive conditional heteroscedasticity models, in models of spatial and longitudinal data, and median regression analysis. The chapter is concluded with some remarks on the place of estimating functions in the history of estimation.

**CONTENTS**

# 1 Prologue: Early Appearances of the Concept of Estimating Function in Statistics

## 1.1 A defining moment in the history of statistics

In the history of any scientific field there is always a defining moment - a moment that arrives with some maturity and when an authoritative figure clearly states the purpose, progress and problems of the field. For statistics, it can be safely argued that, the defining moment arrived in 1922 with the appearance of Fisher's epochal article, " On the Mathematical Foundations of Theoretical Statistics." After discussing the purpose of statistical methods, Fisher (1922, p.313) proclaimed the three fundamental problems in statistics as :

"(1) Problems of Specification. These arise in the choice of the mathematical form of the population.
(2) Problems of Estimation. These involve the choice of methods of calculating from a sample statistical derivates, or as we shall call them statistics, which are designed to estimate the values of the parameters of the hypothetical population.
(3) Problems of Distribution. These include discussions of the distribution of statistics derived from samples, or in general any functions of quantities whose distribution is known."

Fisher did not dwell much on the Problem of Specification, and quickly stated (p.315), " The discussion of theoretical statistics may be regarded as alternating between problems of estimation and problems of distribution." He occupied himself mostly with the problems of estimation and distribution. In terms of estimation, he went on to introduce some of the fundamental concepts, such as, consistency, efficiency and sufficiency. These concepts solely focussed on estimators which are functions of observations alone. Fisher demonstrated that his suggested method of estimation, namely, the maximum likelihood (ML) method [Fisher (1912)] is "superior" to Karl Pearson's method of moments [Pearson (1894, 1902)] in terms of efficiency.

## 1.2 Estimating function approach: A short introduction

In the estimating function (EF) approach to estimation, the focus is on a function that involves both the parameters and the sample, such as $g(y, \theta)$ where $y = (y_1, y_2, \ldots, y_n)$ represent the data and, $\theta$ denotes the parameter. We obtain the estimator say, $\widehat{\theta}$ by solving $g(y, \theta) = 0$, which we will call the estimating equation (EE).

We can impose certain desirable properties on the function $g(y, \theta)$ rather than on the resulting estimator $\widehat{\theta}$. For example, $g(.)$ is unbiased if $E[g(y, \theta)] = 0$; $g(.)$ is a minimum variance unbiased (MVU) EF if $Var[g(y, \theta)]$ is minimum among all unbiased estimating functions (EFs). At the outset, the benefits of focusing on the EFs rather than on the estimators are not so immediate. Following Durbin (1960) let us consider the first-order autoregressive ($AR$) model:

$$y_t = \theta y_{t-1} + u_t; \qquad u_t \sim IID(0, \sigma^2), \qquad t = 1, \ldots, n. \tag{1}$$

In the context of least squares (LS) estimation, we can focus our attention on three quantities. First, the objective function to be minimized with respect to $\theta$, namely

$$\min_{\theta} \sum_{t=2}^{n} (y_t - \theta y_{t-1})^2. \tag{2}$$

Second, the equation for solving the optimization problem,

$$g(y, \theta) = \sum_{t=2}^{n} y_t y_{t-1} - \theta \sum_{t=2}^{n} y_{t-1}^2 = 0, \tag{3}$$

and finally, the estimator $\widehat{\theta}$, itself,

$$\hat{\theta} = \sum_{t=2}^{n} y_t y_{t-1} / \sum_{t=2}^{n} y_{t-1}^2. \tag{4}$$

A major part of the estimation literature is concerned with the properties of the estimators, like $\widehat{\theta}$ such as unbiasedness, consistency and efficiency. The robust approaches to estimation emphasize the objective function, for example another legitimate function that we can minimize is $\sum_{t=2}^{n} |y_t - \theta y_{t-1}|$. However, the function (2) has an extra appeal of being the same objective function under the ML framework with normality assumptions on the errors $u_t$. Durbin (1960) observed that viewing the LS estimators as roots of certain equations such as (3), i.e., working with the first order conditions directly, is much more convenient than studying the objective function like (2) or the estimator $\widehat{\theta}$ in (4). The function $g(y, \theta)$ in (3) is linear in the parameter $\theta$ and $E[g(y, \theta)] = 0$. Durbin (1960) termed $g(y, \widehat{\theta}) = 0$ as *linear unbiased estimating equation*. This is a *finite sample* characterization of the EF $g(y, \theta)$, and it is clear that we cannot attach desirable properties, such as linearity and unbiasedness, to the resulting estimator $\widehat{\theta}$ in (4). Also, as we know, the standard ML estimator (MLE) emphasizes *asymptotic* efficiency rather than finite-sample properties. This simple example illustrates the benefits of focussing on the EFs rather than on the

estimators. As we will see later, many of the standard methods of estimation, such as LS, ML, minimum $\chi^2$ and M-estimation, can be considered as special cases of the EF approach.

The literature on estimating functions and equations is indeed very vast. There are quite a few survey articles, such as, Desmond (1989), Heyde (1989), Bhat (1990), Godambe and Kale (1991), Liang and Zeger (1995), Naik-Nimbalkar (1996), Vinod (1998) and Kale (2001-2002). Several books and edited volumes are also devoted on this subject, for example, see, McLeish and Small (1988), Godambe (1991a), Chen (1992), Basawa, Godambe and Taylor (1997), Heyde (1997) and Mukhopadhyay (2004). However, these papers and books do not cover the very early developments on estimating functions and equations, as discussed below. Also, apart from describing the formal theoretical progress as done in this and the following sections, another aim of this review paper is to explore the usefulness of this estimation technique to econometrics. We attempt to do that in Section 3. While providing the narrative details on some key theoretical developments, we also try to offer some personal perspectives by adding a human element to our narration. It is our experience that students take a greater interest in a subject when they clearly see the historical progress and know more about the personalities involved. The overall aim of this chapter is quite modest; our main purpose is to provide an easy-to-access description of EF approach and its potential applications to econometrics, to attract students' attention to this fascinating research area.

## 1.3 The origin of (optimal) estimating equation/ function and some surprising findings

The idea of using EFs or equations goes a long way back, at least to Karl Pearson's (1894) introduction of the method of moments. To the best of our knowledge, the term "estimating equation" was first used by Yule (1902, p.197). It, however, referred to estimated linear regressions like, $\widehat{y}_i = x_i'\widehat{\beta}$, using the popular notation. Therefore, it is quite different from our notion of an EE or EF. When one reads the literature on EF, the presented history appears to be very unambiguous. With minor reference to Fisher (1935a) and Kimball (1946) in terms of terminology and concepts, the whole development appears to starts from Durbin (1960) and Godambe (1960). However, this temporal clarity veils some of the much earlier, though disconnected, developments. Here we make an attempt to record those historical developments. Of course, it is quite possible that, we still miss certain important work.

Although, quite justifiably, R. A. Fisher [Fisher (1912)] is credited with inventing the ML method of estimation, as we all know, nothing under the sun is completely

new. ML estimation pre-figured many times in earlier works, such as, Edgeworth (1908, 1909) [see Bera and Bilias (2002, fn 9)]. Edgeworth (1909) is a continuation of Edgeworth (1908), where he attempts to prove a Cramér-Rao type inequality, more specifically to show that the posterior mode has the smallest variance. The treatment of 1909 article is more ambitious and the set up is quite general [for an illuminating exposition, that we follow, see Hald (1998, pp.703-705)]. Edgeworth (1909, p.82) stated his objective as, to "determine that function of which the several values, each formed from a large set of observations, hover with minimum dispersion about the true value of some constant represented by a symmetrical function of the observations." He considered the location model with "law of frequency" (probability density function) $f(y-\theta)$ and a class of functions defined by the equation

$$\sum_{i=1}^{n} h(y_i - \widehat{\theta}) = 0, \tag{5}$$

where $h$ is an arbitrary function satisfying $E[h(y-\theta)] = 0$ and the derivative of $h$ at zero, $h'(0) \neq 0$. To approximate the "error" in estimation $e = \widehat{\theta} - \theta$, let us write $h(y_i - \widehat{\theta}) = h(z_i - e)$, where $z_i = y_i - \theta$. Now a Taylor expansion gives

$$0 = \sum_{i=1}^{n} h(y_i - \widehat{\theta}) = \sum_{i=1}^{n} h(z_i - e) = \sum_{i=1}^{n} h(z_i) - e \sum_{i=1}^{n} h'(z_i) + \ldots, \tag{6}$$

and hence a first approximation to the error term $e$ is

$$e = \frac{\sum_{i=1}^{n} h(z_i)}{\sum_{i=1}^{n} h'(z_i)}. \tag{7}$$

Replacing sums by integrals, Edgeworth obtained the asymptotic fluctuations (variance) of $e$ as

$$Var(e) = \frac{1}{n} \frac{\int h^2(y) f(z) dz}{[\int h'(y) f(z) dz]^2} = \frac{1}{n} \frac{P^2}{Q^2}, \quad say, \tag{8}$$

where $f(z)$ denotes the probability density function. His objective was to find the function "$h$" such that $P^2/Q^2$ is minimum. A minimum is secured if $h$ is such that, when it receives an arbitrary variation ($\delta h$), the first term of variation vanishes and the second term is positive. Using Schwartz's inequality, Edgeworth (1909, p.84) proved the positivity of the second term. Let us concentrate only on the first term of the variation obtained by putting $h + (\delta h)$ in place of $h$. Using a simplified notation, we have

$$\begin{aligned}\frac{\int [h+(\delta h)]^2 f dz}{\left[\int [h'+(\delta h')] f dz\right]^2} &= \frac{\int h^2 f dz + 2\int (\delta h) h f dz + \int (\delta h)^2 f dz}{\left[\int h' f dz\right]^2 + 2\int h' f dz \int (\delta h') f dz + \left[\int (\delta h') f dz\right]^2} \\ &= \frac{P^2\left[1+2P^{-2}\int (\delta h) h f dz + \ldots\right]}{Q^2\left[1+2Q^{-1}\int (\delta h') f dz + \ldots\right]} \\ &= \frac{P^2}{Q^2}\left[1+2P^{-2}\int (\delta h) h f dz + \ldots\right]\left[1+2Q^{-1}\int (\delta h') f dz + \ldots\right]^{-1}. \end{aligned} \tag{9}$$

Since $\int (\delta h') f dz = (\delta h) f]_{-\infty}^{\infty} - \int (\delta h) f' dz = -\int (\delta h) f' dz$, an approximation to (9), is given by

$$\frac{P^2}{Q^2}\left[1+2P^{-2}\int (\delta h) h f dz + 2Q^{-1}\int (\delta h) f' dz\right] = \frac{P^2}{Q^2}\left[1+2\int (\delta h)\left\{P^{-2} h f + Q^{-1} f'\right\} dz\right]. \tag{10}$$

Therefore, a necessary condition for minimizing $Var(e)$ in (8), is that the expression in (10) is zero , or equivalently,

$$P^{-2} h f + Q^{-1} f' = 0$$

$$i.e., \qquad h(z) = -\frac{P^2}{Q}.\frac{f'(z)}{f(z)} = c\frac{f'(z)}{f(z)} = c\frac{d\log f(z)}{dz}, \tag{11}$$

where $c$ is a constant. In other words, for minimum variance $h(z)$ should be proportional to the score function. At this optimum $h(z)$,

$$P^2 = \int c^2 \left(\frac{f'}{f}\right)^2 f dz = c^2 \int \left(\frac{f'}{f}\right)^2 f dz = c^2 \mathcal{I},$$

$$Q = \int h' f dz = h f]_{-\infty}^{\infty} - \int f' h dz = 0 - \int f' c.\frac{f'}{f} dz = -c\int \left(\frac{f'}{f}\right)^2 f dz = -c\mathcal{I}, \tag{12}$$

where $\mathcal{I} = \int \left(\frac{f'}{f}\right)^2 f dz = E\left[\left(\frac{f'}{f}\right)^2\right]$ is the standard Fisher's information on each observation. From (8), $Var(e)$ reduces to

$$\frac{1}{n}\frac{P^2}{Q^2} = (n\mathcal{I})^{-1}. \tag{13}$$

Therefore, the asymptotic variance of $\widehat{\theta}$, obtained from the optimal EF, reaches the Cramér-Rao lower bound (CRLB).

Edgeworth did not proceed further with his optimal "estimating function" ap-

proach. He was more interested in proving that the posterior mode (which is the same function of the observations as the ML estimate) has the smallest asymptotic sample variance. Historically, Edgeworth's work has been treated as precursor to Fisher's (1912, 1922) work on the ML method. However, now we can see that it is much more than that. It has the fundamental result of the EF approach. Of course, Edgeworth did not grasp the far reaching implication of his result, and neither did R. A. Fisher (a rare occasion), in the context of his article Fisher (1935a). As noted by several researchers [see for example, Desmond (1989, p.57)] the term "equation of estimation" with its current meaning, first appeared in Fisher (1935a, p.45). What is overlooked in the EF literature is another of Fisher's pathbreaking results which we will discuss shortly. The result is the same as Edgeworth's, but Fisher did it quite elegantly and under a more general set up.

To maintain a chronological order, we now present yet another result by Fisher that may be regarded as the first substantial illustration on the use of EEs. Fisher (1924) wanted to compare ML and minimum $\chi^2$ as methods of estimation. He simply showed that they are asymptotically equivalent by comparing the first order conditions of the two estimation procedures. For him, it was much easier to analyze properties of estimators when focusing on the corresponding EEs rather than on the objective functions or estimators themselves. The same is possibly true even after eight decades.

To illustrate, let us consider the minimum $\chi^2$ objective function,

$$\chi^2(\theta) = \sum_{j=1}^{k} \frac{[n_j - nq_j(\theta)]^2}{nq_j(\theta)}, \tag{14}$$

where $n_j$ is the observed frequency, and $q_j(\theta)$ is the probability of being in the $j$-th class, $j = 1, 2, \ldots, k$ with $\theta = (\theta_1, \theta_2, \ldots, \theta_p)'$ as the unknown parameter vector. Let $n = \sum_{j=1}^{n} n_j$ ; thus $nq_j(\theta)$ is the expected frequency of the $j$-th class. We can write

$$\chi^2(\theta) = \sum_{j=1}^{k} \frac{n_j^2}{nq_j(\theta)} - n.$$

Therefore, the minimum $\chi^2$ estimates will be obtained by solving $\partial\chi^2(\theta)/\partial\theta = 0$, i.e., from

$$\sum_{j=1}^{k} \frac{n_j^2}{[nq_j(\theta)]^2} \frac{\partial q_j(\theta)}{\partial\theta_l} = 0, \qquad l = 1, 2, \ldots, p. \tag{15}$$

To connect these equations to those from Fisher's (1912) ML equations, we note that, since $\sum_{j=1}^{k} q_j(\theta) = 1$, we have $\sum_{j=1}^{k} \partial q_j(\theta)/\partial\theta_l = 0$. Therefore, from (15), the

minimum $\chi^2$ EEs are

$$\sum_{j=1}^{k} \frac{n_j^2 - [nq_j(\theta)]^2}{[nq_j(\theta)]^2} \frac{\partial q_j(\theta)}{\partial \theta_l} = 0, \qquad l = 1, 2, \ldots, p. \tag{16}$$

Under the multinomial framework, Fisher's likelihood function, denoted as $L(\theta)$ is

$$L(\theta) = n! \prod_{j=1}^{k} [(n_j!)^{-1}] \prod_{j=1}^{k} [q_j(\theta)]^{n_j}.$$

Therefore, the log-likelihood function, denoted by $\ell(\theta)$, can be written as

$$\log L(\theta) = \ell(\theta) = \text{ constant } + \sum_{j=1}^{k} n_j \log q_j(\theta).$$

The corresponding ML EEs are $\partial \ell(\theta)/\partial \theta = 0$, i.e.,

$$\sum_{j=1}^{k} \frac{n_j}{q_j(\theta)} \frac{\partial q_j(\theta)}{\partial \theta_l} = 0,$$

or, equivalently,

$$\sum_{j=1}^{k} \frac{[n_j - nq_j(\theta)]}{nq_j(\theta)} \frac{\partial q_j(\theta)}{\partial \theta_l} = 0, \qquad l = 1, 2, \ldots, p. \tag{17}$$

Fisher (1924) argued that the difference between (16) and (17) is of the factor $[n_j + nq_j(\theta)]/nq_j(\theta)$, which tends to the value 2 for large values of $n$ and, therefore, these two methods are asymptotically equivalent. The point we want to emphasize is that to compare estimates from two different methods, Fisher (1924) used the "estimating equations" rather than the estimates themselves. Here let us mention that, although the two EEs (16) and (17) are asymptotically equivalent, there is a fundamental difference. Since $E(n_j) = nq_j(\theta)$ and $E(n_j^2) = nq_j(\theta)[1 - q_j(\theta)] + [nq_j(\theta)]^2$, the EFs corresponding to the minimum $\chi^2$ method are not unbiased, while the EFs for the ML method are. As we will discuss later, unbiasedness of the EF is a very important requirement. Of course, unbiased EF may not lead to unbiased estimator.

Now getting back to Fisher (1935a), for ease of exposition, we replace Fisher's "summation" sign by an integral. Fisher (1935a, p.45) started with an unbiased EF $k(y, \theta)$. Differentiating

$$E[k(y, \theta)] = \int k(y, \theta) f(y, \theta) dy = 0, \tag{18}$$

where $f(y,\theta)$ denotes the density function. Fisher obtained

$$\int \frac{dk(y,\theta)}{d\theta} f(y,\theta)dy = -\int k(y,\theta)\frac{df(y,\theta)}{d\theta}dy. \tag{19}$$

A Taylor series expansion of the sample equation of estimation $\sum_{i=1}^{n} k(y_i,\widehat{\theta}) = 0$, around $\theta$ gives

$$0 = \sum_{i=1}^{n} k(y_i,\theta) + (\widehat{\theta}-\theta)\sum_{i=1}^{n}\frac{dk(y_i,\theta)}{d\theta} + \ldots, \tag{20}$$

i.e., approximately,

$$(\widehat{\theta}-\theta) = -\frac{\sum_{i=1}^{n} k(y_i,\theta)}{\sum_{i=1}^{n}\frac{dk(y_i,\theta)}{d\theta}}. \tag{21}$$

Hence, using (19), the asymptotic variance of $\widehat{\theta}$ is given by

$$\begin{aligned} Var(\widehat{\theta}) &= \frac{\int k^2(y,\theta)f(y,\theta)dy}{n\left[\int \frac{dk(y,\theta)}{d\theta}f(y,\theta)dy\right]^2} \\ &= \frac{\int k^2(y,\theta)f(y,\theta)dy}{n\left[\int k(y,\theta)\frac{df(y,\theta)}{d\theta}dy\right]^2}. \end{aligned} \tag{22}$$

This is same as Edgeworth's equation as given in (8). After obtaining the expression (22), Fisher (1935a, p.46) stated, "We may now apply the calculus of variations or simple differentiation to find the functions of $k$, which will minimize the sampling variance. Since the variance must be stationary for variations of each several values of $k$, the differential coefficients of the numerator and the denominator with respect to $k$, must be proportional for all classes." Thus for the "optimal values of $k$," he obtained

$$k(y,\theta)f(y,\theta) \propto \frac{df(y,\theta)}{d\theta}, \tag{23}$$

which is satisfied by putting

$$k(y,\theta) = \frac{1}{f(y,\theta)}\frac{df(y,\theta)}{d\theta} = \frac{d\log f(y,\theta)}{d\theta}. \tag{24}$$

Fisher then noted that $E[k(y,\theta)] = 0$ is the ML equation, and at the optimum value of $k(y,\theta)$ in (24), the asymptotic variance in (22) reduces to

$$Var(\widehat{\theta}) = \frac{1}{n\left[\int \left(\frac{d\log f(y,\theta)}{d\theta}\right)^2 f(y,\theta)dy\right]} = \frac{1}{n\mathcal{I}(\theta)}, \quad say. \tag{25}$$

Fisher (1935a, p.44) defined $\mathcal{I}(\theta)$ as the amount of information supplied by each observation.

## 1.4 Asymptotically shortest confidence interval using optimal estimating function

Fisher (1935b) developed a theory of fiducial inference by considering a function, say, $g(y,\theta)$ which is pivotal, i.e., its distribution is free of $\theta$. Wilks (1938) utilized this approach for interval estimation [see also Wilks (1962, pp.371-376)]. He considered the pivotal function

$$\sqrt{n}p(y,\theta) = \frac{\sum_{i=1}^n S(y_i,\theta)}{\left[\frac{1}{n}\sum_{i=1}^n S^2(y_i,\theta)\right]^{1/2}}, \tag{26}$$

where $S(y_i,\theta) = \frac{d\log f(y_i,\theta)}{d\theta}$. Under certain regularity condition $\sqrt{n}p(y,\theta) \to^D N(0,1)$, and, hence, asymptotically it is pivotal. Now, denoting $\widehat{\theta}$ as MLE,

$$0 = p(y,\widehat{\theta}) = p(y,\theta) + p'(y,\theta)(\widehat{\theta}-\theta) + \ldots, \tag{27}$$

where $p'(y,\theta) = dp(y,\theta)/d\theta$. Therefore, $\sqrt{n}(\theta-\widehat{\theta})p'(y,\theta)$ is asymptotically equivalent to $\sqrt{n}p(y,\theta)$ and hence distributed as $N(0,1)$ for large enough $n$. Utilizing this result, Wilks (1938) obtained the $(1-\alpha)100\%$ confidence interval for $\theta$ as

$$\lim_{n\to\infty} \Pr\left[-Z_{\alpha/2} \le \sqrt{n}(\theta-\widehat{\theta})p'(y,\theta) \le Z_{\alpha/2}\right] = 1-\alpha, \tag{28}$$

where $Z_{\alpha/2}$ is the upper $\alpha/2$-quantile of the standard normal distribution. Thus we have

$$\lim_{n\to\infty} \Pr\left[\widehat{\theta} - \frac{Z_{\alpha/2}}{\sqrt{n}p'(y,\theta)} \le \theta \le \widehat{\theta} + \frac{Z_{\alpha/2}}{\sqrt{n}p'(y,\theta)}\right] = 1-\alpha. \tag{29}$$

Wilks (1938) further showed that, under certain regularity conditions, the ratio of the squared length of this interval to that of a similar interval using any arbitrary EF, converges in probability to a number that cannot exceed 1. In other words, the asymptotically shortest confidence intervals results when the pivotal function is constructed from the score function $S(y,\theta)$ as in (26). Wald (1942) obtained the same result under a more general framework. Barnard (1973) further explored the advantages of Fisher's approach of formulating the parameter estimation problem in

terms of pivotal quantities.

## 1.5 Sufficient statistical estimating function

As we mentioned in Section 1.1, Fisher (1922) suggested three important criteria of estimation, namely, consistency, efficiency and sufficiency; and of these three, he found the concept of "sufficiency" to be most powerful to advance his ideas on ML estimation. He defined "sufficiency" as (p.310): "A statistic satisfies the criterion of sufficiency when no other statistic which can be calculated from the same sample provides any additional information as to the value of the parameter to be estimated." However, as it is now well known, there are certain distributions for which it is not possible to find nontrivial sufficient statistic(s) for the underlying parameter(s). Kimball (1946) worked with the extreme-value distribution with density function

$$f(y;\theta) = \alpha e^{-\alpha(y-u)} e^{-e^{-\alpha(y-u)}}, \tag{30}$$

where $\alpha$ and $u$ denote parameters. Kimball found the ordinary definition of sufficiency to be inadequate for this distribution; however, he (p.299) "was struck by the fact that certain functions of the data involving one of the parameters could be used to play a very similar role to a set of *sufficient statistic* for determining $\alpha$ and $u$, in spite of the fact that one function involved the value of $\alpha$, and hence was not directly determined by the data, - and hence not a 'statistic'." He argued for a broader definition of sufficiency and introduced a new terminology (p.300) that of the "statistical estimating function." Possibly, this was the first occurrence of the term in the sense currently used in the literature. Kimball (1946), however, acknowledged Wald (1940) who stated (p.290, fn 13), "An 'estimate' is usually a function of the observations not involving any unknown parameters. We designate here as estimates also some functions involving the parameter $\alpha$."

Rao (1945, p.81) also used the term "estimating function" as: "The validity of this (ML) principle arises from the fact that out of a large class of unbiased estimating functions following the normal distribution the function given by maximizing the probability density has the least variance." We see from the context that he essentially meant "estimating function" of sample $y = (y_1, y_2, \ldots, y_n)$ only. However, if we consider Rao's sentence "out of context," he might as well be stating that ML method is based on optimal EF! Kimball (1946) also introduced the concept of a "stable" EF as the one whose *expectation* is constant in the parameter. In the context of errors-in-variables models, Kendall (1951) introduced an "unbiased" EE which led to an biased estimator. Kendall's concept of unbiasedness is very close to that of stability used by Kimball (1946). Kendall (1951, p.21) emphasized, "We must draw

a distinction between an unbiased estimator and an unbiased estimating equation."

Suppose the density function involves $p$ parameters $\theta = (\theta_1, \theta_2, \ldots, \theta_p)'$, and we have a sample $y = (y_1, y_2, \ldots, y_n)$. Kimball (1946, p.302) defined a set statistical functions $g_1(y, \theta), g_2(y, \theta), \ldots, g_p(y, \theta)$ to be sufficient estimating functions (SEFs) if

**(i)** There is a one-to-one correspondence between $(g_1, g_2, \ldots, g_p)$ and $(\theta_1, \theta_2, \ldots, \theta_p)$.

**(ii)** It is possible to express the joint distribution (likelihood function) $f(y_1, y_2, \ldots, y_n; \theta)$ as

$$f(y; \theta) = f(y_1, y_2, \ldots, y_n; \theta) = f_1(g_1, g_2, \ldots, g_p; \theta) f_2(y_1, y_2, \ldots, y_n),$$

where the first factor is purely function of the EFs and parameters, and the second factor is free of the parameters.

Clearly, the requirement (ii) is along the lines of Neyman-Fisher factorization. To illustrate his approach, Kimball considered a distribution with two parameters $\theta_1$ and $\theta_2$, and claimed that the score functions

$$S_{\theta_1}(y; \theta_1, \theta_2) = \frac{\partial \log f(y; \theta_1, \theta_2)}{\partial \theta_1} \tag{31}$$

and

$$S_{\theta_2}(y; \theta_1, \theta_2) = \frac{\partial \log f(y; \theta_1, \theta_2)}{\partial \theta_2} \tag{32}$$

are SEFs according to the above definition. To see this, note that $\log f(y; \theta_1, \theta_2)$ can be expressed as

$$\log f(y; \theta_1, \theta_2) = \int_{\theta_1^0}^{\theta_1} S_{\theta_1}(y; \theta_1, \theta_2) d\theta_1 + \int_{\theta_2^0}^{\theta_2} S_{\theta_2}(y; \theta_1, \theta_2) d\theta_2 + \log f(y; \theta_1^0, \theta_2^0), \tag{33}$$

where $\theta_1^0$ and $\theta_2^0$ are arbitrarily chosen from the parameter space. The first two term in (33) entirely depends on scores $S_{\theta_1}, S_{\theta_2}$ (along with the parameters) and while the third term is free of $\theta_1$ and $\theta_2$.

For the extreme value distribution in (30), Kimball (1946, p.304) showed that

$$g_1(y; \theta) = [\alpha(\overline{y} - u) - C], \tag{34}$$

$$g_2(y; \theta) = \left[\frac{\overline{z}}{z_0} - 1\right], \tag{35}$$

where $C = E[\alpha(\overline{y} - u)]$, $z_i = \exp[-\alpha y_i]$ with mean $\overline{z}$, $z_0 = \exp[-\alpha u]$, are SEFs. Using 57 years of maximum flood data, Kimball (1946) estimated the parameters $\alpha$ and $u$ based on EFs $g_1(y;\theta)$ and $g_2(y;\theta)$, and compared them with the ML estimates. Although, Kimball's approach was very novel, it was not followed up much in the later literature on EF, though Kale (1962) connected sufficiency to the extended CRLB, and Bhapkar (1991) argued that for any given EF, a sufficient statistic can be used to derive a more informative EF. McLeish and Small (1988, Ch.2) discussed ancillarity, sufficiency and projection in the context of EFs and advocated that sufficiency for EFs should be developed in its own right.

To summarize, in the first half of the last century, we notice some very important but rather sporadic and disconnected progress in the EF approach. Though the criteria unbiasedness and sufficiency have been thought of as requirements for an EF, what was missing from all these developments is any notion of *optimality* of the EF. The topic was almost forgotten for several years. It was then rekindled with the appearance of V. P. Godambe's seminal article in 1960, the essence of Godambe (1960) being the introduction of an "optimality" criteria in addition to unbiasedness. This is very much akin to Neyman and Pearson's (1933) theory of hypothesis testing, where they introduced the concept of optimality (through maximization of power) to the earlier somewhat ad hoc significance and likelihood ratio tests. Godambe (1960) introduced optimality through the minimization of the variance of "unbiased estimating functions" for independent samples while, Durbin (1960) did it mainly for the linear unbiased EF for dependent data in the context of AR time series model.

# 2 Basic Theory of Estimating Functions

Fisher (1935a) noted a basic fact of estimation that any procedure for obtaining an estimate of parameter $\theta$ can be regarded as a solution to an equation, like

$$g(y;\theta) = 0, \tag{36}$$

where $g(y;\theta)$ is a function of the observation vector $y = (y_1, y_2, \ldots, y_n)'$ and parameter $\theta$. The traditional approach to estimation imposes conditions on the resulting estimator $\widehat{\theta}$, such as linearity, unbiasedness, consistency, invariance, minimum variance etc. The EF approach shifts the attention from the estimator $\widehat{\theta}$ to the properties of the EF. For example, we will consider an *unbiased* EF instead of an unbiased $\widehat{\theta}$, i.e., we will require

$$E[g(y;\theta)] = 0. \tag{37}$$

The notion of unbiasedness of an EF is an extension of that of an estimator, and it ensures that the root of the equation (36) is close to the true value of the parameter $\theta$ when little random variation is present. When $g(y;\theta)$ has a special form, for instance, $g(y;\theta) = g(y) - \theta$, then $\widehat{\theta} = g(y)$ and an unbiased EF leads to an unbiased estimator. However, in general, the requirement (37) does not necessarily imply unbiasedness of the resulting estimator, though under certain regularity conditions it does imply consistency of the estimator [see Desmond (1997, p.80)]. For more on the role and importance of unbiasedness in EFs, see Yanagimoto and Yamamoto (1991, 1993).

As we discussed in Section 1, the importance of the role of unbiasedness and sufficiency [as in Kimball (1946)] was well recognized. The missing element was a criterion of optimality. Durbin (1960, p.146) stated, "it seems reasonable to develop the idea of unbiased estimating equations with minimum variance" and exploited this idea to derive optimal linear unbiased EFs, reminiscent of the Gauss-Markov theorem. Around the same time, Godambe (1960) started with a class of EFs satisfying certain conditions, which he called regular EFs and devised a procedure to select an optimal EF.

## 2.1 The fundamental result: Godambe (1960) and Durbin (1960)

Godambe's (1960) regular EF $g(y;\theta)$ satisfies the following conditions:

**(i)** $E[g(y;\theta)] = \int g(y;\theta) f(y;\theta) dy = 0$,

**(ii)** $\frac{dg(y;\theta)}{d\theta}$ exists for all $\theta \in \Theta$, where $\Theta$ is the parameter space,

**(iii)** $\int g(y;\theta) f(y;\theta) dy$ is differentiable under the sign of integration,

**(iv)** $E\left[\frac{dg(y;\theta)}{d\theta}\right]^2 > 0$, for all $\theta \in \Theta$,

**(v)** $Var[g(y;\theta)] = E[g^2(y;\theta)] < \infty$.

Godambe (1960) also assumed that the likelihood function $f(y;\theta) = \Pi_{i=1}^{n} f(y_i;\theta)$ satisfies the regularity conditions required for establishing the CRLB. For ease of exposition, we now consider the *scalar* parameter case; EF for the multiparameter case and the presence of nuisance parameters will be discussed in Sections 2.2 and 2.3, respectively. Let $\mathcal{G}$ denotes the class of all regular EFs.

**Definition 2.1:** *A $g^* \in \mathcal{G}$ is said to be optimal if*

$$\frac{E[g^{*2}(y;\theta)]}{\left\{E\left[\frac{dg^*(y;\theta)}{d\theta}\right]\right\}^2} \leq \frac{E[g^2(y;\theta)]}{\left\{E\left[\frac{dg(y;\theta)}{d\theta}\right]\right\}^2}, \tag{38}$$

*for all* $g \in \mathcal{G}$ *and* $\theta \in \Theta$.

Godambe's (1960) justification for this criterion is as follows. First, it is desirable that $g(y;\theta)$ is as close as possible to zero when evaluated at the true value of $\theta$, i.e., we should minimize $Var[g(y;\theta)] = E[g^2(y;\theta)]$, and hence we should have

$$E[g^{*2}(y;\theta)] \leq E[g^2(y;\theta)]. \tag{39}$$

Second, $g(y;\theta+\delta\theta)$ should differ from $E[g(y,\theta)] = 0$ by as large quantity as possible. This is a kind of "sensitivity" requirement, which can also be viewed as an "identification" condition. This translates as $\{E[dg(y;\theta)/d\theta]\}^2$ should be as large as possible, i.e.,

$$\{E[dg^*(y;\theta)/d\theta]\}^2 \geq \{E[dg(y;\theta)/d\theta]\}^2. \tag{40}$$

These two goals (39) and (40) can be accomplished simultaneously by Godambe's criterion in (38). Now we can state and prove Godambe's celebrated result.

**Theorem 2.1:** *For all* $g \in \mathcal{G}$,

$$\frac{E[g^2(y;\theta)]}{\left\{E\left[\frac{dg(y;\theta)}{d\theta}\right]\right\}^2} \geq \frac{1}{E\left[\frac{d\log f(y;\theta)}{d\theta}\right]^2}, \tag{41}$$

*and the equality is attained by the EF* $g^*(y;\theta) = d\log f(y;\theta)/d\theta$.

Here with a slight change of notation we denote $E[d\log f(y;\theta)/d\theta]^2 = n\mathcal{I}(\theta)$ as the Fisher's information contained in the whole sample $y = (y_1, y_2, \ldots, y_n)$. The proof of Theorem 2.1 is very similar to that of the CRLB.

***Proof:*** Differentiating the unbiasedness condition (37) with respect to $\theta$, we obtain

$$\int \frac{dg}{d\theta} f dy + \int g \frac{d\log f}{d\theta} f dy = 0,$$

i.e.,

$$E\left[\frac{dg}{d\theta}\right] = -Cov\left[g, \frac{d\log f}{d\theta}\right]. \tag{42}$$

Here we suppress the arguments of the functions $g(y;\theta)$ and $f(y;\theta)$ for ease of notation. Since

$$\left\{Cov\left[g, \frac{d\log f}{d\theta}\right]\right\}^2 \leq Var[g]Var\left[\frac{d\log f}{d\theta}\right],$$

using (42) we have,

$$\left\{E\left[\frac{dg}{d\theta}\right]\right\}^2 \leq Var[g]E\left[\frac{d\log f}{d\theta}\right]^2$$

and the result follows immediately.

This result was also mentioned by Durbin (1960, p.145), and he acknowledged G. A. Barnard for suggesting "extension to non-linear estimating equations" from his linear EFs. Godambe (1960) was also aware of this as he stated (p.1210): "The author acknowledges with pleasure that G. A. Barnard communicated to the Royal Statistical Society, London, a result similar to the preceding theorem, independently, and at nearly the same time when the paper was written," and he made a reference to Durbin (1960, p.415). Godambe's manuscript was received by the *Annals of Mathematical Statistics* on July 28, 1959, and the revised version on May 17, 1960. Durbin's paper, most possibly the final version, was received by the *Journal of the Royal Statistical Society* on August 1959. It is quite a coincidence that Godambe and Durbin reported "similar" (in fact the same) result "at nearly the same time." To put this result in an historical context, let us recall that Rao (1945) and Crámer (1946) provided the *finite sample* version of Fisher's result that the *asymptotic* variance of a consistent estimator is bounded below by the reciprocal of Fisher information measure. We can view the Godambe-Durbin result as the finite sample version of the Edgeworth (1909) and Fisher (1935a) result [noted in equation (11) and (24)] that asymptotically the score function $d\log f(y;\theta)/d\theta$ is the optimum EF. Therefore, from the Godambe-Durbin result, for the first time we have a *finite sample justification* of the ML method of estimation.

Durbin is well known among econometricians, starting from his celebrated Durbin-Watson test statistics for serial correlation. An account of Durbin's life and work is also available from the ET interview [see Phillips (1988)]. However, Godambe (and his work) is somewhat unfamiliar to econometricians. The only references to his work on EF that we find in econometric textbooks, are in Mittelhammer, Judge and Miller (2000, Ch.11) and Davidson and MacKinnon (2004, pp.369-372). It would not be, therefore, out of place to add a few sentences on V. P. Godambe. Bellhouse (1992) provides a short but illuminating discussion of his life and time. The Statistical Science interview [Thompson (2002)] gives further insights on his work and views on statistical methodologies. Vidyadhar (which means "bearer of wisdom") Godambe was born on June 1, 1926 at Poona, India. He studied sanskrit, philosophy, theoretical physics and mathematics during his undergraduate years. After obtaining his M.Sc. degree (the first batch) from the University of Bombay in 1950, he joined the Bureau of Economics and Statistics in the Government of Bombay (the current state of Maharastra, India). He received his Ph.D from the Imperial College, University

of London under the supervision of George Barnard. He spent a year (1958-59) as Senior Research Fellow at the Indian Statistical Institute, Calcutta, and the seminal 1960 paper was written there. In 1967, just as the Department of Statistics and Actuarial Science at the University of Waterloo, Canada, was being formed, he joined that department. Upon his retirement in 1991, he was awarded the title of Distinguished Professor Emeritus at the same University.

Godambe's 1960 paper, which is just a little over three pages, appears to be well ahead of its time (though in historical context it can be argued to be long overdue given the results of Edgeworth (1909), Fisher (1935a) and CRLB). In Thompson (2002, p.460), Godambe traced his idea on EF way back to 1948, as he recounted: "When the conventional theory of unbiased minimum variance estimation was introduced to me in 1948, my immediate reaction was that 'modal unbiasedness' rather than 'mean unbiasedness,' was a desirable property for an estimate. And from among all the modally unbiased estimates, one should choose the estimate whose distribution has maximum probability at the mode for all parameter values." Godambe (1960) was not "noticed" by others for a long time; it had only *two* citations (excluding Godambe's own) during the period 1961-75, 37 citations during 1976-90 and around 165 during the last fifteen years. This paper played a central role in introducing, crystallizing newer concepts and advancing the EF approach to a full fledged area of research on its own right. It also foretold what to expect from Godambe in terms of his own contribution. Godambe confined his research to survey sampling during much of 1960s; then in the 1970s, he began a fruitful research collaboration on EF with his colleague (Mary) Thompson and, as we will see later, that resulted in a series of important papers. There was also an external factor - (George) Barnard delivered a series of lectures at the University of Waterloo during the academic year 1972-73. As Bellhouse (1992, p.4) noted: "For Godambe the lectures stimulated him to return to the problems of inference using estimating functions or estimating equations. His first results on the theory of optimal estimating functions in the presence of nuisance parameters were obtained with Mary Thompson in an Annals paper in 1974." (James) Durbin did not continue his research on EF that vigorously, and we are aware of only one more published paper on EF by him [see Durbin (1997)].

After somewhat long digression on some personal narration, let us now return to the Godambe-Durbin optimality result. One of the attractive properties of the MLE is that it is invariant under a one-to-one transformation of the parameter, i.e., if $\widehat{\theta}$ is MLE of $\theta$, then MLE of $\varphi \equiv \alpha(\theta)$ with $J = d\alpha/d\theta \neq 0$, is given by $\widehat{\varphi} \equiv \alpha(\widehat{\theta})$. This is due to the fact that

$$\frac{d \log f(y;\theta)}{d\theta} = \frac{d \log f(y;\alpha^{-1}(\varphi))}{d\varphi} J. \tag{43}$$

Optimal EF shares this property of invariance. To see this note that if $g(y;\theta)$ is an unbiased EF for $\theta$, then $g(y;\alpha^{-1}(\varphi)) = g_1(y;\varphi)$ is an unbiased EF for $\varphi$. Let $\widehat{\theta}_g$ and $\widehat{\varphi}_{g_1}$ be the estimates from $g = 0$ and $g_1 = 0$, respectively. Then we have the invariance $\widehat{\varphi}_{g_1} = \alpha(\widehat{\theta}_g)$. Many good estimators, such as the MVU estimator, do not possess the property of invariance. Okuma (1976) provides a useful discussion on invariance of the EF from a different perspective.

There are several ways to represent and interpret the inequality (38). The equations $g(y;\theta) = 0$ and $\widetilde{g}(y;\theta) = cg(y;\theta) = 0$, where $c \neq 0$ is a constant, will lead to the same estimator, say, $\widehat{\theta}$. $Var[\widetilde{g}(y;\theta)] = c^2 Var[g(y;\theta)]$ can, however, be made arbitrarily small and thus the comparison of two EFs based on their variances alone is not meaningful without some standardization. The standardized version of $g \equiv g(y;\theta)$ is defined as

$$g_s = \frac{g}{E\left[\frac{dg}{d\theta}\right]}. \tag{44}$$

Thus we have,

$$Var[g_s] = Var[\widetilde{g}_s] = \frac{E\left[g^2\right]}{\left\{E\left[\frac{dg}{d\theta}\right]\right\}^2}. \tag{45}$$

The Godambe-Durbin optimality result can now be stated as: $g^*$ is optimal in class $\mathcal{G}$ if $g^* \in \mathcal{G}$ and if

$$Var[g_s^*] \leq Var[g_s], \quad \forall g \in \mathcal{G}. \tag{46}$$

The asymptotic properties of an estimator are inherited from the statistical behavior (e.g., variance) of the corresponding EF. A first-order Taylor series expansion of $g(\widehat{\theta}) = 0$ around $\theta$ [as in equation (20) in the context of Fisher (1935a)] gives us

$$\begin{aligned} n^{1/2}(\hat{\theta} - \theta) &\approx -n^{-1/2} g(\theta) \times \left(n^{-1}\frac{dg}{d\theta}\right)^{-1} \\ &\approx -n^{1/2} g(\theta) \times \left\{E\left[\frac{dg}{d\theta}\right]\right\}^{-1}, \end{aligned} \tag{47}$$

i.e., the estimator $\widehat{\theta}$ and the standardized EF $g_s$ in (44) are statistically equivalent asymptotically. Also, a measure of finite sample performance of $g(y;\theta)$ should not conflict with asymptotic properties of $\widehat{\theta}$. Therefore, in order to obtain an estimator with minimum limiting variance, the EF $g$ has to be chosen with minimum variance of its standardized form $g_s$.

Using the variance minimization criterion (46), as a basis for selecting the optimal EFs $g^*$, has some further implications [see, for instance, Bera and Bilias (2001a)]. First, consider the correlation between an unbiased EF$g$ for the parameter $\theta$ and the score function $S(\theta) = d\log f(y;\theta)/d\theta$. In view of the identity in (42) we can write :

$$\begin{aligned}
[Corr(g, S(\theta))]^2 &= \frac{\{E[g, S(\theta)]\}^2}{E[g^2]E[(S(\theta))^2]} \\
&= \frac{\{E[dg/d\theta]\}^2}{E[g^2]} \frac{1}{E[(S(\theta))^2]} \\
&= \frac{1}{Var[g_s]} \frac{1}{Var[S(\theta)]}. \qquad (48)
\end{aligned}$$

Therefore, choosing $g$ with the minimum variance of its standardized version is equivalent to maximizing the correlation of $g$ with the score function. Second, consider the $L_2$ distance of $g_s$ from the standardized score function $S_s(\theta)$. Upon noting that the variance of standardized score is $1/Var[S(\theta)]$ and using (42), we obtain

$$E\left[(g_s - S_s(\theta))^2\right] = Var[g_s] - \frac{1}{Var[S(\theta)]}. \qquad (49)$$

Thus, minimization of the variance of the $g_s$ is equivalent to minimizing the Euclidean distance of $g_s$ from the score function. The two results above certainly highlight the nature of the optimal EF as a best approximation to the score function, which, in general, is unknown.

Kale (1962) independently proved the result in (41), and called it an extension of the CRLB for the variance of an EF $g(y;\theta)$ instead of a statistic (which is a function of sample alone). He also proved that if the variance of $g(y;\theta)$ attains the lower bound given by the extended inequality, then $g(y;\theta)$ is a sufficient EF in the sense of Kimball (1946) (as discussed in Section 1.5). Kale (1962, p.82) expressed his result as

$$Var[g] \geq \frac{\left\{\frac{d\psi}{d\theta} - E\left[\frac{dg}{d\theta}\right]\right\}^2}{\mathcal{I}(\theta)}, \qquad (50)$$

where $E[g(y;\theta)] = \psi(\theta) = \psi$ and $\mathcal{I}(\theta)$ is the Fisher information in the whole sample, $y = (y_1, y_2, \ldots, y_n)$. He also noted that the score function $d\log f(y;\theta)/d\theta$ is a sufficient EF since it attains the extended CRLB. The extended inequality (50) reduces to the standard Cramér-Rao inequality by putting $g(y;\theta) = T(y) - \theta$ and writing

$$Var[T(y)] \geq \frac{1}{\mathcal{I}(\theta)}, \qquad (51)$$

where $T(y)$ is an unbiased estimator for $\theta$. We should however, note that, the CRLB is attained only exceptionally, whereas the optimality of ML equations among EEs holds merely under regularity conditions.

Bhapkar (1972) defined the information contained in an EF $g(y;\theta)$ about $\theta$ by the reciprocal of the variance of the standardized EF $g_s$, i.e.,

$$\mathcal{I}_g(\theta) = \frac{1}{Var[g_s]} = \frac{\left\{E\left[\frac{dg}{d\theta}\right]\right\}^2}{E[g^2]}, \tag{52}$$

and the ratio

$$RE(g) = \frac{\mathcal{I}_g(\theta)}{\mathcal{I}(\theta)}, \tag{53}$$

as the efficiency of the EF $g(y;\theta)$. Therefore, we can rewrite inequality (41) simply as

$$\mathcal{I}_g(\theta) \leq \mathcal{I}(\theta), \tag{54}$$

and hence,

$$RE(g) \leq 1, \tag{55}$$

for all $\theta \in \Theta$ and $g \in \mathcal{G}$. Therefore, $\mathcal{I}(\theta)$ is the maximum amount of information contained in a regular EF $g \in \mathcal{G}$. Let $T \equiv t(y)$ be sufficient for $\theta$ and define $\widetilde{g} \equiv \widetilde{g}(t(y);\theta) = E[g(y;\theta)|t]$, which is a Rao-Blackwellization of the original unbiased EF $g(y;\theta)$. Bhapkar (1972, p.469) showed that

$$\mathcal{I}_g(\theta) \leq \mathcal{I}_{\widetilde{g}}(\theta), \tag{56}$$

with equality iff $g(y;\theta) = \widetilde{g}(t(y);\theta)$. In other words, if we start with a EF that is already a function of the sufficient statistic $T$, there is no room for improvement. Also combining (54) and (56), it is easy to see that

$$\mathcal{I}_g(\theta) \leq \mathcal{I}_{\widetilde{g}}(\theta) \leq \mathcal{I}(\theta) = \mathcal{I}_{g^*}(\theta), \tag{57}$$

where $\mathcal{I}_{g^*}(\theta)$ denotes the information contained in optimal EF $g^*(y;\theta)$.

Wedderburn (1974) observed that from a computational point of view, the only assumptions on a generalized linear model necessary to estimate the model were a specification of the mean and the relationship between mean and variance, without specifying the probability density function. Let us consider a very simple model where the random variable $y$ has mean $\mu$ and variance $V(\mu)$ that may be dependent on the mean. Define the function

$$g(y;\mu) = \frac{\sum_{i=1}^{n}(y_i - \mu)}{V(\mu)}. \tag{58}$$

Wedderburn (1974) noticed that (58) is very close to the true score of all the distributions that belong to the exponential family. In addition, $g(y;\mu)$ has properties similar to those of a score function in the sense that

**(i)** $E[g(y;\mu)] = 0$,

**(ii)** $E[g^2(y;\mu)] = -E[dg(y;\mu)/d\mu]$.

Wedderburn termed $g(y;\mu)$ in (58) as the "quasi-score function," the integral of $g(y;\mu)$ the "quasi-likelihood," and the equation $g(y;\mu) = 0$, the "quasi-likelihood equation." Godambe and Heyde (1987) showed that Wedderburn method can be regarded as a particular case of the optimal EF approach [see also Heyde (1997, pp.21-26) and Desmond (1997, pp. 78-80)]. The attractive feature of the EF approach is that we do not need to assume that the true underlying distribution belongs to the exponential family. Since one good example is worth a thousand theories, we now discuss an example, often used in the context of EF [for example, see Godambe and Kale (1991), Desmond (1997) and Bera and Bilias (2002)].

**Example 2.1:** Let $y_i, i = 1, \ldots, n$ be independent random variables with $E(y_i) = \mu_i(\theta)$ and $Var(y_i) = \sigma_i^2(\theta)$, where $\theta$ is a scalar parameter. The quasi-score approach of Wedderburn (1974) suggests that in the class of linear EFs we should solve

$$g^*(y;\mu) = \sum_{i=1}^{n} \frac{[y_i - \mu_i(\theta)]}{\sigma_i^2(\theta)} \frac{d\mu_i(\theta)}{d\theta} = 0. \tag{59}$$

Under the assumption of normality of $y_i$, the ML equation

$$\frac{d \log f(y;\mu)}{d\theta} = g^*(y;\theta) + \frac{1}{2}\sum_{i=1}^{n} \frac{[y_i - \mu_i(\theta)]^2}{\sigma_i^4(\theta)} \frac{d\sigma_i^2(\theta)}{d\theta} - \frac{1}{2}\sum_{i=1}^{n} \frac{d \log \sigma_i^2(\theta)}{d\theta} = 0, \tag{60}$$

is globally optimal and the estimation based on the quasi-score (59) is inferior. If one is unwilling to assume normality, one could claim that the weighted LS approach that minimizes $\sum_i [y_i - \mu_i(\theta)]^2/\sigma_i^2(\theta)$ and yields the EE

$$w(y;\mu) = g^*(y;\theta) + \frac{1}{2}\sum_{i=1}^{n} \frac{[y_i - \mu_i(\theta)]^2}{\sigma_i^4(\theta)} \frac{d\sigma_i^2(\theta)}{d\theta} = 0 \tag{61}$$

is preferable. However, because of the dependence of the variance on $\theta$, (61) delivers an inconsistent root, in general; see Crowder (1986), McLeish (1984) and Sørensen (1999). The application of a law of large numbers shows that $g^*(y;\theta)$ is stochastically

closer to the score (60) than is $w(y;\theta)$. In a way, the second term in (61) creates a bias in $w(y;\theta)$, and the third term in (60) "corrects" for this bias in the score equation.

Let us now consider the optimal EF for this model. We start with a linear EF of the form

$$g(y;\theta) = \sum_{i=1}^{n}[y_i - \mu_i(\theta)]b_i(\theta), \tag{62}$$

where $b_i(\theta)$'s need to be determined. Its standardized version is

$$\begin{aligned} g_s(y;\theta) &= \frac{g(y;\theta)}{E\left[\frac{dg(y;\theta)}{d\theta}\right]} \\ &= \frac{\sum_{i=1}^{n}[y_i - \mu_i(\theta)]b_i(\theta)}{E\left[-\sum_{i=1}^{n}\frac{d\mu_i(\theta)}{d\theta}b_i(\theta) + \sum_{i=1}^{n}[y_i - \mu_i(\theta)]\frac{db_i(\theta)}{d\theta}\right]} \\ &= -\frac{\sum_{i=1}^{n}[y_i - \mu_i(\theta)]b_i(\theta)}{\sum_{i=1}^{n}\frac{d\mu_i(\theta)}{d\theta}b_i(\theta)}, \end{aligned} \tag{63}$$

whose variance is equal to

$$Var[g_s(y;\theta)] = \frac{\sum_{i=1}^{n}\sigma_i^2(\theta)b_i^2(\theta)}{\left[\sum_{i=1}^{n}\frac{d\mu_i(\theta)}{d\theta}b_i(\theta)\right]^2}. \tag{64}$$

The variance in (64) is minimized at

$$b_i(\theta) \propto \frac{d\mu_i(\theta)}{d\theta}\sigma_i^{-2}(\theta). \tag{65}$$

Using this value of $b_i(\theta)$ in (65), leads to the optimal EF

$$\sum_{i=1}^{n}\frac{[y_i - \mu_i(\theta)]}{\sigma_i^2(\theta)}\frac{d\mu_i(\theta)}{d\theta}, \tag{66}$$

which is identical to that obtained from the Wedderburn quasi-likelihood approach as in equation (59). Now assume that $Var(y_i) = c\sigma_i^2(\theta)$, where $c$ is an unknown positive constant not depending on $\theta$, then under the ML approach, $\theta$ and $c$ cannot be estimated separately. For a specified value $c = c_0$, the ML equation now changes from (60) to

$$\frac{d\log f(y;\mu)}{d\theta} = \frac{g^*(y;\theta)}{c_0} + \frac{1}{2c_0}\sum_{i=1}^{n}\frac{[y_i - \mu_i(\theta)]^2}{\sigma_i^4(\theta)}\frac{d\sigma_i^2(\theta)}{d\theta} - \frac{1}{2}\sum_{i=1}^{n}\frac{d\log\sigma_i^2(\theta)}{d\theta} = 0, \tag{67}$$

with

$$E\left[\frac{d\log f(y;\mu)}{d\theta}\right] = \frac{1}{2}\left(\frac{c}{c_0} - 1\right)\sum_{i=1}^{n}\frac{d\log\sigma_i^2(\theta)}{d\theta}, \tag{68}$$

which is also zero only when $c = c_0$. Thus, the ML equation is biased, as was LS equation in (61). However, the optimal EF in (62) [and also the quasi-score in (59)] remains unaffected by the value of $c$. Therefore, here we have situations in which both the LS and ML methods could be inconsistent while the EF retains its optimality property.

## 2.2 Generalization to the multiparameter case

The extension of the EF approach from the single parameter case to the multiparameter framework with $\theta = (\theta_1, \theta_2, \ldots, \theta_p)'$ is quite natural and straightforward. The basic technique is to replace the scalars by $(p \times 1)$ vectors and variances by $(p \times p)$ variance-covariance matrices. Therefore, instead of presenting all generalizations to the multiparameter case, we will only mention the key results. We start with a $(p \times 1)$ vector EF, $g(y;\theta) = (g_1(y;\theta), g_2(y;\theta), \ldots, g_p(y;\theta))'$ satisfying the regularity conditions stated in Section 2.1 for the single parameter case. Let us denote the class of regular unbiased EFs by $\mathcal{G}$ and define

$$\Sigma_g = Var[g(y;\theta)] = E[g(y;\theta)g'(y;\theta)], \tag{69}$$

and

$$D_g = E\left[\frac{\partial g(y;\theta)}{\partial\theta}\right], \tag{70}$$

both being $(p \times p)$ nonsingular matrices. The standardized vector EF can be written as

$$g_s(y;\theta) = D_g^{-1}g(y;\theta), \tag{71}$$

and hence

$$Var\left[g_s(y;\theta)\right] = D_g^{-1}\Sigma_g D_g^{'-1} = \Sigma_{g_s}, \quad say. \tag{72}$$

Therefore, our objective could be stated as to minimize (maximize) a scalar measure corresponding to $D_g^{-1}\Sigma_g D_g^{'-1}$ $(D_g\Sigma_g^{-1}D_g')$. Bhapkar (1972) defined an optimal EF $g^*$ as follows:

**Definition 2.2:** *A $g^* \in \mathcal{G}$ is said to be optimal if*

$$\begin{aligned} Var[g_s^*] &\leq Var[g_s] && (73)\\ or, \quad \Sigma_{g_s^*} &\leq \Sigma_{g_s} && (74)\\ or, \quad D_{g^*}^{-1}\Sigma_{g^*}D_{g^*}^{'-1} &\leq D_g^{-1}\Sigma_g D_g^{'-1} && (75) \end{aligned}$$

*i.e., the difference of the left hand side matrix from the right hand side matrix is nonnegative definite (nnd) for all $g \in \mathcal{G}$.*

This is the multiparameter counterpart of Definition 2.1, given in Section 2.1. The above criterion is called *matrix optimality* of $g^*$. Unlike in the scalar case, there could be many ways to compare the two matrices, say, in (74) and define optimality of $g^*$; for example, two other ways could be through

**(i)** trace optimality, i.e., $Tr(\Sigma_{g_s^*}) \leq Tr(\Sigma_{g_s})$,

**(ii)** determinant optimality, i.e., $|\Sigma_{g_s^*}| \leq |\Sigma_{g_s}|$.

Chandrasekhar and Kale (1984) proved that these three criteria are equivalent in the sense that if $g^*$ is optimal with respect one criterion then it is also optimal with respect to the remaining two [see also Heyde (1997, pp. 19-21)]. Godambe-Durbin's optimality result can now be presented as: for all $g \in \mathcal{G}$ and $\theta \in \Theta$

$$D_g^{-1}\Sigma_g D_g^{'-1} - \mathcal{I}^{-1} \geq 0, \tag{76}$$

where $\mathcal{I} \equiv \mathcal{I}(\theta) = E[\partial \log f(y;\theta)/\partial\theta\partial\theta']$ is the $(p \times p)$ Fisher information matrix. The equality in (76) holds by the optimal EF $g^*(y;\theta) = \partial \log f(y;\theta)/\partial\theta = S(\theta)$, the $(p \times 1)$ score vector. From (76) it follows that

$$|\Sigma_g| \geq |D_g\mathcal{I}^{-1}D_g'| = \frac{|D_g|^2}{|\mathcal{I}|}. \tag{77}$$

Following the scalar case, Bhapkar (1972) defined the amount of information contained in the EF $g(y;\theta)$ about $\theta$, by

$$\mathcal{I}_g(\theta) = \frac{|D_g|^2}{|\Sigma_g|}. \tag{78}$$

The equality

$$RE(g) = \frac{\mathcal{I}_g(\theta)}{|\mathcal{I}(\theta)|} \tag{79}$$

provides a measure of efficiency of $g$. Clearly $0 \leq RE(g) \leq 1$, and the upper bound is

attained by the score function $S(\theta)$. In the multiparameter case, alternative measures of efficiency can also be defined, such as,

$$\frac{Tr\left(D_g \Sigma_g^{-1} D_g'\right)}{Tr\left(\mathcal{I}\right)}. \tag{80}$$

Both the measures, (79) and (80) reduce to (53) in the scalar case.

## 2.3 Estimating function in the presence of nuisance parameters

In the 1930s, the controversy between Karl Pearson and Fisher spilled over to Jerzy Neyman. Neyman found it difficult to accept the ML method as a *general* method of estimation. As Barnard (1973, p.133) stated, Neyman's objection to ML method arose not because of its failure in unusual pathological cases, but because it seemed to give "wrong" answers for some simple cases. One of the simplest cases is estimation of $\theta_1$ when $Y \sim N(\theta_2, \theta_1)$. The ML method gives a biased estimate for $\theta_1$. A more serious objection to ML approach was raised by Neyman and Scott (1948), who showed that when the number of nuisance parameters increases with the sample size, the MLE of a parameter of interest could be inefficient or even *inconsistent*. Perhaps, for problems involving nuisance parameters, the EF approach has the most potential.

Let us partition the $p \times 1$ parameter vector $\theta$ by $\theta = (\theta_1', \theta_2')' \in \Theta$, where $\theta_1 \in \Theta_1$ is an $r \times 1$ $(r < p)$ vector of unknown parameter of interest, and $\theta_2 \in \Theta_2$ is a $(p-r) \times 1$ vector of "nuisance" or "incidental" parameters. As noted, nuisance parameters can have a major influence on the estimation of parameter of interest.

The problem of estimating a real parameter $\theta_1$, in the presence of nuisance parameter $\theta_2$ was first addressed by Godambe and Thompson (1974), yet another "conceptually clean" and pathbreaking paper - just over three pages long. They first defined a class $\mathcal{G}_1$ of regular unbiased EFs of the form $g(y; \theta_1)$. The generalization of the earlier optimality criteria [see equations (38) and (46)] now defines an optimal EF $g^* \in \mathcal{G}_1$ for which

$$Var[g_s^*] \leq Var[g_s], \tag{81}$$

for all $g \in \mathcal{G}_1$. Taking $\theta_1$ and $\theta_2$ scalars, Godambe and Thompson (1974) showed that under the regularity conditions the function $g^* \in \mathcal{G}_1$, satisfying (81) is given by

$$
\begin{aligned}
g^* &= c_1(\theta_1,\theta_2)\frac{\partial \log f(y;\theta)}{\partial \theta_1} \\
&+ c_2(\theta_1,\theta_2)\left\{\left[\frac{\partial \log f(y;\theta)}{\partial \theta_2}\right]^2 + \frac{\partial^2 \log f(y;\theta)}{\partial \theta_2^2}\right\}, \qquad (82)
\end{aligned}
$$

where $c_1(\theta_1,\theta_2)$ and $c_2(\theta_1,\theta_2)$ are such that the resulting $g^*$ is free of $\theta_2$.

**Example 2.2:** Godambe and Thompson (1974) considered the $N(\theta_2,\theta_1)$ case. Here

$$
f(y;\theta) = \frac{1}{(\sqrt{2\pi\theta_1})^n} e^{-\frac{1}{2\theta_1}\sum_{i=1}^n (y_i-\theta_2)^2}, \qquad (83)
$$

$$
\frac{\partial \log f(y;\theta)}{\partial \theta_1} = -\frac{n}{2\theta_1} + \frac{(n-1)\widehat{\theta}_1 + n(\overline{y}-\theta_2)^2}{2\theta_1^2}, \qquad (84)
$$

$$
\frac{\partial \log f(y;\theta)}{\partial \theta_2} = \frac{n(\overline{y}-\theta_2)}{\theta_1}, \qquad (85)
$$

$$
\frac{\partial^2 \log f(y;\theta)}{\partial \theta_2^2} = -\frac{n}{\theta_1}, \qquad (86)
$$

where $\widehat{\theta}_1 = \sum_{i=1}^n (y_i-\overline{y})^2/(n-1)$ with $\overline{y} = \sum_{i=1}^n y_i/n$. Using (84) − (86), it is easy to see that by choosing $c_1(\theta_1,\theta_2) = 1$ and $c_1(\theta_1,\theta_2) = -1/2n$, we can get a $g^*$ which is free of $\theta_2$, and it is given by

$$
g^* = \frac{n-1}{2\theta_1^2}(\widehat{\theta}_1 - \theta_1), \qquad (87)
$$

and $g^* = 0$ gives an unbiased estimator. In this connection we should mention that Fisher (1912), while proposing his ML method, also produced an unbiased estimator of $\theta_1$ from the mode of the posterior distribution (inverse probability) with a non-informative prior for $\theta_2$. At a later stage, Fisher (1922, p.326) himself did not approve of basing his argument upon the principle of inverse probability.

Towards the end of their paper, without any fanfare, Godambe and Thompson (1974) suggested an EF $g_1(y;\theta_1)$ of the form

$$
g_1(y;\theta_1) = \left[\frac{\partial \log f(y;\theta)}{\partial \theta_1}\middle|\theta_2 = \widehat{\theta}_2\right] - E\left[\frac{\partial \log f(y;\theta)}{\partial \theta_1}\middle|\theta_2 = \widehat{\theta}_2\right], \qquad (88)
$$

where $\widehat{\theta}_2$ is the MLE of $\theta_2$. They attributed their result to George Barnard (through oral communication) who was then visiting their Department of Statistics, University of Waterloo. Eight years later, as we discuss below, in a very influential paper, Lindsay (1982) considered precisely this form of EF and established its importance and usefulness.

**Example 2.3:** Let us consider $y_{ij} = \mu_i + \epsilon_{ij}, \epsilon_{ij} \sim IIDN(0, \theta_1), i = 1, 2, \ldots, k, j = 1, 2$. Here $\theta_2 = (\mu_1, \mu_2, \ldots, \mu_k)'$ is the nuisance parameter vector. Neyman and Scott (1948) used this model for their famous illustration of the inconsistency of MLE of a parameter of interest when the number of nuisance parameter increases with the sample size. The loglikelihood and score functions, respectively, are given by

$$\log f(y;\theta) = -k\log(2\pi) - k\log\theta_1 - \frac{1}{2\theta_1}\sum_{i=1}^{k}\sum_{j=1}^{2}(y_{ij}-\mu_i)^2,$$

$$\frac{\partial \log f(y;\theta)}{\partial \theta_1} = -\frac{k}{\theta_1} + \frac{1}{2\theta_1^2}\sum_{i=1}^{k}\sum_{j=1}^{2}(y_{ij}-\mu_i)^2,$$

$$\frac{\partial \log f(y;\theta)}{\partial \mu_i} = \frac{1}{\theta_1}\sum_{j=1}^{2}(y_{ij}-\mu_i), \quad i = 1, 2, \ldots, k.$$

Therefore,

$$\widehat{\theta}_1 = \frac{1}{2k}\sum_{i=1}^{k}\sum_{j=1}^{2}(y_{ij}-\overline{y}_i)^2, \tag{89}$$

where $\overline{y}_i = \sum_{j=1}^{2} y_{ij}/2$, is the MLE for $\theta_1$. Since $\frac{1}{\theta_1}\sum_{j=1}^{2}(y_{ij}-\overline{y}_i)^2 \sim \chi^2_1$, if we define $z_i = \frac{1}{2}\sum_{j=1}^{2}(y_{ij}-\overline{y}_i)^2$, then $z_i \sim IID(\theta_1/2, \theta_1^2/2)$. By the weak law of large numbers, as $k \to \infty$, $\widehat{\theta}_1 = \sum_{i=1}^{k} z_i/k$ converges to $E(z_i) = \theta_1/2$.

Godambe resolved this inconsistency of the MLE problem by showing that under certain conditions the optimal EF leads to a *conditional* ML approach and provides a consistent estimator for $\theta_1$. The basic problem for the ML approach is that although

$$E[\partial \log f(y;\theta_1,\theta_2)/\partial\theta_1] = 0, \tag{90}$$

$$E[\partial \log f(y;\theta_1,\widehat{\theta}_{2.1})/\partial\theta_1] \neq 0, \tag{91}$$

where $\widehat{\theta}_{2.1}$ is the MLE of $\theta_2$ for fixed $\theta_1$. Therefore, the use of (91) will lead to a biased EF for $\theta_1$. Let $T \equiv t(y)$ be a complete sufficient statistic for the parameter $\theta_2$, for every fixed $\theta_1$ and also assume that $T$ does not involve $\theta_1$. Suppose we can decompose the likelihood function as

$$f(y;\theta) = f(y|t;\theta_1)h(t;\theta_1,\theta_2), \tag{92}$$

where $f(y|t;\theta_1)$ is the conditional pdf of $y$ given $t$, and $h$ is the pdf of $T$. Then Godambe (1976, Theorem 3.2) showed that the "conditional" score function $\frac{\partial \log f(y|t;\theta_1)}{\partial\theta_1}$ gives a *unique optimal* EF.

**Example 2.3:** (*Continued*) It can be shown that $T(y) = (\overline{y}_1, \overline{y}_2, \ldots, \overline{y}_k)'$ is

complete sufficient for $\theta_1 = (\mu_1, \mu_2, \ldots, \mu_k)'$. Using (92) we can show that

$$g^* = \frac{\partial \log f_t(y|t;\theta_1)}{\partial \theta_1} = -\frac{k}{2\theta_1} + \frac{1}{2\theta_1^2}\sum_{i=1}^{k}\sum_{j=1}^{2}(y_{ij} - \overline{y}_i)^2, \tag{93}$$

is the optimal EF. The difference between the score $\partial \log f(y;\theta)/\partial\theta_1$ and the conditional score $g^*$ is quite obvious. Solving $g^* = 0$, we have the solution

$$\widetilde{\theta}_1 = \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{2}(y_{ij} - \overline{y}_i)^2 = \frac{1}{4k}\sum_{i=1}^{k}(y_{i1} - y_{i2})^2 = 2\widehat{\theta}_1, \tag{94}$$

which converges to $\theta_1$, and hence is consistent. This was the Godambe's (1976) solution to the Neyman-Scott (1948) problem through the EF approach.

Godambe's method works well as long as the conditioning statistic $T(y)$ does not involve $\theta_1$, which will be the case when $f(y;\theta)$ has the exponential family structure. However, that will exclude a large class of distributions. To accommodate a general situation, Lindsay (1982) extended Godambe's (1976) conditional score function $\frac{\partial \log f(y|t;\theta_1)}{\partial\theta_1}$ to

$$S_1^*(\theta) = \frac{\partial \log f(y;\theta)}{\partial\theta_1} - E\left[\frac{\partial \log f(y;\theta)}{\partial\theta_1}\middle| t_{\theta_1}\right], \tag{95}$$

where $t_{\theta_1}$ is the minimal sufficient statistic for $\theta_2$ and the notation signifies that $t$ is functionally dependent on $\theta_1$. When $t_{\theta_1} \equiv t$, $S_1^*(\theta)$ reduces to $\partial \log f(y|t;\theta)/\partial\theta_1$. The closeness of (88) and (95) is unmistakable. $S_1^*(\theta)$ which is sometimes also called the effective score, is orthogonal to the space spanned by the sufficient statistic $t_{\theta_1}$. Though $S_1^*(\theta)$ will continue to depend on $\theta_2$, the representation in (95) implies that the dependence on $\theta_2$ is now reduced. For a rigorous discussion of these issues and further results see, for instance, Lindsay and Waterman (1992) and Liang and Zeger (1995).

We end our discussion on the nuisance parameter issues by giving another, though asymptotic, justification on Lindsay's conditional (effective) score (95) following Neyman's (1959) approach to testing in the presence of nuisance parameter [see Bera and Bilias (2001a, 2001b)]. For simplicity, we assume that both $\theta_1$ and $\theta_2$ are scalars. The need to leave the asymptotic distribution of EF unchanged after the substitution of a $\sqrt{n}$-consistent estimate of $\theta_2$ leads to the orthogonalization step: starting from an arbitrary EF $g$, we will regress $g$ on the part of the score for the nuisance parameter $S_2 = \partial \log f(y;\theta)/\partial\theta_2$ and keep the residual. The new EF, will be

$$g - bS_2,$$

where $b$ denotes the regression coefficient. Next, we want to choose $(g - bS_2)$, and therefore $g$, according to Godambe-Durbin optimality criterion. This dictates that the optimal EF, in its standardized form, should have minimum variance. By differentiating the moment condition

$$E[g - bS_2] = 0 \tag{96}$$

with respect to $\theta_1$, we have

$$E\left[\frac{\partial(g - bS_2)}{\partial\theta_1}\right] + Cov[(g - bS_2), S_1] = 0,$$

where $S_1$ is the score for the parameter of interest. Since $(g - bS_2)$ is *orthogonal* to $S_2$, $Cov[(g - bS_2), S_1] = Cov[(g - bS_2), (S_1 - bS_2)]$. Therefore,

$$E\left[\frac{\partial(g - bS_2)}{\partial\theta_1}\right] = -Cov[(g - bS_2), (S_1 - bS_2)], \tag{97}$$

which, using the Cauchy-Schwartz inequality, yields :

$$\begin{aligned}\left\{E\left[\frac{\partial(g - bS_2)}{\partial\theta_1}\right]\right\}^2 &= \{Cov[(g - bS_2), (S_1 - bS_2)]\}^2 \\ &\leq Var(g - bS_2)Var(S_1 - bS_2). \end{aligned} \tag{98}$$

The inequality (98) can be rearranged so that a lower bound for the variance of the standardized EF is formed as

$$\frac{Var[(g - bS_2)]}{\left\{E\left[\frac{\partial(g - bS_2)}{\partial\theta_1}\right]\right\}^2} \geq \frac{1}{Var(S_1 - bS_2)}. \tag{99}$$

The bound is reached when $g = S_1$. Thus the optimal EF in the presence of nuisance parameters is given by the effective score $(S_1 - bS_2)$, where $b = Cov(S_2, S_1)/Var(S_1)$. However, it should be remarked that, in contrast to Godambe's result (Theorem 2.1) that the score is the optimal EFs, our argument in the presence of nuisance parameter holds only asymptotically.

## 2.4 The dependent case and optimal combination of (elementary) estimating functions

As noted in Section 2.1, for some time it was believed that the MM estimators are inefficient compare to the ML estimators. Godambe's (1960) analysis highlights, for the IID case, the equivalence of MM and ML estimation when one replaces a ar-

bitrary moment function with the score function. Much of his analysis also carries over to the case of dependent data. For a general discrete time stochastic process, an optimality criterion for an EF was established in two important papers by Godambe (1985) and Godambe and Thompson (1989) using a "flexible" *conditioning* method. Their flexible set up can cope with the estimation of parameters in dependent data, such as those from time series processes on a real line or a spatial process on a lattice. In this section, we present the theory developed for the optimum combination of (elementary) EEs for the estimation of parameters of stochastic process. In Section 3, we will provide its applications to some widely used models in the applied econometrics and statistics literature.

Let us consider a discrete time stochastic process $\{y_t; t \geq 0\}$ taking values in the real line $\mathbf{R}$. Also let $\mathcal{F} = \{F\}$ be a class of probability distributions on $\mathbf{R}^n$ and $\theta = \theta(F) \in \Theta$, be a real parameter. The objective is to estimate $\theta$ by an estimator $\widehat{\theta}_n$ which is a function of observations $\{y_t; 0 \leq t \leq n\}$. By definition, the EF $g(y_1, ..., y_n; \theta(F))$ is a real valued function of both the observation $\{y_t\}$ and the parameter $\theta$, that satisfies certain regularity conditions (such as square-integrability and differentiability, given in Section 2.1). It is called regular unbiased EF if

$$E_F \{g(y_1, ..., y_n; \theta(F))\} = 0, \quad F \in \mathcal{F}.$$

Among all regular unbiased EFs $g(y_1, ..., y_n; \theta(F))$, $g^*(y_1, ..., y_n; \theta(F))$ is said to be optimum if

$$E_F \left[g(y_1, ..., y_n; \theta(F))^2\right] / \left\{ E_F \left[ \frac{\partial g(y_1, ..., y_n; \theta)}{\partial \theta} \right]_{\theta=\theta(F)} \right\}^2 \tag{100}$$

is minimized $\forall F \in \mathcal{F}$ at $g = g^*$. An estimator $\widehat{\theta}_n$ is obtained by solving

$$g^*(y_1, ..., y_n; \theta(F)) = 0.$$

Suppose that we have the unbiased elementary EFs $\psi_t, t = 1, \ldots, n$ involving $\theta$. The question is what is the best way of combining these $n$ EFs into one EF for estimation of $\theta$. Godambe (1985) restricted his search to the class $\mathcal{L}$ of linear combination of $\psi_t$'s, namely,

$$\left\{ g : g(\theta) = \sum_{t=1}^{n} a_{t-1} \psi_t \right\}, \tag{101}$$

where the coefficient $a_{t-1}$ is a function of $\{y_1, \ldots, y_{t-1}\}$ and $\theta$. Also, the elementary EF $\psi_t$ is such that $E_F[\psi_t | \mathcal{F}_{t-1}^y] = 0$, with $\mathcal{F}_{t-1}^y$ being the $\sigma$-field generated by $\{y_s; s \leq t-1\}$. This further implies that $\forall F \in \mathcal{F}$,

$$E_F[\psi_t \psi_{t'}] = 0, \quad for \quad t \neq t'. \tag{102}$$

i.e., $\psi_t$ and $\psi_{t'}$ are orthogonal. Under this set up a new definition follows. Among all unbiased EFs $g$, an optimal EF $g^*$ is the one that provides the smallest value of

$$E\left[g(y_1, ..., y_n; \theta)^2 | \mathcal{F}_{t-1}^y\right] / \left\{ E\left[ \left( \frac{\partial g(y_1, ..., y_n; \theta)}{\partial \theta} \right) \middle| \mathcal{F}_{t-1}^y \right] \right\}^2. \tag{103}$$

Note that $\mathcal{L}$ is a subset of the class of all unbiased EFs where $\psi_t$ and $a_{t-1}$ are assumed to be differentiable with respect to $\theta, \forall t = 1, \ldots, n$. Now we state and prove Godambe's (1985) result on optimal EF for the dependent case.

**Theorem 2.2:** *Within the class of estimating functions $\mathcal{L}$ defined in* (101)*, the optimal estimating function $g^*$ that minimizes* (103) *is given by $g^*(\theta) = \sum_{t=1}^n \psi_t a_{t-1}^*$ where $a_{t-1}^* = \{E[\frac{\partial \psi_t}{\partial \theta} | \mathcal{F}_{t-1}^y]\} / \{E\left[\psi_t^2 | \mathcal{F}_{t-1}^y\right]\}$.*

***Proof:*** Using the equations (101) and (102), we have

$$E[g^2] = E\left\{ \sum_{t=1}^n a_{t-1}^2 E\left[\psi_t^2 | \mathcal{F}_{t-1}^y\right] \right\} \tag{104}$$

and

$$\begin{aligned} \left\{ E\left[\frac{\partial g}{\partial \theta}\right] \right\}^2 &= \left\{ E \sum_{t=1}^n \left( a_{t-1} E\left[ \frac{\partial \psi_t}{\partial \theta} \middle| \mathcal{F}_{t-1}^y \right] + \left( \frac{\partial a_{t-1}}{\partial \theta} \right) E[\psi_t | \mathcal{F}_{t-1}^y] \right) \right\}^2 \\ &= \left\{ E \sum_{t=1}^n a_{t-1} E\left[ \frac{\partial \psi_t}{\partial \theta} \middle| \mathcal{F}_{t-1}^y \right] \right\}^2, \end{aligned} \tag{105}$$

as $E[\psi_t | \mathcal{F}_{t-1}^y] = 0$. Letting $B = \sum_{t=1}^n a_{t-1} E\left[\frac{\partial \psi_t}{\partial \theta} | \mathcal{F}_{t-1}^y\right]$ and $A^2 = \sum_{t=1}^n a_{t-1}^2 E\left[\psi_t^2 | \mathcal{F}_{t-1}^y\right]$, we have

$$\frac{\left\{E\left[\frac{\partial g}{\partial \theta}\right]\right\}^2}{E\left[g\right]^2} = \frac{\{E[B]\}^2}{E[A^2]} \leq E\left[\frac{B^2}{A^2}\right] \tag{106}$$

by the Cauchy-Schwartz inequality. The equality in (106) holds if $A^2 \propto B$, i.e., if $a_{t-1} = a_{t-1}^*$.

**Example 2.4:** Estimating function (3) for the AR(1) model in (1) can be obtained through Godambe's (1985) approach that sheds light to the distinctive nature of the theory of EF. Here $\psi_t = u_t = y_t - \theta y_{t-1}, t = 2, 3, \ldots, n$ are $n$ elementary EFs, and the issue is how we should combine these $(n-1)$ functions into one to solve for the parameter $\theta$. Let us consider the class of EFs

$$g = \sum_{t=2}^{n} a_{t-1}\psi_t,$$

where the weights $a_{t-1}$ depend only on the conditioning event $(y_1, y_2, \ldots, y_{t-1})$. Theorem 2.2 yields the optimal weights as

$$a^*_{t-1} = \frac{E_{t-1}[\partial(y_t - \theta y_{t-1})/\partial\theta]}{E_{t-1}[(y_t - \theta y_{t-1})^2]} = \frac{-y_{t-1}}{\sigma^2},$$

where $\sigma^2 = Var(u_t)$. Therefore, the optimal EE for $\theta$ is

$$g^* = \sum_{t=2}^{n} y_{t-1}(y_t - \theta y_{t-1}) = 0$$

which is same as (3). Durbin (1960) arrived at the same EF by starting with an unbiased linear EF $g = T_1(y) + \theta T_2(y)$, where $T_1(y)$ and $T_2(y)$ are functions of data $(y_1, y_2, \ldots, y_n)$ only. Then, he imposed a minimum variance requirement on $g$, reminiscent of Gauss-Markov theorem.

Godambe (1985, p.424) also established that in the class of all EFs of the form (101), the partial likelihood score is an optimum EF. For this consider the joint density function of $(y_1, \ldots, y_n)$, involving a parameter of interest $\theta$ and a nuisance parameter $\delta$

$$f(y_1, \ldots, y_n; \theta, \delta) = \prod_{t=1}^{n} f_{t-1}(y_t; \theta, \delta),$$

where $f_{t-1}$ denotes the conditional density of $y_t$ given $y_1, \ldots, y_{t-1} (t = 1, \ldots, n)$. Let $T_t (t = 1, \ldots, n)$ be a minimal sufficient statistic for $\delta$ in the density $f_{t-1}(y_t; \theta, \delta)$, so that $f_{t-1}(y_t|T_t; \theta, \delta) = f_{t-1}(y_t|T_t; \theta)$ is independent of $\delta$. Now by considering the partial likelihood score for $\theta$,

$$\sum_{t=1}^{n} \frac{\partial \log f_{t-1}(y_t|T_t; \theta)}{\partial\theta} = \sum_{t=1}^{n} \psi_t, \quad say. \tag{107}$$

It is easy to see that $E_{t-1}(\psi_t) = 0$ and $E(\psi_t\psi_{t'}) = 0$, $t \neq t' = 1, \ldots, n$. Therefore, from Theorem 2.2, it follows that the optimal EF within the class of linear combination of $\psi_t$'s is $\sum_{t=2}^{n} \psi_t a^*_{t-1}$, where

$$a^*_{t-1} = \frac{E_{t-1}[\partial^2 \log f_{t-1}(y_t|T_t; \theta)/\partial\theta^2]}{E_{t-1}[\partial \log f_{t-1}(y_t|T_t; \theta)/\partial\theta]^2} = -1.$$

This establishes the optimality of the partial likelihood score function in (107).

As we discussed in Section 2.1, in a parametric model the score function provides

the optimum EF; the result of Theorem 2.2 can be extended to a theory of pseudo-score function and the associated Fisher information. Utilizing the pseudo-score function

$$\Psi = -\sum_{t=1}^{n} \psi_t a_{t-1}^*,$$

we derive

$$E\left[\frac{\partial(-\Psi)}{\partial\theta}\right] = E\left[\sum_{t=1}^{n} a_{t-1}^* E_{t-1}\left(\frac{\partial\psi_t}{\partial\theta}\right)\right] = E\left[\sum_{t=1}^{n} a_{t-1}^{*2} E_{t-1}\left(\psi_t^2\right)\right] = E\left[\Psi^2\right]$$

and obtain the EF

$$\left(\sum_{t=1}^{n} \psi_t a_{t-1}^*\right) / \left\{\sum_{t=1}^{n} \left(a_{t-1}^{*2}\right) E_{t-1}\left[\psi_t^2\right]\right\}^{1/2}, \tag{108}$$

which is a standardized martingale. Asymptotically, the density of EF in (108) converges to $N(0,1)$. This suggests the existence of an associated pseudo Fisher information, independent of parameter $\theta$, given by

$$\mathcal{I} = \sum_{t=1}^{n} a_{t-1}^{*2} E_{t-1}(\psi_t^2).$$

Interestingly, one can interpret $\mathcal{I}$ as an unbiased estimate of the variance of $\Psi$.

Another justification of the Godambe optimality criterion is the fact that under standard regularity conditions the EF estimator $\widehat{\theta}_n^*$ that solves the optimal EE $g^*(y_1, ..., y_n; \theta(F)) = 0$, minimizes, at least asymptotically, the mean squared error $E(\widehat{\theta} - \theta)^2$ where $\widehat{\theta}$ is the estimator from $g(y_1, ..., y_n; \theta(F)) = 0$. Also, one can utilize the choice of weights $a_{t-1}^*$ to get the most benefit from any knowledge about the unknown distribution of $\{y_t; t \geq 0\}$, especially when specifications of third and fourth moments are unknown. The suboptimal weights can reduce the efficiency of the estimator significantly without affecting its consistency and asymptotic normality properties.

It is important to note that, the optimal estimation procedure by Godambe (1985) is based on martingale structure with the corresponding filtering method which, in some sense, restricts the nature of the stochastic process. However, Godambe and Thompson (1989) provided an extension of the concept of optimality of such an EF into a general setting using a more "flexible" *conditioning* method which is related to the concept of quasi-likelihood approach. This broadens the applicability of their method to a wider class. Using the same set up with $\mathcal{Y}$ as an arbitrary sample space,

they considered the class of EFs $\psi_j$ which is a real function defined on $\mathcal{Y} \times \Theta$ such that

$$E_F\left[\psi_j\left(y_1, ..., y_n; \theta(F)\right)|\mathcal{Y}_j\right] = 0, \quad F \in \mathcal{F}, \tag{109}$$

where $E_F[.|\mathcal{Y}_j]$ is the expectation under $F$, conditional on $\mathcal{Y}_j$, $\mathcal{Y}_j (j = 1, \ldots, k)$ being a $\sigma$-field generated by a specified partition on the sample space $\mathcal{Y}$. To estimate $\theta$ on the basis of observations $\{y_t\}$ they considered the class of EFs $\mathcal{H} = \{h\}$, where

$$h = \sum_{j=1}^{k} a_j \psi_j$$

and $a_j$ is a real function on $\mathcal{Y} \times \Theta$. The EFs $\psi_j, j = 1, \ldots, k$ satisfying (109) are said to be mutually orthogonal if $E_F(\psi_j \psi_i | \mathcal{Y}_i) = 0$ and $E_F(\psi_i \psi_j | \mathcal{Y}_j) = 0$ for $F \in \mathcal{F}$ and $i \neq j, i, j = 1, \ldots, k$. An estimate of $\theta$ based on the EF $h$ is obtained by solving the equation $h(y_1, ..., y_n; \theta(F)) = 0$. For the optimal EF they defined

$$h^* = \sum_{j=1}^{k} a_j^* \psi_j, \tag{110}$$

where

$$a_j^* = \frac{E_F\left\{\left(\frac{\partial \psi_j}{\partial \theta}\right)_{\theta=\theta(F)} \middle| \mathcal{Y}_j\right\}}{E_F\left\{[\psi_j(y_1, ..., y_n; \theta(F))]^2 | \mathcal{Y}_j\right\}}.$$

The following result, a proof of which is given in Godambe and Thompson (1989, p.140), demonstrates how to construct such an optimal EF.

**Theorem 2.3** *The estimating function $h^*$ of* (110) *is optimum in the class $\mathcal{H}$, if the elementary estimating functions $\psi_j$ are mutually orthogonal.*

The above theorem provides an optimal EF in a wide class of functions $\mathcal{H}$ when the $\psi_j$'s need not necessarily be linear functions of $y_i$'s and can be formed using an optimal orthogonal combination involving the first few moments of $y_i$. In some cases, the function $a_j$ can be the functions of all $y_i$'s except the $y_j$ itself. A similar criterion of optimality, but without the notion of orthogonality, was also used by Crowder (1986) based on optimum quadratic EFs. However, in Crowder (1986), the criterion of optimality is in terms of the asymptotic variance of the estimate whereas, for Godambe (1985), the finite sample optimality criterion for a general stochastic process is for the EF. Also the class of unbiased and orthogonal EFs in Godambe and Thompson (1989) is broader than the class of quadratic EFs. For more on the theory of optimum orthogonal EFs, see Godambe (1991b).

## 2.5 Estimating functions and generalized method of moments

The EF approach to estimation, while very popular among statisticians, has been largely ignored by econometricians who were mainly absorbed by the use of generalized method of moments (GMM). Today it looks as if the two methods produce the same results from the point of view of the user. The EF methodology started by defining a concrete optimality criterion for the choice of *elementary* EFs. In many instances these elementary EFs were essentially what was called in econometrics *conditional moments*; compare Godambe (1985) and Chamberlain (1987). Then, the EF approach went on with the issue of how best to combine these elementary EFs into a number of EEs that equals the number of the unknown parameters of the statistical model. In particular, as we discussed in Section 2.4, Godambe (1985) worked the problem for stochastic processes where the conditioning information set is formed naturally from the past of the process. According to his solution, if we restrict ourselves to linear combinations of the various EFs, then an optimal combination is formed by utilizing weights given in Theorem 2.2. This result was generalized by Godambe and Heyde (1987), who termed the optimal EF as the quasi-score.

In the econometric practice of GMM, the emphasis seem to be on the formation of convenient unconditional moments from the conditional restrictions. Then, the question of optimality usually concerns the optimal choice of the weighting matrix in the objective function for a *given* set of *unconditional* moments. Consider the framework of the generalized linear regression with strictly exogenous regressors. The econometric practice will be to form the unconditional moments that eventually lead to the least squares estimator. The optimal EF approach will point to the first order conditions that correspond to the generalized least squares. It is true that the first approach is adopted by applied researchers who want to avoid making specific assumptions about the variances and covariances of the responses. However, it is certainly useful to know what is the benchmark for optimality.

In econometric literature, a result similar to the one given by Godambe (1985), is now well known and it seems that it was first given by Chamberlain (1987); see also Newey (2004) for more results and examples, and Davidson and MacKinnon [section 17.4 (1993)] for a textbook discussion.

It should be noted that the result in econometric literature was produced from asymptotic considerations by studying the variance matrix of the estimator, while in the statistical literature the focus was on finite sample optimality of the EF.

# 3 Applications

In this Section we apply the optimal EF approach of estimation discussed in previous section to a number of widely used econometric models. First, we demonstrate its applicability to various non-linear time series models and then utilize it for spatial regression model. This is followed by applications to longitudinal data and the median regression model.

The general expression of a non-linear univariate time series model is

$$X_t = \varphi(X_{t-1}, \ldots, X_{t-p}; \epsilon_{t-1}, \ldots, \epsilon_{t-q}; \theta) + \epsilon_t, \quad t \in \mathbf{Z} \tag{111}$$

where $\varphi(.)$ is some known non-linear function with finite dimensional parameter vector $\theta$, $\{\epsilon_t\}$ is strictly white noise, $p, q$ are non-negative integers and $\mathbf{Z}$ denotes the set of all integers. For a description of nonlinear time series models we use, see Teräsvirta (2006). In Sections $3.1 - 3.4$ we discuss the EF approach to estimate $\theta$ for such non-linear time series models that are frequently used, sometimes even as competing models. As long as we express the first two conditional moments of the observed series, the EF theory is readily applicable. Our discussion is valid for observed time series data as well as estimated residual in a regression set up.

## 3.1 Random coefficient autoregressive model

An important class of non-linear time series model is the random coefficient autoregressive (RCA) model for which a fairly extensive theory of estimation exists based on LS and ML procedures [for instance see, Nicholls and Quinn (1982)]. One of the common features of RCA model is the *varying conditional variance* that is similar to the autoregressive conditional heteroscedastic (ARCH) type models [Tsay (1987), Bera and Lee (1993), Granger and Teräsvirta (1993, Ch.4)]. Also since many properties of ARCH model, conditional and unconditional, can be derived directly from the RCA model, the usefulness of the latter becomes more appealing in both a theoretical and an applied context. Therefore, it makes sense to apply an optimal EF approach to obtain a more *efficient* estimate without any distributional assumptions, which has important finite sample properties. The important references on which this section is based, are Thavaneswaran and Abraham (1988), Heyde (1997) and Chandra and Taniguchi (2001).

A stochastic process $\{X_t, t \in \mathbf{Z}\}$ is said to follow a RCA model of order $p$ if it

satisfies

$$\begin{aligned} X_t &= \sum_{i=1}^{p} \theta_{it} X_{t-i} + \epsilon_t \\ &= \sum_{i=1}^{p} (\theta_i + \eta_{it}) X_{t-i} + \epsilon_t, \end{aligned} \tag{112}$$

where $\theta = (\theta_1, \ldots, \theta_p)'$ is the parameter to be estimated, $\eta_{it}$ are random components and $\epsilon_t$ is the innovation term. For model (112), it is customary to define $\mathbf{X}_{t-1} = (X_{t-1}, \ldots, X_{t-p})'$ and make the following assumptions: **(i)** For $t = 1, \ldots, n$, $\{\eta_t = (\eta_{1t}, \ldots, \eta_{pt})'\}$ is a sequence of IID random vector with zero mean and variance $E(\eta_t \eta_t') = \Sigma$, a $p \times p$ matrix, **(ii)** $\{\epsilon_t\}$ is a sequence of IID random variables with $E(\epsilon_t) = 0$ and $E(\epsilon_t^2) = \sigma_\epsilon^2 < \infty$, **(iii)** $\{\eta_t\}$ and $\{\epsilon_t\}$ are mutually independent, **(iv)** $\{\eta_t\}$ and $\{\mathbf{X}_{t-1}\}$ are mutually independent.

For simplicity, let us consider the estimation of the parameter $\theta$ assuming that the nuisance parameters $(\sigma_\epsilon^2, \Sigma)$ are known. Consider a linear class of EFs of the form $g_i = \sum_{t=1}^{n} \psi_t a_{i,t-1}$, with

$$\psi_t = X_t - E[X_t | \mathcal{F}_{t-1}^X] = X_t - \sum_{i=1}^{p} \theta_i X_{t-i} = X_t - \mathbf{X}_{t-1}' \theta.$$

Note that, the information set $\mathcal{F}_t^X$ is now based on $\{\eta_s; s \leq t\}$ and $\{X_s; s \leq t\}$; so $X_{t-1} \in \mathcal{F}_{t-1}^X$ and $E[(\partial \psi_t / \partial \theta) | \mathcal{F}_{t-1}^X] = -\mathbf{X}_{t-1}$. We derive the optimal EF using Theorem 2.3 as

$$g_i^* = \sum_{t=1}^{n} \psi_t a_{i,t-1}^* \tag{113}$$

where $a_{i,t-1}^* = -X_{t-i}/Q_t$ with $Q_t = E[\psi_t^2 | \mathcal{F}_{t-1}^X] = \sigma_\epsilon^2 + \mathbf{X}_{t-1}' \Sigma \mathbf{X}_{t-1}$. Therefore, by solving the EF $g_i^* = 0$ we obtain the optimal estimate of $\theta$ as

$$\widehat{\theta}_n^{EF} = \left[ \sum_{t=p+1}^{n} \left( \frac{\mathbf{X_{t-1}} \mathbf{X_{t-1}}'}{Q_t} \right) \right]^{-1} \left[ \sum_{t=p+1}^{n} \left( \frac{\mathbf{X_{t-1}} X_t}{Q_t} \right) \right]. \tag{114}$$

The scaling factor $Q_t$ is nothing but the conditional variance of the original random process $X_t$ [Bera, Higgins and Lee (1992)], as it can be seen from

$$
\begin{aligned}
Var(X_t|\mathcal{F}_{t-1}^X) &= E\left[\left(\sum_{i=1}^{p} X_{t-i}\eta_{it} + \epsilon_t\right)^2 \middle| \mathcal{F}_{t-1}^X\right] \\
&= E\left[\left\{\left(\sum_{i=1}^{p} X_{t-i}\eta_{it}\right)^2 + 2\epsilon_t\left(\sum_{i=1}^{p} X_{t-i}\eta_{it}\right) + \epsilon_t^2\right\} \middle| \mathcal{F}_{t-1}^X\right] \\
&= \mathbf{X}_{t-1}'\Sigma\mathbf{X}_{t-1} + \sigma_\epsilon^2 \\
&= Q_t. \qquad (115)
\end{aligned}
$$

When $\Sigma = 0$, both LS and EF estimates of $\theta$ becomes the same, but when $\Sigma \neq 0$ $Q_t$ has an ARCH type form that need to be taken into consideration. Also, whether $\Sigma$ is diagonal or a full matrix has important implications for the joint presence of autocorrelation and conditional heteroscedasticity. If $\Sigma$ is diagonal then the scaling factor of the optimal EF will be same as in Engle (1982). However, if $\Sigma$ has non-zero off-diagonal terms, the interpretation of the scaling factor becomes closer to the *asymmetric* ARCH model proposed by Nelson (1991). For diagonal $\Sigma$, we can write $E(\eta_t\eta_t') = \sigma_\eta^2 I_p$, where $I_p$ is the identity matrix of dimension $p$. The optimal estimate for a first-order RCA model based on EF $g_i^* = \sum_{t=1}^n \psi_t a_{i,t-1}^*$ is given by $\widehat{\theta}_n^{EF}$ in (114), with $Q_t = \sigma_\epsilon^2 + \sum_{i=1}^p X_{t-i}^2\sigma_\eta^2$. This is similar to the traditional generalized LS estimator and is an improvement over the naive LS. This estimator was first proposed by Thavaneswaran and Abraham (1988). To implement (114), in the first step LS estimation can be used to obtain initial estimates of $\sigma_\epsilon^2$ and $\sigma_\eta^2$. Then, in the second step plugging in all the relevant information an efficient estimate can be obtained for $\widehat{\theta}_n^{EF}$. For example, if $p = 1$ the optimal EF turns out to be

$$
g_i^* = \sum_{t=2}^{n}\left(\frac{X_{t-1}}{Q_t}\right)\psi_t, \qquad (116)
$$

where $\psi_t = (X_t - \theta X_{t-1})$ and $Q_t = E[\psi_t^2|\mathcal{F}_{t-1}^X] = \sigma_\epsilon^2 + X_{t-1}^2\sigma_\eta^2$. Therefore, the estimator based on $g_i^*$ is

$$
\widehat{\theta}_n^{EF} = \frac{\sum_{t=2}^n a_{t-1}^* X_t}{\sum_{t=2}^n a_{t-1}^* X_{t-1}}, \qquad (117)
$$

with $a_{t-1}^* = -X_{t-1}/(\sigma_\epsilon^2 + X_{t-1}^2\sigma_\eta^2)$. By letting $u_t = \psi_t^2 - \sigma_\epsilon^2 - X_{t-1}^2\sigma_\eta^2$, the estimates of $\widehat{\sigma}_\epsilon^2$ and $\widehat{\sigma}_\eta^2$ can be obtained by minimizing $\sum_{t=2}^n u_t^2$ with respect to $\sigma_\epsilon^2, \sigma_\eta^2$ [Nicholls and Quinn (1982, p.43)], i.e., by regressing $\widehat{\psi}_t^2$ on 1 and $X_{t-1}^2$. The LS estimate $\widehat{\theta}_n^{LS} = (\sum_{t=2}^n X_t X_{t-1})/(\sum_{t=2}^n X_{t-1}^2)$ can be used to obtain $\widehat{\psi}_t^2$.

## 3.2 Threshold autoregressive model

It is widely known that many nonlinear features such as limit cycles and asymmetry can be explained by threshold autoregressive (TAR) models [Tong (1990, Ch.1)], where we assume that the function $\varphi(.)$ in (111) is piecewise linear and allow the parameters to be determined partly by past data. In simplest form of a TAR model for a time series $\{X_t, t \in \mathbf{Z}\}$ is given by

$$X_t = \theta_1 X_{t-1}^+ + \theta_2 X_{t-1}^- + \epsilon_t \quad , \tag{118}$$

where $X_t^+ = \min(X_t, 0), X_t^- = \max(X_t, 0)$ and $\{\epsilon_t\} \sim IID(0, \sigma_\epsilon^2)$. The process is known as "double threshold" if both the conditional mean and variance change with thresholds [Granger (1998)].

For the general expression of model (118), we partition the range of $X_t$ into $k$ parts by the set of ordered values $r_1 < \ldots < r_{k-1}$. If the value $X_t$ lies in the interval $D_j = (r_j, r_{j+1}]$ with $r_j, r_{j+1}$ as threshold values, the $j$th set of parameters is used to generate $X_{t-d}$ $(d < k)$ where $d$ is the delay (lag) parameter. The zero mean threshold AR($p$) (TAR($p$)) process can be expressed as

$$X_t = \sum_{i=1}^{p} \theta_i^j X_{t-i} I(X_{t-d} \in D_j) + \epsilon_t,$$

where $I(.)$ is the indicator function, $\epsilon_t$ is a white noise and $j \in \{1, \ldots, k\}$ is determined by $X_{t-d} \in D_j$. By considering the elementary EF

$$\psi_t = X_t - \sum_{i=1}^{p} \sum_{j=1}^{k} \theta_i^j X_{t-i} I(X_{t-d} \in D_j),$$

and noting that $E(\psi_t^2) = \sigma_\epsilon^2$, the optimal EF for the set of parameters $(\theta_i^j)$ becomes [see Ainkaran (2004)]

$$\begin{aligned} g^*_{\theta_i^j} &= \sum_{t=k+1}^{n} \psi_t a^*_{t-1,\alpha_i} \\ &= -\sum_{t=k+1}^{n} X_{t-i} I(X_{t-d} \in D_j) \left( X_t - \sum_{i=1}^{p} \sum_{j=1}^{k} \theta_i^j X_{t-i} I(X_{t-d} \in D_j) \right) / \sigma_\epsilon^2 \quad . \end{aligned}$$

Solving the corresponding EEs, we obtain the optimal EF estimates for $TAR(p)$ parameters. For a discussion on generalized kernel smoothing estimate using optimal EF approach for threshold models, see Thavaneswaran and Peiris (1996).

## 3.3 Bilinear model

In the bilinear class, we incorporate cross-product terms involving lagged values of the time series and of the disturbance process. This class of models, originally introduced by Granger and Anderson (1978), has many interesting statistical properties and act as competing models with ARCH for *nonlinear dependence* [e.g., see Weiss (1986), Bera and Higgins (1997)]. However, it is important to note that even though both ARCH and bilinear process have similar unconditional moments, conditionally their moment structure is different. The simplest form of bilinear time series $\{X_t, t \in \mathbf{Z}\}$ model is given by

$$X_t = \sum_{i=1}^{r} \sum_{j=1}^{s} \theta_{ij} X_{t-i} \epsilon_{t-j} + \epsilon_t \tag{119}$$

where $\{\epsilon_t\} \sim IID(0, \sigma_\epsilon^2)$ is the innovation that drives the bilinear process and $\{\theta_{ij}, i = 1, \ldots, r, j = 1, \ldots, s\}$ are parameters to be estimated. Here the conditional mean is a nonlinear function of past values of $\{X_t, \epsilon_t\}$ while the conditional variance is constant. This is in contrast with ARCH process, as we will see in Section 3.4, the conditional mean is in general, a constant, but conditional variance is time varying.

The general expression of bilinear model $BL(p, q, r, s)$ can be obtained by adding a linear ARMA component to (119), such as

$$X_t = \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{i=1}^{r} \sum_{j=1}^{s} \theta_{ij} X_{t-i} \epsilon_{t-j} + \epsilon_t + \sum_{j=1}^{q} \beta_j \epsilon_{t-j} \quad . \tag{120}$$

Here, in addition to $\theta_{ij}$, we need to estimate $p+q$ parameters $\alpha_i$ and $\beta_j$. As before let us assume that the conditioning information set $\mathcal{F}_t^X$ is a $\sigma$-field, based on $\{X_r; r \leq t\}$ and $\{\epsilon_r; r \leq t\}$; so $X_{t-1} \in \mathcal{F}_{t-1}^X$ and $E[\epsilon_{t-i}|\mathcal{F}_{t-1}^X] = \epsilon_{t-i}, i \geq 1$. Also, as $\{\epsilon_t\} \sim IID(0, \sigma_\epsilon^2)$, an obvious choice for an elementary EF becomes $\psi_t = X_t - E[X_t|\mathcal{F}_{t-1}^X]$, where

$$E[X_t|\mathcal{F}_{t-1}^X] = \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{i=1}^{r} \sum_{j=1}^{s} \theta_{ij} X_{t-i} \epsilon_{t-j} + \sum_{j=1}^{q} \beta_j \epsilon_{t-j} \quad .$$

Therefore, following Theorem 2.3, the optimal EF is given by the following set of equations [see Ainkaran (2004)]:

$$
\begin{aligned}
g^*_{\alpha_i} &= \sum_{t=m+1}^{n} \psi_t a^*_{t-1,\alpha_i} \quad , \\
g^*_{\beta_j} &= \sum_{t=m+1}^{n} \psi_t a^*_{t-1,\beta_j} \quad , \\
g^*_{\theta_{ij}} &= \sum_{t=m+1}^{n} \psi_t a^*_{t-1,\theta_{ij}} \quad , \qquad (121)
\end{aligned}
$$

where

$$
a^*_{t-1,\alpha_i} = E\left[\frac{\partial \psi_t}{\partial \alpha_i}\middle| \mathcal{F}^X_{t-1}\right] / Q_t = -X_{t-i}/Q_t \quad ,
$$

$$
a^*_{t-1,\beta_j} = E\left[\frac{\partial \psi_t}{\partial \beta_j}\middle| \mathcal{F}^X_{t-1}\right] / Q_t = \left(-\epsilon_{t-j} - \beta_j \frac{\partial \epsilon_{t-j}}{\partial \beta_j} - \sum_{i=1}^{r} \theta_{ij} X_{t-i} \frac{\partial \epsilon_{t-j}}{\partial \beta_j}\right) / Q_t \quad ,
$$

$$
a^*_{t-1,\theta_{ij}} = E\left[\frac{\partial \psi_t}{\partial \theta_{ij}}\middle| \mathcal{F}^X_{t-1}\right] / Q_t = \left(-X_{t-i}\epsilon_{t-j} - \beta_j \frac{\partial \epsilon_{t-j}}{\partial \theta_{ij}} - \theta_{ij} X_{t-i} \frac{\partial \epsilon_{t-j}}{\partial \theta_{ij}}\right) / Q_t \quad ,
$$

$$
Q_t = E[\psi_t^2 | \mathcal{F}^X_{t-1}] = \sigma^2_\epsilon \left(1 + \sum_{j=1}^{q} \beta_j^2 + \sum_{i=1}^{r}\sum_{j=1}^{s} \theta_{ij}^2 X^2_{t-i}\right)
$$

and $m = \max(p, r)$. Solving EEs corresponding to EFs in (121), the estimates of bilinear models $(p + q + rs)$ parameters can be obtained.

## 3.4 ARCH and GARCH models

The ARCH model introduced by Engle (1982), and its various extension, have become arguably the most popular and extensively used financial econometric models [for surveys on this topic, see Bera and Higgins (1993), Bollerslev, Engle and Nelson (1994) and Engle (2002)]. The standard procedure is to estimate an ARCH or GARCH model using ML approach assuming normal or Student's t distribution. However, such assumptions are hard to justify in practice due to the presence of asymmetry and high excess kurtosis in real data. Li and Turtle (2000) and Chandra and Taniguchi (2001) proposed EF method that is free of any distributional assumptions.

A general expression for an ARCH($p$) model is given by

$$X_t = \epsilon_t \sqrt{h_t}, \quad h_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2, \tag{122}$$

where $\{\epsilon_t\} \sim IID(0, \sigma_\epsilon^2)$ with fourth-order cumulant $\kappa_4$ and $\alpha_0 > 0, \alpha_j \geq 0, \forall j = 1, \ldots, p$. A candidate class for unbiased and mutually orthogonal EFs is $\psi_t = X_t^2 - h_t, t = 1, \ldots, n$. The linear combination of which becomes $g_\alpha = \sum_{t=1}^{n} a_t \psi_t$, where the weights $a_t$ are functions of the data and the unknown parameter vector $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_p)'$. Using Theorem 2.3, we can derive the optimal EF as $g_\alpha^* = \sum_{t=1}^{n} a_t^* \psi_t$, where

$$\begin{aligned} a_t^* &= \frac{E\left[\frac{\partial \psi_t}{\partial \alpha} | \mathcal{F}_{t-1}^X\right]}{E[\psi_t^2 | \mathcal{F}_{t-1}^X]} \\ &= -\frac{\partial h_t}{\partial \alpha} / \{E[X_t^4 | \mathcal{F}_{t-1}^X] - h_t^2\} \\ &= -\frac{\partial h_t}{\partial \alpha} / \{(\kappa_4 + 2) h_t^2\}, \end{aligned}$$

and $\mathcal{F}_{t-1}^X$ is the $\sigma$-field generated by $\{X_s; s \leq t-1\}$. Therefore, the optimal EF estimate of the ARCH($p$) parameters turns out to be the solution of

$$g_\alpha^* = -\sum_{t=1}^{n} \frac{\frac{\partial h_t}{\partial \alpha}(X_t^2 - h_t)}{\{(\kappa_4 + 2) h_t^2\}} = 0.$$

Next let us consider ARCH($p$) errors in the context of a linear regression model:

$$y_t = z_t \beta + X_t, \quad X_t | \mathcal{F}_{t-1}^X \sim (0, h_t), \tag{123}$$

where $z_t$ is non-stochastic regressors and $\beta$ represents regression coefficient. The conditional mean from (123) becomes nonzero as $E(y_t | \mathcal{F}_{t-1}^X) = z_t \beta$ and $\mathcal{F}_{t-1}^X$ is now the $\sigma$-field generated by $\{z_t, X_{t-1}, X_{t-2}, \ldots\}$. The objective is to estimate the set of parameters $\alpha$ and $\beta$. Let us denote the skewness and excess kurtosis coefficient as $\gamma_{1t} = \frac{E[(y_t - z_t\beta)^3 | \mathcal{F}_{t-1}^X]}{h_t^{3/2}}$ and $\gamma_{2t} = \frac{E[(y_t - z_t\beta)^4 | \mathcal{F}_{t-1}^X]}{h_t^2} - 3$, respectively, and choose the following two orthogonal EFs $\psi_{1t} = (y_t - z_t\beta)$ and $\psi_{2t} = (y_t - z_t\beta)^2 - h_t - \gamma_{1t} h_t^{1/2}(y_t - z_t\beta)$. Then following Theorem 2.3, the optimal EF becomes $g_{\alpha,\beta}^* = g_1^* + g_2^* = 0$, with

$$
\begin{aligned}
g_1^* &= \sum_{t=1}^{n} \frac{E\left[\frac{\partial \psi_{1t}}{\partial \alpha}|\mathcal{F}_{t-1}^X\right]}{E[\psi_{1t}^2|\mathcal{F}_{t-1}^X]}\psi_{1t} + \sum_{t=1}^{n} \frac{E\left[\frac{\partial \psi_{2t}}{\partial \alpha}|\mathcal{F}_{t-1}^X\right]}{E[\psi_{2t}^2|\mathcal{F}_{t-1}^X]}\psi_{2t} \\
&= -\sum_{t=1}^{n} \frac{\frac{\partial h_t}{\partial \alpha}}{h_t^2(\gamma_{2t}+2-\gamma_{1t}^2)}\psi_{2t},
\end{aligned}
$$

and

$$
\begin{aligned}
g_2^* &= \sum_{t=1}^{n} \frac{E\left[\frac{\partial \psi_{1t}}{\partial \beta}|\mathcal{F}_{t-1}^X\right]}{E[\psi_{1t}^2|\mathcal{F}_{t-1}^X]}\psi_{1t} + \sum_{t=1}^{n} \frac{E\left[\frac{\partial \psi_{2t}}{\partial \beta}|\mathcal{F}_{t-1}^X\right]}{E[\psi_{2t}^2|\mathcal{F}_{t-1}^X]}\psi_{2t} \\
&= -\sum_{t=1}^{n} \frac{\frac{\partial z_t\beta}{\partial \beta}}{h_t}\psi_{1t} - \sum_{t=1}^{n} \frac{h_t^{1/2}\gamma_{1t}\frac{\partial z_t\beta}{\partial \beta} - \frac{\partial h_t}{\partial \beta}}{h_t^2(\gamma_{2t}+2-\gamma_{1t}^2)}\psi_{2t}.
\end{aligned}
$$

The above discussion is also valid for the class of GARCH processes with

$$
h_t = Var(X_t|\mathcal{F}_{t-1}^X) = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + \sum_{j=1}^{q} \delta_j h_{t-j}. \tag{124}
$$

Here, in addition to $\alpha$ and $\beta$, we need to estimate $q$ additional parameters $\delta = (\delta_1, \ldots, \delta_q)'$. It is easy to see that, $E[X_tX_{t-k}] = E[E(X_tX_{t-k}|\mathcal{F}_{t-1}^X)] = 0, \forall k \geq 1$, and hence GARCH errors are uncorrelated.

To illustrate the usefulness of optimal EF approach, we concentrate on a simple GARCH$(1,1)$ process given by $h_t = \alpha_0 + \alpha_1 X_{t-1}^2 + \delta_1 h_{t-1}$. As with the ARCH model, let us choose the same two orthogonal EFs $\psi_{1t}$ and $\psi_{2t}$. Then by denoting $\theta = (\alpha_0, \alpha_1, \delta_1)'$, following Theorem 2.3, we obtain the optimal EF as $g_{\theta,\beta}^* = g_1^* + g_2^* = 0$, where

$$
\begin{aligned}
g_1^* &= \sum_{t=1}^{n} \frac{E\left[\frac{\partial \psi_{1t}}{\partial \theta}|\mathcal{F}_{t-1}^X\right]}{E[\psi_{1t}^2|\mathcal{F}_{t-1}^X]}\psi_{1t} + \sum_{t=1}^{n} \frac{E\left[\frac{\partial \psi_{2t}}{\partial \theta}|\mathcal{F}_{t-1}^X\right]}{E[\psi_{2t}^2|\mathcal{F}_{t-1}^X]}\psi_{2t} \\
&= -\sum_{t=1}^{n} \frac{\frac{\partial h_t}{\partial \theta}}{h_t^2(\gamma_{2t}+2-\gamma_{1t}^2)}\psi_{2t},
\end{aligned}
$$

$$
\begin{aligned}
g_2^* &= \sum_{t=1}^{n} \frac{E\left[\frac{\partial \psi_{1t}}{\partial \beta}|\mathcal{F}_{t-1}^X\right]}{E(\psi_{1t}^2|\mathcal{F}_{t-1}^X)}\psi_{1t} + \sum_{t=1}^{n} \frac{E\left[\frac{\partial \psi_{2t}}{\partial \beta}|\mathcal{F}_{t-1}^X\right]}{E[\psi_{2t}^2|\mathcal{F}_{t-1}^X]}\psi_{2t} \\
&= -\sum_{t=1}^{n} \frac{\frac{\partial z_t\beta}{\partial \beta}}{h_t}\psi_{1t} + \sum_{t=1}^{n} \frac{h_t^{1/2}\gamma_{1t}\frac{\partial z_t\beta}{\partial \beta} - \frac{\partial h_t}{\partial \beta}}{h_t^2(\gamma_{2t}+2-\gamma_{1t}^2)}\psi_{2t}.
\end{aligned}
$$

If we impose conditional normality, i.e., $\gamma_{1t} = 0, \gamma_{2t} = 0$, the optimal EFs become

$$g_1^* = -\sum_{t=1}^{n} \frac{1}{2h_t} \left\{ \frac{\partial h_t}{\partial \theta} \left( \frac{X_t^2}{h_t} - 1 \right) \right\} = 0,$$

$$g_2^* = -\sum_{t=1}^{n} \left( \frac{z_t X_t}{h_t} \right) - \sum_{t=1}^{n} \frac{1}{2h_t} \left\{ \frac{\partial h_t}{\partial \beta} \left( \frac{X_t^2}{h_t} - 1 \right) \right\} = 0$$

which are, as expected, similar to the first-order conditions for MLE under the normality assumption.

## 3.5 Spatial regression model

Recently there has been a considerable interest among economists in the applications of spatial econometric techniques to an increasing number of problems [see Anselin and Bera (1998) and Anselin (2006)]. Due to its unique nature and defining characteristic, no existing method is dominant for modelling spatial data, and operational implementation is still a debatable issue. In this section we discuss the implementation of the optimal EF technique of Section 2.4 in a simple spatial regression set up following Naik-Nimbalkar (1996) [for additional references, see Lele (1997) and Yasui and Lele (1997)].

Consider the following simple simultaneous model known as spatial autoregressive model of first order: $y = \rho W y + \epsilon$, where $W = ((w_{ij}))$ a $n \times n$ weights matrix. For the $i$th observation, the model can be written as

$$y_i = \rho \sum_{j \neq i}^{n} w_{ij} y_j + \epsilon_i, i = 1, ..., n \quad , \tag{125}$$

where we use the sum over "neighbors $j$" of the $i$th cross sectional observation and assume $\epsilon_i \sim IID(0, \sigma_\epsilon^2)$. We are interested in estimating the spatial dependence parameter $\rho$ and the distribution of the error term is not known. It is very difficult to derive optimal EF for a general W matrix. However, assuming $w_{i,i+1} = w_{i,i-1} = 1$ for all $i = 1, ..., n$ with all other $w_{ij} = 0$, we can easily obtain an optimal EF. Define $\psi_i = \epsilon_i \epsilon_{i+1}$, where $\epsilon_i = y_i - \rho(y_{i-1} + y_{i+1})$ and $\epsilon_{i+1} = y_{i+1} - \rho(y_i + y_{i+2})$. The implication becomes clear as due to the independence of $\epsilon_i$'s, $\psi_i, i = 2, \ldots, n-2$, are mutually orthogonal for any trivial conditioning $\sigma$-field. The optimal EF in the class $\{\sum_{i=2}^{n-2} \psi_i a_i\}$ with $a_i$ non-random functions of $\rho$, becomes

$$g^* = \sum_{i=2}^{n-2} \left\{ E\left[\frac{\partial \psi_i}{\partial \rho}\right] / \sigma_\epsilon^4 \right\} \psi_i. \tag{126}$$

Since stationarity implies $E\left[\frac{\partial \psi_i}{\partial \rho}\right]$ is constant, the optimal EE becomes $\sum_{i=2}^{n-2} \psi_i = \sum_{i=2}^{n-2} \epsilon_i \epsilon_{i+1} = 0$, which is basically the weighted LS equation suggested by Ord (1975).

We can generalize the above discussion by using conditional moment functions and exploiting their optimal orthogonal combinations. For example, consider the $\sigma$-field $Q_i = \sigma\{J(i)|i \neq j; i, j = 1, \ldots, n\}$ defined over the information set $J(i)$ which includes all locations other than $i$. Then if we define the following conditional moments

$$E[y_i|Q_i] = m_{1i}(\theta; y_{i-1}, y_{i+1}) = m_{1i},$$

$$Var[y_i|Q_i] = m_{2i}(\theta; y_{i-1}, y_{i+1}) = m_{2i},$$

the possible elementary EFs turn out to be $\psi_i = y_i - m_{1i}(\theta; y_{i-1}, y_{i+1})$, with $E[\psi_i|Q_i] = 0, \forall i = 1, \ldots, n$. But since $\{\psi_i\}$'s are not mutually orthogonal, using Besag's (1974) coding method we can obtain a set of mutually orthogonal EFs as the subclass of functions $\{\psi_i$ for $i$-even$\}$ and $\{\psi_i$ for $i$-odd$\}$. Therefore, the optimal combination of EFs becomes

$$g_1^* = \sum_{i=odd} (y_i - m_{1i}) \frac{1}{m_{2i}} \left[\frac{\partial m_{1i}}{\partial \theta}\right] \tag{127}$$

and

$$g_2^* = \sum_{i=even} (y_i - m_{1i}) \frac{1}{m_{2i}} \left[\frac{\partial m_{1i}}{\partial \theta}\right]. \tag{128}$$

Then, under the assumption of strong stationarity of the underlying process, the optimal linear combination of $g_1^*$ and $g_2^*$ will be in the class $\mathcal{G} = \{ag_1^* + bg_2^*\}$, with $a, b$ being real functions of $\theta$. This class of optimum EFs is more meaningful if either $a$ and $b$ are known, or $a = b$, suggesting the optimal equation as $g_1^* + g_2^* = 0$; this is basically the same as the equation obtained from maximizing Besag's pseudo-likelihood (Besag 1974, 1977).

Therefore, we can see that the usefulness of EF optimality in both simultaneous and conditional spatial models can be interpreted as the existing methods of Ord (1975) and Besag (1974, 1977). Interestingly, the underlying notion of orthogonality is not unique and can be achieved by many alternative ways as discussed above, i.e., by constructing different sub-lattices such that one is independent of other and then

reversing the procedure and combining all separate estimates. For future research, Godambe's flexible approach can be generalized to accommodate higher dimensional spatial autoregressive process with non-stochastic regressors and a general form of simultaneous or conditional specification.

## 3.6 Longitudinal data analysis

The Generalized EE (GEE) approach was devised by Liang and Zeger (1986) to deal with longitudinal data. In longitudinal data, we are presented with repeated measurements on different cross sectional units over time. It is typically assumed that the cross sectional units are independent, but the time series data on the same subject are positively correlated. The GEE methodology was formulated especially from the need to handle discrete type data where no Gaussian likelihood seemed to be appropriate. At the end, the approach looks like an extension of the Wedderburn's (1974) quasi-likelihood to a class of correlated data, which models only the mean and the variance of the responses instead the full joint distribution.

To establish notation, suppose the (balanced) panel data set consists of responses $y_{it}, \quad i = 1, 2, \ldots, n; \;\; t = 1, 2, \ldots, T$ on $n$ units over a $T$ periods. The $nT \times 1$ vector $\mathbf{y} = (y_{11}, \ldots, y_{1T}, \ldots, y_{n1}, \ldots, y_{nT})'$ has a corresponding mean model $\boldsymbol{\mu} = (\mu_{11}, \ldots, \mu_{1T}, \ldots, \mu_{n1}, \ldots, \mu_{nT})'$ and by assumption has a variance-covariance matrix $\mathbf{V}$ with block diagonal structure

$$\mathbf{V} = diag(\mathbf{V_1}, \mathbf{V_2}, \ldots, \mathbf{V_n}).$$

Also, we will assume that $\mathbf{V_i} = \mathbf{V_i}(\mu_{i1}, \ldots, \mu_{iT}, \lambda_i)$ for $i = 1, 2, \ldots, n$, where $\lambda_i$ is a parameter characterizing variance and correlation components. In addition each mean $\mu_{it} = \mu_{it}(\theta)$ depends on a $p \times 1$ coefficient vector $\theta$.

From the family of EFs

$$\mathbf{\{A(y - \mu)\}} \tag{129}$$

where $\mathbf{A} = nT \times nT$, the quasi-score EF, i.e., the optimal EF, is given by

$$\boldsymbol{\dot{\mu}'} \mathbf{V^{-1}(y - \mu)}$$

or, by exploiting the block-diagonal structure of the covariance matrix $\mathbf{V}$,

$$\boldsymbol{U(\theta)} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{\dot{\mu}_i'} \mathbf{V_i^{-1}(y_i - \mu_i)}, \tag{130}$$

where $\mathbf{y_i} = (y_{i1}, \ldots, y_{iT})$, $\boldsymbol{\mu_i} = (\mu_{i1}, \ldots, \mu_{iT})$, and $\boldsymbol{\dot{\mu}_i} = \partial \boldsymbol{\mu_i} / \partial \theta$. The GEE is based

on the EF $\mathbf{U}(\theta)$ at (130), and the estimator is obtained by finding the root of (130).

One characteristic of the GEE methodology is the use of the so-called "working" covariance matrix in place of the generally unknown matrix $\mathbf{V}$. Even if the "working" covariance matrices are misspecified, Liang and Zeger (1986) show that the GEE estimator will be consistent although generally inefficient. When a consistent estimator of true $\mathbf{V}$ is utilized, then the estimator from (130) is efficient. The consistency of the estimator requires only correct specification of the mean functions $\mu_{it}$.

As it has been presented in the literature, GEE corresponds to the random effects models but it treats the variance components $\lambda_i$ as nuisance parameters. The advantage of this approach is that it can handle in a unified way a variety of types of data as continuous data, discrete data, or count data. For a detailed review of the literature and related references see Fitzmaurice, Laird and Rotnitzky (1993).

## 3.7 Median regression model

Consider the median regression model proposed by Koenker and Basset (1978):

$$y_i = \mu_i(\beta) + \epsilon_i, \qquad i = 1, 2, \ldots, n$$

where $\epsilon_i$ is a random variable with median zero and marginal pdf $f_i$. Let

$$f_i(0) = \frac{1}{\phi}\gamma(\mu_i),$$

be the pdf of $y_i$ at $\mu_i$, where $\phi > 0$ is a scale parameter, and $\gamma$ is considered a known function. We will assume the regularity condition that $\gamma(\mu_i) > 0$, which is needed for the median to be unique. For the later use we will denote the $n \times 1$ vector of medians of $y_i$'s by $\mu = (\mu_1(\beta), \ldots, \mu_n(\beta))'$. Jung (1996) analyzed the estimation of median regression models using the approach of Wedderburn (1974) and Godambe and Heyde (1987). In the following we use Jung's framework and notation; a similar analysis is given by Godambe (2001).

For estimation of $\beta$, the $p \times 1$ vector parameter, we start from the $n$ elementary EFs

$$\left\{ I(y_i - \mu_i(\beta) \geq 0) - \frac{1}{2} \right\}, \qquad i = 1, 2, \ldots, n; \tag{131}$$

which clearly have zero expectation. Let $V$ denote the $n \times n$ variance-covariance matrix of the elementary EFs; the $n$ diagonal elements of $V$ equal $1/4$.

The $n$ elementary EFs can be combined linearly in an optimal way by using the

theory developed by Godambe and Heyde (1987). Consider any $n \times p$ matrix $H$ of rank $p$. The unbiasedness of the elementary EFs ensures that $p \times 1$ EF

$$U_H(\beta) = \phi^{-1} H' \begin{pmatrix} I(y_1 - \mu_1(\beta) \geq 0) - \frac{1}{2} \\ \vdots \\ I(y_n - \mu_n(\beta) \geq 0) - \frac{1}{2} \end{pmatrix}$$

will deliver a consistent estimator of $\beta$.

The choice of weighting matrix $H' = D'\Gamma V^{-1}$, where $D = \partial\mu/\partial\beta$, and $\Gamma = diag\{\gamma(\mu_1), \ldots, \gamma(\mu_n)\}$, yields the optimal EF $U_{opt}$ within the class of linear combinations of (131). By solving the system of $p$ equations $U_{opt}(\beta) = 0$, we obtain the so called quasi-likelihood estimator $\hat{\beta}$. If the true model is double exponential, $\hat{\beta}$ is the MLE. We note that the use of optimal EF theory makes clear from the outset the role of the density $f_i$ of $y_i$'s. Optimality dictates weighting the elementary EFs (131) in a way that is reminiscent of weighted LS. In case of identically and independently distributed random variables, the density is constant and it falls out of the picture. The optimum EF reduces to

$$U_{opt} = D' \begin{pmatrix} I(y_1 - \mu_1(\beta) \geq 0) - \frac{1}{2} \\ \vdots \\ I(y_n - \mu_n(\beta) \geq 0) - \frac{1}{2} \end{pmatrix}. \tag{132}$$

Furthermore, in case of the linear median regression model, $\mu_i(\beta) = x_i'\beta$, the resulting system of EEs (132) is the familiar sum of cross products of the $x_i$'s with the elementary EF's (131).

The advantage of the optimal EF approach to estimation of the median regression model, is that the form of optimal EEs $U_{opt}(\beta) = 0$ allows for a wide variety of data structures, such as dependent or heteroscedastic data, offering a unified treatment. In addition, due to the invariance of medians to monotone transformations, we can handle censored [Powell (1984)] or binary data [Manski (1975)] as well.

# 4 Epilogue

It was the 1930s. The conflict between the two statistical giants Karl Pearson and R. A. Fisher was at its height. One issue of their heated arguments was the relative merits of the MM and ML approaches. "I am even ready to adopt new methods," Karl Pearson wrote to Fisher on August 28, 1935, "if they are quicker and more exact than the old. Now I do not suppose you spend much, if any, time in fitting frequency curves; nevertheless I should like to have your method of fitting them to

observations, which avoids the 'traditional but inefficient method of fitting them by moments.' (*Annals of Eugenics* Vol VI p.252) It would aid me in many inquiries, if you would let me know the more efficient way." On August 30, 1935, Fisher sent a prompt reply, "The fullest examination of the method of moments in fitting the Pearsonian curves is in a paper 'On the mathematical foundations of theoretical statistics,' *Phil. Trans. A*, ccxxii. 309-368. High efficiencies are only obtained in the neighborhood of normal curve. Efficient equations of estimation may always be obtained by the maximum likelihood." The acrimonious debate culminated in two final papers. Karl Pearson in one of his very last papers that was published in June 1936 issue of *Biometrika* after he passed away on April 26, 1936, began with the italized and striking line, "*Wasting your time fitting curves by moments, eh* ?" Fisher, not to be outdone, sent an equally scathing reply. After his step by step rebuttal to Pearson's (1936) arguments, Fisher (1937, p.317), now feeling free after Pearson's death, bluntly stated : "So long as 'fitting curves by moments' stands in the way of students' obtaining proper experience of these other activities, all of which require time and practice, so long will it be judged with increasing confidence to be waste of time." MM was basically swept away by ML revolution; Fisher and his method came out to be winner from this battle. For several decades chapters were devoted to ML method in statistics (and econometrics) textbooks, while MM had only scant mentions. However, now it appears that after all MM did not lose the war, and econometricians can take credit in reviving the MM approach through GMM.

Looking back at the Fisher-Pearson conflict after nearly seven decades on the light of Godambe's EF approach, much of the sharpness of their debate is lost, as Desmond (1997, pp.116-117) noted, "One of the advantage of estimating functions framework is that the apparent dichotomy between these two methods (MM and ML) is nullified and it is possible to see these methods as lying within a unifying framework of continuum, ranging from weak second-order assumption to fully specified parametric models." It is indeed ironic that EF method which is essentially a Pearsonian-type moment-based approach provides, as we discussed in Section 2, a *finite sample* justification to Fisher's *asymptotically* efficient ML method. Apart from the potential practical applications some of which are discussed in Section 3, this unifying feature of the EF method is very attractive along with its philosophical and foundational approach.

Of course, we could not do full justice to the proliferations of papers written in theory and applications of EFs. For instance, we did not cover the topic of hypothesis testing based on EFs. McLeish and Small (1988, p.10) argued that EFs can be regarded as vehicles for more general focus on inference than simple estimation, and

they preferred to call these functions "inference functions." For discussion on tests utilizing EFs, see Basawa (1985, 1991), Hall and Mathiason (1990), Thavaneswaran (1991), Bhat (1996) and Heyde (1997, Ch.9). A very related area to EF method of estimation is the empirical likelihood (EL) approach. The link between EF and EL method and how to combine different sources of information on parameters are discussed in Qin and Lawless (1994) and Owen (2001, pp.39-42, 51-55). We have tried to project EF method emphasizing its finite sample justification. However, the consistency and asymptotic normality of the resulting estimator are also important issues and good references on these are Crowder (1986), Heyde (1997, Ch.12) and Sørensen (1999).

To conclude, in this chapter we have reviewed the important phases in the development of EF method which now appears to have at least a century-old history. We have stressed the historical continuity in our discussion. It appears that regarding the choice of estimation techniques, we are now back to the Pearsonian MM paradigm which now looks more useful than ever after a very long devotion to Fisher's ML approach. Given that economic theory provides characterization of the stochastic laws mostly in terms of moment restrictions and EF approach is a sufficiently flexible moment-based method, usefulness of this estimation technique looks very promising in econometric applications.

# 5 References


Ainkaran, P., 2004. Analysis of some linear and nonlinear time series models, unpublished thesis, School of Mathematics and Statistics, University of Sydney.

Anselin, L., 2006. Spatial econometrics. (This volume).

Anselin, L. and Bera, A. K., 1998. Spatial dependence in linear regression models with an introduction to spatial econometrics. In: A. Ullah and D.E.A. Giles (Editors), *Handbook of Applied Economic Statistics*, 237- 289.

Barnard, G. A., 1973. Maximum likelihood and nuisance parameters, Sankhya: The Indian Journal of Statistics, A, 35, 133-138.

Basawa, I. V., 1985. Neyman-LeCam tests based on estimating functions. In: *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Keifer*, 2, (Wadsworth, Monterey), 811-826.

Basawa, I. V., 1991. Generalized score tests for composite hypothesis. In: *Estimating Functions*, V. P. Godambe (Editor), (Oxford University Press, Oxford), 121-131.

Basawa, I. V., Godambe, V. P. and Taylor, R. L., (Editors), 1997. *Selected Proceedings of the Symposium on Estimating Functions*, (Institute of Mathematical Statistics, Lecture Notes - Monograph Series), Vol. 32.

Bellhouse, D. R., 1992. The life and times of V. P. Godambe. In: J. Chen (Editor), *Recent concepts in Statistical Inference*, Proceedings of a Symposium in Honor of Prof. V. P. Godambe, (University of Waterloo, Canada), 1-5.

Bera, A. K. and Bilias, Y., 2001a. On some optimality properties of Fisher-Rao score function in testing and estimation, Communications in Statistics - Theory and Methods, 30 , 1533-1559.

Bera, A. K. and Bilias, Y., 2001b. Rao's score, Neyman's $C(\alpha)$ and Silvey's LM tests: an essay on historical developments and some new results, Journal of Statistical Planning and Inference, 97, 9-44.

Bera, A. K. and Bilias, Y., 2002. The MM, ME, ML, EL, EF and GMM approaches to estimation: A synthesis, Journal of Econometrics, 107, 51-86.

Bera, A. K. and Higgins, M. L., 1993. ARCH Models: properties, estimation and testing, Journal of Economic Surveys, 7, 305-366.

Bera, A. K. and Higgins, M. L., 1997. ARCH and bilinearity as competing models for nonlinear dependence, Journal of Business and Economic Statistics, 15, 43-50.

Bera, A. K., Higgins, M., L and Lee, S., 1992. Interaction between autocorrelation and conditional heteroscedasticity: A random coefficient approach, Journal of Business and Economic Statistics, 10, 133-142.

Bera, A. K. and Lee, S., 1993. Information matrix test, parameter heterogeneity and ARCH: A synthesis, Review of Economic Studies, 60, 229-240.

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems (with discussion), Journal of Royal Statistical Society, Series B, 36, 192-236.

Besag, J., 1977. Efficiency of pseudo-likelihood estimators for simple gaussian fields, Biometrika, 64, 616-618.

Bhapkar, V. P., 1972. On a measure of efficiency of an estimating equation, Sankhya: The Indian Journal of Statistics, A, 34, 467-472.

Bhapkar, V. P., 1991. Sufficiency, ancillarity and information in estimating functions. In: *Estimating Functions*, V. P. Godambe (Editor), (Oxford University Press, New York), 241-254.

Bhat, B. R., 1990. Optimal statistical estimating functions, Paper presented at the Indian Science Congress, Cochin.

Bhat, B. R., 1996. Tests based on estimating functions. In: Prakasa Rao, B. L. S., Bhat, B. R., (Editors), *Stochastic Processes and Statistical Inference*, (New Age International Publishers, New Delhi), 20-38.

Bollerslev, T., Engle, R. F., and Nelson, D. B., 1994. ARCH models. In: *Handbook of Econometrics*, R. F. Engle and D. McFadden (Editors), (Amsterdam: North-Holland), Vol. 4, 2959-3038.

Chamberlain, G., 1987. Asymptotic efficiency in estimation with conditional moment restrictions, Journal of Econometrics, 34, 305-334.

Chandrasekhar, B. and Kale, B. K., 1984. Unbiased statistical estimating functions in presence of nuisance parameters, Journal of Statistical Planning and Inference, 9, 45-54.

Chandra, A. S. and Taniguchi, M., 2001. Estimating functions for non-linear time series models, Annals of the Institute of Statistical Mathematics, 53, 125-141.

Chen, J., (Editor)., 1992. *Recent Concepts in Statistical Inference*. Proceedings of a Symposium in Honor of Prof. V. P. Godambe, (University of Waterloo, Canada).

Cramér, H., 1946. *Mathematical Methods of Statistics*, (Princeton University Press, Princeton, NJ).

Crowder, M., 1986. On consistency and inconsistency of estimating equations, Econometric Theory, 2, 305-330.

Davidson, D. and MacKinnon, J. G., 1993. *Estimation and Inference in Econometrics*, (Oxford University Press, New York).

Davidson, D. and MacKinnon, J. G., 2004. *Econometric Theory and Methods*, (Oxford University Press, New York).

Desmond, A. F., 1989. The theory of estimating equations. In: S. Kotz, N. L. Johnson, and C. B. Read (Editors), *Encyclopedia of Statistical Sciences*, Supplement Volume, (Wiley, New York), 56-59.

Desmond, A. F., 1997. Optimal estimating functions, quasi-likelihood and statistical modelling (with discussion), Journal of Statistical Planning and Inference, 60, 77-121.

Durbin, J., 1960. Estimation of parameters in time-series regression models, Journal of the Royal Statistical Society 22, Series B, 139-153.

Durbin, J., 1997. Optimal estimating equations for state vectors in non-Gaussian and nonlinear space time series models. In: I. V. Basawa, V. P. Godambe and R. L. Taylor (Editors), *Selected Proceedings of the Symposium on Estimating Functions*, (IMS Lecture Note- Monograph Series), Vol. 32, 285-291.

Edgeworth, F. Y., 1908. On the probable errors of frequency-constants, Journal of the Royal Statistical Society, 71, 381-397, 499-512, 651-678.

Edgeworth, F. Y., 1909. Addendum on "Probable errors of frequency-constants", Journal of the Royal Statistical Society, 72, 81-90.

Engle, R. F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation, Econometrica, 50, 987-1007.

Engle, R. F., 2002. New frontiers for ARCH models, Journal of Applied Econometrics, 17, 425-446.

Fitzmaurice, G. M., Laird, N. M. and Rotnitzky, A. G., 1993. Regression models for discrete longitudinal responses (with discussion), Statistical Science, 8, 284-309.

Fisher, R. A., 1912. On an absolute criterion for fitting frequency curves, Messenger of Mathematics, 41, 155-160.

Fisher, R. A., 1922. On the mathematical foundations of theoretical statistics, Philosophical Transactions of the Royal Society of London, 222, Series A, 309-368.

Fisher, R. A., 1924. The conditions under which $\chi^2$ measures the discrepancy between observation and hypothesis, Journal of the Royal Statistical Society, 87, 442-450.

Fisher, R. A., 1935a. The logic of inductive inference, Journal of the Royal Statistical Society, 98, 39-54.

Fisher, R. A., 1935b. The fiducial argument in statistical inference, Annals of Eugenics, 6, 391-396.

Fisher, R. A., 1937. Professor Karl Pearson and the method of moments, Annals of Eugenics, 7, 303-318.

Godambe, V. P., 1960. An optimum property of regular maximum likelihood estimation, Annals of Mathematical Statistics, 31, 1208-1212.

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

Godambe, V. P., 1984. On ancillarity and Fisher information in the presence of nuisance parameter, Biometrika, 71, 626-629.

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

Godambe, V. P., (Editor)., 1991a. *Estimating Functions* (Oxford Science Publications, New York).

Godambe, V. P., 1991b. Orthogonality of estimating functions and nuisance parameters, Biometrika, 78, 143-151.

Godambe, V. P., 2001. Estimation of Median: Quasi-Likelihood and Optimum Estimating Functions, Working Paper 2001-01, Department of Statistics and Actuarial Science, University of Waterloo, Canada.

Godambe, V. P., Heyde, C. C., 1987. Quasi-likelihood and optimal estimation, International Statistical Review, 55, 231-244.

Godambe, V. P. and Kale, B. K., 1991. Estimating functions: An overview. In: *Estimating Functions*, V. P. Godambe (Editor), (Oxford University Press, Oxford), 3-20.

Godambe, V. P. and Thompson, M. E., 1974. Estimating equations in the presence of a nuisance parameter, Annals of Statistics, 2, 568-571.

Godambe, V. P. and Thompson, M. E., 1978. Some aspects of the theory of estimating equations, Journal of Statistical Planning and Inference, 2, 95-104.

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

Godambe, V. P. and Thompson, M. E., 1989. An extension of quasi-likelihood estimation, Journal of Statistical Planning and Inference, 22, 137-152.

Granger, C. W. J., 1998. Overview of nonlinear time series specification in Economics, NSF Symposium on Nonlinear Time Series Models, University of California, Berkeley.

Granger, C. W. J. and Anderson, A. P., 1978. *An Introduction to Bilinear Time Series Models*, (Göttingen: Vandenhoeck and Ruprecht).

Granger, C. W. J. and Teräsvirta, T., 1993. *Modelling Nonlinear Economic Relationships*, (Oxford University Press, New York).

Hald, A., 1998. *A History of Mathematical Statistics: From 1750 to 1930*, (John Wiley and Sons, New York).

Hall, W. J. and Mathiason, D. J., 1990. On large- sample estimation and testing in parametric models, International Statistical Review, 58, 77-97.

Heyde, C. C., 1989. Quasi-likelihood and optimality of estimating functions: some current unifying themes, Bulletin International Statistical Institute, Book 1, 19-29.

Heyde, C. C., 1997. *Quasi-likelihood and Its Applications: A General Approach to Optimal Parameter Estimation*, (Springer, New York).

Jung, S. H., 1996. Quasi-likelihood for median regression models, Journal of the American Statistical Association, 91, 251-257.

Kale, B. K., 1962. An extension of the Cramér-Rao inequality for statistical estimation functions, Skandinaviske Akturietidskrift, 45, 80-89.

Kale, B. K., 2001-2002. Estimating functions and equations, Journal of the Indian Society for Probability and Statistics, 6, 1-27.

Kendall, M. G., 1951. Regression, structure and functional relationship - I, Biometrika, 38, 11-25.

Kimball, B. F., 1946. Sufficient statistical estimation functions for the parameters of the distribution of maximum values, Annals of Mathematical Statistics, 17, 299-309.

Koenker, R. and Bassett, G., 1978. Regression quantiles, Econometrica, 46, 33-50.

Lele, S., 1994. Estimating functions in chaotic systems, Journal of the American Statistical Association, 89, 512-516.

Lele, S., 1997. Estimating functions for semivariogram estimation. In: I. V. Basawa, V. P. Godambe and R. L. Taylor (Editors), *Selected Proceedings of the Symposium on Estimating Functions*, (IMS Lecture Note- Monograph Series), Vol. 32, 381-396.

Li, D. X., Turtle, H. J., 2000. Semiparametric ARCH models: An estimating function approach, Journal of Business and Economic Statistics, 18, 174-186.

Liang, K. Y. and Zeger, S. L., 1986. Longitudinal data analysis using generalized linear models, Biometrika, 73, 13-22.

Liang, K. Y. and Zeger, S. L., 1995. Inference based on estimating functions in the presence of nuisance parameter, Statistical Science, 10, 158-173.

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

Lindsay, B. G. and Waterman, R. P., 1992. Extending Godambe's method in nuisance parameter problems. In: J. Chen (Editor), *Recent concepts in Statistical Inference*, Proceedings of a Symposium in Honor of Prof. V. P. Godambe, (University of Waterloo, Canada).

Manski, C. F., 1975. Maximum score estimation of the stochastic utility model of choice, Journal of Econometrics, 3, 205-228.

McLeish, D. L., 1984. Estimation for aggregate models: The aggregate markov chain, Canadian Journal of Statistics, 12, 265-282.

McLeish, D. L. and Small, C. G., 1988. *The Theory and Applications of Statistical Inference Functions*, Lecture Notes in Statistics, 44, (Springer-Verlag, New York).

Mittelhammer, R. C., Judge, G. G. and Miller, D. J., 2000. *Econometric Foundations* , (Cambridge University Press, Cambridge).

Mukhopadhyay, P., 2004. *An Introduction to Estimating Functions*, (Narosa Publishing House, New Delhi).

Naik-Nimbalkar, U. V., 1996. Estimating functions for stochastic processes. In: Prakasa Rao, B. L. S., Bhat, B. R., (Editors), *Stochastic Processes and Statistical Inference*, (New Age International Publishers, New Delhi), 52-72.

Nelson, D. B., 1991. Conditional heteroscedasticity in asset returns: A new approach, Econometrica, 59, 347-370.

Newey, W. K., 2004. Efficient semiparametric estimation via moment restrictions, Econometrica, 72, 1877-1897.

Neyman, J., 1959. Optimal asymptotic test of composite statistical hypothesis. In: Grenander, U. (Editors), *Probability and Statistics, the Herald Cramér Volume*, (Uppsala: Almqvist and Wiksell), 213-234.

Neyman, J. and Pearson, E., 1933. On the problem of the most efficient tests of statistical hypothesis, Philosophical Transactions of the Royal Society A, 231, 289-337.

Neyman, J., and Scott, E. L., 1948. Consistent estimates based on partially consistent observations, Econometrica, 16, 1-32.

Nicholls, D. F and Quinn, B. G., 1982. *Random Coefficient Autoregressive Models: An Introduction*, (Springer-Verlag, New York).

Okuma, A., 1976. On invariance of estimating equations, Bulletin of Kyushu Institute of Technology, 23, 11-16.

Ord, J. K., 1975. Estimation methods for models of spatial interaction, Journal of the American Statistical Association, 70, 120-126.

Owen, A. B., 2001. *Empirical Likelihood*, Monographs on Statistics and Applied Probability Series, (Chapman and Hall/CRC Press).

Pearson, K., 1894. Contribution to the mathematical theory of evolution, Philosophical Transactions of the Royal Society of London 185, Series A, 71-110.

Pearson, K., 1902. On the systematic fitting of curves to observations and measurements, Parts I and II, Biometrika 1, 265-303; 2, 1-23.

Pearson, K., 1936. Method of moments and method of maximum likelihood, Biometrika, 28, 34-59.

Phillips, P. C. B., 1988. The ET interview: Professor James Durbin, Econometric Theory, 4, 125-157.

Powell, J. L., 1984. Least absolute deviations estimation for the censored regression model, Journal of Econometrics, 25, 303-325.

Qin, J. and Lawless, J., 1994. Empirical likelihood and general estimating equations, Annals of Statistics, 22, 300-325.

Rao, C. R., 1945. Information and accuracy attainable in the estimation of statistical parameters, Bulletin of Calcutta Mathematical Society, 37, 81-91.

Sørensen, M., 1999. On asymptotics of estimating functions, Brazilian Journal of Probability and Statistics, 13, 111-136.

Teräsvirta, T., 2006. Univariate nonlinear time series models. (This volume).

Thavaneswaran, A., 1991. Tests based on optimal estimate. In: *Estimating Functions*, Godambe (Editor), (Oxford University Press, New York), 189-197.

Thavaneswaran, A. and Abraham, B., 1988. Estimation for non-linear time series models using estimating equations, Journal of Time Series Analysis, 9, 99-108.

Thavaneswaran, A. and Peiris, S., 1996. Nonparametric estimation for some nonlinear models, Statistics and Probability Letters, 28, 227-233.

Thompson, M. E., 2002. A conversation with V. P. Godambe, Statistical Science, 17, 458-466.

Tong, H., 1990. *Nonlinear Time Series: A Dynamical System Approach*, (Oxford University Press, Oxford).

Tsay, R. S., 1987. Conditional heteroscedastic time series models, Journal of the American Statistical Association, 82, 590-604.

Vinod, H. D., 1998. Foundations of statistical inference based on numerical roots of robust pivot functions, Journal of Econometrics, 86, 387-396.

Wald, A., 1940. The fitting of straight lines if both variables are subject to error, Annals of Mathematical Statistics, 11, 284-300.

Wald, A., 1942. Asymptotically shortest confidence intervals, Annals of Mathematical Statistics, 13, 127-137.

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

Weiss, A. A., 1986. ARCH and bilinear time series models: Comparison and combinations, Journal of Business and Economics Statistics, 4, 59-70.

Wilks, S. S., 1938. Shortest average confidence intervals from large samples, Annals of Mathematical Statistics, 9, 166-175.

Wilks, S. S., 1962. *Mathematical Statistics*, (John Wiley and Sons, New York).

Yanagimoto, T. and Yamamoto, E., 1991. The role of unbiasedness in estimating equations. In: *Estimating Functions*, Godambe (Editor), (Oxford University Press, New York), 89-101.

Yanagimoto, T. and Yamamoto, E., 1993. A criterion of sensitivity of an estimating function, Communications in Statistics - Theory and Methods, 22, 451-460.

Yasui, Y. and Lele, S., 1997. A regression method for spatial disease rates: An estimating function approach, Journal of the American Statistical Association, 92, 21-32.

Yule, G. U., 1902. Mendel's laws and their probable relations to intra-racial heredity, New Phytologist, 1, 193-207.