10/01/85

CONDITIONAL SCORES AND OPTIMAL SCORES FOR GENERALIZED LINEAR MEASUREMENT-ERROR MODELS

by

Leonard A. Stefanski
Department of Economic and Social Statistics
Cornell University
Ithaca, New York 14853

and

Raymond J. Carroll
Department of Statistics
University of North Carolina
Chapel Hill, North Carolina 27514

SUMMARY

In this paper we study estimation of the parameters of generalized linear models in canonical form when the explanatory vector is measured with independent normal error. For the functional case, i.e., when the explanatory vectors are fixed constants, unbiased score functions are obtained by conditioning on certain sufficient statistics. This work generalizes results obtained by the authors (Stefanski & Carroll, 1986) for logistic regression. In the case that the explanatory vectors are independent and identically distributed with unknown distribution, efficient score functions are obtained using the theory developed in Begun *et al.* (1983). Related results can be found in Bickel & Ritov (1986).

Some key words: Conditional score function; Efficient score function; Functional model; Generalized linear model; Measurement error; Structural model.

1. INTRODUCTION

Given a covariate p-vector u assume that Y has the density

$$h_Y(y;\theta,u) = \exp\left\{\frac{y(\alpha+\beta^T u)-b(\alpha+\beta^T u)}{a(\phi)} + c(y,\phi)\right\} \qquad (1.1)$$

with respect to a σ-finite measure $m(\cdot)$; in (1.1) $\theta^T = (\alpha,\beta^T,\phi)$ and $a(\cdot),b(\cdot)$ and $c(\cdot,\cdot)$ are known functions. The density (1.1) is that of a generalized linear model in canonical form (McCullagh & Nelder, 1983, Ch. 2). Suppose now that u cannot be observed but that M independent measurements $X=(X_1,\ldots,X_M)$ of u are available. When measurement error is normally distributed the matrix X has density

$$h_X(x;\theta,u) = \prod_{j=1}^{M} \frac{(2\pi)^{-p/2}}{|\bar{\Omega}|^{\frac{1}{2}}} \exp\left\{-\tfrac{1}{2}(x_j - u)^T \bar{\Omega}^{-1}(x_j - u)\right\}. \qquad (1.2)$$

Together (1.1) and (1.2) define a generalized linear measurement-error model with normal measurement error. If for a sample (Y_i,X_i) $(i=1,\ldots,n)$ the covariables (u_i) are unknown constants, a functional model is obtained; if (u_i) are independent and identically distributed random vectors from some unknown distribution, a structural model is obtained (Kendall & Stuart, 1979, Chapter 29). In this paper the problem of deriving unbiased scores for θ in both functional and structural models is studied.

There is a vast literature on this problem in the special case that (1.1) is a normal density. This dates back to Adcock (1878) and has been reviewed by Anderson (1976); see also Moran (1971). Recently there has been considerable interest in nonlinear measurement-error models; see Prentice (1982), Wolter & Fuller (1982a, 1982b), Carroll *et al.* (1984), Stefanski (1985) and Stefanski & Carroll (1986).

The density (1.1) includes normal, Poisson, logistic and gamma regression models. The key feature these models have in common is the existence of a natural sufficient statistic for u when all other parameters are fixed. The same is true of the normal density in (1.2). In fact (1.2) could be replaced with any density possessing a natural sufficient statistic for u when other parameters are fixed and much of the following theory holds with little or no modification. However, in the framework of measurement-error models no other assumption on the error distribution is more palatable than that of normality and thus the added generality is sacrificed for a reduction in notational complexity.

In Section 2 functional models are studied and unbiased score functions for estimating θ in the presence of the unknown u_i's are presented. This work generalizes and extends results of Stefanski & Carroll (1986) for logistic regression. Structural models are studied in Section 3 and efficient score functions for estimating θ in the presence of the unknown distribution for u are identified. These results are obtained using the theory of efficient estimation developed by Begun *et al.* (1983). Other work in this area includes that of Bickel & Ritov (1986).

In the case that the covariates $u_1,\ldots,u_n$ are observed without error the maximum likelihood estimator of θ maximizes

$$\sum_{i=1}^{n} \log h_Y(Y_i;\theta,u_i)$$

with respect to θ. Let $\bar{X}_i$ be the mean of the M measurements of u_i; that value of θ which maximizes

$$\sum_{i=1}^{n} \log h_Y(Y_i;\theta,\bar{X}_i)$$

will be referred to as the naive estimator. This estimator is usually inconsistent (Stefanski, 1985) although when $\bar{\Omega}/M$ is small its bias will be small.

2. FUNCTIONAL MODELS

2.1 *The functional likelihood*

Consider the functional version of model (1.1) & (1.2). In this section the case M = 1 is studied under the additional assumption that

$$\bar{\Omega}/a(\phi) = \Omega \text{ (known)}. \tag{2.1}$$

Throughout this section the random variables (Y_i,X_i), $(i=1,\ldots,n)$ are independent but not identically distributed since their distributions depend on the true regressors u_i, which vary with i. However, for notational convenience the subscript i will be dropped when referring to (Y_i,X_i) in those situations where it causes no confusion. Under (1.1), (1.2) and (2.1) the joint density of (Y,X) takes the form

$$h_{Y,X}(y,x;\theta,u) = h_Y(y;\theta,u)h_X(x;\theta,u). \tag{2.2}$$

For a set of n observations the log-likelihood is

$$L(\theta,u_1,\ldots,u_n) = \sum_{i=1}^{n} \log\{h_{Y,X}(Y_i,X_i;\theta,u_i)\} \tag{2.3}$$

In the case that Y is normally distributed it is known that under (2.1) maximizing (2.3) with respect to $(\alpha,\beta^T,\phi,u_1,\ldots,u_n)$ results in consistent estimators of the regression coefficients α and β (Gleser, 1981). For any model other than the normal, the task of maximizing (2.3) with respect to its n+p+2 parameters is formidable and not likely to be undertaken. More importantly it is not generally true that maximizing (2.3) produces consistent estimators. It follows from results in the first author's University of North Carolina Ph.D. thesis that in the case of logistic regression the functional maximum likelihood estimator of (α,β) is not consistent under assumption (2.1); see also Stefanski & Carroll (1986). The unwieldy functional likelihood, and its failure to produce consistent estimators in some important cases point to the need for an alternative theory of estimation which is now pursued.

2.2 *Unbiased score functions*

In this section unbiased score functions for the functional model are obtained by conditioning on certain sufficient statistics. Note that (2.2) can be written as

$$h_{Y,X}(y,x;\theta,u) = q(\delta,\theta,u)r(y,x,\theta) \tag{2.4}$$

where

$$q(\delta,\theta,u) = \exp\left\{ \frac{u^T\Omega^{-1}\delta}{a(\phi)} - \frac{u^T\Omega^{-1}u + 2b(\alpha+\beta^T u)}{2a(\phi)} \right\}; \tag{2.5}$$

$$r(y,x,\theta) = \exp\left\{ \frac{2\alpha y - x^T\Omega^{-1}x}{2a(\phi)} + C^*(y,\phi) \right\};$$

$$\delta = \delta(y,x,\theta) = x + y\Omega\beta;$$

$$C^*(y,\phi) = c(y,\phi) - (\tfrac{1}{2})\log[\{2\pi a(\phi)\}^P|\Omega|].$$

Thus viewing u as a parameter and α, β and ϕ as fixed, the statistic

$$\Delta = \Delta(Y,X,\theta) = X + Y\Omega\beta \tag{2.6}$$

is sufficient for u. As a consequence, the distribution of $Y|\Delta$ depends only on the observed variables Y and X and θ, but not on u. From this conditional distribution it is possible to derive unbiased estimating equations for θ which are independent of u.

Let $h_{Y|\Delta}(y|\delta;\theta)$ denote the conditional distribution of $Y|\Delta = \delta$. To find $h_{Y|\Delta}$ note that the Jacobian of the transformation which takes (Y,X) into $(Y,X+Y\Omega\beta)$ has a determinant of one. Thus

$$pr(Y=y,\ \Delta=\delta)dm(y)d\delta = pr(Y=y,\ X=\delta-y\Omega\beta)dm(y)d\delta$$

and after some routine calculations one finds

$$h_{Y|\Delta}(y|\delta;\theta) = \frac{J(y,\delta,\theta)}{\int J(y,\delta,\theta)dm(y)} \tag{2.7}$$

where

$$J(y,\delta,\theta) = \exp\left\{\frac{2y(\alpha+\delta^T\beta) - y^2\beta^T\Omega\beta}{2\alpha(\phi)} + c(y,\phi)\right\}. \tag{2.8}$$

Define

$$S(\eta,\beta,\phi) = \int \exp\left\{y\eta - \frac{y^2\beta^T\Omega\beta}{2a(\phi)} + c(y,\phi)\right\}dm(y).$$

This allows (2.7) to be written as

$$h_{Y|\Delta}(y|\delta;\theta) = \exp\left[y\eta - \frac{y^2\beta^T\Omega\beta}{2\alpha(\phi)} + c(y,\phi) - \log\{S(\eta,\beta,\phi)\}\right] \tag{2.9}$$

when $\eta = (\alpha + \delta^T\beta)/a(\phi)$.

Note that (2.9) is an exponential-family density with Y as the natural sufficient statistic for η. Thus moments of $Y|\Delta=\delta$ can be computed from the partial derivatives of $S(\eta,\beta,\phi)$ with respect to η, e.g.,

$$\begin{aligned} E_\theta(Y|\Delta=\delta) &= (\partial/\partial\eta)\log\{S(\eta,\beta,\phi)\}|_{\eta=(\alpha+\beta^T\delta)/a(\phi)}; \\ \mathrm{var}_\theta(Y|\Delta=\delta) &= (\partial^2/\partial\eta^2)\log\{S(\eta,\beta,\phi)\}|_{\eta=(\alpha+\beta^T\delta)/a(\phi)}. \end{aligned} \tag{2.10}$$

Since $h_{Y|\Delta}$ is an exponential-family density it is true that

$$\int \dot{h}_{Y|\Delta}(y|\delta;\theta)dm(y) = 0, \tag{2.11}$$

where

$$\dot{h}_{Y|\Delta}(y|\delta;\theta) = (\partial/\partial\theta)h_{Y|\Delta}(y|\delta;\theta).$$

Thus defining

$$\psi_S(y,x,\theta) = \left.\frac{\dot{h}_{Y|\Delta}(y|\delta;\theta)}{h_{Y|\Delta}(y|\delta;\theta)}\right|_{\delta=x+y\Omega\beta}, \tag{2.12}$$

it follows that $\psi_S(\cdot,\cdot,\cdot)$ is unbiased for θ, i.e.,

$$E_\theta\{\psi_S(Y,X,\theta)\} = E_\theta[E_\theta\{\psi_S(Y,X,\theta)|\Delta\}] = 0.$$

The inner conditional expectation is zero by virtue of (2.11).

The score ψ_S will be called the *sufficiency score* and any estimator $\hat{\theta}_S = (\hat{\alpha}_S, \hat{\beta}_S^T, \hat{\phi}_S)^T$ which satisfies

$$\sum_{i=1}^{n} \psi_S(Y_i, X_i, \hat{\theta}_S) = 0 \qquad (2.13)$$

will be called a *sufficiency estimator*.

Consider the density in (2.4) and let $\dot{h}_{Y,X} = (\partial/\partial\theta)h_{Y,X}$. Note that

$$\frac{\dot{h}_{Y,X}(y,x;\theta,u)}{h_{Y,X}(y,x;\theta,u)} - E\left\{\frac{\dot{h}_{Y,X}(Y,X;\theta,u)}{h_{Y,X}(Y,X;\theta,u)}\middle|\Delta = \delta\right\}$$

$$= \begin{bmatrix} \{y - E(Y|\Delta=\delta)\}/a(\phi) \\ \{y - E(Y|\Delta=\delta)\}u/a(\phi) \\ r_\phi(y,x,\theta) - E\{r_\phi(Y,X,\theta)|\Delta=\delta\} \end{bmatrix}$$

where

$$r_\phi(y,x,\theta) = \frac{\partial C^*(y,x,\theta)}{\partial\phi} - \left\{\frac{2\alpha y - x^T\Omega^{-1}x}{2\alpha^2(\phi)}\right\} a'(\phi).$$

As the expression in brackets above depends on the unknown covariate u only as a 'weight' this suggests the class of score functions

$$\psi_C(y,x,\theta) = \begin{bmatrix} \{y-E(Y|\Delta=\delta)\}/a(\phi) \\ \{y-E(y|\Delta=\delta)\}\Omega t(\delta) \\ r_\phi(y,x,\theta)-E\{r_\phi(Y,X,\theta)|\Delta=\delta\} \end{bmatrix}_{\delta=x+y\Omega\beta} \qquad (2.14)$$

indexed by the vector-valued function $t(\cdot)$. The score (2.14) will be called a *conditional score* following Lindsay (1980, 1982, 1983). Some natural choices for $t(\delta)$ might be $t(\delta)=\delta$ and $t(\delta)=E_\theta(X|\Delta=\delta)$. Note that since X is unbiased for u and Δ is sufficient for u the latter choice corresponds to replacing u by its uniformly minimum variance unbiased estimator. Also since

$$E_\theta(X|\Delta=\delta) = \delta - E_\theta(Y|\Delta=\delta)\Omega\beta \qquad (2.15)$$

only the conditional moments of $Y|\Delta$ are needed to find $E_\theta(X|\Delta=\delta)$. More will be said on appropriate choices for $t(\cdot)$ in Section 3.3.

Any estimator $\hat{\theta}_C$ which satisfies

$$\sum_{i=1}^{n} \psi_C(Y_i, X_i, \hat{\theta}_C) = 0 \qquad (2.16)$$

will be called a *conditional estimator*.

The estimating equations in (2.12) and (2.14) are both unbiased. Although it should be possible to show that under reasonable conditions there always exist consistent sequences of estimators $\hat{\theta}_S$ and $\hat{\theta}_C$ satisfying (2.13) and (2.16) respectively, it is not generally true that (2.13) and (2.16) define $\hat{\theta}_S$ and $\hat{\theta}_C$ uniquely. More importantly, there can exist sequences of solutions to (2.13) and (2.16) which are not consistent, and thus care must be taken when defining $\hat{\theta}_S$ and $\hat{\theta}_C$. In practice a couple of solutions to this dilemma are possible. The first consists of defining the estimators $\hat{\theta}_S$ and $\hat{\theta}_C$ as the solutions to (2.13) and (2.16) which are closest to the naive estimator introduced in Section 1. This rule is justifiable when measurement error is small, however it can break down when measurement error is large. This is discussed in greater detail for the normal model in the next section. The second solution entails doing one or two steps of a Newton-Raphson iteration of (2.13) and (2.16) starting from the naive estimator. Again this is generally appropriate only when the measurement is small. However, in some realistic sampling situations, Stefanski & Carroll (1986) show that such an approach substantially improves upon the naive estimator in their study of measurement error in logistic regression. Finally, preliminary work by the authors suggests that it is possible to deconvolute the empirical distribution function of the observed X_i's to obtain an estimator of the empirical distribution function of the u_i's, which under regularity conditions can be used to construct consistent estimators for the functional model. These estimators can then be used to uniquely define the more manageable

M-estimators, $\hat{\theta}_S$ and $\hat{\theta}_C$.

When consistent sequences of solutions to (2.13) and (2.16) are obtained the asymptotic distributions of $\hat{\theta}_S$ and $\hat{\theta}_C$ are easily derived since both are M-estimators; see Huber (1967).

2.3 *Normal, logistic and Poisson regression*

In this section the strengths and limitations of the estimation theory are illustrated by studying it in three particular generalized linear models.

Consider first the case in which Y has a normal distribution with mean $\alpha + \beta^T u$ and variance σ^2. For this model $\phi = \sigma^2$, $a(\phi) = \phi$ and $m(\cdot)$ is Lebesque measure. Using (2.7) one finds that the distribution of $Y|\Delta = \delta$ is normal with variance $\sigma^2/(1 + \beta^T\Omega\beta)$ and mean μ where

$$\mu = \frac{\alpha + \beta^T\delta}{1 + \beta^T\Omega\beta} . \tag{2.17}$$

Corresponding to (2.12) one finds

$$\psi_S(y,x,\theta) = \left[\begin{array}{c} + \frac{1}{\sigma^2}(y-\mu) \\ \frac{\Omega\beta}{1+\beta^T\Omega\beta} - \frac{1}{\sigma^2}\left\{(y-\mu)^2\Omega\beta-(y-\mu)(\delta-2\mu\Omega\beta)\right\} \\ \frac{-1}{2\sigma^2} + \frac{(y-\mu)^2(1+\beta^T\Omega\beta)}{2\sigma^4} \end{array} \right]_{\delta=x+y\Omega\beta}$$

where μ is defined in (2.17). Define

$$\Delta_i^* = (I+\Omega\beta\beta^T)^{-1}\{\Delta(Y_i,X_i,\theta)-\alpha\Omega\beta\}$$

where $\Delta(\cdot,\cdot,\cdot)$ is given by (2.6) and consider the equations

$$\sum_{i=1}^{n}(Y_i-\alpha-\beta^T\Delta_i^*)\begin{pmatrix}1\\ \Delta_i^*\end{pmatrix} = 0 \tag{2.18}$$

$$\sigma^2 = \frac{1 + \beta^T\Omega\beta}{n} \sum_{i=1}^{n}(Y_i-\mu)^2 .$$

It is a simple matter to show that every solution to (2.18) is also a solution to $\sum\psi_S(Y_i,X_i,\theta) = 0$, i.e., any solution to (2.18) is a sufficiency estimator. The similarity of (2.18) to the usual normal equations is readily apparent. However, keep in mind that Δ_i^* depends on α and β and thus (2.18) is nonlinear in the parameters.

Note that Δ_i^* is also a sufficient statistic for u_i when α, β, and σ^2 are fixed. Because of (2.18) and the fact that given Δ_i^*, Y_i is normal with mean $\alpha + \beta^T\Delta_i^*$, Δ_i^* will be called a *conjugate sufficient statistic*. Also since Δ_i^* is the functional maximum likelihood estimator for u_i in this model (Gleser, 1981), equation (2.18) shows that the functional maximum likelihood estimator is a sufficiency estimator.

From (2.18) it follows that $\hat{\alpha}_S = \bar{Y} - \hat{\beta}_S^T\bar{X}$ and using this it is possible to deduce that $\hat{\beta}_S$ satisfies

$$-\hat{\beta}_S^T\left(\sum_{i=1}^n Y_{i*}X_{i*}\right)\Omega\hat{\beta}_S + \sum_{i=1}^n\left(Y_{i*}^2\Omega - X_{i*}X_{i*}^T\right)\hat{\beta}_S + \sum_{i=1}^n Y_{i*}X_{i*} = 0 \qquad (2.19)$$

where $Y_{i*} = Y_i - \bar{Y}$, $X_{i*} = X_i - \bar{X}$.

Consider (2.19) for the case $p = 1$, i.e., $\hat{\beta}_S$ is a scalar. This quadratic equation has two real roots (Kendall & Stuart, 1979, Chapter 29); unfortunately the sufficiency principle does not indicate which root is appropriate. Had the equations (2.18) been derived as the gradient of the functional log-likelihood the appropriate root would have been dictated by the maximizing principle.

In the previous section it was suggested that in the case of multiple solutions to (2.13) and (2.16) to pick that solution closest to the naive estimator and that this selection rule would work as long as the measurement error variance was small. In this particular case the two roots of

(2.19) converge to β_o and $-\sigma^2/(\beta_o\tau^2)$, where $\tau^2 = \Omega\sigma^2$ is the measurement error variance. The naive estimator converges to $\sigma_u^2\beta_o/(\sigma_u^2+\tau^2)$ where σ_u^2 is the limiting value of the sample variance of the true u_i's. Thus the suggested selection rule will asymptotically choose the right root whenever

$$|\beta_o|\left|1 - \frac{\sigma_u^2}{\sigma_u^2 + \tau^2}\right| < \left|\frac{\sigma_u^2}{\sigma_u^2 + \tau^2}\beta_o + \frac{\sigma^2}{\tau^2\beta_o}\right| .$$

This inequality is satisfied if

$$2\tau^2 < \sigma_u^2 + (\sigma^2/\beta_o^2) + \left\{\left(\sigma_u^2 + \frac{\sigma^2}{\beta_o^2}\right)^2 + \frac{4\sigma^2\sigma_u^2}{\beta_o^2}\right\}^{\frac{1}{2}} .$$

The infimum of the right hand side above with respect to the ratio σ^2/β_o^2 is $2\sigma_u^2$. Thus whenever $\tau^2 < \sigma_u^2$ the selection rule works no matter what the values of σ^2 and β_o^2; however, if $\tau^2 > \sigma_u^2$ and σ^2/β_o^2 is sufficiently small then the selection rule chooses the wrong root. This is encouraging for it is unusual to have measurement error so large that $\tau^2 \geq \sigma_u^2$.

To gain some additional insight into the performance of the sufficiency estimator suppose that $u_1,\ldots,u_n$ are independent normal variates with mean μ_u and variance σ_u^2, i.e. assume a structural model. In this case, (Kendall & Stuart, 1979, Chapter 29) the structural and functional maximum likelihood estimators are the same, and in light of the previous discussion this common estimator is also a sufficiency estimator. Thus in this particular case the sufficiency score is an efficient score.

Finally for the normal model $E_\theta(Y_i|\Delta_i) = \alpha + \beta^T\Delta_i^*$ and from (2.15)

$$E_\theta(X_i|\Delta_i) = \Delta_i^* .$$

Thus ψ_S and ψ_C define the same estimators, i.e., $\hat{\theta}_S = \hat{\theta}_C$, when $t(\delta)=E_\theta(X|\Delta=\delta)$.

Now consider logistic regression in which

$$pr_\theta(Y=1|u) = F(\alpha + \beta^T u), \qquad F(t) = (1 + e^{-t})^{-1}.$$

For this model $a(\phi) \equiv 1$ and $m(\cdot)$ is counting measure on $\{0,1\}$.

Using (2.7) one obtains

$$pr_\theta(Y=1|\Delta=\delta) = F\{\alpha+(\delta-\tfrac{1}{2}\Omega\beta)^T\beta\} \tag{2.20}$$

and corresponding to (2.12) is the logistic sufficiency score

$$\psi_S(y,x,\theta) = \left[y - F\{\alpha+(\delta-\tfrac{1}{2}\Omega\beta)^T\beta\}\right]\begin{pmatrix}1\\ \delta-\Omega\beta\end{pmatrix}\Bigg|_{\delta=x+y\Omega\beta}; \tag{2.21}$$

and setting $\sum\psi_S(Y_i,X_i,\theta) = 0$ results in the equivalent equations

$$\sum_{i=1}^{n}\{Y_i - F(\alpha + \beta^T\Delta_i^*)\}\begin{pmatrix}1\\ \Delta_i^*\end{pmatrix} = 0. \tag{2.22}$$

where $\Delta_i^* = \Delta_i - \tfrac{1}{2}\Omega\beta$; note that Δ_i^* is a conjugate sufficient statistic. Stefanski & Carroll (1986) introduced this estimator and show in a Monte Carlo study that in spite of the possibility of multiple solutions to (2.22), a modified one-step version of $(\hat\alpha_S,\hat\beta_S^T)^T$ starting from the naive estimator, performed well in some realistic sampling situations. Unlike the normal model the logistic sufficiency estimator does not correspond to the functional maximum likelihood estimator, which in this case is not consistent; see the first author's University of North Carolina Ph.D. thesis and Stefanski & Carroll. (1986). In Section 3.3 it is shown that the logistic sufficiency score is optimal for a particular structural model.

For logistic regression it is not true that $\hat\theta_S = \hat\theta_C$, when $t(\delta)=E(X|\Delta=\delta)$. Indeed with $E_\theta(Y|\Delta=\delta)$ given by (2.20),

$$E(X|\Delta=\delta) = \delta - F\{\alpha+(\delta-\tfrac{1}{2}\Omega\beta)^T\beta\}\Omega\beta$$

and corresponding to (2.16) are the equations

$$\sum_{i=1}^{n}\{Y_i - F(\alpha + \beta^T\Delta_i^*)\}\begin{Bmatrix}1\\ \Delta_i^* + \left\{\frac{1}{2} - F(\alpha + \beta^T\Delta^*)\right\}\Omega\beta\end{Bmatrix} = 0.$$

The final model considered is that of Poisson regression in which

$$pr_\theta(Y=k|u) = (k!)^{-1}\exp\{k(\alpha+\beta^T u) - \exp(\alpha+\beta^T u)\}.$$

For this model $a(\phi) \equiv 1$ and $m(\cdot)$ is counting measure on $\{0,1,\ldots\}$.

From (2.7) it follows that

$$pr_\theta(Y=k|\Delta=\delta) = \frac{(k!)^{-1}\exp\{k(\alpha+\beta^T\delta) - k^2\beta^T\Omega\beta/2\}}{\sum_{j=0}^{\infty}(j!)^{-1}\exp\{j(\alpha+\beta^T\delta) - j^2\beta^T\Omega\beta/2\}} . \qquad (2.23)$$

Since (2.23) has no closed form the sufficiency scores ψ_S and ψ_C are quite messy and are not given. Note that there is no conjugate sufficient statistic. Also, as in logistic regression, the estimators $\hat{\theta}_S$ and $\hat{\theta}_C$ are not equal for this model when $t(\delta)=E(X|\Delta=\delta)$.

The conditional distribution (2.23) is more typical of generalized linear models than are those from the logistic and normal models. Since in (2.8) the factor $y^2\beta^T\Omega\beta/2a(\phi)$ appears in the exponent it is only in special cases that the denominator of (2.7) can be obtained in closed form. Thus implementation of the sufficiency estimators will often require numerical integration or summation.

3. STRUCTURAL MODELS

3.1 *The structural likelihood*

In this section the model studied is the structural version of (1.1) & (1.2), , i.e., $u_1,\ldots,u_n$ are independent and identically distributed observations with unknown density $g_U(u)$. Since it should cause no confusion the subscript U on $g_U(u)$ is omitted. The density g is an element of *G*, a family of densities with respect to the measure $\nu(\cdot)$. As in Section 2, it is assumed that M=1 along with the identifiability condition (2.1). Under these conditions the joint density of (Y,X) is

$$f_{Y,X}(y,x;\theta,g) = \int h_{Y,X}(y,x;\theta,u)g(u)d\nu(u) \qquad (3.1)$$

where $h_{Y,X}$ is defined in (2.2).

Let $\dot{f}_{Y,X}(y,x;\theta,g) = (\partial/\partial\theta)f_{Y,X}(y,x;\theta,g)$ and assume that

$$\dot{f}_{Y,X}(y,x;\theta,g) = \int \dot{h}_{Y,X}(y,x;\theta,u)g(u)d\nu(u)$$

where,

$$\dot{h}_{Y,X}(y,x;\theta,u) = (\partial/\partial\theta)h_{Y,X}(y,x;\theta,u)$$

i.e., assume that differentiation and integration can be interchanged in (3.1). If $g(\cdot)$ were known then the efficient score for θ would be

$$\bar{\dot{\ell}}(y,x,\theta,g) = \frac{\dot{f}_{Y,X}(y,x;\theta,g)}{f_{Y,X}(y,x;\theta,g)} ,$$

and the information available in (Y,X) for estimating θ would be

$$I = E(\bar{\dot{\ell}}\bar{\dot{\ell}}^T).$$

Throughout this section interest lies in estimating θ when g and hence $\bar{\dot{\ell}}$ are unknown. Note that both the sufficiency and conditional scores of Section 2 are unbiased for the structural model (3.1) also. Attention therefore is directed to the problem of finding efficient score functions.

3.2 *Efficient score functions and information bounds.*

Efficient score functions for estimation of $\theta = (\alpha,\beta^T,\phi)^T$ in the presence of the nuisance function $g(\cdot)$ are now derived. As with the theory in Section 2 the existence of certain sufficient statistics plays a key role here. The derivation draws heavily on the results of Begun, Hall, Huang & Wellner (1983); see also Pfanzagl (1982, Chapter 14). The structural model studied here is a generalization of a model considered by Bickel & Ritov (1986). Whereas they study simple linear regression under a number of conditions, including that of replicated measurements and our assumption (2.1), we consider the more general model only under the latter

assumption. However, the approach used here to derive efficient scores extends quite naturally to the case of replicated measurements when (2.1) is not assumed.

In the following μ denotes the product measure of $m(\cdot)\times$ Lebesque measure on p-dimensional Euclidean space. As in Begun *et al.* (1983) let $L^2(\mu)$ and $L^2(\nu)$ denote the L^2-spaces of square-integrable functions with respect to the measures μ and ν respectively. Norms and inner products on these spaces are denoted by $\|\cdot\|_\mu$ and $\langle\cdot,\cdot\rangle_\mu$, and $\|\cdot\|_\nu$ and $\langle\cdot,\cdot\rangle_\nu$.

The theory of Begun *et al.* (1983) requires *Hellinger-differentiability* of the square root of (3.1) with respect to (θ,g); see their Definition 2.1. It is assumed here, and can be proven under regularity conditions, that $f_{Y,X}^{\frac{1}{2}}(y,x;\theta,g)$ satisfies condition (2.1) of Begun *et al.* and hence is Hellinger-differentiable. Its *differential*, for sequences (θ_n,g_n) satisfying $\|\theta_n-\theta\| + \|g_n^{\frac{1}{2}}-g^{\frac{1}{2}}\|_\nu$ converging to zero, is given by

$$\rho^T(\theta_n-\theta) + A(g_n^{\frac{1}{2}}-g^{\frac{1}{2}})$$

where

$$\rho = (\tfrac{1}{2})f_{Y,X}^{\frac{1}{2}}(y,x;\theta,g)\dot{\ell}(y,x,\theta,g),$$

and the linear operator A taking $L^2(\nu)$ into $L^2(\mu)$ is defined for Γ in $L^2(\nu)$ via

$$A\Gamma = \frac{\int h_{Y,X}(y,x;\theta,u)\Gamma(u)d\nu(u)}{2f_{Y,X}^{\frac{1}{2}}(y,x;\theta,g)}.$$

When necessary to indicate dependence on (y,x,θ,g), ρ is written $\rho(y,x,\theta,g)$ and $A\Gamma$ as $A\Gamma(y,x,\theta,g)$.

The key result of Begun *et al.* (1983) used here is that when g is unknown the efficient score for θ is

$$\dot{\ell}(y,x,\theta,g) = \frac{2(\rho_\theta - A\Gamma^*)}{f_{Y,X}^{\frac{1}{2}}(y,x;\theta,g)} \qquad (3.2)$$

where the $L^2(\nu)$ function Γ^* satisfies

$$\langle \rho_\theta - A\Gamma^*, A\Gamma \rangle_\mu = 0 \qquad (3.3)$$

for all functions Γ obtained as $L^2(\nu)$ limits of sequences (Γ_n) of the form

$$\Gamma_n = n^{\frac{1}{2}}(g_n - g).$$

In terms of expectations (3.3) becomes

$$(\tfrac{1}{2})E_{\theta,g}\left[\left\{\dot{\ell}(Y,X,\theta,g) - \frac{2A\Gamma^*(Y,X,\theta,g)}{f^{\frac{1}{2}}_{Y,X}(Y,X;\theta,g)}\right\}\left\{\frac{2A\Gamma(Y,X,\theta,g)}{f^{\frac{1}{2}}_{Y,X}(Y,X;\theta,g)}\right\}\right] = 0. \qquad (3.4)$$

Note that for any $\Gamma \in L^2(\nu)$

$$\frac{2A\Gamma(Y,X,\theta,g)}{f^{\frac{1}{2}}_{Y,X}(Y,X;\theta,g)} = \frac{\int h_{Y,X}(Y,X;\theta,u)\Gamma(u)d\nu(u)}{\int h_{Y,X}(Y,X;\theta,u)g(u)d\nu(u)}$$

and thus in view of (2.4)

$$\frac{2A\Gamma(Y,X,\theta,g)}{f^{\frac{1}{2}}_{Y,X}(Y,X,\theta,g)} = \frac{\int q(\Delta,\theta,u)\Gamma(u)d\nu(u)}{\int q(\Delta,\theta,u)g(u)d\nu(u)} \qquad (3.5)$$

where $\Delta = X + Y\Omega\beta$ is defined in (2.6). The important fact here is that the right hand side of (3.5) depends on (Y,X) only through the complete sufficient statistic Δ irrespective of Γ; this is a consequence of the sufficiency of Δ for u, when u is regarded as a parameter. It follows now that (3.3) holds for all Γ when

$$\frac{2A\Gamma^*(Y,X,\theta,g)}{f^{\frac{1}{2}}_{Y,X}(Y,X;\theta,g)} = E_{\theta,g}\{\dot{\ell}(Y,X,\theta,g)|\Delta\}\ .$$

Thus the efficient score given by (3.2) is

$$\ell(y,x,\theta,g) = \dot{\ell}(y,x,\theta,g) - E_{\theta,g}\{\dot{\ell}(Y,X,\theta,g)|\Delta=\delta\}\Big|_{\delta=x+y\Omega\beta},$$

and the "information" available in (Y,X) for estimating θ in the presence of g is

$$I_* = E\{\ell(Y,X,\theta,g)\ell^T(Y,X,\theta,g)\}; \qquad (3.6)$$

see Equation (3.4) of Begun *et al.* (1983).

To compute $\ell(y,x,\theta,g)$ let $q^*(\theta,u) = q\{\delta(y,x,\theta),\theta,u\}$ where $q(\delta,\theta,u)$ and $\delta(y,x,\theta)$ are given in (2.5). Then using (2.4)

$$f_{Y,X}(y,x;\theta,g) = \int q^*(\theta,u)r(y,x,\theta)g(u)d\nu(u),$$

and thus

$$\bar{\ell}(y,x,\theta,g) = \frac{\int\left(\frac{\partial q^*}{\partial\theta}r + q\frac{\partial r}{\partial\theta}\right)gd\nu}{\int q^* r g d\nu}$$

$$= \frac{\int\frac{\partial q^*}{\partial\theta}gd\nu}{\int q^* g d\nu} + \frac{\frac{\partial r}{\partial\theta}}{r} \quad .$$

Now since

$$\frac{\partial q^*}{\partial\theta} = \begin{bmatrix} \frac{-b'(\alpha+u^T\beta)}{a(\phi)} \\ \frac{\{y - b'(\alpha+u^T\beta)\}u}{a(\phi)} \\ \left\{\frac{u^T\Omega^{-1}u + 2b(\alpha+u^T\beta) - 2u^T\Omega^{-1}\delta(y,x,\theta)}{2a^2(\phi)}\right\}a'(\phi) \end{bmatrix} q^*(\theta,u)$$

and

$$\frac{\partial r}{\partial\theta} = \begin{bmatrix} \frac{y}{a(\phi)} \\ 0 \\ r_\phi(y,x,\theta) \end{bmatrix} r(y,x,\theta)$$

where

$$r_\phi(y,x,\theta) = \frac{\partial C^*(y,\phi)}{\partial\phi} - \left\{\frac{2\alpha y - x^T\Omega^{-1}x}{2a^2(\phi)}\right\}a'(\phi),$$

$\bar{\ell}(y,x,\theta,g)$ can be written as

$$\bar{\ell}(y,x,\theta,g) = a^{-1}(\phi)\begin{bmatrix} y - f_1(\delta,\theta,g) \\ yR(\delta,\theta,g) - f_2(\delta,\theta,g) \\ a(\phi)r_\phi(y,x,\theta) - f_3(\delta,\theta,g) \end{bmatrix}_{\delta=x+y\Omega\beta}$$

where the scalar-valued functions f_1 and f_3, and the p-vector-valued functions R and f_2 depend on (y,x) only through $\delta = x+y\Omega\beta$. It follows that

$$\ell(y,x,\theta,g) = \bar{\ell}(y,x,\theta,g) - E\{\bar{\ell}(Y,X,\theta,g)|\Delta=\delta\}$$

$$= \begin{bmatrix} \{y - E(Y|\Delta=\delta)\}/a(\phi) \\ \{y - E(Y|\Delta=\delta)\}R(\delta,\theta,g) \\ r_\phi(y,x,\theta) - E\{r_\phi(Y,X,\theta)|\Delta=\delta\} \end{bmatrix}_{\delta=x+y\Omega\beta}. \qquad (3.7)$$

Define

$$w(\gamma) = \int q(\gamma,\theta,u)g(u)d\nu(u); \qquad (3.8)$$

and

$$\dot{w}(\gamma) = (\partial/\partial\gamma)w(\gamma) = \frac{1}{a(\phi)} \int \Omega^{-1}uq(\gamma,\theta,u)g(u)d\nu(u);$$

Now the function $R(\delta,\theta,g)$ appearing in (3.7) is given by

$$R(\delta,\theta,g) = \frac{\Omega\dot{w}(\delta)}{w(\delta)} . \qquad (3.9)$$

Using the relation $x = \delta - y\Omega\beta$

$$r_\phi(y,x,\theta) = \frac{\partial C^*(y,\phi)}{\partial\phi} - \left\{ \frac{2\alpha y - (\delta-y\Omega\beta)^T\Omega^{-1}(\delta-y\Omega\beta)}{2a^2(\phi)} \right\} a'(\phi)$$

and thus (3.7) involves only expectations of functions of $Y|\Delta=\delta$.

3.3. *Efficiency of the sufficiency and conditional scores in a structural setting*

In the discussion of the normal linear functional model in Section 2.3, it was deduced that the sufficiency score is equivalent to the efficient score for the structural version of this model when the true predictors $(u_1,\ldots,u_n)$ are themselves normally distributed. A similar result for logistic regression is now derived. Compare (2.21) to the logistic efficient score given by

$$\ell(y,x,\theta,g) = [y-F\{\alpha+(\delta-\tfrac{1}{2}\Omega\beta)^T\beta\}]\left\{\frac{1}{R(\delta,\theta,g)}\right\}_{\delta=x+y\Omega\beta}; \qquad (3.10)$$

equation (3.10) is just (3.7) for the case of logistic regression. For (2.21) and (3.10) to be equivalent the function $R(\delta,\theta,g)$ must be linear in δ. Since by (3.9)

$$R(\delta,\theta,g) = \frac{\Omega\dot{w}(\delta)}{w(\delta)},$$

this means that $\log\{w(\delta)\}$ must be a quadratic form in δ, call it $Q(\delta)$, i.e., using (3.8) with $q(\delta,\theta,u)$ chosen accordingly for logistic regression,

$$\exp\{Q(\delta)\} = \int \exp\{u^T\Omega^{-1}\delta - \frac{u^T\Omega^{-1}u}{2}\}\frac{g(u)}{1+\exp(\alpha+\beta^T u)}\,du.$$

Now using a moment-generating-characteristic-function argument it follows that

$$\exp\left\{-\frac{u^T\Omega^{-1}u}{2}\right\}\frac{g(u)}{1+\exp(\alpha+\beta^T u)}$$

must be proportional to a p-variate normal distribution. This means that $g(u)$ must be a mixture of two p-variate normal distributions with different means and common covariance matrix. The picture is now clear; the sufficiency score (2.21) is efficient in a structural setting only when (Y,U) satisfy the assumptions of the normal discrimination model,

$$\text{pr}(Y=1)=\pi_1, \quad U|Y=y \sim N(\mu_y,\Psi).$$

Of course if all of this information were known a priori then the linear discriminant, $\alpha+\beta^T u$, would most likely be estimated using the full likelihood as opposed to using logistic regression, see Efron (1975) and Michalik & Tripathi (1980).

A theorem is now proved which indicates when the conditional score ψ_C defined in (2.14) is the efficient score in some structural setting. This provides some insight into appropriate choices for $t(\cdot)$ when choosing a conditional score (2.14).

THEOREM. *The conditional score* ψ_C *is the efficient score in a structural model for some density* $g(\cdot)$ *and some measure* $\nu(\cdot)$ *if and only if there exists a real-valued function* $T(\cdot)$ *such that* $t(\delta)=(\partial/\partial\delta)T(\delta)$ *where* $\exp[T\{a(\phi)\Omega\delta\}]$ *is a moment generating function for some probability density with respect to* $\nu(\cdot)$.

PROOF. Assume that ψ_C is the efficient score in a structural model with density $g(\cdot)$ and measure $\nu(\cdot)$. Then comparing (2.14) and (3.7) it follows that

$$t(\delta) = \frac{\dot{w}(\delta)}{w(\delta)}$$

where $w(\delta)$ is given by (3.8). Let $T(\delta) = \log\{w(\delta)\}-k$ for a constant k to be determined later. Clearly $(\partial/\partial\delta)T(\delta)=t(\delta)$ and furthermore, using (3.8),

$$\exp\{T(\delta)\} = \int \exp\left\{\frac{u^T\Omega^{-1}\delta}{a(\phi)}\right\} \exp\left\{k - \frac{u^T\Omega^{-1}u + 2b(\alpha+\beta^T u)}{2a(\phi)}\right\} g(u)d\nu(u).$$

Thus with k chosen accordingly $M(\delta)=\exp[T\{a(\phi)\Omega\delta\}]$ is a moment generating function of the density

$$\exp\left\{k - \frac{u^T\Omega^{-1}u + 2b(\alpha+\beta^T u)}{2a(\phi)}\right\} g(u)$$

with respect to $\nu(\cdot)$.

The steps in this argument can be reversed to prove the theorem in the other direction. ////

(2.14), which is known to be efficient for some structural model. The theorem indicates that the class of appropriate functions $t(\cdot)$ is fairly restricted.

3.4. *Efficient estimation.*

Since the efficient score in (3.7) depends on the unknown density $g(\cdot)$, it is not readily apparent how one constructs a sequence of estimators with asymptotically minimum variance. Begun *et al.* (1983) suggest in general solving

$$\sum_{i=1}^{n} \dot{\ell}(Y_i, X_i, \hat{\theta}, \hat{g}) = 0 \qquad (3.11)$$

where $\hat{g}(\cdot)$ is some suitable initial estimator of $g(\cdot)$. Since the empirical distribution function, $\hat{F}_n$, of the observed X's converges to the convolution of G with a normal distribution function it should in theory be able to deconvolute $\hat{F}_n$ to obtain consistent estimators of G which would then be smoothed to obtain estimators of $g(\cdot)$. In practice this is quite difficult and technical problems might arise when p>1. Also given a $\hat{g}(\cdot)$ it is still possible that (3.11) will have multiple solutions, not all yielding consistent sequences.

This last problem can be avoided if a root-n consistent preliminary estimator, $\tilde{\theta}$, is available. Again let $\hat{g}(\cdot)$ be an estimator of $g(\cdot)$ and define

$$\hat{\theta} = \tilde{\theta} + \hat{I}_*^{-1} n^{-\frac{1}{2}} \sum_{i=1}^{n} \dot{\ell}(Y_i, X_i, \tilde{\theta}, \hat{g})$$

where $\hat{I}_*$ is an estimator of I_*, e.g.,

$$\hat{I}_* = - n^{-1} \sum_{i=1}^{n} \dot{\ell}(Y_i, X_i, \tilde{\theta}, \hat{g})$$

and

$$\dot{\ell}(y,x,\theta,g) = (\partial/\partial\theta)\ell(y,x,\theta,g).$$

Then $\hat{\theta}$ will generally be asymptotically efficient provided $\hat{I}_*$ and $\hat{g}(\cdot)$ are good estimates of I_* and $g(\cdot)$ respectively. This approach still requires an estimator $\hat{g}(\cdot)$ of $g(\cdot)$.

Note that $\ell(y,x,\theta,g)$ depends on $g(\cdot)$ only through the function $w(\cdot)$ in (3.8) and its derivative. In work in progress the authors are investigating a one-step construction of an asymptotically efficient estimator which estimates $w(\cdot)$ directly, avoiding the intermediate step of estimating $g(\cdot)$.

4.0 CONCLUDING REMARKS

In conclusion we reiterate that the assumption of normal errors, (1.2), is not crucial to the theory developed herein; the existence of a complete sufficient statistic for u when regarded as a parameter is crucial. The situation in which (2.1) is replaced with an assumption of replicated measurements, i.e., $m > 1$ in (1.2), is conceptually no different than when (1.2) is assumed with the exception that both $\bar{\Omega}$ and ϕ can now be estimated; thus there will be an additional $p(p+1)/2$ - dimensional component to all the scores.

Although no distributional assumptions on the measurement errors is more reasonable than that of normality it is still an unverifiable assumption unless replicate measurements are made. The sufficiency, conditional and efficient scores lose their unbiasedness when the assumption of normal errors is erroneous. Thus when measurement error is nonnormal, estimates derived from these scores will generally be biased and the bias will generally not be computable. Approximations to the bias can probably

be obtained using the small-measurement asymptotics employed by Stefanski (1985) although we have not attempted these calculations.

ACKNOWLEDGEMENTS

The work of R. J. Carroll was supported by the U.S. Air Force Office of Scientific Research.

References

Adcock, R. J. (1878). A Problem in least squares. *The Analyst* **5**, 53-4.

Anderson, T. W. (1976). Estimation of linear functional relationships (with discussion). *J. Roy. Statist. Soc. Ser.B* **38**, 1-36.

Begun, J. M. Hall, W. J., Hwang, W. M. & Wellner, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.* **11**, 432-52.

Bickel, P. J. & Ritov, Y. (1986). Efficient estimation in the errors-in-variables model. *Ann. Statist.* to appear.

Carroll, R. J., Spiegelman, C. H., Lan, K. K., Bailey, K. T., & Abbott, R. D. (1984.) On errors-in-variables for binary models. *Biometrika* **71**, 19-25.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *J. Amer. Statist. Assoc.* **70**, 892-98.

Gleser, L. J. (1981). Estimation in a multivariate 'errors-in-variables' regression model: large sample results. *Ann. Statist.* **9**, 24-44.

Huber, P. J. (1967). The behavior of maximum liklihood estimators under nonstandard conditions.. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 1.* Ed. L. M. LeCam & J. Neyman, 221-33, University of California Press.

Kendall, M. G. & Stuart, A. (1979). *The Advanced Theory of Statistics, 2.* London: Griffin.

Lindsay, B. G. (1980). Nuisance parameters, mixture models, and the efficiency of partial likelihood estimators. *Philos. Trans. Roy. Soc. London Ser. A* **296**, 639-65.

Lindsay, B. G. (1982). Conditional score functions: some optimality results. *Biometrika* **69**, 503-12.

Lindsay, B. G. (1983). Efficiency of the conditional score in a mixture setting. *Ann. Statist.* **11**, 486-97.

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

Michalik, J. E. & Tripathi, R. C. (1980). The effect of errors in diagnosis and measurement on the estimation of the probability of an event. *J. Amer. Statist. Assoc.* **75**, 713-21.

Moran, P. (1971). Estimating structural and functional relationshps. *J. Multi. Anal.* **1**, 232-55.

Pfanzagl, J. (1982). *Contributions to a General Asymptotic Statistical Theory.* New York: Springer-Verlag.

Prentice, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* **69**, 331-42.

Stefanski, L. A. (1985). The effects of measurement error on parameter estimation. *Biometrika* to appear.

Stefanski, L. A. & Carroll, R. J. (1986). Covariate measurement error in logistic regression. *Ann. Statist.* to appear.

Wolter, J. M. & Fuller, W. A. (1982a). Estimation of nonlinear errors-in-var iables models. *Ann. Statist.* **10**, 539-48.

Wolter, J. M. & Fuller, W. A. (1982b). Estimation of the quadratic errors-in-variables model. *Biometrika* **69**, 175-82.