The Annals of Statistics
1996, Vol. 24, No. 3, 1283–1297

A MINIMAX APPROACH TO CONSISTENCY AND EFFICIENCY FOR ESTIMATING EQUATIONS[1]

BY BING LI

Pennsylvania State University

The consistency of estimating equations has been studied, in the main, along the lines of Cramér's classical argument, which only asserts the existence of consistent solutions. The statement similar to that of Doob and Wald, which identifies the consistent solutions, has not yet been established. The obstacle is that the solutions of estimating equations cannot in general be defined as the maximum of likelihood functions. In this paper we demonstrate that the consistent solutions can be identified as the minimax of a function R, whose properties resemble those of a log likelihood ratio, but which exists in a much wider context. Furthermore, since we do not need R to be differentiable, the minimax is consistent even when the estimating equation does not exist. In this respect, the minimax is a new estimator. We first convey the idea by focusing on the quasi-likelihood estimate, and then indicate its full generality by providing a set of sufficient conditions for consistency and studying a number of important cases. Efficiency will also be verified.

1. Introduction. The quasi-likelihood estimate is defined as the solution to the quasi-likelihood equation, the optimal estimating equation constructed under the assumptions of the first two moments and the differentiability of the mean function. See Wedderburn (1974) and McCullagh (1983). Let $X^T = (X_1, \ldots, X_n)$ be random observations with joint distribution P_θ for some p-dimensional parameter θ in some parameter space Θ. About the family of distributions $\{P_\theta\colon\ \theta \in \Theta\}$, we only know of the first two moments $\mu_\theta = (\mu_{1\theta}, \ldots, \mu_{n\theta})^T = E_\theta X$ and $V_\theta = \{V_{ij\theta}\colon\ i, j = 1, \ldots, n\} = \mathrm{cov}_\theta(X_i, X_j)$. The quasi-likelihood equation is defined to be

$$q(\theta, X) = \dot{\mu}_\theta^T V_\theta^{-1}(X - \mu_\theta) = 0, \tag{1}$$

where $\dot{\mu}_\theta$ is the $n \times p$-dimensional derivative matrix. The function on the left, usually called the quasi-score, is so constructed as to contain the greatest amount of information among the class of all linear and unbiased estimating equations in terms of Godambe (1960). See Jarrett (1984), McLeish (1984), Godambe and Heyde (1987) and McLeish and Small (1992).

If the quasi-score is the gradient vector of a potential function, in other words, if $\partial Q(\theta, X)/\partial\theta = q(\theta, X)$ for some $Q(\theta, X)$, then $Q(\theta, X)$ is defined to be the quasi-likelihood, and its global maximum is the quasi-likelihood

Received November 1993; revised July 1995.
[1]Research supported by NSF Grant DMS-93-06738.
AMS 1991 *subject classifications*. Primary 62J12; secondary 62A10, 62F12.
Key words and phrases. Quasi-likelihood estimation, Doob–Wald approach to consistency, estimating equations.

estimate. In general, however, the quasi-score may not be the gradient of any potential function, in which cases the quasi-likelihood is not defined. See McCullagh and Nelder (1989) and McCullagh (1990). Hence, in general, the quasi-likelihood estimate is only the solution to an estimating equation and not the maximum of an objective function. Two questions arise: (1) If there is more than one solution to the quasi-likelihood equation, which are consistent? (2) If the mean μ_θ is not a differentiable function or if the quasi-likelihood equation has no solution in the parameter space, how should we draw sensible inference based on the mean-variance assumption?

To make the point clearer, let us draw an analogy with the two main approaches to consistency in the classical theory of maximum likelihood estimation. By the first approach [Cramér (1946)], one exhibits the existence of a sequence of consistent solutions to the likelihood equation. Since it employs only the properties of the likelihood equation, this approach can be applied generally to estimating equations which need not correspond to likelihood functions. By the second approach [Doob (1934); Wald (1949); Wolfowitz (1949)], one demonstrates that the global maximum of the likelihood function is consistent. The advantages of the Doob–Wald approach are (i) the consistency does not depend on the differentiability of the likelihood function or the existence of the likelihood equation and (ii) when the likelihood equation does exist, we know which solutions are consistent in case there are more than one of them. However, since it appeals to the properties of likelihood functions, which in general have no direct correspondence for estimating equations, the Doob–Wald approach seems difficult to apply generally to estimating equations.

A key property of the likelihood function used in the Doob–Wald approach is the inequality

$$E_{\theta_0}\{\log(dP_\eta/dP_{\theta_0})\} < 0 \quad \text{for all } \eta \text{ in } \Theta,\ \eta \neq \theta_0. \tag{2}$$

Cox and Hinkley (1974) gave the consistency argument of Doob and Wald a concise and accurate summarization: The defining property of the maximum likelihood estimate,

$$\log p_{\hat\theta}(X) \geq \log p_\eta(X), \tag{3}$$

is incompatible with (2) unless $\hat\theta$ converges to θ_0. If this incompatibility could be derived without the use of a likelihood function, then it seems one might be able to obtain a Doob–Wald type result for general estimating equations.

Our starting point is a function introduced in Li (1993), which is derived from the projection of an approximate log likelihood ratio. It takes the simple form

$$R(\theta,\eta) = \tfrac{1}{2}(\mu_\eta - \mu_\theta)^T V_\theta^{-1}(X - \mu_\theta) + \tfrac{1}{2}(\mu_\eta - \mu_\theta)^T V_\eta^{-1}(X - \mu_\eta), \tag{4}$$

where the dependence of R on the data X and the sample size n is suppressed, and θ and η are two parameter values in Θ. The following properties motivate the idea of the present paper:

$$\text{(i)}\quad R(\theta,\eta) = -R(\eta,\theta), \qquad \text{(ii)}\quad E_{\theta_0}\{R(\theta_0,\eta)\} < 0 \quad \text{for all } \eta \neq \theta_0. \tag{5}$$

From these it follows that

$$\sup_{\eta\in\Theta} E_{\theta_0}\{R(\theta_0,\eta)\} = \inf_{\theta\in\Theta}\sup_{\eta\in\Theta} E_{\theta_0}\{R(\theta,\eta)\}. \tag{6}$$

This leads us naturally to estimating θ_0 by the minimax $\hat{\theta}$ of $R(\theta,\eta)$. In symbols, $\hat{\theta}$ is any parameter value that satisfies the relation

$$\sup_{\eta\in\Theta}\{R(\hat{\theta},\eta)\} = \inf_{\theta\in\Theta}\sup_{\eta\in\Theta}\{R(\theta,\eta)\}. \tag{7}$$

The idea of the minimax approach to the consistency of quasi-likelihood estimation, which will be the focus of the present paper, can be summarized as: (6) and (7) are incompatible unless $\hat{\theta}$ converges to θ_0. This approach makes no differentiability assumptions. Consequently, the consistency does not depend on the existence of the quasi-likelihood equation (1). Furthermore, it will be shown that all the minimax points of $R(\theta,\eta)$ are consistent; this avoids the ambiguity that occurs when there are multiple solutions to the quasi-likelihood equation and when the quasi-score does not integrate to a potential function. Under mild assumptions, the minimax $\hat{\theta}$ has the same efficiency as the quasi-likelihood estimate. Under further mild assumptions, $\hat{\theta}$ is necessarily a solution to the quasi-likelihood equation (1). In either case, the minimax approach provides us with a specific estimate that is consistent and efficient, rather than merely indicative of the existence of such an estimate.

Evidently the maximum likelihood estimate itself can be considered as the minimax of the log likelihood ratio, because (3) is equivalent to

$$\sup_{\eta\in\Theta}\{\log p_\eta(X) - \log p_{\hat{\theta}}(X)\} = \inf_{\theta\in\Theta}\sup_{\eta\in\Theta}\{\log p_\eta(X) - \log p_\theta(X)\}. \tag{8}$$

Indeed, for any function $f_\theta(X)$, the minimax point of $f_\eta(X) - f_\theta(X)$ is simply the maximum point of $f_\theta(X)$. However, for general estimating equations, it is often impossible to find a suitable function $L(\theta,\eta)$ that can be decomposed as $f_\eta(X) - f_\theta(X)$ and that satisfies inequality (2). The functions similar to (4), on the other hand, can be constructed for fairly general classes of estimating equations. In this sense, the minimax approach generalizes the idea of Doob and Wald.

From (5) and the weak law of large numbers it follows that

$$\begin{aligned} &P\{R(\theta_0,\eta) < R(\theta_0,\theta_0)\} \to 1, \\ &P\{R(\theta,\theta_0) > R(\theta_0,\theta_0)\} \to 1 \quad \text{for each } \theta \neq \theta_0,\ \eta \neq \theta_0. \end{aligned} \tag{9}$$

See Li (1993). Using these inequalities we can distinguish, with probability tending to 1, two sequences of solutions $\{\hat{\theta}_{1n}\}$ and $\{\hat{\theta}_{2n}\}$, provided that one converges to the correct parameter value and the other converges to an incorrect value. This pointwise comparison is the first step toward the consistency of the minimax, but the latter is a stronger statement, which requires the global properties of R. To achieve this we appeal to the weak

convergence of the random function $n^{-1}\{R - E(R)\}$. In this respect, the method used here is closely related to that of Wong (1986), in which, along with other theoretical development, the consistency of the maximum partial likelihood estimate is established.

The consistency of estimating equations and that in the context of generalized linear models have been studied by a number of researchers. Fahrmeir and Kaufmann (1985) studied the consistency of the maximum likelihood estimate based on generalized linear models. For natural link function, they proved the consistency of the maximum likelihood estimate; for nonnatural link functions, they proved the asymptotic existence of a sequence of consistent solutions to the likelihood equation. Crowder (1986) studied the behavior of all existing solutions of an estimating equation. In particular, he gave sufficient conditions for a set of parameter values to contain all the existing solutions with probability tending to 1, and for a set to contain no solution with positive probability in the limit. There has also been research to tackle the problem of the nonexistence of the quasi-likelihood function. McCullagh (1990) demonstrated that when the quasi-likelihood equation has multiple solutions, the confidence interval based on the score test may be misleading. He suggested the possibility of constructing a quasi-likelihood function by decomposing the nonconservative quasi-score function into a conservative part and a residue part. Firth and Harris (1991) observed a similar phenomenon in their multiplicative random effect model and used the profile quasi-score function for inference purposes. McLeish and Small (1992) introduced the projection of the likelihood ratio onto the Hilbert space spanned by the products of the observations and studied its properties. Along the lines of the discussions of McCullagh (1990), Li and McCullagh (1994) constructed a conservative estimating function that is nearest to the quasi-score in terms of a metric associated with a prior distribution, and whose potential function belongs to the linear exponential family. They also looked into the possibility of incorporating prior information into the quasi-likelihood estimate using this potential function.

The rest of the paper will be organized as follows. We will show in Section 2 that all the minimax points of $R(\theta, \eta)$ are consistent and, in Section 3, that they are efficient. In Section 4, we demonstrate under certain conditions that the minimax points are solutions of the quasi-likelihood equation and we study the further relations between these two approaches. In Section 5 a number of possibilities of extending the minimax approach to other classes of estimating equations are explored.

For simplicity, we refer to a point at which a function obtains its minimax value as a minimax of that function, and a point at which a function achieves its maximum value as a maximum of that function. We refer to the corresponding values of the functions as the minimax value and the maximum value. The maximum or minimax will always mean the global maximum or minimax. To reduce the number of indices, without further specifications, the expectation E and probability P are always evaluated at the true parameter value, which is denoted by θ_0. Perhaps it is helpful to mention that the

quasi-likelihood function $Q(\theta, X)$, if it exists, is different from the underlying distribution P, the latter being unknown except for its first two moments.

Finally, the word "efficiency" is used in a semiparametric sense. Thus when we say that the quasi-likelihood estimate is efficient, we mean it is so among the solutions of all linear and unbiased estimating equations, and when we say that the minimax of R is efficient, we mean it is as efficient as the quasi-likelihood estimate. The term "minimax approach" also needs clarification. Another suitable name is "the saddle point approach," which is not used in order to avoid possible confusion with the term "saddle point approximation." It should, however, be understood, that the minimax is not in the sense of decision theory; in particular, the minimax approach is not related to what is referred to as the minimax–asymptotic variance method in the literature.

2. The consistency of the minimax. In order to prove the consistency of the minimax $\hat{\theta}$ of $R(\theta, \eta)$, we must go beyond the pointwise comparison (9). In particular, we need to make probability statements about the entire function R. We first state and explain the assumptions under which the sequence of random functions $\{n^{-1}(R - ER)\}$ converges in probability to 0 uniformly.

With θ_0 fixed, we denote $E\{R(\theta, \theta_0)\}$ by $J_n(\theta)$ and $R(\theta, \theta_0) - J_n(\theta)$ by $M_n(\theta)$. For now we assume Θ to be compact. Since $n^{-1}M_n(\theta)$ is a weighted average of independent random variables with 0 expectations, by the weak law of large numbers, for each θ, $n^{-1}M_n(\theta) \to_P 0$ under θ_0. The condition for this is very mild [Serfling (1980), page 27] and we do not count it as an assumption. We assume (i) the sequence of random functions $\{n^{-1}M_n(\theta)$: $n = 1, 2, \ldots\}$ is stochastic equicontinuous in Θ [see Pollard (1984), page 139] and (ii) for any compact subset G of Θ that does not contain θ_0,

$$\liminf_{n\to\infty} \inf_{\theta\in G} \{n^{-1}J_n(\theta)\} > 0.$$

It is a standard result, as can be verified by applying the Arzelà–Ascoli theorem, that assumption (i), together with the fact that $n^{-1}M_n(\theta^\dagger) \to_P 0$ for some $\theta^\dagger$ in Θ, implies that the sequence $\{n^{-1}M_n(\theta)\}$ is tight in $C(\Theta)$. See Billingsley [(1968), page 55]. By Prohorov's theorem, the sequence of functions $\{n^{-1}M_n(\theta)\}$, which we know converges in probability to 0 pointwise in Θ, also converges weakly in $C(\Theta)$ to the constant function 0. Hence

$$\sup_{\theta\in\Theta} n^{-1}|M_n(\theta)| \to_P 0.$$

Also notice that $J_n(\theta)$ in assumption (ii) is simply the quadratic form

$$J_n(\theta) = (1/2)(\mu_\theta - \mu_{\theta_0})^T V_\theta^{-1}(\mu_\theta - \mu_{\theta_0}).$$

THEOREM 1. *Let Θ be compact. Then, under assumptions* (i) *and* (ii) *made in the last paragraph, any parameter value $\hat{\theta}$ that satisfies the minimax relation* (7) *is a consistent estimate of θ_0.*

PROOF. Let O_{θ_0} be an open ball centered at θ_0. The gist of the argument is that, on the one hand, there exists a positive number δ' for which

$$\lim_{n\to\infty} P\left\{ n^{-1} \inf_{\theta\notin O_{\theta_0}} \sup_{\eta\in\Theta} R(\theta,\eta) > \delta' \right\} = 1 \tag{10}$$

and, on the other, for any positive number δ,

$$\lim_{n\to\infty} P\left\{ n^{-1} \inf_{\theta\in\Theta} \sup_{\eta\in\Theta} R(\theta,\eta) < \delta \right\} = 1. \tag{11}$$

To see this, let A be a subset of Θ, which we will take to be either O_{θ_0} or $\Theta \setminus O_{\theta_0}$. Then

$$\inf_{\theta\in A} n^{-1}R(\theta,\theta_0) \geq \inf_{\theta\in A} n^{-1}M_n(\theta) + \inf_{\theta\in A} n^{-1}J_n(\theta).$$

Since the first term on the right is sandwiched between $\pm n^{-1}\sup_{\theta\in A}|M_n(\theta)|$, it converges to 0 in probability. This implies, for any $\delta > 0$,

$$\lim_{n\to\infty} P\left\{ \inf_{\theta\in A} n^{-1}R(\theta,\theta_0) \geq \inf_{\theta\in A} n^{-1}M_n(\theta) + \inf_{\theta\in A} n^{-1}J_n(\theta), \left| \inf_{\theta\in A} n^{-1}M_n(\theta) \right| < \delta \right\} = 1. \tag{12}$$

Take $A = \Theta$. Then $\inf_{\theta\in A} n^{-1}J_n(\theta) = 0$. Hence (12) implies that, with probability tending to 1, $\inf_{\theta\in\Theta} n^{-1}R(\theta,\theta_0) > -\delta$. However, this event is equivalent to $\sup_{\eta\in\Theta} n^{-1}R(\theta_0,\eta) < \delta$, whose probability is apparently not greater than that in (11). This proves (11). Take $A = \Theta \setminus O_{\theta_0}$. Then, by assumption (ii), $\inf_{\theta\in A} n^{-1}J_n(\theta) > 0$. Denote this number by $3\delta'$ and let the δ in (12) be δ'. Then (12) implies that

$$\lim_{n\to\infty} P\left\{ \inf_{\theta\notin O_{\theta_0}} n^{-1}R(\theta,\theta_0) > \delta' \right\} = 1. \tag{13}$$

However, since $\inf_{\theta\notin O_{\theta_0}} R(\theta,\theta_0) \leq \inf_{\theta\notin O_{\theta_0}} \sup_{\eta\in\Theta} R(\theta,\eta)$, (13) implies (10).

Next, let $\hat\theta$ be any minimax solution of $R(\theta,\eta)$. Suppose, contrary to the assertion of the theorem, that $\limsup_{n\to\infty} P(\hat\theta \notin O_{\theta_0}) > 0$. Then,

$$\begin{aligned} &\limsup_{n\to\infty} P\left\{ \inf_{\theta\in\Theta} \sup_{\eta\in\Theta} n^{-1}R(\theta,\eta) \geq \delta' \right\} \\ &\quad\geq \limsup_{n\to\infty} P\left\{ \inf_{\theta\in\Theta} \sup_{\eta\in\Theta} R(\theta,\eta) = \inf_{\theta\notin O_{\theta_0}} \sup_{\eta\in\Theta} R(\theta,\eta); \inf_{\theta\notin O_{\theta_0}} \sup_{\eta\in\Theta} n^{-1}R(\theta,\eta) > \delta' \right\}. \end{aligned}$$

By (10) the limit on the right-hand side reduces to the left side of the next string of inequalities:

$$\limsup_{n\to\infty} P\left\{ \inf_{\theta\in\Theta} \sup_{\eta\in\Theta} R(\theta,\eta) = \inf_{\theta\notin O_{\theta_0}} \sup_{\eta\in\Theta} R(\theta,\eta) \right\} \geq \limsup_{n\to\infty} P\left(\hat\theta \notin O_{\theta_0} \right) > 0.$$

This contradicts (11). □

To cover more general parameter spaces, we follow the discussions of Wald (1949) and Wong (1986) and assume that the parameter space is essentially compact. The proof of the next corollary is a simple extension of Theorem 1 and will be omitted.

COROLLARY 1. *Suppose that* Θ *contains a compact subset* K *whose interior contains* θ_0, *and there is a positive number* δ *such that*

$$P\left\{n^{-1}\sup_{\eta\in\Theta\setminus K} R(\theta_0,\eta) < -\delta\right\} \to 1.$$

Suppose that the conditions of Theorem 1 *are satisfied for* K. *Then any minimax solution of* R *on* Θ *is consistent*.

The following example shows that the minimax solution of $R(\theta,\eta)$ sometimes gives a more reasonable, less ambiguous answer to an estimation problem than does the solution to the quasi-likelihood equation.

EXAMPLE 1. Let $X_1,\ldots,X_n$ be independent random variables each taking values on the closed interval $[0,\theta]$. Suppose that $E_\theta X_i = \theta/2$ for each i. Let the variance $\sigma_\theta^2 = \mathrm{var}_\theta(X_i)$ be an unknown function of θ. Assume that σ_θ^2 is finite and positive for all possible θ. Observe that

$$R(\theta,\eta) = 8^{-1}n(\eta-\theta)\left\{\sigma_\theta^{-2}(2\overline{X}-\theta) + \sigma_\eta^{-2}(2\overline{X}-\eta)\right\}.$$

Here, the parameter space is $[0,\infty)$, and the sample space varies with θ, so that $X_{(n)} \equiv \max\{X_1,\ldots,X_n\} \le \theta$. For further discussions of this type of problems in a parametric setting, see Woodroofe (1972).

First, consider the case $2\overline{X} < X_{(n)}$. Notice that $R(X_{(n)},\eta) < 0$ for all $\eta \neq X_{(n)}$, so

$$\sup_{\eta\in\Theta} R(X_{(n)},\eta) = R(X_{(n)},X_{(n)}) = 0.$$

If $\theta \neq X_{(n)}$ then $R(\theta,\eta^\dagger) > 0$ for each $X_{(n)} < \eta^\dagger < \theta$, so $\sup_{\eta\in\Theta} R(\theta,\eta) > 0$ for $\theta \neq X_{(n)}$. Therefore,

$$\sup_{\eta\in\Theta} R(X_{(n)},\eta) = \inf_{\theta\in\Theta}\sup_{\eta\in\Theta} R(\theta,\eta) \quad \text{if } 2\overline{X} < X_{(n)}. \tag{14}$$

Next, let $2\overline{X} > X_{(n)}$. Notice that $R(2\overline{X},\eta) < 0$ if $\eta \neq 2\overline{X}$, so $\sup_{\eta\in\Theta} R(2\overline{X},\eta) = R(2\overline{X},2\overline{X}) = 0$. If $X_{(n)} < \theta < 2\overline{X}$, then $R(\theta,\eta^\dagger) > 0$ for all $\theta < \eta^\dagger < 2\overline{X}$, so $\sup_{\eta\in\Theta} R(\theta,\eta) > 0$. Finally, if $\theta > 2\overline{X}$, then $R(\theta,\eta^\dagger) > 0$ for all $2\overline{X} < \eta^\dagger < \theta$, so $\sup_{\eta\in\Theta} R(\theta,\eta) > 0$. Therefore,

$$\sup_{\eta\in\Theta} R(2\overline{X},\eta) = \inf_{\theta\in\Theta}\sup_{\eta\in\Theta} R(\theta,\eta) \quad \text{if } 2\overline{X} > X_{(n)}. \tag{15}$$

Combining (14) and (15), it is seen that the minimax solution $\hat\theta$ of $R(\theta,\eta)$ is $\max\{2\overline{X}, X_{(n)}\}$. This is a reasonably good estimate of θ considering that we have only made an assumption about the mean.

In the meantime, the solution of $q(\theta, X) = 0$ violates the relation $X_{(n)} \le \theta$ as frequently as $2\bar{X}$ is less than $X_{(n)}$.

3. The efficiency of the minimax. We now demonstrate that $\hat{\theta}$ is efficient among all solutions of linear and unbiased estimating equations. As a preparation, we first show that any $\tilde{\theta}$ that maximizes $R(\hat{\theta}, \eta)$ as a function of η is also consistent. We assume that $R(\theta, \eta)$ is continuously differentiable with respect to either of its arguments. We denote $n^{-1} \sup_{\eta \in \Theta} \|\partial R(\theta, \eta)/\partial\theta\|$ by $B(\theta, X)$, where $\|\cdot\|$ is the Euclidean norm in R^p. Since $\partial R(\theta, \eta)/\partial\theta$ is continuous and Θ is compact, the function $B(\theta, X)$ is defined and finite.

LEMMA 1. *Suppose that the assumptions of Theorem* 1 *hold and in some* O_{θ_0}, $B(\theta, X)$ *is bounded by a random variable* $B_0(X)$, *which does not depend on* θ, *and satisfies* $\limsup_{n\to\infty} EB_0(X) < \infty$. *Let* $\hat{\theta}$ *be a minimax solution of* $R(\theta, \eta)$. *Then*, *under the assumptions made in the last paragraph*, *any* $\tilde{\theta} = \tilde{\theta}(X)$ *that satisfies* $R(\hat{\theta}, \tilde{\theta}) = \sup_{\eta\in\Theta} R(\hat{\theta}, \eta)$ *is consistent*.

PROOF. Suppose that there is an open ball O_{θ_0} about θ_0 such that $\limsup_{n\to\infty} P(\tilde{\theta} \notin O_{\theta_0}) > 0$. Then $\limsup_{n\to\infty} P\{\sup_{\eta\notin O_{\theta_0}} R(\hat{\theta}, \eta) = \sup_{\eta\in\Theta} R(\hat{\theta}, \eta)\} > 0$. Hence

$$\limsup_{n\to\infty} P\left\{ \sup_{\eta\notin O_{\theta_0}} R\big(\hat{\theta}, \eta\big) \ge 0 \right\} > 0. \tag{16}$$

By the Taylor theorem,

$$R\big(\hat{\theta}, \eta\big) = R(\theta_0, \eta) + \big(\hat{\theta} - \theta_0\big)^T \frac{\partial R}{\partial\theta}(\theta^\dagger, \eta)$$

for some $\theta^\dagger$ satisfying $\|\theta^\dagger - \theta_0\| \le \|\hat{\theta} - \theta_0\|$. By the Cauchy–Schwarz inequality,

$$\begin{aligned}\sup_{\eta\notin O_{\theta_0}} n^{-1}R\big(\hat{\theta}, \eta\big) &\le \sup_{\eta\notin O_{\theta_0}} n^{-1}R(\theta_0, \eta) + \|\hat{\theta} - \theta_0\|B(\theta^\dagger, X) \\ &\le \sup_{\eta\notin O_{\theta_0}} n^{-1}R(\theta_0, \eta) + \|\hat{\theta} - \theta_0\|B_0(X).\end{aligned}$$

Since $B_0(X)$ is bounded in probability, $\|\hat{\theta} - \theta_0\|B_0(X) \to_P 0$. By (13), $P\{n^{-1} \sup_{\eta\in\Theta\setminus O_{\theta_0}} R(\theta_0, \eta) < -\delta\} \to 1$ for some $\delta > 0$. Hence, as $n \to \infty$,

$$\begin{aligned}&P\left\{ n^{-1} \sup_{\eta\notin O_{\theta_0}} R(\theta_0, \eta) < 0 \right\} \\ &\quad\ge P\left\{ n^{-1} \sup_{\eta\notin O_{\theta_0}} R(\theta_0, \eta) + \|\hat{\theta} - \theta_0\| \times B_0(X) < 0 \right\} \\ &\quad\ge P\left\{ n^{-1} \sup_{\eta\notin O_{\theta_0}} R(\theta_0, \eta) < -\delta, \|\hat{\theta} - \theta_0\| \times B_0(X) < \delta \right\} \to 1.\end{aligned}$$

This contradicts (16). □

Write $-E\{\partial q(\theta_0, X)/\partial\theta\}$ as $I(\theta_0)$. It is well known that if θ^* is the solution to any linear and unbiased estimating equation, then $\sqrt{n}(\theta^* - \theta_0)$

cannot have asymptotic variance lower than $\{I(\theta_0)/n\}^{-1}$ in terms of the Loewner's ordering of matrices. See McCullagh (1983, 1990). Hence to prove the efficiency of the minimax $\hat{\theta}$, it suffices to show that $n^{-1/2}I(\theta_0)(\hat{\theta} - \theta_0)$ tends in distribution to a standard p-dimensional normal random vector. In fact more is true: as will be seen shortly, both $\hat{\theta}$ and $\tilde{\theta}$ are efficient and they are asymptotically linearly dependent. To simplify the notation, we will abbreviate $(\partial^{i+j}R/\partial\theta^i\,\partial\eta^j)(\theta, \eta)$ as $R_{ij}(\theta, \eta)$, $i, j = 0, 1, 2, 3$. For example, $R_{21}(\theta, \eta) = (\partial^3 R/\partial\theta^2\,\partial\eta)(\theta, \eta)$. If we evaluate these derivatives at $(\theta, \eta) = (\theta_0, \theta_0)$, then they will be further abbreviated as $R_{ij}(\theta_0)$. For example, $R_{21}(\theta_0) = (\partial^3 R/\partial\theta^2\,\partial\eta)(\theta_0, \theta_0)$. By antisymmetry of R, $R_{ij}(\theta_0) = -R_{ji}(\theta_0)$. Also notice that $R_{11}(\theta_0) = 0$. We shall assume that the central limit theorem can be applied to $q(\theta_0, X)$.

THEOREM 2. *Suppose*:

(a) *The conditions in Theorem* 1 *are satisfied*.

(b) *Both* $\tilde{\theta}$ *and* $\hat{\theta}$ *are in the interior of* Θ.

(c) *In a neighborhood* $O^*_{\theta_0}$ *of* (θ_0, θ_0) *in* $\Theta \times \Theta$, *the partial derivatives* $\{R_{ij}(\theta, \eta),\ i, j \geq 0,\ i + j \leq 3\}$ *exists and the sequences*

$$\left\{n^{-1}R_{ij}(\theta, \eta)\colon n = 1, 2, \ldots\right\}, \qquad 0 \leq i, j \leq 3,\ i + j = 3,$$

are tight in $C(G^*_{\theta_0})$, $G^*_{\theta_0}$ *being the closure of* $O^*_{\theta_0}$.

Then:

(i) *Both* $n^{-1/2}I(\theta_0)(\hat{\theta} - \theta_0)$ *and* $n^{-1/2}I(\theta_0)(\tilde{\theta} - \theta_0)$ *converge in law to the standard p-dimensional normal random vector*.

(ii) *The asymptotic correlation coefficient between* $\hat{\theta}$ *and* $\tilde{\theta}$ *equals* 1.

That $\hat{\theta}$ and $\tilde{\theta}$ become linearly dependent as the sample size tends to infinity is not surprising: Otherwise it would be possible to combine the two estimates to achieve higher information, exceeding the information bound for quasi-likelihood estimation. Also notice that if θ_0 is in the interior of Θ, the probability that $\tilde{\theta}$ and $\hat{\theta}$ are both in the interior of Θ tends to 1.

PROOF OF THEOREM 2. The consistency of $\hat{\theta}$ and $\tilde{\theta}$, as given in Theorem 1 and Lemma 2, together with the tightness condition (17), suggests that the following approximation is valid:

$$\begin{aligned}\begin{pmatrix}0\\0\end{pmatrix} &= n^{-1/2}\begin{pmatrix}R_{10}(\hat{\theta}, \tilde{\theta})\\ R_{01}(\hat{\theta}, \tilde{\theta})\end{pmatrix}\\ &= n^{-1/2}\begin{pmatrix}R_{10}(\theta_0)\\ R_{01}(\theta_0)\end{pmatrix}\\ &\quad + \left\{n^{-1}\begin{pmatrix}R_{20}(\theta_0) & 0\\ 0 & R_{02}(\theta_0)\end{pmatrix} + o_p(1)\right\}\begin{pmatrix}n^{1/2}(\hat{\theta} - \theta_0)\\ n^{1/2}(\tilde{\theta} - \theta_0)\end{pmatrix},\end{aligned}$$

where the zeros on the left are $p \times 1$ vectors and on the right are $p \times p$ matrix blocks, and $o_p(1)$ is a $2p \times 2p$ matrix. By Li (1993), $ER_{20}(\theta_0) = I(\theta_0)$ and $R_{10}(\theta_0) = -q(\theta_0, X)$. It follows that

$$\begin{pmatrix} n^{1/2}(\hat{\theta} - \theta_0) \\ n^{1/2}(\tilde{\theta} - \theta_0) \end{pmatrix} = n \begin{pmatrix} I^{-1}(\theta_0) & 0 \\ 0 & -I^{-1}(\theta_0) \end{pmatrix} \begin{pmatrix} -n^{-1/2} q(\theta_0, X) \\ n^{-1/2} q(\theta_0, X) \end{pmatrix} + o_p(1).$$

Now applying the central limit theorem to the quasi-score function $q(\theta_0, X)$, and noticing that $Eq(\theta_0, X) = 0$ and $-E\{\partial q(\theta_0, X)/\partial\theta\} = I(\theta_0)$, we find that $n^{1/2}((\hat{\theta} - \theta_0)^T, (\tilde{\theta} - \theta_0)^T)^T$ is asymptotically normally distributed with mean zero and covariance matrix

$$n \begin{pmatrix} I(\theta_0) & 0 \\ 0 & -I(\theta_0) \end{pmatrix}^{-1} \begin{pmatrix} I(\theta_0) & -I(\theta_0) \\ -I(\theta_0) & I(\theta_0) \end{pmatrix} \begin{pmatrix} I(\theta_0) & 0 \\ 0 & -I(\theta_0) \end{pmatrix}^{-1}$$
$$= n \begin{pmatrix} I^{-1}(\theta_0) & I^{-1}(\theta_0) \\ I^{-1}(\theta_0) & I^{-1}(\theta_0) \end{pmatrix}.$$

This proves the theorem. □

4. The relation between the minimax and the quasi-likelihood estimate. In this section we study the relation between the minimax $\hat{\theta}$ and the solutions to the quasi-likelihood equations. We will demonstrate that if the condition

$$\inf_{\theta \in \Theta} \sup_{\eta \in \Theta} R(\theta, \eta) = \sup_{\eta \in \Theta} \sup_{\theta \in \Theta} R(\theta, \eta) \tag{17}$$

holds, then the minimax is necessarily a solution of (1). Thus the minimax specifies a consistent solution. Since, by antisymmetry,

$$\inf_{\theta \in \Theta} \sup_{\eta \in \Theta} R(\theta, \eta) = \inf_{\theta \in \Theta} \sup_{\eta \in \Theta} \{-R(\eta, \theta)\}$$
$$= -\sup_{\theta \in \Theta} \inf_{\eta \in \Theta} \{R(\eta, \theta)\} = -\sup_{\eta \in \Theta} \inf_{\theta \in \Theta} \{R(\theta, \eta)\},$$

condition (17) is equivalent to $\inf_{\theta \in \Theta} \sup_{\eta \in \Theta} R(\theta, \eta) = 0$. This condition holds quite generally in practice. If $\theta > \theta_0$, $R(\theta, \eta)$ is most likely maximized at some $\eta = \tilde{\theta} < \theta$ and the maximum value is positive; if $\theta < \theta_0$, then $R(\theta, \eta)$ is most likely maximized at some $\eta = \tilde{\theta} > \theta$ and the maximum value is also positive. Hence, if the maximum $\eta = \tilde{\theta}(\theta)$ of $R(\theta, \eta)$ moves continuously as θ moves from the left to the right of θ_0, the curve $(\theta, \tilde{\theta}(\theta))$ should cross the line $\eta = \theta$ at some $(\hat{\theta}, \hat{\theta})$, which is necessarily a minimax. Since $R(\hat{\theta}, \hat{\theta}) = 0$, (17) is satisfied. Of course, if the mode $\eta = \tilde{\theta}(\theta)$ moves discontinuously, as would be the case if, as θ moves near θ_0, a point which does not share the same side of the line $\eta = \theta$ with the mode suddenly becomes a mode, (17) may fail. If it fails, the minimax $\hat{\theta}$ need not be a solution to (1). However, even in these cases, $\hat{\theta}$ is efficient, as is guaranteed by Theorem 2.

THEOREM 3(a). *Suppose that the minimax* $\hat{\theta}$ *of* $R(\theta, \eta)$ *is in the interior of the parameter space* Θ, *that* $R(\theta, \eta)$ *is differentiable with respect to either of its arguments and that condition* (17) *holds. Then* $\hat{\theta}$ *is a solution to the quasi-likelihood equation.*

PROOF. Since, by (17), for all η, $R(\hat{\theta}, \eta) \le 0 = R(\hat{\theta}, \hat{\theta})$, the function $R(\hat{\theta}, \cdot)$ is maximized at $\hat{\theta}$. Hence $R_{01}(\hat{\theta}, \hat{\theta}) = 0$. However, by computation, $R_{01}(\hat{\theta}, \hat{\theta}) = q(\hat{\theta}, X)$. Thus $\hat{\theta}$ satisfies (1). □

Under stronger regularity conditions, we can obtain further relations between the minimax approach and the quasi-likelihood approach. For each θ, let $T(\theta, X)$ be a global maximum of $R(\theta, \cdot)$. Suppose that $T(\theta, X)$ is a differentiable function of θ and that $T(\hat{\theta}, X) = \hat{\theta}$. These are reasonable assumptions in the light of the discussions that precede Theorem 3(a). Let $l(\theta, X) = R(\theta, T(\theta, X))$. We now describe the relation between the function $l(\theta, X)$ and the quasi-score $q(\theta, X)$.

THEOREM 3(b). *Suppose that the assumptions made in the last paragraph hold.*

(i) *If* $R(\theta, \eta)$ *is twice differentiable, then*

$$\frac{\partial^2 l(\hat{\theta}, X)}{\partial \theta^2} = -\frac{\partial q(\hat{\theta}, X)}{\partial \theta}.$$

(ii) *If the function* $R(\theta, \eta)$ *is thrice differentiable and, for a fixed* θ, *there is an open neighborhood* O_θ *with closure* G_θ *such that the sequence of functions* $\{n^{-1}\partial^3 R(\theta, \cdot)/\partial\theta\,\partial\eta^2 \colon n = 1, 2, \ldots\}$, *is tight in* $C(G_\theta)$, *then*

$$\frac{\partial l(\theta, X)}{\partial \theta} = -q(\theta, X)\{1 + O_p(n^{-1/2})\} \quad \textit{under } P_\theta.$$

Under P_θ, $T(\theta, X)$ is consistent and efficient. Consistency can be proved along the lines of Lemma 1, and efficiency can be proved by the Taylor expansion. The details will be omitted. That $T(\theta, X)$ is consistent and efficient is itself not of practical interest, for it is not a statistic. However, as will be seen shortly, the fact that $\|T(\theta, X) - \theta\| = O_p(n^{-1/2})$ serves as a bridge that relates the functions $l(\theta, X)$ and $q(\theta, X)$.

PROOF OF THEOREM 3(b). (i) Since $R_{01}\{\theta, T(\theta, X)\} = 0$ for all θ,

$$\frac{\partial^2 l(\theta, X)}{\partial \theta^2} = R_{20}\{\theta, T(\theta, X)\} + R_{11}\{\theta, T(\theta, X)\}\frac{\partial T(\theta, X)}{\partial \theta}.$$

By computation, $R_{11}(\theta, \theta) = 0$ for all θ. Hence the second term vanishes once we substitute $\theta = \hat{\theta}$ and evoke $T(\hat{\theta}, X) = \hat{\theta}$. By Theorem 1 of Li (1993), the first term on the right equals $-\partial q(\hat{\theta}, X)/\partial\theta$.

(ii) Since $R_{01}\{\theta, T(\theta, X)\} = 0$, $\partial l(\theta, X)/\partial\theta = R_{10}\{\theta, T(\theta, X)\}$. By the Taylor theorem,

$$R_{10}\{\theta, T(\theta, X)\} = R_{10}(\theta, \theta) + R_{11}(\theta, \theta)\{T(\theta, X) - \theta\} + \tfrac{1}{2}\{T(\theta, X) - \theta\}^T R_{12}(\theta, T^\dagger)\{T(\theta, X) - \theta\}$$

for some $T^\dagger$ satisfying $\|T^\dagger - \theta\| \le \|T(\theta, X) - \theta\|$. By Theorem 1 of Li (1993), the first term on the right-hand side is $-q(\theta, X)$. The second term is zero because $R_{11}(\theta, \theta) = 0$. The third term is $O_p(1)$ because, by the discussion preceding the theorem, $\sqrt{n}\{T(\theta, X) - \theta\}$ is $O_p(1)$, and by the tightness assumption, $n^{-1}R_{12}(\theta, T^\dagger)$ is $O_p(1)$. □

EXAMPLE 2. Let $X_1, \ldots, X_n$ be independent observations with $E_\theta X_i = \theta$ and $\mathrm{var}_\theta(X_i) = \theta$ for $\theta > 0$. By simple calculation, we find that

$$-l(\theta, X) = \frac{n}{2}\left\{1 - \left(\frac{\bar{X}}{2\theta - \bar{X}}\right)^{1/2}\right\}(\bar{X} - \theta) + \frac{n}{2}\left\{\left(\frac{2\theta - \bar{X}}{\bar{X}}\right)^{1/2} - 1\right\}\left\{\bar{X} - \theta\left(\frac{\bar{X}}{2\theta - \bar{X}}\right)^{1/2}\right\}.$$

The quasi-likelihood function is $Q(\theta, X) = n(\bar{X}\log\theta - \theta) + \text{constant}$. Both $-l(\theta, X)$ and $Q(\theta, X)$ are maximized at $\theta = \bar{X}$, as is predicted by Theorem 3(a). The first four derivatives of $-l(\theta, X)$ at $\bar{X}$ are 0, -1, $3/10$ and $-3/20$, respectively, and those for $Q(\theta, X)$ are 0, -1, $1/5$ and $-3/50$, respectively. At the maximum, the two functions have the same second derivative, as asserted by Theorem 3(b). Their next two derivatives have the same signs and their third derivatives are quite close to each other. It is interesting to compare the estimating equation derived from $l(\theta, X)$ with the quasi-score. Differentiating $l(\theta, X)$ with respect to θ, we obtain

$$-\frac{\partial l(\theta, X)}{\partial\theta} = \frac{n(\bar{X} - \theta)}{\theta + (\theta - \bar{X})}\{1 - A(\theta, X)\},$$

where

$$A(\theta, X) = \frac{2}{(2\theta - \bar{X})^{1/2}}\frac{\bar{X}(2\theta - \bar{X})^{1/2} + \bar{X}^{3/2} - 2\theta\bar{X}^{1/2}}{\bar{X}^{1/2}(2\theta - \bar{X})^{1/2} + \bar{X}}.$$

More generally, suppose that $E_\theta X_i = \theta$ and $\mathrm{var}_\theta(X_i) = \phi\theta^\alpha$, $\phi > 0$ being the dispersion parameter, and $\alpha > 0$. Then $T(\theta, X)$ is one of the solutions for η to the algebraic equation

$$(\bar{X} - \theta)\theta^{-\alpha}\eta^{\alpha+1} + (\alpha - 2)\eta^2 + (1 - \alpha)(\bar{X} + \theta)\eta + \bar{X}\alpha\theta = 0.$$

If $\alpha = 0$, then $T(\theta, X)$ equals $\bar{X}$ and $-l(\theta, X)$ recovers the normal likelihood $(\bar{X} - \theta)^2/(2\phi)$.

5. Remarks on generalizations. The minimax approach can be extended to cover fairly general classes of estimating equations. We first give a set of sufficient conditions for the consistency of the minimax.

1. The parameter space Θ is essentially compact in the sense of Corollary 1.
2. There is a continuous antisymmetric function R: $\Theta \times \Theta \times \mathscr{X} \to R^1$ which satisfies the following conditions.
3. The sequence of random functions $\{n^{-1}\{R(\theta, \theta_0) - ER(\theta, \theta_0)\}: n = 1, 2, \ldots\}$ is stochastically equicontinuous in Θ.
4. Under θ_0, the weak law of large numbers applies to $n^{-1}\{R(\theta, \theta_0) - ER(\theta, \theta_0)\}$.
5. Let $J(\theta, \eta) = E\{R(\theta, \eta)\}$. For each compact subset G of Θ that does not contain θ_0,
$$\liminf_{n \to \infty} \inf_{\theta \in G} \{n^{-1} J(\theta, \theta_0)\} > 0.$$

For many estimating equations $g(\theta, X)$, the construction of $R(\theta, \eta)$ resembles that described in the previous sections, and entails the additional relations

$$R_{01}(\theta, \theta) = g(\theta, X), \qquad R_{02}(\theta, \theta) = \partial g(\theta, X)/\partial\theta, \tag{18}$$

so that the minimax solution of R is as efficient as the consistent solutions of g.

CASE 1 (Higher order optimal estimating equations). If we know higher moments, we can construct a higher order optimal estimating equation similarly as one constructs the quasi-score. See Jarrett (1984), Crowder (1987) and Godambe and Thompson (1989). The function $R(\theta, \eta)$ can be constructed similarly and it retains the properties (5) and (18); see Li (1993). Under the additional regularity conditions listed above, the consistency and efficiency of the minimax solution of R can be established.

CASE 2 [Martingale estimating equations (discrete)]. McLeish and Small (1988) discussed the following method. Let $\{X_n: n = 1, 2, \ldots\}$ be a sequence of random observations. We want to make inference about some parameters θ based on the first two conditional moments. Let $\mu_i(\theta)$ and $V_i(\theta)$ be the conditional mean and variance of X_i given $X_1, \ldots, X_{i-1}$. Then the estimating function

$$g(\theta, X) = \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \theta}\right)^T \{V_i(\theta)\}^{-1}\{X_i - \mu_i(\theta)\}$$

is optimal among the "conditionally linear" estimating functions of the form $\sum_{i=1}^{n} a_i(\theta)\{X_i - \mu_i(\theta)\}$, where $a_i(\theta)$ is a p-dimensional vector, whose components may depend on $\{X_1, \ldots, X_{i-1}\}$. Consider the function $R(\theta, \eta) = \sum_{i=1}^{n} R_i(\theta, \eta)$, where

$$\begin{aligned} R_i(\theta, \eta) = {} & 2^{-1}\{\mu_i(\eta) - \mu_i(\theta)\}^T V_i^{-1}(\theta)\{X_i - \mu_i(\theta)\} \\ & + 2^{-1}\{\mu_i(\eta) - \mu_i(\theta)\}^T V_i^{-1}(\eta)\{X_i - \mu_i(\eta)\}. \end{aligned}$$

In this case, we let $J(\theta, \theta_0)$ be the accumulative information

$$J(\theta, \theta_0) = \sum_{i=1}^{n} E\{R_i(\theta, \theta_0)|X_1, \ldots, X_{i-1}\}.$$

The assumptions 4 and 5 should be replaced by the following assumption about the accumulative information:

4′. For each compact subset G of Θ that does not contain θ_0, there is a $\delta > 0$ for which

$$P\left\{n^{-1} \inf_{\theta \in G} J(\theta, \theta_0) > \delta\right\} \to 1$$

and

$$n^{-2} J(\theta, \theta_0) \to 0 \quad \text{in probability for all } \theta \in G.$$

This condition is an analogue of Wong's condition made on the accumulative Kullback–Leibler information of the partial likelihood [Wong (1986)].

CASE 3 (Optimal linear combination of marginal estimating equations). Suppose that, for each observation X_i, there is a preferable unbiased estimating equation $g_i(\theta, X_i)$, which, for example, may be the marginal likelihood score for X_i. Sometimes it may be more realistic or more convenient to assume the first two joint moments of the marginal estimating equations than the joint distribution. Assuming that $E_\eta\{g_i(\theta, X_i)\}$ and $\text{cov}_\eta\{g_i(\theta, X_i), g_j(\theta, X_j)\}$ are known for each θ and η in Θ, the optimal linear combination of these equations [in terms of Godambe (1960)] is

$$g(\theta, X) = \sum_{i=1}^{n} \sum_{j=1}^{n} E_\theta\left\{-\frac{\partial g_i(\theta, X_i)}{\partial \theta}\right\}^T \times \left[\text{cov}_\theta\{g_i(\theta, X_i), g_j(\theta, X_j)\}\right]^{-1} g_j(\theta, X_j).$$

The function $R(\theta, \eta)$ can be defined as $R_0(\theta, \eta) - R_0(\eta, \theta)$, where

$$R_0(\theta, \eta) = \sum_{i=1}^{n} \sum_{j=1}^{n} \{E_\eta g_i(\theta, X_i)\}^T \times \left[\text{cov}_\theta\{g_i(\theta, X_i), g_j(\theta, X_j)\}\right]^{-1} g_j(\theta, X_j).$$

It is easy to see that conditions in (5) and condition (18) are satisfied. Thus, under mild assumptions the minimax approach also applies to this case.

Acknowledgments. I would like to thank a referee and an Associate Editor for their very helpful comments.

REFERENCES

BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York.
COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

CRAMÉR, H. (1946). *Mathematical Methods of Statistics*. Princeton Univ. Press.
CROWDER, M. (1986). On consistency and inconsistency of estimating equations. *Econometric Theory* **3** 305–330.
CROWDER, M. (1987). On linear and quadratic estimating functions. *Biometrika* **74** 591–597.
DOOB, J. S. (1934). Probability and statistics. *Trans. Amer. Math. Soc.* **36** 759–775.
FAHRMEIR, L. and KAUFMANN, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann. Statist.* **13** 342–368.
FIRTH, D. and HARRIS, I. R. (1991). Quasi-likelihood for multiplicative random effects. *Biometrika* **78** 545–555.
GODAMBE, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.* **31** 1208–1211.
GODAMBE, V. P. and HEYDE, C. C. (1987). Quasi-likelihood and optimal estimation. *Internat. Statist. Rev.* **55** 231–244.
GODAMBE, V. P. and THOMPSON, M. E. (1989). An extension of quasi-likelihood estimation (with discussion). *J. Statist. Plann. Inference* **22** 137–172.
JARRETT, R. G. (1984). Bounds and expansions for Fisher information when moments are known. *Biometrika* **74** 233–245.
LI, B. (1993). A deviance function for the quasi likelihood method. *Biometrika* **80** 741–753.
LI, B. and MCCULLAGH, P. (1994). Potential functions and conservative estimating equations. *Ann. Statist.* **22** 340–356.
MCCULLAGH, P. (1983). Quasi-likelihood functions. *Ann. Statist.* **11** 59–67.
MCCULLAGH, P. (1990). Quasi-likelihood and estimating functions. In *Statistical Theory and Modelling: In Honour of Sir David Cox* (D. V. Hinkley, N. Reid and E. J. Snell, eds.). Chapman and Hall, London.
MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London.
MCLEISH, D. L. (1984). Estimation for aggregate models: the aggregate Markov chain. *Canad. J. Statist.* **12** 265–282.
MCLEISH, D. L. and SMALL, C. G. (1988). *The Theory and Applications of Statistical Inference Functions. Lecture Notes in Statist.* **44**. Springer, New York.
MCLEISH, D. L. and SMALL, C. G. (1992). A projected likelihood function for semiparametric models. *Biometrika* **79** 93–102.
POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.
SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.
WALD, A. (1949). Note on the consistency of maximum likelihood estimate. *Ann. Math. Statist.* **20** 595–601.
WEDDERBURN, R. W. M. (1974). Quasi-likelihood, generalized linear models, and the Gauss–Newton method. *Biometrika* **61** 439–447.
WOLFOWITZ, J. (1949). On Wald's proof of the consistency of the maximum likelihood estimate. *Ann. Math. Statist.* **20** 601–602.
WONG, W. H. (1986). Theory of partial likelihood. *Ann. Statist.* **14** 88–123.
WOODROOFE, M. (1972). Maximum likelihood estimation of a translation parameter of a truncated distribution. *Ann. Math. Statist.* **43** 113–122.

DEPARTMENT OF STATISTICS
PENNSYLVANIA STATE UNIVERSITY
410 CLASSROOM BUILDING
UNIVERSITY PARK, PENNSYLVANIA 16802
E-MAIL: bing@stat.psu.edu