CONDITIONALLY UNBIASED BOUNDED INFLUENCE ESTIMATION IN GENERAL REGRESSION MODELS, WITH APPLICATIONS TO GENERALIZED LINEAR MODELS

Hans R. Künsch[1]
Leonard A. Stefanski[2]
Raymond J. Carroll[3]

[1] Seminar für Statistik, ETH-Zentrum, CH-8092 Zürich, Switzerland

[2] Department of Statistics, North Carolina State University, Raleigh, NC 27695

[3] Department of Statistics, Texas A & M University, College Station, TX 77843

AUTHORS' FOOTNOTE

Hans R. Künsch is Assistant Professor, Seminar für Statistik, ETH-Zentrum, CH-8092 Zürich, Switzerland. Leonard A. Stefanski is Assistant Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695. Raymond J. Carroll is Professor and Head, Department of Statistics, Texas A&M University, College Station, TX 77843. The work of Stefanski has been supported by the National Science Foundation. The work of Carroll has been supported by the Air Force Office of Scientific Research. The authors wish to thank two referees for their helpful and thought provoking comments which improved the clarity of presentation.

ABSTRACT

In this paper we study robust estimation in general models for the dependence of a response y on an explanatory vector x. We extend previous work on bounded influence estimators in linear regression. Second we construct optimal bounded influence estimators for generalized linear models. We consider the class of estimators defined by an estimating equation with a conditionally unbiased score function given the desgin. The resulting estimators are said to be conditionally Fisher-consistent. Ordinary least squares in linear regression has this property as does the Mallows type bounded influence estimator. The Schweppe class does not have a conditionally unbiased score function if the errors are asymmetric. For generalized linear models, the optimal conditionally Fisher-consistent estimators are computationally simpler than the unconditional ones proposed by Stefanski, Carroll and Ruppert (1986) because the centering constant can be given in explicit form. The optimal score function contains an unknown auxiliary nuisance matrix B. In contrast to the estimator of Stefanski et al. (1986) estimation of B has in our case asymptotically no effect on the distribution of the estimator. Two examples using logistic regression are discussed in detail. It is shown that robust estimation can identify outliers even in situations where the model is close to being indeterminate.

Key Words and Phrases: Robust Regression, Bounded Influence, Asymptotic Bias, Generalized Linear Models, Linear Regression, Logistic Regression.

1.INTRODUCTION

The basic generalized regression model states that given the values of a p-dimensional explanatory variable x, the response y has a distribution function $P_\theta(y|x)$. We are interested in estimating the parameter θ from N independent observations (y_i, x_i). In such a general model, M-estimators are defined implicitly by the equation

$$\sum_{i=1}^{N} \psi(y_i, x_i, \hat{\theta}_N) = 0. \tag{1.1}$$

In equation (1.1), θ and ψ have the same dimension. Of course, maximum likelihood estimators are (non-robust) M-estimators. Assume for the moment that x is also a random variable with distribution function F. In order that $\hat{\theta}_N$ be consistent, standard theory requires that the estimating equation (1.1) be *unbiased*, i.e.,

$$E_\theta(\psi(y, x, \theta)) = \int\int \psi(y, x, \theta) P_\theta(dy \mid x) F(dx) = 0 \text{ for all } \theta. \tag{1.2}$$

Requiring (1.2) is the same as saying that ψ is Fisher consistent. Certainly, Fisher consistency is a minimal requirement, but in linear and generalized linear regression it is too weak and even somewhat unpalatable because it involves the distribution of the predictors x. After all, the x_i may not be random variables even. What statisticians do in practice is to condition on the observed values of x. We say that an M-estimator is *conditionally Fisher consistent* if it satisfies

$$E_\theta(\psi(y, x, \theta) \mid x) = \int\int \psi(y, x, \theta) P_\theta(dy \mid x) = 0 \text{ for all } \theta \text{ and } x. \tag{1.3}$$

In linear and generalized linear regression, maximum likelihood estimators are conditionally Fisher consistent whenever the distribution of x does not depend on θ.

Conditional Fisher consistency is an appealing concept because it does not depend on the x's being random and, even if they are, it does not involve the distribution of the x's. This does not mean that the properties of conditionally Fisher consistent M-estimators are independent of the design of course; just remember the formula for the covariance of least squares estimates. Similarly the influence function and the sensitivity to be defined below will also depend on the design.

In the linear model with symmetric errors, essentially all M-estimators in the literature (including least squares) satisfy (1.3). However, as we show in section 2, there are some bounded influence M-estimators which are not conditionally Fisher consistent when the errors are asymmetric. This is particularly true of the class popularly known as Schweppe-type estimators. The

Mallows-type estimates, including ordinary M-estimators, are Fisher consistent. For the definition of these two types, see Krasker and Welsch (1982) or Hampel, Ronchetti, Rousseeuw and Stahel (1986, pp. 315-316).

In generalized linear models, it is also possible to define Schweppe- and Mallows-type estimators, see Stefanski et al. (1986) for the former and Pregibon (1981) for the latter (in logistic regression). The optimal robust estimators of Stefanski et al. (1986) are not conditionally Fisher consistent, although they do satisfy (1.2). These estimators are difficult to compute, and even when computable their asymptotic distributions are very difficult to understand because of a dependence upon an auxilliary nuisance matrix B. In sections 3 and 4, we study optimal robust conditionally Fisher consistent estimates for generalized linear models. There is a practical payoff to restriction on this narrower class. In contrast to the estimates of Stefanski et al. (1986), our estimates are relatively easy to compute and the asymptotic distribution theory is straightforward. In section 5, we apply our methods to two data sets involving logistic regression.

We now review some general results and definitions from robust statistics, see Hampel et al. (1986). The influence function of an M-estimator is

$$IC_{\psi}(y,x,\theta) = D(\psi,\theta)^{-1}\psi(y,x,\theta); \tag{1.4}$$

$$D(\psi,\theta) = -\frac{\partial}{\partial\beta}\int\int\psi(y,x,\beta)P_{\theta}(dy \mid x)F(dx)|_{\beta=\theta}. \tag{1.5}$$

The influence function measures the effect of an infinitesimal contamination at (y,x) standardized by the mass of the contamination. It thus gives an approximation to the effect of the inclusion or deletion of a single observation. Moreover it gives the asymptotic covariance matrix. Under regularity conditions, $N^{1/2}(\hat{\theta}_N - \theta)$ is asymptotically normally distributed with mean zero and covariance matrix $V(\psi,\theta)$, also written $V(\psi)$,

$$V(\psi,\theta) = E_{\theta}[IC_{\psi}(y,x,\theta)IC_{\psi}(y,x,\theta)^T] = D(\psi,\theta)^{-1}W(\psi,\theta)D(\psi,\theta)^{-T} \tag{1.6}$$

where

$$W(\psi,\theta) = E_{\theta}[\psi(y,x,\theta)\psi(y,x,\theta)^T]. \tag{1.7}$$

The main idea of bounded influence estimation is to define a bound on the influence function (1.4), and then find an estimator which has small variance subject to a chosen bound. The usual operational difficulty is that the influence function is a vector, and we need to reduce this to a scalar measure. This scalar is called the *sensitivity* of the influence function. This problem is not too different from what happens with regression diagnostics. For example, consider case deletion diagnostics, where we try to understand the effect of deleting an observation on the parameter estimates. Changes in parameter estimates are necessarily p-dimensional. The beauty of a diagnostic such as Cook's distance (Cook and Weisberg, 1982) is that the p-dimensional change in the parameters is summarized by a scalar. The same issue arises in bounded influence regression, and as in deletion diagnostics we have to decide upon a summary measure. Just as we can define diagnostics such as Cook's distance or DFFITS, we can define different methods of measuring the sensitivity of the influence function. The most common method, used with success by Krasker and Welsch (1982), is the *self-standardized sensitivity* defined as

$$s(\psi)^2 = sup_{y,x} sup_{\lambda \neq 0} \frac{(\lambda^T IC_\psi)^2}{\lambda^T V(\psi)\lambda} = sup_{y,x} \psi(y,x,\theta)^T W(\psi,\theta)^{-1} \psi(y,x,\theta). \tag{1.8}$$

This definition of sensitivity measures the maximal influence an observation can have on a linear combination of parameters, with a standardization by the asymptotic standard deviation of this linear combination. Integrating (1.8) and taking the trace shows that $s(\psi)^2 \geq p$. Other measures of sensitivity are considered by Hampel et al. (1986,p.317) and Giltinan et al. (1986). For example, the latter authors consider bounding the influence on predicted values rather than on the parameter estimates. The estimators resulting from the definition of sensitivity used by Giltinan et al. (1986) tend to downweight suspect observations much more severely than do those considered here.

A referee has raised the question about the meaning of the parameter θ if deviations from the model are considered. For instance in logistic regression one cannot have errors with fatter tails, but this is not the only deviation robustness protects for. Whenever the true distribution is a *small* deviation from the parametric model with $\theta = \theta_0$, then the robust estimator is asymptotically close to that value θ_0. We may consider for instance a gross error model with the amount of contamination depending on the regressor x : The conditional distribution of y given x is $P_\theta(dy \mid x)$ with probability $1 - \epsilon(x)$ and arbitrary with probability $\epsilon(x)$. This is a small deviation if the total proportion of contamination $E[\epsilon(x)]$ is small. It makes sense also for logistic regression and the meaning of the parameter is clear here.

A somewhat different concept is the requirement that the estimators should not change much if a few observations are included or deleted. This is clearly desirable for any type of data analysis. Because of the interpretation of the influence function given above, our robust estimators should be less affected by the inclusion or deletion of a few observations than the classical ones.

2. LINEAR REGRESSION WITH ASYMMETRIC ERRORS

The purpose of this section is to show that the differences between (1.2) and (1.3) are relevant also in linear regression if the errors are not symmetric. Write $\theta^T = (\theta_0, \theta_1^T)$ and consider the model

$$Y_i = \theta_0 + X_i^T\theta_1 + U_i. \tag{2.1}$$

where $\{U_i\}$ are independent and identically distributed with common density g having no point of symmetry. The usual design matrix is assumed to be of full rank. The vector θ_1 is of length (p-1). The most important proposals for M-estimators in linear regression are the Mallows and Schweppe types

$$\psi\ (y.x.\theta) = w(x)\ \varphi(y-\theta_0-x^T\theta_1)(1x^T)^T; \tag{2.2}$$

$$\psi\ (y.x.\theta) = w(x)\ \varphi((y-\theta_0-x^T\theta_1)\ /\ w(x))(1\ x^T)^T. \tag{2.3}$$

Here w(x) is a scalar weight function and φ is a scalar function. Without additional assumptions on g. θ_0 is not identifiable. but the following result gives conditions for consistency of θ_1. For related results. see Carroll (1979).

<u>THEOREM 2.1</u> :

i) For the Mallows-form (2.2). there is a constant τ such that $\psi(y.x.\theta_0+\tau.\theta_1)$ satisfies (1.3).

ii) If the X_i are symmetrically distributed with center $c \in \mathbb{R}^{p-1}$ and if $w(x-c) \equiv w(-x+c)$, then there is a constant τ such that for the Schweppe form (2.3) $\psi(y.x.\theta_0+\tau.\theta_1)$ satisfies (1.2), but in general not (1.3).

<u>PROOF</u>: i) is obvious by defining τ as the solution of $\int \psi(u-\tau)g(u)du = 0$. In case ii) define τ as the solution of

$$\iint w(x)\ \varphi((u-\tau)/w(x))\ g(u)\ F(dx)du = 0.$$

This is just (1.2) for the first component of ψ. For the other components we observe that

$$\iint x_j\ w(x)\ \varphi((u-\tau)/w(x))\ g(u)\ F(dx)du =$$

$$\iint (x_j-c_j)w(x)\ \varphi((u-\tau)/w(x))\ g(u)\ F(dx)du = 0$$

by symmetry, so (1.2) holds. Equation (1.3) does not hold in general, because the solution τ of

$$\int \varphi((u-\tau)/w(x))\ g(u)du = 0$$

depends in most cases on x. □

Thus, the Mallows form has the advantage of being a conditionally unbiased estimating equation even when the $\{X_i\}$ or $\{U_i\}$ are not symmetrically distributed.

3. GENERALIZED LINEAR MODELS

We consider a generalized linear model with canonical link function

$$P_\theta(dy|x) = \exp\{y\ x^T\theta - G(x^T\theta) - S(y)\}\ \mu(dy)\ , \qquad (3.1)$$

see McCullagh & Nelder (1983). If g is the derivative of G, the likelihood score functior is

$$\ell(y,x,\theta) = \{y-g(x^T\theta)\}x. \qquad (3.2)$$

Note that (3.2) satisfies (1.3), so that the score is conditionally unbiased. Because ℓ is proportional to x, the influence is unbounded, i.e., $s(\ell) = \infty$.

We are looking here for M-estimators satisfying (1.3) and $s(\psi) \leq b$ which minimize $V(\psi)$ in some sense. Motivated by a general principle for constructing optimal robust estimators satisfying (1.2) (Hampel et al. (1986), Section 4.3a), we consider the following score function:

$$\psi_{cond}(y,x,\theta,B) = d(y,x,\theta,B)\ w_b(|d(y,x,\theta,B)|(x^TB^{-1}x)^{\frac{1}{2}})\ x, \qquad (3.3)$$

where

$$d(y,x,\theta,B) = y-g(x^T\theta) - c(x^T\theta,\ b/(x^T\ B^{-1}x)^{\frac{1}{2}})$$

and

$$w_b(a) = H_b(a)/a.$$

where H_b is the Huber function $H_b(a) = \max(-b,\min(a,b))$.

We work within the context of the Schweppe-type, although related results are obtainable for the Mallows-type as in Stefanski et al (19S6). The major difference is that w_b in (3.3) would factor into two parts. The first depends only on x and is of the form $w_1((x^TB^{-1}x)^{1/2})$. The other depends only on

$$d(y,x,\theta) = y - g(x^T\theta) - c(x^T\theta,b/(x^TB^{-1}x)^{1/2})$$

and is of the form $w_2(|d(y,x,\theta)|)$.

The scalar function c and the matrix B in (3.3) will be chosen so that the side conditions (1.3) and $s(\psi_{cond}) = b$ are satisfied. By the definition of ψ_{cond}, (1.3) holds if and only if for all β and all $a > 0$,

$$\int(y-g(\beta)-c(\beta,a))\ w_a(|y-g(\beta)-c(\beta,a)|)\exp(y\ \beta-G(\beta)-S(y))\ \mu(dy) = 0. \tag{3.4}$$

First we discuss the existence of a solution to (3.4).

<u>LEMMA 3.1</u> For any $a > 0$ and β, there is a solution $c = c(\beta,a)$ to (3.4).

<u>PROOF</u>: For fixed y, β, a, the function

$$c \longrightarrow (y - g(\beta) - c)\ w_a(|y - g(\beta) - c|)$$

is continuous, bounded and monotone nonincreasing with limits ± a. Hence the existence follows from dominated convergence and the intermediate value theorem.

□

A practical advantage here is that often the function c can be calculated in closed form. This is particularly important compared to the optimal ψ satisfying (1.2), where c is a vector (depending only on θ) whose computation is quite difficult, see Stefanski, et. al (1986), Section 2.4. Here are two examples where $c(\beta,a)$ can be calculated explicitly.

<u>Example 3.1</u> : <u>Logistic Regression</u> Here μ puts equal mass at 0 and 1, $S(y)=0$, and $G(\beta) = \log\{1 + \exp(\beta)\}$. Write $p = \exp(\beta)/(1 + \exp(\beta))$ and $q = 1-p$. It is easily checked that

$$c(\beta,a) = \begin{cases} ap/q - p & \text{if } \beta < 0 \text{ and } a < q \\ q - aq/p & \text{if } \beta > 0 \text{ and } a < p \\ 0 & \text{otherwise} \end{cases}$$

satisfies (3.4).

<u>Example 3.2</u> : <u>Negative Exponential Regression</u> Here μ is Lebesgue-measure on $[0,\infty)$, $G(\beta) = -\log(-\beta)$, $S(y) = 0$ and $\beta < 0$. Two cases occur. If the bound is large, the Huberization in ψ_{cond} is one-sided (for large y's only),

whereas for small a's both large and small y's will be Huberized. It can be checked by straightforward calculations that the cutting point between the two cases is given by the equation $e^{2\beta a} = 1 + \beta a$. so that $\beta a \approx -0.797$. In the former case $c(\beta.a) = -\beta^{-1}$ times the smaller solution of $\exp(x+\beta a-1) = x$ and in the latter case $c(\beta.a) = -\beta^{-1}(1 + \log(\beta a/(\exp(\beta a) - \exp(-\beta a))))$.

Turn now to the matrix B. We note first that the estimator is conditionally Fisher consistent and has bounded influence for any choice of B. However if we want $s(\psi_{cond}) = b$, B will depend on both the design and θ. In linear regression B depends on the design but not on θ. It follows from the definition of ψ_{cond}, that $s(\psi_{cond}) = b$ provided

$$E_\theta[\psi_{cond}(y, x, \theta, B)\psi_{cond}(y, x, \theta, B)^T] = B \tag{3.5}$$

Equation (3.5) is used to define $B = B(\theta, F)$. Because $s(\psi)^2 \geq p$ a necessary condition for (3.5) to have a solution is $b^2 \geq p$, but we do not know if it is also sufficient.

The estimators we have defined are intuitively appealing because they downweight observations according to their leverage and "outlyingness".. It is reasonable to ask if they satisfy any optimality criterion. The discussion of optimality within a bounded influence class started with Krasker and Welsch (1982), but the results of Ruppert (1985) suggest that there is no estimator which has uniformly smallest covariance subject to a bound on the influence. The best known optimality result seems to be that of Stefanski, et al. (1986). It is not completely satisfactory because the criterion to be minimized depends already on the solution. Nevertheless it implies that no other estimator satisfying the same bound on $s(\psi)$ can have a uniformly smaller covariance. We can achieve the same optimality result within the class of conditionally Fisher consistent estimates. We state this in the following theorem.

THEOREM 3.1: Suppose that for a given b, (3.5) has a solution $B(\theta)$. Then ψ_{cond} minimizes trace $(V(\psi)V(\psi_{cond})^{-1})$ among all ψ which satisfy both (1.3) and

$$sup_{y,x} IC_\psi^T V(\psi_{cond})^{-1} IC_\psi \leq b^2.$$

Theorem 3.1 is a corollary of the following analogy to Theorem 1 of Stefanski et al. (1986). Note that Theorem 3.2 below also applies to any kind of model with explanatory variables.

THEOREM 3.2 : Let $\ell(y.x.\theta)$ be the likelihood score function. Define the score function

$$\psi_{cond}(y.x.\theta) = (\ell-c)\min\left(1.b\,/\{(\ell-c)^T B^{-1}(\ell-c)\}^{1/2}\right). \qquad (3.6)$$

where $c = c(x.\theta)$ and $B = B(\theta)$ are assumed to exist and satisfy

$$E(\psi_{cond}(y.x.\theta)\,|x)) = 0$$

$$E(\psi_{cond}(y.x.\theta)\ \psi_{cond}(y.x.\theta)^T) = B.$$

Then (3.6) minimizes $tr(V(\psi)V(\psi_{cond})^{-1})$ among all ψ satisfying (1.3) and

$$\sup_{(y.x)}\ IC_\psi\ V(\psi_{cond})^{-1}\ IC_\psi \le b^2.$$

With the exception of multiplication by a constant matrix. ψ_{cond} is unique almost surely.

PROOF OF THEOREM 3.2 The proof is almost identical to that of Theorem 1 in Stefanski et al. (1986), once one notes that for any conditionally unbiased score function ψ, $E[c(x,\theta)\psi(y,x,\theta)] = E[c(x,\theta)\ E(\psi(y,x,\theta)\,|x)] = 0$. □

The computational simplicity of the conditional Fisher consistent estimator is not particular to the canonical model (3.1). For instance, consider a generalized linear model with arbitrary link function h , i.e. $x^T\theta$ in (3.1) is replaced by $h(x^T\theta)$. Then we have to replace in (3.3) $d(y,x,\theta,B)$ by

$$h'(x^T\theta)\{y - g(h(x^T\theta)) - c(h(x^T\theta).b/((x^TB^{-1}x)^{1/2}\ |\ h'(x^T\theta)|))\}.$$

where $c(\beta.a)$ is still defined by (3.4).

In applications, the distribution F of the $\{X_i\}$ is unknown. It is common to replace F by its empirical distribution. From (3.3) and (3.5), this means that we solve

$$\sum_{i=1}^{N} \psi_{\text{cond}}(y_i . x_i . \hat{\theta}_N . \hat{B}_N) = 0. \tag{3.7}$$

$$N^{-1} \sum_{i=1}^{N} x_i x_i^T \, v(x_i^T \hat{\theta}_N . \, b/(x_i^T \hat{B}_N^{-1} x_i)^{1/2}) = \hat{B}_N. \tag{3.8}$$

where

$$v(\beta . a) = \int (y-g(\beta)-c(\beta . a))^2 \, w^2(y . \beta . a) \exp(y\beta - G(\beta) - S(y)) \, \mu(dy). \tag{3.9}$$

$$w(y, \beta, a) = \min(1, a/|y-g(\beta)-c(\beta, a)|) \tag{3.10}$$

In many applications, improved protection against outliers through higher breakdown points can be achieved by the use of redescending ψ functions, see Rousseeuw (1984). In equations (3.3), (3.4) and (3.10) the Huber function H_b could be replaced by any of the redescenders such as Hampel's three part function or the Tukey biweight. The calculation of $c(\beta . a)$ is of the same complexity as with the Huber function. The breakdown properties and efficiency of such estimates remain to be studied.

4. THE EFFECT OF ESTIMATING THE MATRIX B

In the last section we derived the estimator defined by (3.7)-(3.8) as an approximation to the optimal estimator which uses $\psi_{cond}(y,x,\theta,B(\theta))$. We may consider (3.7) and (3.8) as an M-estimator for both θ and a nuisance parameter B. The ψ-function defining this M-estimator is

$$(\psi_{cond}(y,x,\theta,B)^T, \chi(x,\theta,B)^T)^T, \text{ where}$$
$$\chi(x,\theta,B) = x\, x^T v(x^T\theta,\ b/(x^TB^{-1}x)^{1/2}) - B.$$

The influence function of this estimator is (compare (1.4) and (1.5))

$$IC_{\psi,\chi}(y,x,\theta,B) = D_{\psi,\chi}(\theta)^{-1}(\psi_{cond}(y,x,\theta,B)^T,\ \chi(x,\theta,B)^T)^T, \tag{4.1}$$

where

$$D_{\psi,\chi} = \begin{bmatrix} -\frac{\partial}{\partial\beta} E_\theta\left[\psi_{cond}(y,x,\beta,B)\right]\Big|_{\beta=\theta} & -\frac{\partial}{\partial A} E_\theta\left[\psi_{cond}(y,x,\theta,A)\right]\Big|_{A=B(\theta)} \\ -\frac{\partial}{\partial\beta} E_\theta\left[\chi(x,\beta,B)\right]\Big|_{\beta=\theta} & -\frac{\partial}{\partial A} E_\theta\left[\chi(x,\theta,A)\right]\Big|_{A=B(\theta)} \end{bmatrix} \tag{4.2}$$

By the definition of ψ_{cond} and $c(\beta,a)$ in (3.3) and (3.4), $\psi_{cond}(y,x,\theta,A)$ satisfies (1.3) for arbitrary A. Hence, $E_\theta[\psi_{cond}(y,x,\theta,A)] = 0$ for all A and the upper right block of $D_{\psi,\chi}$ is zero. This means that the θ part of the influence function for (3.7) and (3.8) is equal to

$$\left(-\frac{\partial}{\partial\beta}E_\theta[\psi_{cond}(y,x,\beta,B)]\ \Big|_{\beta=\theta}\right)^{-1} \psi_{cond}(y,x,\theta,B). \tag{4.3}$$

On the other hand, the influence function for the optimal $\psi_0(y,x,\theta) = \psi_{cond}(y,x,\theta,B(\theta))$ is also equal to (4.3) because,by the same argument.

$$D_{\psi_0}(\theta) = -\frac{\partial}{\partial\beta} E_\theta[\psi_{cond}(y.x.\beta.B)] \Big|_{\beta=\theta} - \frac{\partial}{\partial\Lambda} E_\theta[\psi_{cond}(y.x.\beta.\Lambda)] \Big|_{\Lambda=B} \frac{\partial}{\partial\theta}B(\theta)$$

$$= -\frac{\partial}{\partial\beta} E_\theta[\psi_{cond}(y.x.\beta.B)] \Big|_{\beta=\theta} .$$

We have thus shown

THEOREM 4.1 : The θ part of the influence function in the case that θ and B are simultaneously estimated by (3.7) and (3.8) is the same as the influence function in the case that θ alone is estimated using the optimal $\psi_0(y.x.\theta) = \psi_{cond}(y,x,\theta,B(\theta))$. As a consequence, the asymptotic covariance matrix of $\hat{\theta}_N$ is the same in both cases.

REMARKS :

i) $\hat{\theta}_N$ and $\hat{B}_N$ are not asymptotically independent: $E_\theta[\psi_{cond}\ \chi^T] = 0$ by (1.3), but $\frac{\partial}{\partial\beta} E_\theta[\chi(x.\beta.B)]\Big|_{\beta=\theta} \neq 0$ in general.

ii) Because in linear regression with symmetric errors χ does not depend on θ, an analogue to Theorem 4.1 is obvious. In addition, estimation of the scale of the errors does not change the asymptotic covariance either, and $\hat{\theta}_N$ is asymptotically independent of all nuisance parameters.

iii) From the finite sample interpretation of the influence function, (4.3) means the following: to the first order of approximation the change in $\hat{\theta}_N$ caused by adding or deleting an observation at (x,y) is

$$\left(\sum_{i=1}^{N} \frac{\partial}{\partial\beta} E_{\hat{\theta}_N}[\psi_{cond}(y.x_i\beta.\hat{B}_N)|x_i]\Big|_{\beta=\hat{\theta}_N}\right)^{-1} \psi_{cond}(y.x.\hat{\theta}_N.\hat{B}_N).$$

i.e. the change in $\hat{B}_N$ has approximately no effect on the change in $\hat{\theta}_N$. In this sense the estimator (3.7)-(3.8) is reasonably stable.

(iv) For the Fisher-consistent estimator (2.12)-(2.13) of Stefanski, et al. (1986), there is no analogue of Theorem 4.1. The θ part of the influence function is in general a linear combination of ψ_{BI}, $E_\theta[\psi_{BI} \mid x]$ and $E_\theta[\psi_{BI}\psi_{BI}^T|x]-B$ because all blocks in D are in general different from zero.

(v) As both referees have pointed out, estimation of B makes no difference to asymptotic arguments, but almost certainly will have some effects in small samples. The analogue to ordinary M-estimation in linear regression is the problem of simultaneous estimation of scale, say by MAD or Proposal 2, which does have some effect on small sample properties. One way to investigate this difference, at least in principle, is through the use of second order expansions. Such expansions are extremely tedious even for trying to understand the effect of scale in linear regression, and they are likely to be prohibitively difficult and complex for understanding the effect of estimating B. We even doubt if formal second order expansion gives a much better approximation to the small sample effects. Small sample asymptotics (see Hampel et al., 1986, Sec. 8.5) looks more promising, but it is even harder. In any case Theorem 4.1 suggests that the small sample effect of estimating B will be smaller for our estimator than for the one studied by Stefanski et al. (1986). In the Example 5.2 below we check how good the approximation described in Remark iii) is in practice.

5. EXAMPLES

To illustrate our estimators we consider two examples of logistic regression which have appeared elsewhere in the context of robust regression; see Pregibon (1981,1982) and Stefanski et al. (1986). Both data sets are difficult because they contain outliers and without these outliers they are rather close to indeterminacy. The first example is particularly extreme whereas in the second it is still possible to fit a meaningful model.

Example 5.1: Skin Vaso-Constriction Data (Pregibon, 1981 and 1982).

These data consist of 39 observations on three variables, the occurrence of vaso-constriction in the skin of the digits and the rate and volume of air inspired. The model to be fit regresses the occurrence of vaso-constriction on the logarithms of the remaining two variables. In our work we took Rate 32 = 0.30, see Pregibon (1981).

Pregibon (1981) established that observations #4 and #18 are enormously influential in determining the maximum likelihood fit. It is not as evident from his analysis that without these two observations the model is nearly indeterminate, i.e. almost perfect discrimination is possible or, in other words, the data are very close to being nonoverlapping (Santner and Duffy, 1986); and any estimation procedure will necessarily reflect this near indeterminacy through the estimated standard errors. In such a situation we cannot expect any estimator, robust or otherwise, to produce a completely satisfactory model (for none exists) and the main advantage of a robust procedure lies in its diagnostic capabilities. The situation is similar to one which occurs in linear regression when the $X'X$ matrix approaches singularity upon the removal of one (or a few) observations, see for example Draper and Smith (1981,

p. 258) or Chatterjee and Hadi (1986, Fig. 2). Although the data analyst should be aware of these points, he cannot hope to obtain meaningful parameter estimates under these circumstances; this fact would be reflected by dramatic increases in standard errors when these points are downweighted or removed.

Table 1 contains parameter estimates, estimated standard errors, and weights for three robust fits, two using the Huber weight function with choices of $bp^{-\frac{1}{2}}$ equal to 3.7 and 3.2; and a "biased" analysis performed using the Huber weight function with $bp^{-\frac{1}{2}} = 3.7$ and setting $c(\beta, \alpha) = 0$. Results from the maximum likelihood fit are given also. When we attempted a fit using the Hampel function, observations 4 and 18 were immediately assigned zero weight and computational difficulties arose as a consequence of the near indeterminacy.

Three comments are worth making. First, as b increases the weights assigned observations 4 and 18 decrease rapidly, clearly indicating their anomolous nature. This fact would also manifest itself in a simple residual plot, see Stefanski et al. (1986). All other observations received weight one in all cases. Second, the estimated standard errors increase significantly as observations 4 and 18 are downweighted. Although some loss of efficiency is to be expected with the use of robust methods, the sizeable increase in standard errors for these data reflects the problem with indeterminacy mentioned above. Finally the choice of b is quite crucial, but this is not surprising in view of the particular nature of the data. The biased estimator with $c \equiv 0$ seems to be more robust than the conditional unbiased one with the same b . We have no explanation for this.

Example 5.2: Food Stamp Data (Stefanski et al., 1986).

For these data the response indicates participation in the Federal Food Stamp program and the predictor variables employed include two dichotomous variables, tenancy and supplemental income, and a logarithmic transformation of monthly income, log(monthly income +1). The data consist of observations on 150 persons of which 24 participated in the program.

Table 2 displays results from a number of robust fits as well as for maximum likelihood estimation. In computing the Hampel estimator a concession was made for computational convenience; rather than solving (3.4) to define $c(\beta, \alpha)$ we chose to use the same formula as in Example 3.1. The conclusions drawn by Stefanski et al. (1986) apply equally well to the estimators here. The two observations #5 and #66 are most influential for the maximum likelihood estimator. As b decreases, these observations are downweighted. This results in an increase in perceived significance of the monthly income accompanied by a decrease of the importance of supplemental income. Besides these two outliers, there are seven other atypical observations which are downweighted, but to a smaller extent.

Unlike in the previous example the estimated standard errors remain relatively stable suggesting a greater degree of overlap (Santner and Duffy, 1986) in the data. However a closer look shows that there are only six persons with tenancy participating in the footstamp program. Once these are eliminated the parameter for tenancy cannot be estimated any more. Moreover without tenancy as a predictor the data become completely separated after elimination of 17 observations. All the nine downweighted observations belong to this group, so any reasonable estimator has to use these outliers to some extent.

In Table 3 we give the changes in the estimates due to deletion of observation #5 or #66 and compare it with the change predicted by the influence function.Although we used a robust estimator the changes are rather big. This is due to the pecularity of the data set mentioned above. Still the changes are much smaller than for the MLE. For instance, if observation #5 is deleted, the MLE for the coefficient of log (monthly income +1) changes by 0.73 compared to 0.26 with our estimator. For B kept constant the predictions by the influence function are excellent. By the results of Section 4 the influence function predicts no effect of reestimating B. In this example this is not quite true, but as a first order approximation it is acceptable, in particular because the effective sample size is maybe not 150 but between 17 and 150.

Our estimators are suitable for inferences based on the majority of the data. Moreover they can also be used as a diagnostic with the following strategy suggested by the examples. Choose a rather large b (say $b = 5p^{\frac{1}{2}}$) and decrease it (for instance in steps by $0.5p^{\frac{1}{2}}$). Looking at the weights allows one to identify outliers. At the same time it should be checked how close the data are to indeterminacy, a possible indication being how fast the estimated standard errors change, but it would be interesting to have other criteria. In this way one can either fit a meaningful model to the good observations or identify the data set as problematic.

6. CONCLUSIONS

Conditionally unbiased score functions are appealing because their definition does not depend on the distribution of the predictors. In the context of robustness, there is an optimality theory for this class analogous to that already developed for unconditionally unbiased score functions. The optimal estimator depends on the unknown distribution of the predictors and one has therefore to estimate a matrix B . However, consistency holds for any B and asymptotically the uncertainty about B does not matter. In addition, conditionally unbiased score functions are often far easier to define. Although ignoring the bias and setting $c \equiv 0$ turned out not to matter much in the examples considered, one can construct situations where this bias is large. With our estimator, we avoid this problem with little additional complexity.

REFERENCES

Carroll, R.J. (1979). Estimating variances of robust estimators when the errors are asymmetric. Journal of the American Statistical Association 74, 674-679.

Chatterjee, S. and Hadi, A.S. (1986). Influential observations, high leverage points, and outliers in linear regression. Statistical Science 1, 379-93.

Cook, R.D. and Weisberg, S. (1982). Residuals and Influence in Regression. Chapman and Hall, New York

Draper, N.R. and Smith, H. (1981). Applied Regression Analysis. 2nd ed., Wiley, New York.

Giltinan,D.M., Carroll, R.J., & Ruppert, D. (1986). Some new methods for weighted regression when there are possible outliers. Technometrics 28, 219-230.

Hampel, F.R., Ronchetti, E., Rousseeuw, P.J. & Stahel, W. (1986). Robust Statistics: The Approach Based on Influence Functions. John Wiley & Sons, New York.

Krasker, W.S. and Welsch, R.E. (1982). Efficient bounded-influence regression estimation. Journal of the American Statistical Association 77, 595-604.

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

Pregibon, D. (1981). Logistic regression diagnostics. Annals of Statistics 9, 705-724.

Pregibon, D. (1982). Resistant fits for some commonly used logistic model with applications. Biometrics 38, 485-98.

Rousseeuw, P.J. (1984). Least median of squares regression. Journal of the American Statistical Association 79, 871-880.

Ruppert, D. (1985). On the bounded-influence regression estimator of Krasker and Welsch. Journal of the American Statistical Association 80, 205-208.

Santner, T.J. and Duffy, D.E. (1986). A note on A. Albert and J.A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. Biometrika 73, 755-758.

Stefanski, L.A., Carroll, R.J. & Ruppert, D. (1986). Optimally bounded score functions for generalized linear models with applications to logistic regression. Biometrika 73, 413-425.

Table 1

Maximum likelihood and robust estimators for the skin vaso-constriction data. For selected observations, the weights w_b in equation (3.3) are computed.

	MLE $b = \infty$	Huber $c(\beta,a)=0$ $b=3.7\ p^{1/2}$	Huber Conditional Unbiased $b=3.7p^{1/2}$	Huber Conditional Unbiased $b=3.2p^{1/2}$
Intercept	-2.92 (1.29)	-5.71 (2.45)	-2.98 (1.35)	-6.41 (2.84)
log(volume)	5.22 (1.93)	9.13 (3.73)	5.27 (1.93)	9.98 (4.38)
log(rate)	4.63 (1.79)	8.09 (3.31)	4.67 (1.86)	8.85 (3.82)
Weights				
#4		0.38	>.80	0.25
#18		0.44	>.80	0.29

Table 2

Maximum likelihood and robust estimators for the food stamp data. For selected observations, the weights w_b in equation (3.3) are computed.

		Huber	Huber	Huber	Hampel
	MLE	$c(\beta,a)=0$	Conditional	Conditional	Conditional
			Unbiased	Unbiased	Unbiased
	$b=\infty$	$b=3.5p^{1/2}$	$b=3.5p^{1/2}$	$b=2.75p^{1/2}$	bends at(3,7,16)$p^{1/2}$
Intercept	0.93	4.26	4.51	5.49	6.00
	(1.62)	(2.55)	(2.54)	(2.66)	(2.76)
Tenancy	-1.85	-1.85	-1.78	-1.76	-1.80
	(.53)	(.54)	(.54)	(.51)	(.54)
Supplemental Income	0.90	0.75	0.74	0.62	0.70
	(.50)	(.52)	(.51)	(.52)	(.52)
Log(1+MI), MI=Monthly Income	-0.33	-0.89	-0.93	-1.10	-1.18
	(.27)	(.43)	(.43)	(.45)	(.47)
Weights					
#5		0.21	0.16	0.13	0.0
#66		0.76	0.60	0.41	0.54

Table 3

Effect of deletion of selected observations in the food stamp data: Changes in the estimates divided by the estimated standard deviations. Conditional unbiased, Huber type, $b = 2.75p^{1/2}$.

a) Deletion of No. 5

	B constant	B reestimated	Approximation by the influence function
intercept	0.48	0.61	0.47
tenancy	-0.05	-0.04	-0.04
suppl. income	-0.08	-0.06	-0.08
log(1+monthly income)	-0.51	-0.62	-0.47

b) Deletion of No. 66

	B constant	B reestimated	Approximation by the influence function
intercept	0.37	0.65	0.32
tenancy	0.26	0.29	0.24
suppl. income	0.05	-0.10	0.05
log (1+monthly income)	-0.40	-0.67	-0.33