Inference Based on Estimating Functions in the Presence of Nuisance Parameters

Kung-Yee Liang and Scott L. Zeger

Abstract. In many studies, the scientific objective can be formulated in terms of a statistical model indexed by parameters, only some of which are of scientific interest. The other "nuisance parameters" are required to complete the specification of the probability mechanism but are not of intrinsic value in themselves.

It is well known that nuisance parameters can have a profound impact on inference. Many approaches have been proposed to eliminate or reduce their impact. In this paper, we consider two situations: where the likelihood is completely specified; and where only a part of the random mechanism can be reasonably assumed. In either case, we examine methods for dealing with nuisance parameters from the vantage point of parameter estimating functions. To establish a context, we begin with a review of the basic concepts and limitations of optimal estimating functions. We introduce a hierarchy of orthogonality conditions for estimating functions that helps to characterize the sensitivity of inferences to nuisance parameters. It applies to both the fully and partly parametric cases. Throughout the paper, we rely on examples to illustrate the main ideas.

Key words and phrases: Conditional score function; estimating function; nuisance parameter; optimality; orthogonality.

1. INTRODUCTION

In many statistical problems, the likelihood of the data $f(\mathbf{y};\boldsymbol{\theta},\boldsymbol{\phi})$ depends on parameters $(\boldsymbol{\theta},\boldsymbol{\phi})$, where a subset $\boldsymbol{\theta}$ is of scientific interest and the remainder $\boldsymbol{\phi}$ is not. The parameters $\boldsymbol{\phi}$ are often referred to as "nuisance" or "incidental" parameters (e.g., McCullagh and Nelder, 1989, page 245) as their values are usually needed to make inferences about $\boldsymbol{\theta}$ even though they have little scientific import of their own.

It is useful to distinguish two different situations based upon the degree of interaction between $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. The first situation is when the very scientific interpretations of the parameters of interest $\boldsymbol{\theta}$ change with the value of the nuisance parameters $\boldsymbol{\phi}$. An example is the Box–Cox family of regression models

$$y_i^\phi = \mathbf{x}_i'\boldsymbol{\theta} + \varepsilon_i,$$

where ε_i are independent normal variates with mean 0 and variance σ^2. Here the interpretation of the regression coefficients $\boldsymbol{\theta}$ changes with the nuisance parameter ϕ as the unit of the response variable y_i^ϕ changes. A detailed discussion of this model and the associated issues for inference is presented in Hinkley and Runger (1984). The second situation, when the interpretation of $\boldsymbol{\theta}$ does not change with the value of $\boldsymbol{\phi}$, is more common and the focus of this paper.

The following examples will be used throughout the paper to examine the effects of nuisance parameters.

EXAMPLE 1.1. *Longitudinal studies.* Epidemiologic and clinical studies often collect repeated observations of the same response variable over time for many subjects (Diggle, Liang and Zeger, 1994). Consider one such study with repeated binary responses in which the scientific objective may be represented by a regression model as logit $\mathbb{E}(y_{ij}) = \mathbf{x}_{ij}'\theta$, $j = 1, \ldots, n_i$, $i = 1, \ldots, K$, where K is the number of subjects and n_i the number of observations for the ith subject. Here $\mathbf{x}_{ij}$ is a $p \times 1$

Kung-Yee Liang and Scott L. Zeger are Professors, Department of Biostatistics, School of Public Health, Johns Hopkins University, 615 N. Wolfe Street, Baltimore, Maryland 21205.

vector of covariates thought to be related to the binary response y_{ij}. It is unlikely that the responses from the same subject are independent. Thus one must also model the $n_i \times n_i$ covariance matrix $\Sigma(\boldsymbol{\theta}, \boldsymbol{\phi})$ of $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})'$. Note the dependence of Σ on nuisance parameters $\boldsymbol{\phi}$ which characterize the correlation among the responses from the same subject. Also note that the interpretation of $\boldsymbol{\theta}$ remains the same regardless of the value of $\boldsymbol{\phi}$. Failing to acknowledge the presence of $\boldsymbol{\phi}$ may lead to spurious inferences about $\boldsymbol{\theta}$. For likelihood-based inference for $\boldsymbol{\theta}$, the complete probability mechanism must be specified so that additional higher-order nuisance parameters may also be involved. See Liang and Zeger (1986), Zhao and Prentice (1990) and Fitzmaurice, Laird and Rotnitzky (1993).

EXAMPLE 1.2. *Case-control studies.* The case-control design is commonly used in epidemiologic studies of disease etiology. In this setting, individuals with the prespecified disease, known as the *cases*, and individuals who are disease-free, the *controls*, are recruited. Information on risk factors which are suspected of causing the disease is ascertained from both disease groups. Frequently, individuals are stratified into subgroups based upon demographic or other variables which are related to both the disease and the risk factors of interest. Collection of information on such "confounding" variables is essential for valid inference on the etiology of the disease. Let y_{i1} and y_{i2} be the number of cases and controls, respectively, from the ith stratum, $i = 1, \ldots, K$, exposed to a binary risk factor. The conventional assumption is that y_{ij} follows a binomial distribution with parameters n_{ij} and μ_{ij}. Here μ_{ij} is the probability a person in stratum i with case status j was exposed to the binary risk factor and n_{i1} and n_{i2} represent the number of cases and controls, respectively, in the ith group. The odds ratio parameter of interest

$$\theta = \mu_{i1}(1 - \mu_{i2})/\{\mu_{i2}(1 - \mu_{i1})\},$$

characterizes the strength of the relationship between the risk factor and the risk of the disease. It takes the value 1 when there is no association. The nuisance parameters in this example are $\{\phi_i = \mu_{i2}, i = 1, \ldots, K\}$, the exposure probabilities of the controls in each of K strata. Again the magnitude of these nuisance parameters does not alter the interpretation of θ, the common odds ratio. In Section 4, an example with similar data structure will be given in which the binomial assumption must be relaxed.

EXAMPLE 1.3. *Teratologic experiments.* In the typical teratologic experiment, pregnant animals (e.g., rats) are randomized to receive varied doses of a chemical and sacrificed prior to the end of gestation or pregnancy. Each fetus is examined and a binary response indicating the presence or absence of a particular malformation is recorded. The scientific objective is to determine whether the risk of malformation μ increases with the teratogen dose x, as characterized, for example, by the parameter θ_1 in the model logit $\mu = \theta_0 + \theta_1 x$. It has long been recognized that teratologic data often include a so-called litter effect whereby fetuses from the same litter tend to respond more alike than fetuses from different litters. In the absence of a litter effect, a binomial model for y, the number of malformed fetuses, would be adequate. In the presence of a litter effect, additional parameters $\boldsymbol{\phi}$ which characterize this "extra-binomial variation" are needed. One model is to assume that y follows a beta-binomial distribution whose variance has the form

$$\mathrm{Var}(y) = n\mu(1 - \mu)\{1 + (n - 1)\phi\}.$$

A litter effect is typically reflected by a positive value of ϕ.

EXAMPLE 1.4. *Proportional hazards model.* In clinical trials, one main concern is whether a new treatment prolongs a patient's life or time to recurrence of disease relative to the standard treatments. Rather than comparing the mean times for different treatment groups, it may be preferable to compare hazard functions $\lambda(t)$ which characterize the instantaneous probability of death among those who were alive t units of time after the randomization of treatments to patients. If y represents the survival time, then $\lambda(t) = \lim_{dt\to 0} \Pr(y \in (t, t + dt)|y > t)/dt$. The celebrated proportional hazards model (Cox, 1972) takes the form

$$\lambda(t; \mathbf{x}) = \lambda_0(t)\exp(\boldsymbol{\theta}'\mathbf{x}),$$

where $\mathbf{x}$ is the vector of covariates and $\lambda_0(t)$ is by definition the hazard function among those with $\mathbf{x} = 0$. An element of $\boldsymbol{\theta}$ is the logarithm of the hazard ratio between groups that differ by one unit in its $\mathbf{x}$, all other explanatory variables held fixed. This interpretation is the same irrespective of the shape of the infinite-dimensional $\lambda_0(t)$. An interesting feature of this model is that while the difference of hazard functions is modelled parametrically through $\boldsymbol{\theta}$, the nuisance part of the model $\lambda_0(t)$ is left totally unspecified.

It is well known that nuisance parameters can seriously compromise likelihood inference, particu-

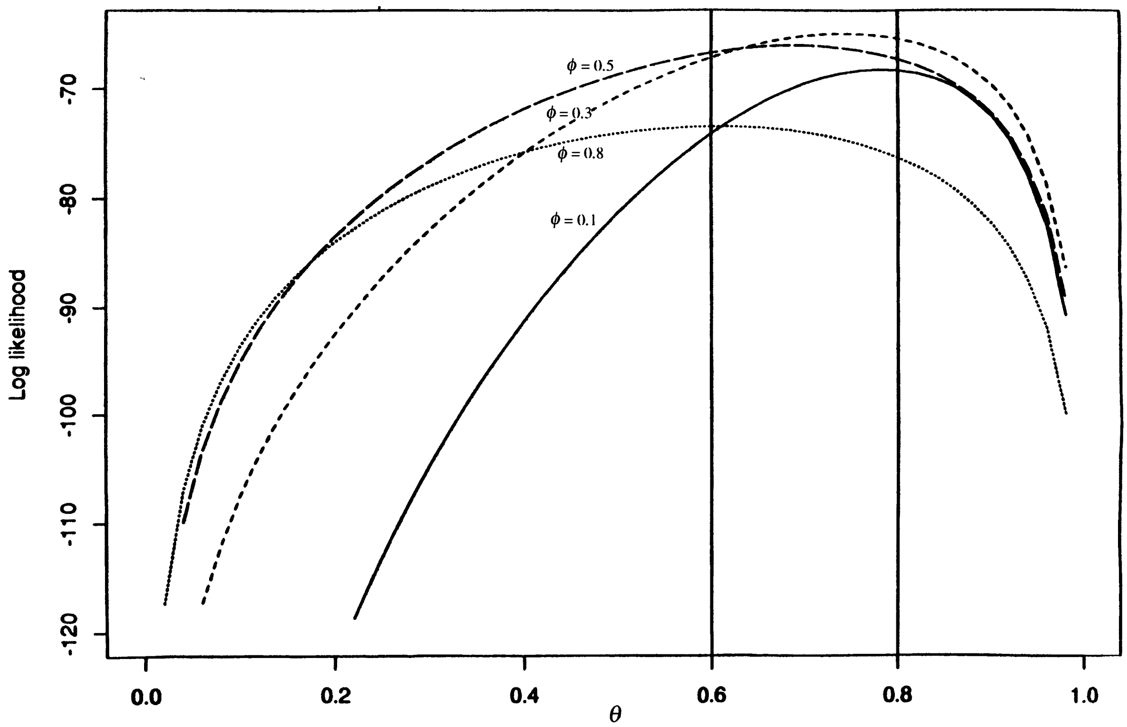

FIG. 1. *Beta-binomial log-likelihoods for the exposed group based on the data reported by Weil* (1970).

larly when their number grows with the sample size (Neyman and Scott, 1948). An example is the estimation of a common odds ratio θ in the case-control study, Example 1.2, with $n_{i1} = n_{i2} = 1$ for all i. As $K \to \infty$, the maximum likelihood estimate (mle) converges to θ^2 rather than θ. While nuisance parameters are especially problematic when there are relatively many, they can also diminish the quality of inferences in the more common situation where there are fewer, as the following example shows.

EXAMPLE 1.3 (Continued). Weil (1970) reported the data from a teratologic experiment in which 32 pregnant rats were randomly assigned to either a control or teratogenic chemical exposure group. The response of interest was the number of pups in each mother's litter that survived a 21-day lactation period. Let θ be the survival probability for a pup in the exposed group, and let ϕ represent the intralitter correlation coefficient in the beta-binomial model. Figure 1 gives plots of the beta-binomial log-likelihoods against θ for selected values of ϕ. Different values of ϕ lead to different conclusions as to which θ is favored by the data. For example, letting $L(\theta, \phi)$ be the likelihood function, $L(0.8, 0.8)/L(0.6, 0.8) = \exp(-2.89)$ suggesting that $\theta = 0.6$ is strongly favored over $\theta = 0.8$ when $\phi = 0.8$, whereas $L(0.8, 0.1)/L(0.6, 0.1) = \exp(5.82)$ strongly supports the opposite conclusion.

Numerous likelihood-based solutions to the nuisance parameter problem have been developed (e.g., Kalbfleisch and Sprott, 1970; Basu, 1977). The $\boldsymbol{\phi}$ can sometimes be eliminated from the likelihood by finding a subset of the data whose distribution is independent of $\boldsymbol{\phi}$. This so-called marginal likelihood approach has particular application in estimating variance components in the general linear mixed model (e.g., Harville, 1977). An alternate approach is to use an "integrated likelihood" obtained by assuming a prior distribution for $\boldsymbol{\phi}$ and basing $\boldsymbol{\theta}$ inferences on the $\boldsymbol{\phi}$ integral of the product of the likelihood and this prior (Kalbfleisch and Sprott, 1970). This partly Bayes strategy suggests the fully Bayesian approach as well where priors for both $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ would be specified. A commonly used likelihood-based strategy is to identify a sufficient statistic $\mathbf{T}(\mathbf{y})$ for $\boldsymbol{\phi}$ for fixed $\boldsymbol{\theta}$ and base $\boldsymbol{\theta}$ inferences on the conditional distribution of $\mathbf{y}$ given $\mathbf{T} = \mathbf{t}$ (e.g., McCullagh and Nelder, 1989, Chapter 7). This "conditional-likelihood" strategy is especially useful for exponential families.

Each of these approaches to reducing or eliminating the effect of $\boldsymbol{\phi}$ on inferences about $\boldsymbol{\theta}$ requires that the full probability mechanism of the data be specified. An alternate strategy to specifying the complete probability mechanism is to combine the data and unknown parameters in an estimating function, the form of which requires specification of only a part of the probability distribution. To illustrate for the case-control example (1.2), we might consider the estimating equation

$$\sum_{i=1}^{K}(n_{i1}+n_{i2})^{-1}\{y_{i1}(n_{i2}-y_{i2})-\theta y_{i2}(n_{i1}-y_{i1})\}=0,$$

which depends only on the odds ratio parameter θ and the data. Note the solution of this equation is the well-known Mantel–Haenszel estimator for θ, which is highly competitive with the mle in terms of both finite sample and asymptotic properties (Breslow, 1981).

Formally, an estimating function $\mathbf{g}$ is a function of the data and unknown parameters $\boldsymbol{\theta}$ such that an estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is obtained as its root, that is, $\mathbf{g}(\mathbf{y}, \hat{\boldsymbol{\theta}}) = 0$. An unbiased estimating function has the property $\mathbb{E}(\mathbf{g}(\mathbf{y}, \boldsymbol{\theta}); \boldsymbol{\theta}, \boldsymbol{\phi}) = 0$ for all $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$.

The main objective of this paper is to examine the use of estimating functions for reducing the influence of nuisance parameters in problems like Examples 1.1–1.4 commonly encountered in health research. In Section 2, we review the history and theory of estimating functions. Section 3 discusses the difficulties nuisance parameters cause for the elementary optimality theory of Section 2. Section 3 focuses on the use of conditional score functions (Lindsay, 1982) which enjoy a hierarchy of orthogonality properties for estimating functions in the parametric case. These properties form the basis for our discussion in Section 4 of estimating functions in which the full likelihood is not specified.

2. BACKGROUND ON ESTIMATING FUNCTIONS

2.1 Historical Perspective

The modern interest in estimating functions is in actuality a return to the earliest roots of statistics. Before the advent of objective functions such as the sum of squares (Legendre, 1805) and the likelihood (Fisher, 1925), scientists grappled with how to design estimating functions which combined observations and unknowns of interest.

For example, Stigler (1986) tells the story of Tobias Mayer, a mid-seventeenth-century cartographer and astronomer who studied the motions of the moon. He repeatedly observed the positions of lunar features and derived, through approximations to the motion equations, a linear system with three unknowns $\boldsymbol{\theta}$ about the characteristics of the moon's motion. Mayer's problem was that he had 27 sets of observations and hence 27 equations with only 3 unknowns. His solution was to divide the 27 equations into 3 sets of 9, to average equations within a set and then finally to solve the resulting 3 equations in 3 unknowns as was standard mathematical practice.

This example illustrates the early use of estimating functions to estimate parameters of scientific interest. The key issue was how to combine independent equations, each of which related an observation made with error to parameters.

With the advent of simple objective functions such as squared deviations (Legendre, 1805) and likelihood (Fisher, 1925), statistics turned to new methods of optimal estimation. Two strands of optimality theory evolved. The first is represented by the Gauss–Markov theorem (e.g., Bickel and Doksum, 1977), which identifies as best the linear unbiased estimator that has minimum variance. Gauss–Markov optimality is a finite sample property defined in terms of the first two moments. The second strand, advocated by Fisher (1925) is the theory of maximum likelihood estimation. The mle by definition optimizes the likelihood but also is asymptotically unbiased and minimum-varianced given regularity conditions. These two strands connect in that they yield the same estimator for the linear model with Gaussian errors.

The method of moments, advocated by Karl Pearson, is a precursor of the modern estimating function approach. Here, selected empirical moments are equated to their expectations which are assumed to depend on parameters of scientific interest $\boldsymbol{\theta}$. The resulting estimating functions are then solved for $\hat{\boldsymbol{\theta}}$. By aggregating the data into sample moments first and then forming estimating functions, the weighting of observations that brings efficiency to the modern approach is lost. Nevertheless, the focus on a few characteristics of the probability distribution without specification of the rest is a precursor of the methods described here.

Kimball (1946) introduced the modern definition of an estimating function for inference about parameters of the extreme value distribution. He defined the "stability" and "sufficiency" of estimating functions in somewhat obscure terms. There appears to have been little direct follow-up to this line of research.

The modern line did follow closely from papers by M. G. Kendall (1951) and Durbin (1960) on linear regression with stochastic predictor variables. Durbin studied the time series model (our notation)

$$y_t = \theta y_{t-1} + \varepsilon_t, \quad \varepsilon_t \text{ iid } N(0, \phi), \; t = 1, \ldots, K,$$

where the initial value y_0 is assumed to be known. Here maximum likelihood leads to

$$\hat{\theta} = \frac{\sum_{t=1}^{K} y_{t-1}y_t}{\sum_{t=1}^{K} y_{t-1}^2}.$$

Since the predictor variable y_{t-1} is random, $\hat{\theta}$ is not a linear function of the responses and hence does not enjoy the Gauss–Markov finite sample optimality it would if the predictors were known constants. Durbin therefore turned his attention to the estimating function for θ

$$(2.1) \qquad g(\mathbf{y}, \theta) = \sum_{t=1}^{K} y_{t-1} y_t - \theta \sum_{t=1}^{K} y_{t-1}^2 = 0,$$

which is a linear function of θ. He referred to (2.1) as an "unbiased linear estimating equation" and derived the analogue of the Gauss–Markov theory for unbiased estimating functions.

Also in 1960, Godambe published the first in a series of papers that form the basis for the modern theory of optimal estimating functions. The 1960 paper generalized the Durbin result by showing that even nonlinear score equations are optimal estimating functions by a finite sample criterion. The major ideas of this and related work are summarized below.

2.2 Optimal Estimating Functions

Let $y_1, \ldots, y_K$ be independent observations with density function $f(\cdot; \theta, \phi)$, where for the moment θ and ϕ are each assumed to be scalar. Let $g(y_i, \theta)$ be an unbiased estimating function of the data and parameter of interest such that $\mathbb{E}(g(y_i, \theta); \theta, \phi) = 0$ for all θ and ϕ. Let

$$g(\mathbf{y}, \theta) = \sum_{i=1}^{K} g(y_i, \theta).$$

Godambe (1960) defined the optimal unbiased estimating function as the minimizer of

$$(2.2) \qquad S_K = \mathbb{E}\left[\left(\frac{g(\mathbf{y}, \theta)}{\mathbb{E}(\partial g(\mathbf{y}, \theta)/\partial \theta)}\right)^2\right].$$

The numerator of (2.2) is the variance of g. The denominator is the square of the average gradient of g. It is intuitively appealing that the optimal g has small variance and be on average as steep as possible near the true θ because these characteristics determine the asymptotic variance of $\hat{\theta}$. Godambe (1960) showed that in the case with no nuisance parameters (i.e., ϕ known), S_K is minimized by the score function which is defined as the θ derivative of the logarithm of the likelihood function.

Godambe (1976) considered the nuisance parameter case. Suppose t is a complete, sufficient statistic for the nuisance parameter ϕ for fixed θ and that the density for $\mathbf{y}$ factors

$$f(\mathbf{y}; \theta, \phi) = f(\mathbf{y} | t, \theta)\, f(t; \theta, \phi).$$

Here the conditional distribution of the data given t is independent of $\boldsymbol{\phi}$ for all θ. Then he showed that the conditional score function $\partial \log f(\mathbf{y}|t; \theta)/\partial \theta$ is the optimal estimating function for θ; that is, it minimizes (2.2).

One of the most important examples of an optimal estimating function from the perspective of health research is the "quasi-score" function proposed by Wedderburn (1974). To trace its development, consider independent response y_i with $\mu_i = \mathbb{E}(y_i)$ and associated explanatory variables $\mathbf{x}_i$, $i = 1, \ldots, K$. Nelder and Wedderburn (1972) unified regression methodology for discrete and continuous responses under the class of generalized linear models (GLM's). A GLM is specified by its systematic part, $h(\mu) = \mathbf{x}'\boldsymbol{\theta}$ for known "link function" h, and its random part,

$$f(y; \gamma, \phi) \propto \exp\left\{\frac{y\gamma - b(\gamma)}{a(\phi)} + c(y, \phi)\right\}$$

for known functions $a(\phi)$, $b(\gamma)$ and $c(y, \phi)$. Here $\gamma = \gamma(\boldsymbol{\theta})$. Special cases include the normal linear regression model, Poisson regression, logistic regression and many parametric survival models. For a detailed discussion on GLM's, see McCullagh and Nelder (1989, Chapter 2).

The score function for $\boldsymbol{\theta}$ in the GLM has the form

$$(2.3) \qquad \sum_{i=1}^{K} \frac{\partial \mu_i(\boldsymbol{\theta})'}{\partial \boldsymbol{\theta}} v_i^{-1}(y_i - \mu_i(\boldsymbol{\theta})) = 0,$$

where $v_i = \operatorname{Var}(y_i) = a(\phi) b''(\gamma)$. Note that the nuisance parameter ϕ appears in (2.3) as a proportional factor and may be ignored when solving the equation.

Wedderburn (1974) pointed out that the GLM score equation (2.3) depends only on μ_i and v_i, the mean and the variance of y_i. He suggested that (2.3) could be solved for $\hat{\boldsymbol{\theta}}$ with arbitrary functional forms of μ and v including those which did not correspond to a particular member of the GLM family. He coined the term "quasilikelihood" for the integral of (2.3), which need not be a proper likelihood function. Godambe and Heyde (1987) showed that the estimating function in (2.3), now known as the quasi-score function, minimizes (2.2) among unbiased estimating functions which are linear in the data, that is, take the form $\sum_i a_i(\boldsymbol{\theta})(y_i - \mu_i(\boldsymbol{\theta}))$. Hence the quasi-score function is an example of an optimal estimating function.

It is helpful at this stage to sketch a proof of the optimality for (2.3) in the scalar case. Extension to the multivariate cases is straightforward and omitted.

For an unbiased estimating function of the form $\sum_i a_i(\theta)(y_i - \mu_i(\theta))$, S_K in (2.2) reduces to

$$\frac{\sum_{i=1}^K a_i^2 v_i}{(\sum_{i=1}^K a_i(\partial\mu_i/\partial\boldsymbol{\theta}))^2}$$

$$= \frac{\sum_{i=1}^K (a_i\sqrt{v_i})^2}{\left(\sum_{i=1}^K (a_i\sqrt{v_i})\cdot((\partial\mu_i/\partial\theta)/\sqrt{v_i})\right)^2}.$$

For the quasi-score function $a_i^*(\theta) = (\partial\mu_i(\theta)/\partial\theta)' v_i^{-1}$ and hence

$$S_K^* = \frac{1}{\left(\sum_{i=1}^K ((\partial\mu_i/\partial\theta)/\sqrt{v_i})\right)^2}.$$

The optimality of the quasi-score function, that is, $S_K^* \leq S_K$ for any choice of $a_i(\theta)$, is a direct consequence of Schwarz's inequality; see Lindsay (1982) for a more thorough proof.

Several authors, including Firth (1987), Crowder (1987) and Godambe and Thompson (1989), have studied an extension of the quasi-score function that includes both y_i and $w_i = (y_i - \mu_i)^2$, $i = 1, \ldots, K$, the so-called quadratic estimating functions. In the regression case, it can easily be shown that the optimal quadratic estimating function for $\boldsymbol{\theta}$ is

$$(2.4) \qquad \sum_{i=1}^K \frac{\partial\mathbb{E}(y_i, w_i)'}{\partial\boldsymbol{\theta}} \operatorname{Var}(y_i, w_i)^{-1} \cdot \begin{pmatrix} y_i - \mathbb{E}(y_i) \\ w_i - \mathbb{E}(w_i) \end{pmatrix} = 0.$$

Note that the variance matrix for $(y_i, w_i)'$ involves third and fourth cumulants, which must be assumed known for the finite sample optimality theory to apply. We will return to this issue in Section 3.

2.3 Role of Unbiasedness

The theory of optimal estimating functions in Section 2.2 takes as a basic assumption that $\mathbb{E}(g(y, \theta); \theta, \phi) = 0$ for all θ and ϕ, that is, $g(y, \theta)$ is an unbiased estimating function. What is the role of this assumption? First, under regularity conditions, unbiased equations have roots which are consistent estimators. To see this, given independent observations $y_1, \ldots, y_K$; let $g_i = g(y_i, \theta)$ and define $g_K = K^{-1}\sum_{i=1}^K g_i$. If the g_i are unbiased and $\mathbb{E}(g^2) < \infty$, then g_K converges almost surely to 0, its expectation at the true θ. For g continuous and a one-to-one function of θ^* in the neighborhood of the true value θ, $\hat{\theta}_K = g_K^{-1}(0) \to \boldsymbol{\theta}$ since $g_K \to 0$ at θ. The problem with maximum likelihood estimation in the presence of nuisance parameters is partly attributable to this bias of the score equation. The score is the derivative of the profile likelihood $\ell(\theta, \hat{\boldsymbol{\phi}}_\theta)$, where $\hat{\boldsymbol{\phi}}_\theta$ is the maximum likelihood estimator for ϕ for fixed θ. While $\mathbb{E}(\partial\ell(\theta, \phi)/\partial\theta) = 0$, it is not true that $\mathbb{E}(\partial\ell(\theta, \hat{\phi}_\theta)/\partial\theta) = 0$. With infinitely many nuisance parameters, the score function can fail to converge to 0 and hence give an inconsistent estimate as its root. On the other hand, the conditional score function is unbiased as discussed in Section 3, which partly explains its superior performance with many nuisance parameters.

Also, there is substantial finite sample evidence from Monte Carlo studies showing the improved performance of unbiased estimating functions over, for example, the analogous score function; see Breslow (1981) for case-control studies and Liang and Hanfelt (1994) for teratological experiments. Section 3 presents a further example for the Weibull distribution.

2.4 Limitations

Up to this point, we have reviewed the basic ideas of optimal estimating functions. Two of their major limitations must also be considered. First, optimality is ascribed to the estimating function, but scientists and other practitioners are concerned about estimators. As Crowder (1989) put it: "This is like admiring the pram rather than the baby." Godambe's optimality criterion is equivalent to asymptotic not finite sample optimality of the estimator (Durbin, 1960). In the context of quasilikelihood, McCullagh (1983) showed that the solution of (2.3) is asymptotically unbiased and has minimum variance among solutions of estimating functions linear in the data. Thus the limitation of estimating functions noted above may not be an issue as such if one is willing to appeal to a large sample argument.

Second, nuisance parameters can compromise the optimality theory. While Godambe (1976) demonstrated that the conditional score function is optimal by criterion (2.2), he had to assume the existence of a complete sufficient statistic for ϕ that did not depend on θ. Such a statistic can be found for exponential family distributions but more generally $t = t(\theta)$. In this case, the conditional score function for θ depends on ϕ as well and hence is only locally optimal at the true ϕ (Lindsay, 1982). The quasi-score function in (2.3) can also suffer this limitation. For example, if the over-dispersion parameter ϕ depends on one of the x covariates, such as treatment assignment, $a(\phi)$ will not factor out of the sum over subjects in (2.3).

In the next section, we discuss in more detail the difficulty nuisance parameters cause for the optimality theory of estimating functions. In addition, we will examine some desirable properties that conditional score functions possess which serve as guidelines about how to handle nuisance parameters in the absence of a fully specified likelihood.

3. OPTIMAL ESTIMATING FUNCTIONS AND NUISANCE PARAMETERS

In Section 2, we reviewed the optimality theory for estimating functions when there are no nuisance parameters or when the data follow an exponential family distribution. In this section we consider the more general case focusing on (i) how the optimality theory may break down in the presence of nuisance parameters and (ii) alternate approaches to estimating θ in the presence of nuisance parameters.

3.1 Parametric Case

Recall that in the absence of nuisance parameters, that is, when $\boldsymbol{\phi}$ is known, the score function for $\boldsymbol{\theta}$,

$$U_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{\partial \log L(\boldsymbol{\theta}, \boldsymbol{\phi})}{\partial \boldsymbol{\theta}},$$

is optimal among unbiased estimating functions for $\boldsymbol{\theta}$. The use of $U_{\boldsymbol{\theta}}$ becomes complicated when $\boldsymbol{\phi}$ is unknown, as $U_{\boldsymbol{\theta}}$ may not be unbiased unless evaluated at the true $\boldsymbol{\phi}$ value, $\boldsymbol{\phi}_0$, that is, in general $\mathbb{E}(U_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\phi}); \boldsymbol{\theta}, \boldsymbol{\phi}_0) \neq 0$. One might expect to resolve this problem by replacing $\boldsymbol{\phi}$ with an estimate $\hat{\boldsymbol{\phi}}$ in $U_{\boldsymbol{\theta}}$. In fact, under regularity conditions, the mle $\hat{\boldsymbol{\theta}}$ is the solution of $U_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}) = 0$, where $\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\phi}$ for fixed $\boldsymbol{\theta}$ (Richards, 1961). However, while it is true in general that $\mathbb{E}(U_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\phi}); \boldsymbol{\theta}, \boldsymbol{\phi}) = 0$, it is not true that $\mathbb{E}(U_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}); \boldsymbol{\theta}, \boldsymbol{\phi}) = 0$, so that $\hat{\boldsymbol{\theta}}$ is in general a solution of a biased estimating function for $\boldsymbol{\theta}$.

EXAMPLE 3.1. Let y_i, $i = 1, \ldots, K$, be independent observations from a Weibull distribution with a shape parameter θ and scale parameter ϕ, so that the likelihood function has the form

$$L(\theta, \phi) \propto \prod_{i=1}^{K} \theta\phi y_i^{\theta-1} \exp(-\phi y_i^{\theta}).$$

The score function for θ is

$$U_\theta(\theta, \phi) = \frac{K}{\theta} + \sum_{i=1}^{K} \log y_i - \phi \sum_{i=1}^{K} y_i^{\theta} \log y_i.$$

The maximum likelihood estimate $\hat{\phi}_\theta$ for ϕ at fixed θ is $\hat{\phi}_\theta = K/\sum_i y_i^\theta$. Evaluated at $\hat{\phi}_\theta$, the expected score equation for θ is

$$\mathbb{E}(U_\theta(\theta, \hat{\phi}_\theta); \theta, \phi) = \frac{K}{\theta}(1 + A_K - KB_K),$$

where

$$A_K = (K-1)\sum_{j=0}^{K-2} \binom{K-2}{j} \frac{(-1)^{j+1}}{(j+1)^2},$$
(3.1)
$$B_K = (K-1)\sum_{j=0}^{K-2} \binom{K-2}{j} \frac{(-1)^{j+1}}{(j+2)^2}.$$

Figure 2 shows plots of $\mathbb{E}(U_\theta(\theta, \hat{\phi}_\theta))/K$ versus θ for $K = 10$, 20, 30, 40 and 50. The expected bias of

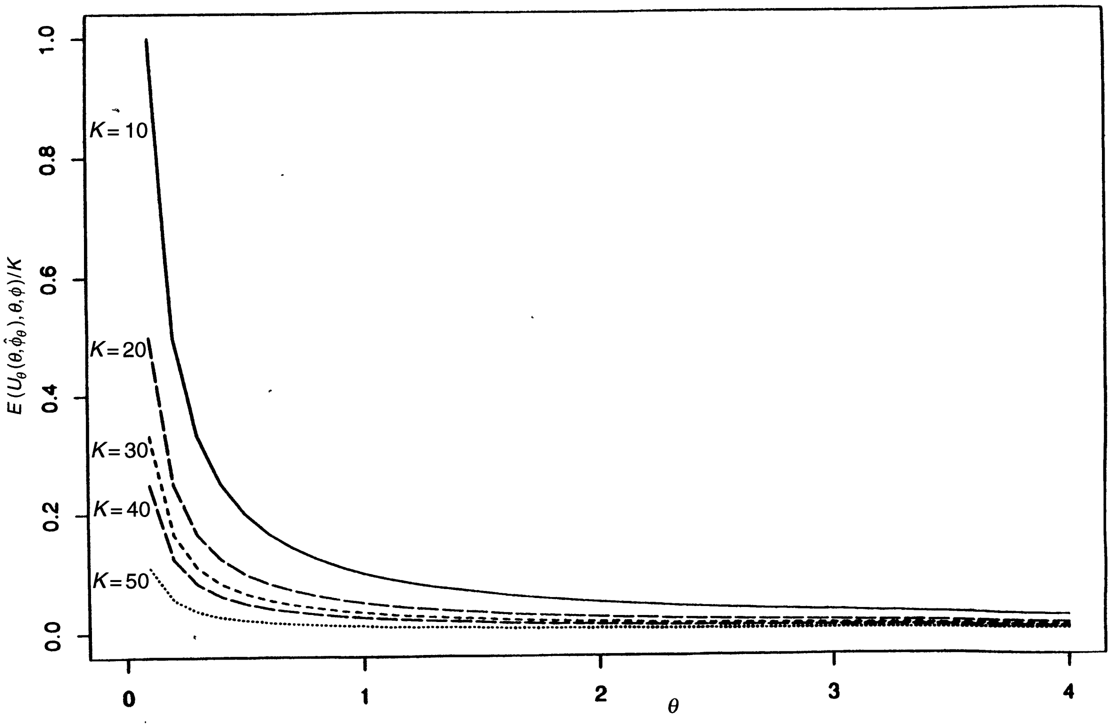

FIG. 2. *Expected bias in Weibull profile score of size K.*

$U_\theta(\theta, \hat{\phi}_\theta)$ per observation can be substantial when K and θ are small. We will return to this example later to examine the bias of $\hat{\theta}$, the mle of θ, induced by this bias in $U_\theta(\theta, \hat{\phi}_\theta)$.

The bias that results in the score equation when nuisance parameters are replaced by estimates leads us to consider alternate estimating functions. One strategy is to consider a function which is the $\boldsymbol{\theta}$-derivative of a proper log-likelihood function which depends upon $\boldsymbol{\theta}$ only. One such candidate is the marginal score function based upon a marginal likelihood defined in Section 1. Formally, $\mathbf{g}(\mathbf{y}; \boldsymbol{\theta})$ is a marginal score function for $\boldsymbol{\theta}$ if

$$\mathbf{g}(\mathbf{y}; \boldsymbol{\theta}) = \frac{\partial \log f(\mathbf{s}(\mathbf{y}); \boldsymbol{\theta})}{\partial \boldsymbol{\theta}},$$

where $\mathbf{s}(\mathbf{y})$ is a function of the data $\mathbf{y}$ whose distribution depends on $\boldsymbol{\theta}$ only.

EXAMPLE 3.2. Let $\mathbf{y}$ be a $K \times 1$ vector of Gaussian observations with mean $\mathbf{x}'\boldsymbol{\phi}$ and covariance matrix $\Sigma = \Sigma(\boldsymbol{\theta})$. It has been shown that the distribution of $\mathbf{s} = A\mathbf{y}$, where A is a $(K - p) \times K$ matrix and $p = \dim(\boldsymbol{\phi})$, will be independent of $\boldsymbol{\theta}$ and of choice of A so long as $\mathbb{E}(A\mathbf{y}) = 0$ (Harville, 1977). One such choice is any $K - p$ rows of $I - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'$ with rank $= K - p$. Note that the corresponding likelihood based on $A\mathbf{y}$ is also known as the restricted maximum likelihood for $\boldsymbol{\theta}$.

There are two limitations with marginal score functions. First, there is no general guidance as to how one finds the proper statistics. Second, there is no known theory to support the optimality of the marginal score functions.

An alternate strategy that avoids these limitations is the use of conditional score functions. Specifically, let $\mathbf{t} = \mathbf{t}(\mathbf{y})$ be sufficient for $\boldsymbol{\phi}$ for given $\boldsymbol{\theta}$. The conditional score function for $\boldsymbol{\theta}$ defined as $\partial \log f(\mathbf{y}|\mathbf{t}; \boldsymbol{\theta})/\partial\boldsymbol{\theta}$ is optimal among unbiased estimating functions for $\boldsymbol{\theta}$ if $\mathbf{t}$ is complete (Godambe, 1976); see Section 2.2.

EXAMPLE 1.2 (Continued). In this example, the likelihood function is proportional to a product of $2K$ binomial distributions indexed by $\theta, \phi_1, \dots, \phi_K$. For given θ, the complete and sufficient statistics for ϕ_i are simply $t_i = y_{i1} + y_{i2}$, the total number of exposed individuals in the ith stratum, $i = 1, \dots, K$. This is a direct consequence of $\sum_i y_{i1}$ and $(t_1, \dots, t_K)$ being jointly sufficient for an exponential family with $\theta, \phi_1, \dots, \phi_K$ as the canonical parameters. The conditional score function has the simple form

$$\sum_i (y_i - \mathbb{E}(y_i | t_i \theta)),$$

which by definition is independent of $\{\phi_i\}$ and, obviously, unbiased for all θ.

Thus far, the conditioning statistic $\mathbf{t}$ is assumed to be functionally independent of $\boldsymbol{\theta}$, which is indeed the case when f is from an exponential family of the form

$$f(\mathbf{y}; \boldsymbol{\theta}, \boldsymbol{\phi}) \propto \exp\{\boldsymbol{\theta}'\mathbf{s} + \boldsymbol{\phi}'\mathbf{t} - h(\boldsymbol{\theta}, \boldsymbol{\phi})\}, \tag{3.2}$$

where both $\mathbf{s}$ and $\mathbf{t}$ are functions of $\mathbf{y}$. However, this is a special case that rules out many important probability models including, for example, Efron's (1986) double exponential models, generalized linear models with errors-in-variables (Stefanski and Carroll, 1987) and curved exponential family distributions (Efron, 1975). More generally, $\mathbf{t}$, the minimal sufficient statistic for $\boldsymbol{\phi}$ for fixed $\boldsymbol{\theta}$, is a function of $\boldsymbol{\theta}$ as well, that is, $\mathbf{t} = \mathbf{t}_{\boldsymbol{\theta}}$ (McCullagh and Nelder, 1989, Chapter 7). In this situation, the conditional distribution of $\mathbf{y}$ given $\mathbf{t}_{\boldsymbol{\theta}}$ depends on $\boldsymbol{\phi}$ except when $\boldsymbol{\theta}$ is equal to its true value. Lindsay (1982) extends the concept of conditional score functions in this more general situation by defining

$$\begin{aligned} U_c(\boldsymbol{\theta}, \boldsymbol{\phi}) = {} & \frac{\partial \log(\boldsymbol{\theta}, \boldsymbol{\phi})}{\partial \boldsymbol{\theta}} \\ & - \mathbb{E}\left(\frac{\partial \log L(\boldsymbol{\theta}, \boldsymbol{\phi})}{\partial \theta} \middle| \mathbf{t}_{\boldsymbol{\theta}} \right), \end{aligned} \tag{3.3}$$

which reduces to $\partial \log f(\mathbf{y}|\mathbf{t}; \boldsymbol{\theta})/\partial\boldsymbol{\theta}$ when $\mathbf{t}_{\boldsymbol{\theta}} = \mathbf{t}$. Due to the dependence of (3.3) on $\boldsymbol{\phi}$, the conditional score function is only locally optimal at the true $\boldsymbol{\phi}$ (Lindsay, 1982). Despite the fact that the optimality theory of estimating functions breaks down in the presence of nuisance parameters, one may argue that the conditional score function is preferable to the conventional score function $U_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ as follows. Note that U_c is, by definition, orthogonal to the space spanned by the sufficient statistic $\mathbf{t}_{\boldsymbol{\theta}}$, which unfortunately, in this case, depends on $\boldsymbol{\theta}$ and hence induces the continued dependence of $U_c(\boldsymbol{\theta}, \boldsymbol{\phi})$ on $\boldsymbol{\phi}$. However, the representation in (3.3) implies the following properties indicative of the reduced dependence of $U_c(\boldsymbol{\theta}, \boldsymbol{\phi})$ which are not shared by $U_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\phi})$. These are listed so that (3.3) implies (a), which implies (b), which implies (c), which implies (d):

(a) $\mathbb{E}(U_c(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}); \boldsymbol{\theta}, \boldsymbol{\phi}) = 0$ for all $\boldsymbol{\theta}, \boldsymbol{\phi}$ and any $\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}$ which is a function of $\mathbf{t}_{\boldsymbol{\theta}}$;
(b) $\mathbb{E}(U_c(\boldsymbol{\theta}, \boldsymbol{\phi}^*); \boldsymbol{\theta}, \boldsymbol{\phi}) = 0$ for all $\boldsymbol{\theta}, \boldsymbol{\phi}$ and $\boldsymbol{\phi}^*$;
(c) $\mathbb{E}(\partial U_c(\boldsymbol{\theta}, \boldsymbol{\phi}^*)/\partial\boldsymbol{\phi}^*; \boldsymbol{\theta}, \boldsymbol{\phi}) = 0$ for all $\boldsymbol{\theta}, \boldsymbol{\phi}$ and $\boldsymbol{\phi}^*$;
(d) $\mathrm{Cov}(U_c(\boldsymbol{\theta}, \boldsymbol{\phi}), \partial \log L(\boldsymbol{\theta}, \boldsymbol{\phi})/\partial\boldsymbol{\phi}; \boldsymbol{\theta}, \boldsymbol{\phi}) = 0$ for all $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$.

Property (a) states that the conditional score function is unbiased when suitable estimates of $\boldsymbol{\phi}$ are inserted (Lindsay, 1982). Property (b) states that the unbiasedness of the conditional score function is also preserved if evaluated at the incorrect value $\boldsymbol{\phi}^*$ for the nuisance parameter. Hence, for example, $\hat{\boldsymbol{\theta}}(\boldsymbol{\phi}^*)$, the root of $U_c(\boldsymbol{\theta}, \boldsymbol{\phi}^*)$ is consistent for any $\boldsymbol{\phi}^*$. Property (c) can be viewed as an asymptotic version of (b) because (c) implies that $\mathbb{E}(U_c(\boldsymbol{\theta}, \tilde{\boldsymbol{\phi}}); \boldsymbol{\theta}, \boldsymbol{\phi}) = 0$ not for any nuisance parameter value, but for those $\tilde{\boldsymbol{\phi}}_{\boldsymbol{\theta}}$ which are $\sqrt{K}$-consistent, that is, for which $\sqrt{K}(\tilde{\boldsymbol{\phi}} - \boldsymbol{\phi}) = O_p(1)$. The final property (d) implies that $\mathbb{E}(U_c(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}); \boldsymbol{\theta}, \boldsymbol{\phi}) = 0$ for $\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}$, the maximum likelihood estimator of $\boldsymbol{\phi}$ for fixed $\boldsymbol{\theta}$. Note that property (d) has been used by Godambe (1991b) to define the orthogonality between two sets of estimating functions.

To illustrate these ideas, consider a class of distributions for $\mathbf{y}$ of the form (Liang and Tsou, 1992)

$$(3.4) \quad f(\mathbf{y}; \boldsymbol{\theta}, \boldsymbol{\phi}) \propto \exp\{\boldsymbol{\theta}'\mathbf{s} + \boldsymbol{\phi}'\mathbf{t}_{\boldsymbol{\theta}} - h(\boldsymbol{\theta}, \boldsymbol{\phi})\},$$

where $\mathbf{s} = \mathbf{s}(\mathbf{y})$ and $\mathbf{t}_{\boldsymbol{\theta}} = \mathbf{t}(\mathbf{y}, \boldsymbol{\theta})$. This class of distributions includes the following as special cases: (i) the exponential family in (3.2) with $\mathbf{t}_{\boldsymbol{\theta}} = \mathbf{t}$; (ii) the subexponential model by Lindsay (1982) with $\mathbf{s} = 0$; (iii) the double exponential family by Efron (1986); (iv) the generalized linear models with Gaussian covariate errors (Stefanski and Carroll, 1987); (v) the Weibull distribution discussed in Example 3.1; and (vi) the variance component model in Example 3.2. The conditional score function is

$$(3.5) \quad \mathbf{U}_c(\boldsymbol{\theta}, \boldsymbol{\phi}) = \mathbf{s} - \mathbb{E}(\mathbf{s} \mid \mathbf{t}_{\boldsymbol{\theta}}) + \boldsymbol{\phi}\{\mathbf{t}'_{\boldsymbol{\theta}} - \mathbb{E}(\mathbf{t}'_{\boldsymbol{\theta}} \mid \mathbf{t}_{\boldsymbol{\theta}})\},$$

where $\mathbf{t}'_{\boldsymbol{\theta}}$ is the derivative of $\mathbf{t}_{\boldsymbol{\theta}}$ with respect to $\boldsymbol{\theta}$. Thus, the conditional score function does depend on $\boldsymbol{\phi}$, which appears only as a weight associated with the third term in (3.5). Properties (a), (b) and (c) are obvious from (3.5), whereas for (d) we note that $\partial \log L(\boldsymbol{\theta}, \boldsymbol{\phi})/\partial\boldsymbol{\phi} = \mathbf{t}_{\boldsymbol{\theta}} - \partial h/\partial\boldsymbol{\phi}$.

EXAMPLE 3.1 (Continued). The Weibull distribution can be seen as a special case of (3.5) with $s = \sum_i \log y_i$, $t_\theta = -\sum_i y_i^\theta$ and $h(\theta, \phi) = K \log(\theta\phi)$. The conditional score function has the form

$$U_c(\theta, \phi)$$
$$= \frac{1}{\theta}\left\{\sum_{i=1}^K \log y_i^\theta - K(A_K + \log t_\theta)\right\}$$
$$-\frac{\phi}{\theta}\left\{\sum_{i=1}^K y_i^\theta \log y_i^\theta K t_\theta\left(B_K + \frac{1}{K}\log t_\theta\right)\right\},$$

where A_K and B_K are given in (3.1). To compare the conditional score estimator of θ with the maximum likelihood estimator, we conducted a simulation study where, for each selected θ and ϕ, 1,000 replicates of $K = 30$ independent Weibull observations were generated. Table 1 gives empirical estimates of $\mathbb{E}\{U_\theta(\theta, \phi_0)\}$, $\mathbb{E}\{U_\theta(\theta, \hat{\phi}_\theta)\}$, $\mathbb{E}\{U_c(\theta, \phi_0)\}$ and $\mathbb{E}\{U_c(\theta, \hat{\phi}_\theta)\}$ and biases of estimates of θ obtained by solving these equations. Here ϕ_0 is the true value of ϕ. It is clear that the impact of estimating ϕ is far greater for U_θ than it is for U_c. The absolute value of $\hat{\mathbb{E}}(U_\theta(\theta, \hat{\phi}_\theta))$ is 5 to 10 times as large as that of $\hat{\mathbb{E}}\{U_\theta(\theta, \phi_0)\}$, whereas there is little discrepancy between $\hat{\mathbb{E}}\{U_c(\theta, \hat{\phi}_\theta)\}$ and $\hat{\mathbb{E}}\{U_c(\theta, \phi_0)\}$, both of which are close to $\hat{\mathbb{E}}\{U_\theta(\theta, \phi_0)\}$. This phenomenon transmits directly to biases of corresponding estimators. Overall, the bias of $\hat{\theta}_c$, the solution of $U_c(\theta, \hat{\phi}_\theta) = 0$, is half that of $\hat{\theta}$, the maximum likelihood estimate. We also note that the variance of $\hat{\theta}_c$ is comparable to that of $\hat{\theta}$ although this is not shown in the table. Several researchers, including Cox and Reid (1987), Liang (1987), Ferguson, Reid and Cox (1991) and Lindsay and Waterman (1992), have investigated strategies for approximating the conditional score function in this more general case and cases where the complete sufficient statistics for $\boldsymbol{\phi}$ for fixed $\boldsymbol{\theta}$ are difficult to find.

TABLE 1
Empirical estimates of expectations and biases in associated estimators of parameter from the Weibull distribution: upper entry, $\phi_0 = 1$; lower entry, $\phi_0 = 2$

θ	$\hat{\mathbb{E}}(U_\theta(\theta, \phi_0))$	$\hat{\mathbb{E}}(U_\theta(\theta, \hat{\phi}_\theta))$	$\hat{\mathbb{E}}(U_c(\theta, \phi_0))$	$\mathbb{E}(U_c(\theta, \hat{\phi}_\theta))$
1	0.088	1.119	0.135	0.119
	(0.238)*	(0.225)	(0.225)	(0.225)
	−0.295	0.647	−0.341	−0.353
	(0.231)	(0.221)	(0.221)	(0.221)
2	0.103	0.605	0.100	0.105
	(0.117)	(0.108)	(0.109)	(0.108)
	0.092	0.601	0.097	0.101
	(0.111)	(0.106)	(0.106)	(0.106)
4	−0.039	0.193	−0.055	−0.057
	(0.058)	(0.054)	(0.054)	(0.054)
	−0.045	0.209	−0.035	−0.041
	(0.057)	(0.054)	(0.054)	(0.054)
		Bias of estimator		
1	0.032	0.053	0.031	0.031
	0.017	0.042	0.020	0.020
2	0.066	0.106	0.062	0.063
	0.051	0.105	0.062	0.062
4	0.101	0.174	0.089	0.088
	0.075	0.179	0.093	0.093

*Standard error of the empirical estimate of expectations

3.2 Quasilikelihood

As indicated in Section 2.2, a major application of the theory of optimal estimating functions is the quasi-score function. While in this case the variance of y_i depends on a nuisance parameter $\boldsymbol{\phi}$, the quasi-score function in (2.3), $\mathbf{g}(\mathbf{y}; \boldsymbol{\theta})$, does not since the variance is assumed to take the form $a(\boldsymbol{\phi})V(\mu_i(\boldsymbol{\theta}))$

so that $a(\boldsymbol{\phi})$ factors out of $\mathbf{g}$. More generally, $\mathrm{Var}(y_i)$ does not factor and is expressed as $V_i(\mu_i, \boldsymbol{\phi})$. An example is the beta-binomial variance from Example 1.3. Here, the quasi-score function depends on the nuisance parameter. One needs to estimate $\boldsymbol{\phi}$ in order to solve it. Consequently, the quasi-score function for $\boldsymbol{\theta}$ is only locally optimal at the true $\boldsymbol{\phi}$ value. Just as $g_i = y_i - \mu_i(\boldsymbol{\theta})$ may be used as the basis for the inference on the mean parameters, it seems natural to consider $w_i = (y_i - \mu_i(\boldsymbol{\theta}))^2 - V_i(\mu_i, \boldsymbol{\phi})$ for the inference on the variance parameters, $\boldsymbol{\phi}$. Following the proof laid out in Section 2.2, one can show that the estimating function in (2.4) is optimal among a class of joint estimating functions for $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ which are a linear combination of $(g_i, w_i)'$, $i = 1, \ldots, K$.

This joint optimality approach is subject to the following criticisms. First, to compute the estimating function in (2.4), one needs to know the third and fourth moments of the y_i's (Godambe and Thompson, 1989). Such knowledge is seldom available (e.g., Firth, 1989). Second, the $\boldsymbol{\theta}$ component in (2.4), call it $U_{\boldsymbol{\theta}}$, is seen as a linear combination of $y_i - \mu_i$ and $(y_i - \mu_i)^2$ rather than a function of $y_i - \mu_i$ alone as in (2.3). Thus the unbiasedness of $U_{\boldsymbol{\theta}}$ and hence the consistency of the corresponding estimator of $\boldsymbol{\theta}$ depends on whether $\mathrm{Var}(y_i) = V_i(\mu_i, \boldsymbol{\phi})$, that is, whether one has correctly specified the variance function of y_i as characterized by $\boldsymbol{\phi}$. This is precisely the situation one tries to avoid, namely, a potentially strong effect on inference for parameters of interest due to the presence of nuisance parameters. Specifically, none of the properties (a)–(d) stated in Section 3.1 is fulfilled by $U_{\boldsymbol{\theta}}$ unless $V_i = \mathrm{Var}(y_i)$.

It is worth noting that the use of quadratic estimating functions was originally motivated on the grounds that a portion of the information on $\boldsymbol{\theta}$ may not be fully captured by y_i, as $\mathrm{Var}(y_i)$ depends on the mean as well (Crowder, 1987; Firth, 1987).

EXAMPLE 3.3. Consider K independent observations y_i with mean $\theta_1 + \theta_2 x_i$, $i = 1, \ldots, K$, and variance assumed to be constant, say, ϕ. The scientific focus is on estimation of $\theta = (\theta_1, \theta_2)$. The optimal estimating functions for $(\theta_1, \theta_2, \phi)$ are

$$\mathbf{U}^* = \sum_{i=1}^{K} \begin{pmatrix} -1 & 0 \\ -x_i & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \phi & \phi_3 \\ \phi_3 & \phi_4 - \phi^2 \end{pmatrix}^{-1} \cdot \begin{pmatrix} y_i - \theta_1 - \theta_2 x_i \\ (y_i - \theta_1 - \theta_2 x_i)^2 - \phi \end{pmatrix}, \tag{3.6}$$

where ϕ_3 and ϕ_4 are the central third and fourth moments of the y's. For ease of discussion, we assume ϕ_3 and ϕ_4 are known to the investigators. We further assume that $x_i = 1$ if $i = 1, \ldots, K/2$, and equals 0 otherwise and that the variances from these two subpopulations are different, with ϕ_1 and ϕ_0 being the true variances, respectively. While it is difficult to derive the solution of $U^* = 0$ in closed form, asymptotically one may show that it converges to $(\theta_1^*, \theta_2^*, \phi^*)$, which is the solution to the system

$$-(\phi_4 - \phi^{*2})\left(\delta_1 + \frac{\delta_2}{2}\right) + \phi_3\left(\frac{\phi_1 + \phi_0}{2} - \phi^* + \frac{\delta_1^2 + (\delta_1 + \delta_2)^2}{2}\right) = 0,$$

$$-(\phi_4 - \phi^{*2})(\delta_1 + \delta_2) + \phi_3(\phi_1 - \phi^* + (\delta_1 + \delta_2)^2) = 0,$$

$$\phi_3\left(\delta_1 + \frac{\delta_2}{2}\right) - \phi^*\left(\frac{\phi_1 + \phi_0}{2} - \phi^* + \frac{\delta_1^2 + (\delta_1 + \delta_2)^2}{2}\right) = 0,$$

where $\delta_1 = \theta_{10} - \theta_1^*$, $\delta_2 = \theta_{20} - \theta_2^*$ and θ_{10} and θ_{20} are the true values of θ_1 and θ_2, respectively. Figure 3 shows the bias δ_2 of $\hat{\theta}_2$ for a range of parameter values for ϕ_1, ϕ_3 and ϕ_4 when $\phi_0 = 1$. Substantial bias can result when the true distribution of the y's departs from being Gaussian as measured by the skewness ϕ_3 and kurtosis ϕ_4. Also evident from this figure is that the bias increases as the discrepancy between ϕ_0 and ϕ_1 increases.

3.3 Summary

In this section, we reviewed the difficulty nuisance parameters cause for the optimality theory discussed in Section 2. Basically, the optimal estimating function for parameters of interest, either the conditional score function in the parametric case or the quasi-score function in the quasilikelihood setting, depends upon nuisance parameters as well. Consequently, the optimality only holds locally at the true $\boldsymbol{\phi}$ values and no global optimality can be claimed.

Nevertheless, we identified in Section 3.1 several desired orthogonality properties possessed by the conditional score function despite its dependence on $\boldsymbol{\phi}$. The tradeoff is that one needs to specify fully the likelihood function for the y's. While no distributional assumption is needed for the quasilikelihood, we argued in Section 3.2 that the jointly optimal estimating equations can lead to inconsistent estimates for $\boldsymbol{\theta}$ when the variance function is misspecified. In the next section, we discuss, in the absence of the likelihood function, ways to derive locally optimal estimating functions which possess the orthogonality properties in Section 3.1 in settings which include the quasilikelihood as a special case.

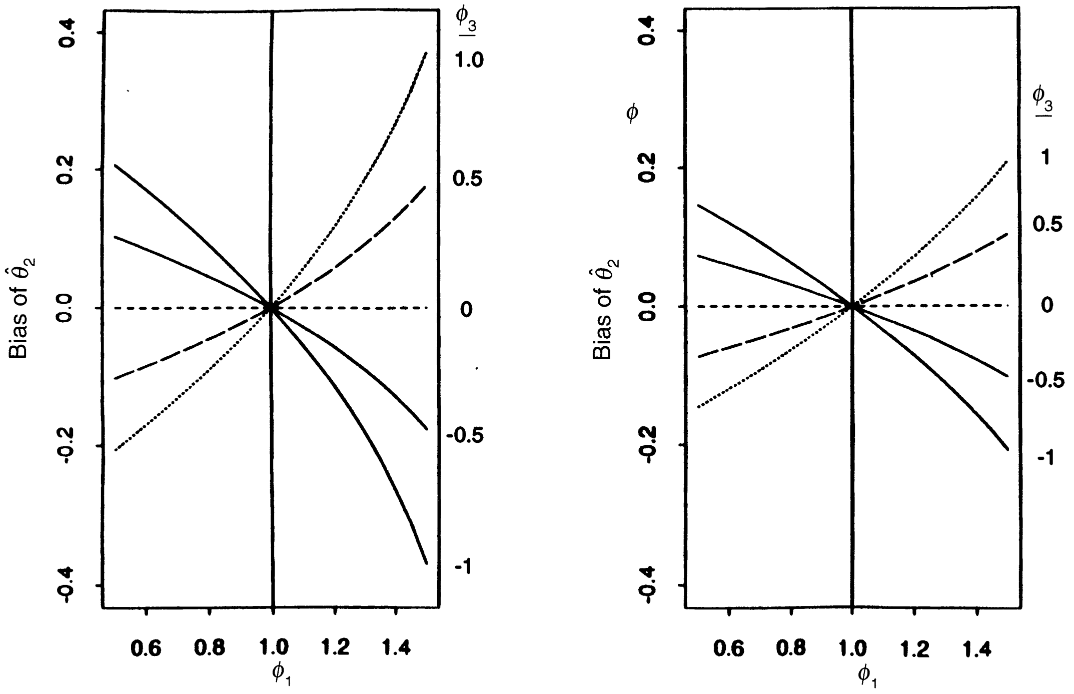

FIG. 3. *Bias of $\hat{\theta}_2$ for a range of parameter values in ϕ_1, ϕ_3 and ϕ_4: left, $\phi_4 = 3.0$; right, $\phi_4 = 4.0$.*

4. HANDLING NUISANCE PARAMETERS THROUGH ESTIMATING FUNCTIONS

Very often one faces a problem where either the investigators are uncertain about the complete random mechanism or the probability distribution is too complicated to write down explicitly. Examples include data from longitudinal studies (e.g., Liang and Zeger, 1986), from complicated stochastic processes (Azzalini, 1984) and from spatial processes encountered in plant ecology (Besag, 1974). It is worth pointing out that in each of these examples, the parameters of interest $\boldsymbol{\theta}$ are well defined, reflecting the scientific interest, even though they may not completely specify the distribution of the data $\mathbf{y}$. The primary issue to be addressed in this section is how to make inference for $\boldsymbol{\theta}$ in the absence of a likelihood function. We are especially interested in controlling the effects of nuisance parameters $\boldsymbol{\phi}$ in this semiparametric situation. Here, "semiparametric" refers to situations where only the quantities of interest are modeled parametrically (e.g., through regression), whereas the nuisance part $\boldsymbol{\phi}$ may be finite or infinite dimensional.

4.1 Optimal Estimating Functions

We assume throughout the section the existence of unbiased estimating equations $\{\mathbf{g}_i = \mathbf{g}_i(\mathbf{y}, \boldsymbol{\theta});\ i = 1, \dots, K\}$ such that $\mathbb{E}(\mathbf{g}_i; \boldsymbol{\theta}, \boldsymbol{\phi}) = 0$ for all $\boldsymbol{\theta}, \boldsymbol{\phi}$ and i and that the $\mathbf{g}_i$ are uncorrelated with each other. Note that the dimension of $\mathbf{g}$ may be different than that of $\boldsymbol{\theta}$, for example, $g_i = y_i - \mathbb{E}(y_i; \boldsymbol{\theta}) = y_i - x_i'\boldsymbol{\theta}$ is a scalar whereas $\boldsymbol{\theta}$ is p-dimensional. Very often it is easier to verify the unbiasedness of $\mathbf{g}_i$ through checking $\mathbb{E}(\mathbf{g}_i \mid \mathbf{A}_i) = 0$ for some statistics $\mathbf{A}_i$. The choice of $\mathbf{A}_i$, preferably to have maximum dimension (McCullagh and Nelder, 1989, Section 9.4), varies from case to case. In a stochastic process problem where $\mathbf{g}_i = \mathbf{g}_i(y_1, \dots, y_i; \boldsymbol{\theta})$ we typically choose $\mathbf{A}_i = (y_1, \dots, y_{i-1})$, the history. The zero-correlation between $\mathbf{g}_i$ and $\mathbf{g}_j$, $j < i = 1, \dots, K$, in this case is also easily seen since $\mathrm{Cov}(\mathbf{g}_i, \mathbf{g}_j) = \mathbb{E}(\mathbf{g}_i\mathbf{g}_j) = \mathbb{E}\{\mathbf{g}_j\mathbb{E}(\mathbf{g}_i \mid \mathbf{A}_i)\} = 0$. In Example 1.2 with K independent 2×2 tables, a natural choice of $\mathbf{A}_i$ is $t_i = y_{i1} + y_{i2}$, the total number of exposures in the ith stratum subgroup. In the case where there is no such conditioning event available, $\mathbf{A}_i$ will be taken as a null set.

To combine these K uncorrelated estimating functions for $\boldsymbol{\theta}$ into a single one, a simple strategy is to consider a linear combination of the $\mathbf{g}_i$. One main advantage when $\mathbf{g}_i$ is conditionally unbiased [i.e., $\mathbb{E}(\mathbf{g}_i \mid \mathbf{A}_i) = 0$] is that one may now consider a broader class of unbiased estimating functions in which the weight associated with g_i is allowed to be a function of $\mathbf{A}_i$ rather than a constant only, that is,

$$\sum_{i=1}^{K} a_i(\boldsymbol{\theta}, \mathbf{A}_i)\mathbf{g}_i.$$

Following the proof in Section 2.2, the linear combination with corresponding root that has minimum asymptotic variance is

$$(4.1)\qquad \mathbf{g} = \sum_{i=1}^{K} \mathbb{E}\left(\frac{\partial g_i}{\partial \boldsymbol{\theta}} \middle| \mathbf{A}_i \right)' \operatorname{Var}(\mathbf{g}_i \mid \mathbf{A}_i)^{-1} \mathbf{g}_i.$$

The weight associated with each $\mathbf{g}_i$ is a product of two terms. The first term, $\operatorname{Var}(\mathbf{g}_i \mid \mathbf{A}_i)^{-1}$, is used to downweight those $\mathbf{g}_i$ with greater degree of uncertainty. The other term, $\mathbb{E}(\partial \mathbf{g}_i / \partial \boldsymbol{\theta} \mid \mathbf{A}_i)$, transforms the space spanned by the data to the parameter space. Lindsay (1982) considered some desirable properties of $\mathbf{g}$ where $\mathbb{E}(\partial \mathbf{g}_i / \partial \boldsymbol{\theta} \mid \mathbf{A}_i) = -\operatorname{Var}(\mathbf{g}_i \mid \mathbf{A}_i)$. He called this situation *information unbiased* and noted that the optimal $\mathbf{g}$ is simply $\sum_i \mathbf{g}_i$ in this case.

EXAMPLE 1.1 (Continued). In the situation where the scientific objective may be characterized by the regression model $\mathbb{E}(\mathbf{y}_i) = \boldsymbol{\mu}_i(\boldsymbol{\theta})$, where $\mathbf{y}_i$ is an $n_i \times 1$ vector of responses, the use of $\mathbf{g}_i = \mathbf{y}_i - \boldsymbol{\mu}_i$ leads to

$$\mathbf{g} = \sum_{i=1}^{K} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}} \right)' \operatorname{Cov}^{-1}(\mathbf{y}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta})).$$

This was referred to as the generalized estimating equation (GEE) by Liang and Zeger (1986). Note that the dimension of $\mathbf{g}_i$ varies from subject to subject in this case and that when $n_i = 1$ for all i, $\mathbf{g}$ reduces to (2.3), the quasi-score function.

Prentice (1988) suggested a second-order system of equations where if $\mathbf{w}_i = (y_{i1}, \ldots, y_{in_i}, y_{i1}y_{i2}, \ldots, y_{in_i-1}y_{in_i})$, then $\mathbf{g}_i = \mathbf{w}_i - \mathbb{E}(\mathbf{w}_i)$. Here the variance of $\mathbf{g}_i$ depends upon third- and fourth-order cumulants so that the optimal second-order equations are less easily applied than simpler alternatives.

EXAMPLE 1.2 (Continued). While the conditional score function has been shown to be optimal, it is complicated to compute. As an alternative, one may consider the estimating function $g_i = y_{i1}(n_{i2} - y_{i2}) - \theta y_{i2}(n_{i1} - y_{i1})$, $i = 1, \ldots, K$. The unbiasedness of g_i is obvious as

$$\mathbb{E}(g_i) = n_{i1}n_{i2}\mu_{i1}(1 - \mu_{i2}) - \theta n_{i1}n_{i2}\mu_{i2}(1 - \mu_{i1}) = 0.$$

In fact it has been shown that $\mathbb{E}(g_i \mid \mathbf{A}_i) = 0$ for $\mathbf{A}_i = y_{i1} + y_{i2}$ (Mantel and Hankey, 1975). It is worth noting that the conditional unbiasedness shown here and many other cases can be done using completeness and sufficiency of $\mathbf{A}_i$, if it exists. The computation of $B_i(\theta) = \mathbb{E}(\partial g_i / \partial \theta \mid \mathbf{A}_i)\operatorname{Var}(g_i \mid \mathbf{A}_i)^{-1}$, which depends on θ only, involves the calculation of the four moments of the extended hypergeometric distribution. However, simplification occurs when evaluated at $\theta = 1$, that is, $B_i(1) = 1/(n_i + m_i)$ (Yanagimoto, 1989; McCullagh, 1991). Interestingly, the solution of $\sum_i B_i(1)g_i = 0$ gives rise to the well-known Mantel–Haenszel estimator (Mantel and Haenszel, 1959). Note that this estimating function approach can easily accommodate regression analysis for odds ratio [e.g., $\theta_i = \theta_i(\mathbf{x}_i; \boldsymbol{\theta})$], where $\mathbf{x}_i$ represent confounding variables to be controlled.

The estimating function approach may be applied to situations where the binomial assumption for the y's no longer holds. This occurs in studies of familial aggregation for diseases where y_{i1} is the numbers of affected among n_{i1} relatives of the ith sampled case and y_{i2} is the numbers of affected among n_{i2} relatives of the matched control (Liang, 1985). The parameter of interest θ measures the degree of familial aggregation. The binomial assumption is clearly violated because individuals from the same family are correlated. Hence, the conditional score method is not applicable. In the absence of a likelihood function to work with, we may appeal to the same g_i defined earlier which still has zero expectation for each $i = 1, \ldots, K$, the number of sampled cases. The trade-off in this situation is that there is no obvious candidate for the conditioning event $\mathbf{A}_i$ such as $y_{i1} + y_{i2}$ when the binomial assumption fails to hold. Consequently, the computation of $\mathbb{E}(\partial g_i / \partial \theta)$ and $\operatorname{Var}(g_i)$ involves higher moments of the y's which may depend on additional parameters such as those describing the within-family correlations.

EXAMPLE 1.3 (Continued). Largely for mathematical convenience, the beta-binomial distribution is used to account for the litter effect in teratologic experiments. The biologic justification and especially the finite sample performance of likelihood inference under the beta-binomial distribution have recently been challenged (e.g., Kupper, Portier, Hogan and Yamamoto, 1986; Liang and Hanfelt, 1994). Note that the conditional score method does not work in this case since the minimal sufficient statistic for $\boldsymbol{\phi}, \mathbf{t}_{\boldsymbol{\theta}}$ is $\mathbf{y}$ itself. Consequently, $U_c(\boldsymbol{\theta}, \boldsymbol{\phi}) = 0$. Since the scientific objective can be realized through a regression model for the means, one alternative is to use the quasi-score function with $g_i = y_i - \mu_i(\boldsymbol{\theta})$ as the basis for inference. Through extensive simulation studies, Liang and Hanfelt (1994) found that, for the configurations being considered, this approach outperformed the mle based on the beta-binomial likelihood even when the data were generated from the beta-binomial distribution.

EXAMPLE 1.4 (Continued). Special attention is needed for the proportional hazards model since $\boldsymbol{\phi}$ is of infinite dimension. To establish the notation, let $t_{(1)} < t_{(2)} < \cdots < t_{(K)}$ be the distinct failure times observed among a sample of n subjects. Frequently, $n > K$ because some subjects are censored, that is, are still event-free at the end of the study or lost to follow-up during the study. For each i, we consider $\mathbf{g}_i = \mathbf{x}_{(i)} - \mathbb{E}(\mathbf{x}_{(i)} \,|\, \mathbf{A}_i)$, where the random variable $\mathbf{x}_{(i)}$ is the vector of covariates of the subject failed at $t_{(i)}$ and $\mathbf{A}_i$ represents the information on the history of the covariates and censoring prior to time $t_{(i)}$ plus the information that a subject fails at $t_{(i)}$. With this conditioning set, we choose

$$\mathbf{g}_i = \mathbf{x}_{(i)} - \frac{\sum_{\ell \in R_{(i)}} \mathbf{x}_\ell \exp(\mathbf{x}_\ell' \boldsymbol{\theta})}{\sum_{\ell \in R_{(i)}} \exp(\mathbf{x}_\ell' \boldsymbol{\theta})},$$

which depends on $\boldsymbol{\theta}$ only. Here $R_{(i)}$ is the set of subjects who were at risk prior to $t_{(i)}$. It is easy to verify that $\mathrm{Cov}(\mathbf{g}_i \,|\, \mathbf{A}_i)^{-1} = -\mathbb{E}(\partial \mathbf{g}_i / \partial \boldsymbol{\theta} \,|\, \mathbf{A}_i)$ and hence $\mathbf{g}_i$ is information unbiased, that is, $\mathbf{g} = \sum_i \mathbf{g}_i$. Of course, this $\mathbf{g}$ is the well known partial score function (Cox, 1975).

4.2 Handling Nuisance Parameters

So far we have avoided consideration of the possible dependence of $\mathbf{g}$ on additional parameters $\boldsymbol{\phi}$. The reason for this dependency on $\boldsymbol{\phi}$ is that, while the $\mathbf{g}_i$ are chosen to be functionally independent of $\boldsymbol{\phi}$, the distribution of $\mathbf{g}_i$ in general depends on $\boldsymbol{\phi}$.

There are two approaches by which this dependence can be eliminated. In the first, we choose $\mathbf{g}_i$ whose distribution, conditional on $\mathbf{A}_i$, depends on $\boldsymbol{\theta}$ only. It is then obvious that $\mathbf{g}$ in (4.1) depends functionally on $\boldsymbol{\theta}$ not on $\boldsymbol{\phi}$. Such $\mathbf{g}_i$ are called *pivotal* and were considered by Morton (1981). While desirable, this approach has found limited usage as, except in special cases such as location-shift models, it is difficult to find such pivotal quantities without the full knowledge of the distribution for $\mathbf{y}$. As an application of this approach, we note that the distribution of $\mathbf{g}_i$ given $\mathbf{A}_i$ in Example 1.4 depends on $\boldsymbol{\theta}$ only.

The second approach relies upon finding $\mathbf{g}_i$'s which are information unbiased, that is, for which $\mathbb{E}(\partial \mathbf{g}_i / \partial \boldsymbol{\theta} \,|\, \mathbf{A}_i) = \mathrm{Cov}(\mathbf{g}_i \,|\, \mathbf{A}_i)^{-1}$ so that $\mathbf{g} = \sum_i \mathbf{g}_i$, independent of $\boldsymbol{\phi}$. The conditional score function in exponential families (3.2) and the partial score function in Example 1.4 are special cases. A limitation of this approach is the requirement that the dimension of $\mathbf{g}_i$ be the same as that of $\boldsymbol{\theta}$. In addition, just as for the first approach, there are no general rules for finding such $\mathbf{g}_i$ without full knowledge on the distribution of $\mathbf{y}$. Furthermore, the finding of such $\mathbf{g}_i$ may be at the expense of losing efficiency, which is sometimes considerable.

More generally, the distribution of $\mathbf{g}$ in (4.1) will depend upon $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. However, we argue that the impact of the nuisance parameters on $\mathbf{g}$ and on the corresponding solution of $\mathbf{g} = 0$ is small. This is because the three orthogonality properties (b), (c) and (d) enjoyed by the conditional score function $U_c(\boldsymbol{\theta}, \boldsymbol{\phi})$ are shared by $\mathbf{g}$ in (4.1) as well; that is, we have that the following hold:

(b) $\mathbb{E}(\mathbf{g}(\boldsymbol{\theta}, \boldsymbol{\phi}^*); \boldsymbol{\theta}, \boldsymbol{\phi}) = 0$ for all $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ and $\boldsymbol{\phi}^*$;
(c) $\mathbb{E}(\partial \mathbf{g}(\boldsymbol{\theta}, \boldsymbol{\phi}^*) / \partial \boldsymbol{\phi}^*; \boldsymbol{\theta}, \boldsymbol{\phi}) = 0$ for all $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ and $\boldsymbol{\phi}^*$,
(d) $\mathrm{Cov}(\mathbf{g}(\boldsymbol{\theta}, \boldsymbol{\phi}), \partial \log f(y; \boldsymbol{\theta}, \boldsymbol{\phi}) / \partial \boldsymbol{\phi}) = 0$ for all $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$.

The orthogonality property (a) would be satisfied if $\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}$ is a function of $\mathbf{A}_i$ which is not null. To complete the process, we assume the existence of $\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}$, which is a $\sqrt{K}$-consistent estimator of $\boldsymbol{\phi}$, that is, $\sqrt{K}(\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}} - \boldsymbol{\phi}) = O_p(1)$. One may estimate $\boldsymbol{\theta}$ by iterating until convergence between solving $g(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}) = 0$ and updating $\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}$. There are three important implications of this approach. First, the choice among $\sqrt{K}$-consistent estimators of $\boldsymbol{\phi}$ is irrelevant, at least when K is large; that is, the asymptotic variance of $\hat{\theta}$ is the same as when $\boldsymbol{\phi}$ is known. To see this, note that, under regularity conditions,

$$\begin{aligned} \frac{g(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}})}{\sqrt{K}} = {} & \frac{g(\boldsymbol{\theta}, \boldsymbol{\phi})}{\sqrt{K}} \\ & + \frac{\partial g(\boldsymbol{\theta}, \boldsymbol{\phi}) / \partial \boldsymbol{\phi}}{K} \sqrt{K}(\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}} - \boldsymbol{\phi}) \\ & + o_p(\sqrt{K}). \end{aligned}$$

The asymptotic equivalence between $g(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}})/\sqrt{K}$ and $g(\boldsymbol{\theta}, \boldsymbol{\phi})/\sqrt{K}$ is established as, according to property (c), $K^{-1}\ \partial g(\boldsymbol{\theta}, \boldsymbol{\phi}) / \partial \phi$ converges to zero as $K \to \infty$, while $\sqrt{K}(\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}} - \boldsymbol{\phi}) = O_p(1)$.

Second, to pursue this asymptotic analysis further, we note that, for any $j \geq 2$,

$$\frac{\partial^{(j)} \mathbf{g}(\boldsymbol{\theta}, \boldsymbol{\phi}) / \partial \boldsymbol{\phi}^{(j)}}{\partial^{(j)} U_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\phi}) / \partial \boldsymbol{\phi}^{(j)}} = o_p(1),$$

due to property (c*):

(c*) $\mathbb{E}(\partial \mathbf{g}^{(j)}(\boldsymbol{\theta}, \boldsymbol{\phi}) / \partial \boldsymbol{\phi}^{(j)}; \boldsymbol{\theta}, \boldsymbol{\phi}^*) = 0$ for all $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ and $\boldsymbol{\phi}^*$.

Thus, the bias of $\mathbf{g}(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}})$ with $\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}$ plugged into the estimating function is diminished at a faster rate than that of $U_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}})$, the ordinary score function evaluated at $\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}$.

The third point is that this approach enjoys a certain robustness, which we now describe. Assuming that $\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}$ can be derived as the solution of $\sum_i \mathbb{E}((\partial \mathbf{w}_i / \partial \boldsymbol{\phi}) \,|\, \mathbf{A}_i^*) \mathrm{Cov}^{-1}(\mathbf{w}_i \,|\, \mathbf{A}_i^*) \mathbf{w}_i = 0$ for some estimating functions $\mathbf{w}_i(y_i, \boldsymbol{\theta}, \boldsymbol{\phi})$ and conditioning event $\mathbf{A}_i^*$. The estimating procedure described above is formally equivalent to jointly solving

$$\sum_{i=1}^{K}\begin{pmatrix}\mathbb{E}\left(\frac{\partial \mathbf{g}_i}{\partial\boldsymbol{\theta}}\,\middle|\,\mathbf{A}_i\right) & 0\\ 0 & \mathbb{E}\left(\frac{\partial \mathbf{w}_i}{\partial\boldsymbol{\phi}}\,\middle|\,\mathbf{A}_i^*\right)\end{pmatrix} \cdot\begin{pmatrix}\mathrm{Cov}^{-1}(\mathbf{g}_i \mid \mathbf{A}_i) & 0\\ 0 & \mathrm{Cov}^{-1}(\mathbf{w}_i \mid \mathbf{A}_i^*)\end{pmatrix} \cdot\begin{pmatrix}\mathbf{g}_i\\ \mathbf{w}_i\end{pmatrix}=0. \tag{4.2}$$

In words, when solving (4.2), $\mathbf{g}_i$ and $\mathbf{w}_i$ are weighted as if they are independent of each other. Consequently, even if the assumption on how $\boldsymbol{\phi}$ describes the distribution of the $\mathbf{y}$'s is misspecified, the solution remains consistent and its asymptotic variance is unaltered. This is because the $\boldsymbol{\theta}$-component of the estimating equations in (4.2) involves $\mathbf{g}_i$ only; in fact it is identical to (4.1). On the other hand, this form of model robustness is achieved at the expense of losing efficiency if the assumption on $\boldsymbol{\phi}$ is indeed incorrect. While the efficiency loss for $\hat{\boldsymbol{\theta}}$ can only be examined case by case, our experience has been that the gain in robustness by adopting this approach is far greater than the loss in efficiency.

Returning to the quasilikelihood setting with $g_i = y_i - \mu_i(\boldsymbol{\theta})$ and $w_i = (y_i - \mu_i)^2 - V_i(\mu_i; \boldsymbol{\theta}, \boldsymbol{\phi})$, the discussion raised above amounts to the comparison between the quasi-score function method and the quadratic estimating function method in (2.4) as discussed by Crowder (1987) and Firth (1987). The approach described in this subsection, which is the quasi-score method, is robust in that it is consistent regardless of the correct specification of V_i as the variance of y_i, a property not shared by the quadratic method. On the other hand this approach is less efficient compared to the quadratic estimating method, as it does not utilize the information about the mean parameters $\boldsymbol{\theta}$ in the second moments. Note that the generalized estimating equation methods discussed in Example 1.1 of Section 4.1 are multivariate analogues of the quasi-score function and the quadratic estimating function, and hence the trade-off between bias and precision discussed above applies here as well.

Example 3.3 (Continued). The estimating function in (4.2) in this case would be the same as in (3.6) except ϕ_3 is taken as zero. Consequently, the estimator for $\boldsymbol{\theta} = (\theta_1, \theta_2)$ is the conventional least-squares estimator. This estimator remains unbiased even if the constant variance assumption is violated. The variance of $\hat{\theta}_2$ derived from solving (3.6) is

$$V(\phi_1, \phi_3, \phi_4) = \frac{\phi(\phi_4 - \phi^2) - \phi_3^2}{K(\phi_4 - \phi^2)\lambda_1\lambda_0},$$

where λ_j is the proportion of subjects with $x_i = j$, $j = 0, 1$. The efficiency of θ_2 estimation when using (4.2) relative to using (3.6) is then

$$\frac{V(\phi_1, \phi_3, \phi_4)}{V(\phi_1, \phi_3 = 0, \phi_4)} = 1 - \frac{\phi_3^2}{(\phi_4 - \phi^2)\phi}. \tag{4.3}$$

While one can always find pathological cases to make the ratio in (4.3) arbitrarily small, the efficiency of the estimator obtained from (4.2) is relatively high in most realistic situations. This observation reflects our experience from analyzing binary longitudinal data and from finite sample and asymptotic efficiency studies (Liang and Zeger, 1986; Liang, Zeger and Qaqish, 1992).

5. DISCUSSION

We have used this opportunity to review the topic of statistical inference using estimation function in the presence of nuisance parameters. We have argued that estimating functions are one tool for minimizing the influence of nuisance parameters. This can occur in two ways. The most fundamental is that the use of estimating functions allows us to specify only that part of the probability mechanism that is of scientific interest. Hence, we avoid a second part and its associated nuisance parameters. For example, in the logistic model for longitudinal data, we need only specify the mean and covariance of the repeated observations for each individual. Higher-order moments which are not of scientific interest are not modelled explicitly. Second, we can design estimating functions so that their solutions are influenced as little as possible by nuisance parameters. In this spirit, we advocate, in larger samples, the use of the estimating equation $\mathbf{g}$ defined in (4.1) or, equivalently, (4.2) because asymptotic inferences about $\boldsymbol{\theta}$ using $\mathbf{g}$ are not affected by the incorrect specification of the model for $\boldsymbol{\phi}$. An example is the choice between quasilikelihood and quadratic estimating function estimators when regression coefficients are the scientific focus. The increase in efficiency that might be gained through use of the quadratic equations (Firth, 1987) may not in the great majority of problems justify their increased sensitivity to the correct specification of the variance structure. As Tukey (1986) has pointed out, it is increasingly difficult to estimate higher moments from data. Hence assumptions about higher-order moments or about their relationships with lower moments are increasingly difficult to verify.

ACKNOWLEDGMENTS

This work was supported in part by Grant GM49909 from the National Institutes of Health.

REFERENCES

Azzalini, A. (1984). Estimation and hypothesis testing for collections of autoregressive time series. *Biometrika* **71** 85–90.

Basu, D. (1977). On the elimination of nuisance parameters. *J. Amer. Statist. Assoc.* **72** 355–366.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B* **36** 192–236.

Bickel, P. J. and Doksum, K. A. (1977). *Mathematical Statistics*. Holden-Day, San Francisco.

Breslow, N. E. (1981). Odds ratio estimators when the data are sparse. *Biometrika* **68** 73–84.

Cox, D. R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc. Ser. B* **34** 187–200.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62** 269–276.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. Roy. Statist. Soc. Ser. B* **49** 1–39.

Crowder, M. J. (1987). On linear and quadratic estimating functions. *Biometrika* **74** 591–597.

Crowder, M. J. (1989). Comment on "An extension of quasi-likelihood estimation" by V. P. Godambe and M. E. Thompson. *J. Statist. Plann. Inference* **22** 167–168.

Diggle, P., Liang, K.-Y. and Zeger, S. L. (1994). *Analysis of Longitudinal Data*. Oxford Univ. Press.

Durbin, J. (1960). Estimation of parameters in time-series regression models. *Biometrika* **47** 139–153.

Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Statist.* **3** 1189–1242.

Efron, B. (1986). Double exponential families and their use in generalized linear regression. *J. Amer. Statist. Assoc.* **81** 709–721.

Ferguson, H., Reid, N. and Cox, D. R. (1991). Estimating equations from modified profile likelihood. In *Estimating Functions* (V. P. Godambe, ed.) 279–293. Oxford Univ. Press.

Firth, D. (1987). On the efficiency of quasi-likelihood estimation. *Biometrika* **74** 233–245.

Firth, D. (1989). Comment on "An extension of quasi-likelihood estimation" by V. P. Godambe and M. E. Thompson. *J. Statist. Plann. Inference* **22** 168–169.

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

Fitzmaurice, G. M., Laird, N. M. and Rotnitzky, A. G. (1993). Regression models for discrete longitudinal responses (with discussion). *Statist. Sci.* **8** 284–309.

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

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

Godambe, V. P. (1991a). Estimating functions: an overview. In *Estimating Functions* (V. P. Godambe, ed.) 3–20. Oxford Univ. Press.

Godambe, V. P. (1991b). Orthogonality of estimating functions and nusiance parameters. *Biometrika* **78** 143–151.

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. *J. Statist. Plann. Inference* **22** 137–152.

Harville, D. (1977). Maximum likelihood estimation of variance components and related problems. *J. Amer. Statist. Assoc.* **72** 320–340.

Hinkley, D. V. and Runger, G. (1984). The analysis of transformed data (with discussion). *J. Amer. Statist. Assoc.* **79** 302–320.

Kalbfleisch, J. D. and Sprott, D. A. (1970). Application of likelihood methods to models involving a large number of nuisance parameters (with discussion). *J. Roy. Statist. Soc. Ser. B* **32** 175–208.

Kendall, M. G. (1951). Regression, structure and functional relationship, part I. *Biometrika* **38** 11–25.

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

Kupper, L. L., Portier, C., Hogan, M. and Yamamoto, F. (1986). The impact of litter effects on dose-response modeling in teratology. *Biometrics* **42** 85–98.

Legendre, A. M. (1805). *Nouvelles Méthodes pour la Détermination des Orbites des Comètes*. Courcier, Paris.

Liang, K.-Y. (1985). Odds ratio inference with dependent data. *Biometrika* **72** 678–682.

Liang, K.-Y. (1987). Estimating functions and approximate conditional likelihood. *Biometrika* **74** 695–702.

Liang, K.-Y. and Hanfelt, J. (1994). On the use of the quasi-likelihood method in teratological experiments. *Biometrics*. **50** 872–880.

Liang, K.-Y. and Tsou, D. (1992). Empirical Bayes and conditional inference with many nuisance parameters. *Biometrika* **79** 261–270.

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

Liang, K.-Y., Zeger, S. L. and Qaqish, B. (1992). Multivariate regression analyses for categorical data (with discussion). *J. Roy. Statist. Soc. Ser. B* **54** 3–40.

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

Lindsay, B. and Waterman, R. P. (1992). Extending Godambe's method in nuisance parameter problems. Technical report, Pennsylvania State Univ.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* **22** 719–748.

Mantel, N. and Hankey, W. (1975). The odds ratio of a 2 × 2 contingency table. *Amer. Statist.* **29** 143–145.

McCullagh, P. (1983). Quasi-likelihood functions. *Ann. Statist.* **11** 59–67.

McCullagh, P. (1991). 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.) 265–286. Chapman and Hall, London.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, London.

Morton, R. (1981). Efficiency of estimating equations and the use of pivots. *Biometrika* **68** 227–233.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *J. Roy. Statist. Soc. Ser. A* **135** 370–384.

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

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44** 1033–1048.

Richards, F. S. G. (1961). A method of maximum likelihood estimation. *J. Roy. Statist. Soc. Ser. B* **23** 469–476.

STEFANSKI, L. A. and CARROLL, R. J. (1987). Conditional scores and optimal scores for generalized linear measurement error models. *Biometrika* **74** 703–716.

STIGLER, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900.* Harvard Univ. Press.

TUKEY, J. W. (1986). Sunset salvo. *Amer. Statist.* **40** 72–76.

WEDDERBURN, R. W. M. (1974). Quasi-likelihood functions, generalized linear models and the Gaussian method. *Biometrika* **61** 439–447.

WEIL, C. S. (1970). Selection of the valid numbers of sampling units and a consideration of their combination in toxicological studies involving reproduction, teratogenesis or carcinogenesis. *Food and Cosmetic Toxicology* **8** 177–182.

YANAGIMOTO, T. (1989). Combining moment estimates of a parameter common through strata. *J. Statist. Plann. Inference* **25** 187–198.

ZEGER, S. L. and LIANG, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42** 121–130.

ZHAO, L. P. and PRENTICE, R. L. (1990). Correlated binary regression using a generalized quadratic model. *Biometrika* **77** 642–648.

Comment

V. P. Godambe

It is indeed very insightful on the part of the editors to put the two papers, one of Reid and the other of Liang and Zeger, together for discussion. For, at first sight, the two papers have little in common. By and large, the first paper has a parametric setup, the other a semiparametric one. Yet the subject matters of the two papers have deeper links which remain to be explored. On one hand, we have results concerning profile likelihood primarily based on parametric models (cf. Cox and Reid, 1987), and on the other hand, we have results based on semiparametric models utilizing optimal estimating function theory. How to compare these two sets of results? This stimulating question has remained largely uninvestigated. Among some exceptions are included the demonstrations of Cox's partial likelihood (Cox, 1975) as the optimal estimating function for a semiparametric model (Godambe, 1985) and similar optimality of the score function obtained from the Cox–Reid (Cox and Reid, 1987) profile likelihood (Godambe, 1991b). Possibly other discussants will provide other examples. Further related comments are given in my discussion of the paper by Liang and Zeger, to follow.

I liked both the papers. However, due to time constraints I will restrict my additional comments only to one paper (Liang and Zeger). I do hope that the two papers and their discussion would stimulate further research in the problem area (briefly mentioned above) implied by the papers.

Liang and Zeger have a lucid style of presentation. With properly selected examples they first illustrate how the existence of nuisance parameters can affect inference about the parameter of interest. Using the same examples they later demonstrate how the effect of the nuisance parameters can be reduced or eliminated using estimating function theory. All this is accomplished at a common level of understanding. This paper therefore has both scientific and pedagogical value.

The following comments are meant to clarify and emphasize some points in the paper which perhaps have not received enough attention.

In Section 2.4, the authors state that a major limitation of estimating function theory is that it ascribes optimality to the estimating function, while scientists and practitioners are concerned about estimators. They quote Crowder's remark "This is like admiring the pram rather than the baby" (Crowder, 1989), from the discussion of the paper of Godambe and Thompson (1989); these authors' reply to Crowder, not reproduced in the present paper, is given below with some elaboration. I hope this will remove some misunderstanding about an important aspect of the subject.

How good is the estimate? Conventionally the question is answered in terms of the "error" of the estimator. Now the concept of error is somewhat complicated and does not admit a simple definition. Certainly error is not just a root of an arbitrary (unbiased or nearly so) estimate of variance. In parametric inference, however, the practice is fairly clear. For a parametric model, the error is derived from the natural estimate of the variance of the score function. The error is the inverse of the square root of observed Fisher information (Efron

V. P. Godambe is Distinguished Professor Emeritus and Adjunct Professor, Statistics and Actuarial Sciences Department, University of Waterloo, Waterloo N2L 3G1, Ontario, Canada.