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.

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

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

and Hinkley, 1978). Similarly, in the example of the "odds ratio" treated semiparametrically in Section 4.1 of the paper the error is derived from the estimating equation yielding the estimate (Breslow, 1981; Yanagimoto, 1989). This practice is formalized and extended by the theory of estimating functions. Consider the confidence intervals, $\hat{\theta} \pm$ const. (error), where the estimate $\hat{\theta}$ is obtained from the estimating equation $g(\hat{\theta}) = 0$. Here a more direct way of obtaining confidence intervals is by inverting the distribution of the standardized version (cf. Godambe, 1991b, equation 40) of the estimating function g, around $\hat{\theta}$. These intervals, compared to the former ones, are easier to compute. For the distribution of the standardized estimating function is generally more manageable than that of the corresponding estimator, particularly for large samples. An important related result here is that if g^* is the optimal estimating function and $g^*(\hat{\theta}) = 0$, then, for large samples, shortest confidence intervals are obtained by inverting the distribution of g^* around $\hat{\theta}$ (Godambe and Heyde, 1987). Similar optimalities results obtain for testing hypotheses (Mantel and Godambe, 1993). In this sense, the finite sample optimality of an estimating function carries over to large sample optimality of the resulting inference.

Thus, for large samples, the distinction between "estimating function optimality" and "estimator optimality," agreeing with the authors, "may not be an issue as such." There is one exception, however, where the former optimality is relevant while the latter is not. This is when we have Bayesian prior knowledge. I will comment on this topic later at the end.

The following are a few remarks about the transition from parametric to semiparametric models. This, roughly speaking, takes place in Section 4 of the paper. Here some more discussion of the basic concept of conditioning could be helpful. Although a common role played by conditioning in both parametric and semiparametric models is that of removing nuisance or unwanted parameters, nevertheless there is a distinction. In the parametric case, there can possibly be such a thing as a conditionally optimal (perhaps just locally) estimating function. This possibility in semiparametric setup would be far rarer because of the availability of a variety of conditionings based on a variety of partitionings of the sample space. Thus, for instance, it should be emphasized that optimality of the estimating function g in the authors' (4.1) is in general "unconditional." The exceptions here, such as when $A_i = A$ for all i, would imply a very restricted semiparametric or a parametric model. We resort to conditioning, not only to try to avoid unwanted parameters, but also to enhance the unconditional efficiency of the estimating function (Godambe, 1976, 1985; Godambe and Thompson, 1989). Unfortunately, McCullagh and Nelder (1989, Section 9.4), to whom the authors refer in Section 4.1, do not emphasize the above aspect of conditioning.

The unconditional optimality of estimating functions becomes all the more essential when there is a semiparametric Bayesian component to the set of elementary estimating functions. For instance, when θ_0 and v_0 are known values of the prior mean and variance of θ, the unconditionally optimum estimating function is given by $g + (\theta - \theta_0)/v_0$, where g is given by the authors' (4.1) (Godambe and Thompson, 1989; Godambe, 1994). Here no conditional optimality is available. Also in this case, the asymptotic optimality of the estimate mentioned previously is not of much relevance. Actually, in this case, the finite sample (unconditional) optimality relates the optimum estimating function directly to the derivative of the logarithm of the posterior density.