ROBUST M-ESTIMATORS

by

Franco Peracchi*

Department of Economics
University of Califonia
Los Angeles

May 1988
UCLA Working Paper No. 477

* This paper is based on a Chapter of my Ph.D. dissertation at Princeton University. I am very grateful to my advisors, Angus Deaton, Whitney Newey and Elvezio Ronchetti, for their generous guidance. I also thank Michele Boldrin, Hashem Pesaran and Sunil Sharma for helpful comments.

Abstract

This paper summarizes from an econometric perspective the influence function approach to robust estimation of parametric models. Hampel's optimality results for M-estimators with a bounded influence function is generalized to allow for arbitrary choices of the asymptotic efficiency criterion and the norm of the influence function. Further extensions to various cases of practical interst are also considered.

JEL No. 211

Keywords: M-estimators, Robust estimation, Influence Function.

1. Introduction

The basic aim of robust statistics is to develop procedures (estimators and tests) that perform well when the assumed model is correctly specified, while being relatively insensitive to small departures from the model assumptions. Robust and semi-parametric statistics are quite distinct. Robust statistics makes use of parametric assumptions, thereby attaining high efficiency at the assumed model, but deals in a formal way with the fact that these are almost never literally true. On the other hand, semi-parametric statistics places only weak, regularity-type assumptions on the statistical model, but does not address the question of what happens when these are not exactly true.

The definition of robustness that is probably most satisfactory theoretically is due to Hampel (1968, 1971), and formalizes the notion that a statistic T_N, indexed by the sample size N, is robust if small changes in the distribution of the observations have only small effects on the distribution of T_N [1]. More precisely, let $\mathcal{L}_F(T_N)$ be the distribution of T_N when F is the underlying distribution of the observations. Then the sequence $\{T_N\}$ is called qualitatively robust at the distribution F, if, for large enough N, the mapping $F \rightarrow \mathcal{L}_F(T_N)$ is continuous at F with respect to the topology of weak convergence. In Hampel's definition, the topology of weak convergence is induced by the Prohorov metric, which captures three main types of deviations from the model: (i) the occurrence of gross errors, (ii) the effects of rounding and grouping of the observation, (iii) the fact that the assumed model is at best an approximation to the unknown 'true' model.

The two main approaches to robust statistics are Huber's minimax

approach [Huber (1964, 1981)] and Hampel's infinitesimal approach [Hampel (1968), Hampel et al. (1986)]. Both approaches assume a parametric model for the observations and then try to construct procedures that 'do well' over a 'neighborhood' of the assumed model. The optimality problem of parametric statistics is modified by introducing, in addition to the classical consistency and efficiency requirements, a robustness condition that refers to the behavior of a procedure in a neighborhood of the assumed model. The exact formulation of the optimality problem depends on the class of procedures, the type of neighborhoods of the assumed model, the efficiency criterion and the robustness condition. In general, the optimal solution entails a trade-off between efficiency and robustness [2].

Huber's approach is to consider a certain neighborhood of the assumed parametric model, and then to safeguard within that neighborhood in a minimax sense. As an example, consider the problem of estimating the mean of a distribution that is only known to be in a given neighborhood of symmetric distributions around the Gaussian model. Since no bias arises, the performance of an estimator can simply be measured by its asymptotic variance. Huber (1964) represented this situation as a game in which 'Nature' picks a distribution so as to minimize the Fisher information in the chosen neighborhood, and the econometrician picks an estimator so as to minimize the maximum asymptotic variance. Huber was able to characterize the 'least informative' distribution. The minimax strategy for the econometrician is the maximum likelihood (ML) estimator based on the least informative distribution. This estimator, called the Huber estimator of location, is essentially a trimmed mean with the trimming proportion depending on the data and the size of the neighborhood. The generalization to the classical linear regression model

is straightforward [see Huber (1973)].

Huber's approach is difficult to extend to the case of non-symmetric distributions, and even in the symmetric case it leads to estimators that are inadmissible [3]. We shall therefore follow the infinitesimal approach to robustness, by focusing on the asymptotic behavior of an estimator in an infinitesimal neighborhood of a given model. The influence function (IF) [Hampel (1968, 1974)] provides a useful description of this behavior. On the one hand, the IF measures the effect on the asymptotic value of an estimator of an arbitrarily small contamination of the assumed model for the observations. On the other hand, it provides information on the asymptotic variance of the estimator.

In this paper we consider a general class of estimators, namely the class of M- (or generalized ML) estimators, defined as roots of an implicit equation. This class of estimators includes most econometric estimators, such as ML, least squares, least absolute deviations, generalized method of moments and certain minimum distance estimators. For M-estimators, Hampel's definition of qualitative robustness is equivalent to the IF being bounded and continuous. An estimator is called B- (or bias-) robust if it has a bounded IF. B- robustness is a desirable property, for it ensures protection against the bias that may arise because of small failures of the model assumptions. Moreover, the sup-norm of the IF, called the estimator's sensitivity, provides a natural quantitative measure of robustness.

An estimator that is B-robust will generally be less efficient than the ML estimator based on a correctly specified model. This is because B-robustness requires the IF to be bounded, while efficiency requires the IF to be equal to the likelihood score (up to a linear transformation).

Since the latter is often unbounded, B-robustness necessarily implies a loss of efficiency. The infinitesimal approach aims at minimizing this efficiency loss by constructing procedures that are 'optimally robust', i.e. asymptotically efficient among all procedures with a bounded IF. An optimality result for M-estimators with a bounded IF was first proved by Hampel (1968). Given a parametric model indexed by a one-dimensional parameter θ, he constructed an optimal B-robust estimator of θ by censoring and recentering the likelihood score so as to satisfy the B-robustness constraint and ensure consistency at the assumed model. The resulting estimator can be interpreted as a weighted ML estimator. In the location model with Gaussian errors, this estimator coincides with Huber's minimax estimator.

The generalization of Hampel's result to the case of multi-dimensional estimators is not trivial. In this paper we show that given an arbitrary mean square error (MSE) criterion and a bound on the gross-error-sensitivity, it is possible under certain conditions to construct a B-robust estimator that is consistent and has minimum asymptotic MSE at the assumed model among all the consistent estimators that satisfy the same sensitivity bound. Such an estimator is obtained by applying to the likelihood score a matrix weight function that depends on the metrics in which the MSE criterion and the gross-error-sensitivity are defined. Our results contain as special cases the ones of Krasker (1980) and Krasker and Welsch (1982) for the linear regression model, Fraiman (1983) for the non-linear regression model, Stefanski, Ruppert and Carroll (1986) for the logit model, and Hampel et al. (1986). The statistical problem is formulated explicitly as a minimum norm problem in a Hilbert space. This clarifies the Lagrange multiplier interpretation of some of the elements

of the solution and allows generalizations to three cases of practical interest. The first case is when the IF of an estimator is the product of two components, and we allow for different sensitivity bounds on each of them. The second is when the distribution of the observations can be factorized as the product of a conditional and a marginal density, but we only care about misspecification of the former. The third is the partitioned parameter case, when we allow for different sensitivity bounds on each subset of estimates.

The rest of the paper is organized as follows. Section 2 introduces the statistical model and the class of M-estimators. Section 3 collects a few results that relate the asymptotic properties of an estimator to the properties of its IF. Section 4 presents our main result. Section 5 contains a number of generalizations. Section 6 contains some final remarks.

The following notation will be used. F_0 denotes the true distribution function (d.f.) of a single observation and E_0 denotes expectations taken with respect to F_0. Expectations taken with respect to some d.f. in a parametric family $\{F_\theta\}$ are denoted by E_θ, and expectations taken with respect to any other d.f. F by E_F. $\|x\| = (x'x)^{1/2}$ denotes the Euclidean norm of a finite dimensional vector x, and $\|x\|_B = (x'B\,x)^{1/2}$ the norm of x in the metric of the positive definite (p.d.) matrix B.

2. The statistical model and the class of M-estimators

Let $z_1,\ldots,z_N$ be a sequence of independently and identically distributed (i.i.d.) [4] random vectors, with values in a known subset $\mathcal{Z}$ of $\mathbb{R}^m$, and

common d.f. F_0. The d.f. F_0 is typically unknown, but prior information may be available to justify the assumption that F_0 belongs to some set $\mathcal{F}$ of d.f.'s over $\mathcal{Z}$, called the assumed model. $\mathcal{F}$ is either a family of d.f.'s indexed by a p-dimensional parameter θ, i.e. $\mathcal{F} = \{F_\theta: \theta\in\Theta\}$, or a set of d.f.'s satisfying a number $s \geq p$ of moment restrictions, i.e. $\mathcal{F} = \{F: E_F\ \psi(z,\theta) = 0,\ \theta\in\Theta\}$, where $\psi: \mathbb{R}^m \times \Theta \rightarrow \mathbb{R}^s$. In either case, the parameter space Θ is assumed to be an open subset of $\mathbb{R}^p$. For F_0 to be identifiable, it is further necessary to assume that there is a unique $\theta_0\in\Theta$ such that either $F_0 = F_{\theta_0}$ or $E_0\ \psi(z,\theta_0) = 0$.

In fact, the assumed model $\mathcal{F}$ may be misspecified, in the sense that it may not contain F_0. In this paper we focus on the case when the assumed model is only approximately true. This is modelled by assuming that $F_0 = (1 - \varepsilon)\ F + \varepsilon\ G$, where $F\in\mathcal{F}$, G is some other unknown d.f. and $\varepsilon\in[0,1]$ [5]. When ε is small, this is called an 'ε-contamination model', and formalizes the notion that the assumed model may be adequate for the majority but not all the observations. In particular, when G is a d.f. with mass concentrated at a point $z\in\mathcal{Z}$, we obtain the so-called 'gross-error model'. This is a convenient way of modelling the occurrence of outliers or gross errors.

An M-estimator $\hat{\theta}_N$ of the unknown parameter θ_0 is a root of an implicit equation of the form

$$N^{-1} \Sigma_{n=1}^{N} \eta_N(z_n,\theta) = 0, \tag{1}$$

where $\eta_N(z,\theta)$, called the score function associated with $\hat{\theta}_N$, is some function mapping $\mathcal{Z} \times \mathbb{R}^p$ into $\mathbb{R}^p$. If (1) has multiple roots, we assume

that some selection rule has been specified. Equation (1) often arises as the necessary condition for optimality in problems of the form [6]

$$\text{Max}_{\theta \in \Theta_0} \; N^{-1} \sum_{n=1}^{N} \rho_N(z_n, \theta), \tag{2}$$

where Θ_0 is a subset of $\mathbb{R}^p$ assumed to contain θ_0 in its interior, and $\rho_N(z,\theta)$ is some real valued function defined on $\mathcal{Z} \times \mathbb{R}^p$ and almost everywhere continuously differentiable with respect to θ. In this case $\eta_N(z,\theta) = (\partial/\partial\theta)\, \rho_N(z,\theta)$, except possibly for a countable set of points.

The class of M-estimators is very large and includes most common econometric estimators. ML, pseudo-ML [see e.g. Gourieroux, Monfort and Trognon (1984)], least squares, least absolute deviations, generalized method of moments (GMM) [see e.g. Burguete, Gallant and Souza (1982) and Hansen (1982)], and the minimum distance estimators of Malinvaud (1970) and Chamberlain (1982) are all members of this class. For example, if $\mathcal{F}$ is a regular parametric model with density $f(z,\theta)$ and likelihood score $s(z,\theta) = (\partial/\partial\theta) \ln f(z,\theta)$, then $\eta_N(z,\theta) = s(z,\theta)$ and $\hat{\theta}_N$ is a ML estimator. If $\mathcal{F} = \{F: E_F\, \psi(z,\theta) = 0\}$, $\tilde{Q}_N$ is an s×s p.d. matrix, and

$$\rho_N(z,\theta) = - [N^{-1} \sum_{n=1}^{N} \psi(z_n,\theta)]'\, \tilde{Q}_N^{-1}\, \psi(z,\theta),$$

then $\hat{\theta}_N$ is a GMM estimator, obtained by minimizing the squared norm of the sample average of $\psi(z,\theta)$ in the metric of the (possibly data dependent) matrix $\tilde{Q}_N^{-1}$. When $\psi(z,\theta)$ is differentiable with respect to θ, the GMM score function is proportional to $\eta(z,\theta) = \tilde{P}_N(\theta)'\, \tilde{Q}_N^{-1}\, \psi(z,\theta)$, where $\tilde{P}_N(\theta) = N^{-1} \sum_{n=1}^{N} (\partial/\partial\theta')\, \psi(z_n,\theta)$ is an s×p matrix. Thus, in the GMM

case the score function depends on the sample size N through the matrices $\tilde{Q}_N$ and $\tilde{P}_N(\theta)$. However, when both $\tilde{Q}_N$ and $\tilde{P}_N$ converge almost surely (a.s.) to p.d. matrices, $\eta_N(z,\theta)$ converges a.s. to a function that does not depend on N.

Consistency and asymptotic normality of M-estimators can be established under very general conditions [7]. Moreover, as we shall see, estimators that are not in this class are often asymptotically equivalent to M-estimators.

3. The influence function and related concepts

Because of the independence assumption, the order of the observations in the sample does not matter. One can therefore replace functions of the observations by statistical functionals, i.e. functionals defined over a set of d.f.'s. In particular, if the score function $\eta_N(z,\theta)$ does not depend on N, equation (1) and the given selection rule implicitly define a functional $\hat{\theta}$: $F_N \rightarrow \hat{\theta}_N = \hat{\theta}(F_N)$, where F_N denotes the empirical d.f. of the observations. More generally, if the score function $\eta_N(z,\theta)$ has an a.s. limit $\eta(z,\theta)$, then we can associate with $\eta(z,\theta)$ the functional $\hat{\theta}$: $F \rightarrow \hat{\theta}(F)$ implicitly defined by

$$\int_{\mathcal{Z}} \eta(z,\hat{\theta}(F))\, dF(z) = 0, \tag{3}$$

for all $F \in \mathcal{F}'$. We assume that the domain $\mathcal{F}'$ of the functional is a convex set containing the true d.f. F_0, the assumed model $\mathcal{F}$ and all empirical d.f.'s over $\mathcal{Z}$. Convexity of $\mathcal{F}'$ is needed because the statistical

functional $\hat{\theta}$ is to be evaluated at 'ε-contamination models' of the form $(1 - \varepsilon)\ F + \varepsilon\ G$, $\varepsilon \in [0,1]$. We also assume that the function $\eta(z,\theta)$ is continuously differentiable with respect to θ, except possibly at a countable number of points, and the matrix $P(\hat{\theta},F) = - E_F\ (\partial/\partial\theta)\ \eta(z,\hat{\theta}(F))$ is finite and p.d. for all d.f.'s F in a neighborhood of F_0.

In the case of a statistical functional it is more reasonable to adopt a definition of consistency that differs slightly from the usual definition of weak consistency [see e.g. Cox and Hinkley (1974)].

DEFINITION 1: $\hat{\theta}$ is called Fisher-consistent for θ_0 if $\hat{\theta}(F_0) = \theta_0$ whenever $F_0 \in \mathcal{F}$.

Thus, an M-estimator defined by (3) is Fisher consistent for θ_0 whenever the assumed model $\mathcal{F}$ is correctly specified and the equation $E_0\ \eta(z,\theta) = 0$ has a unique root at $\theta = \theta_0$. For ML estimators of regular parametric models, the latter condition corresponds to the standard assumption that the likelihood function has a unique global maximum at θ_0. If the assumed model is misspecified, $\hat{\theta}(F_0)$ defines the population parameter estimated by $\hat{\theta}$.

Given a statistical functional $\hat{\theta}$, it is natural to investigate its continuity and differentiability. In the remainder of this Section we present rather informally a number of results that relate these properties to the asymptotic behavior of $\hat{\theta}$.

A statistical functional $\hat{\theta}$ need not be linear. However, when it is linear, that is, $\hat{\theta}(F) = E_F\ \psi(z)$ for some function ψ, its asymptotic properties are easy to establish. This suggests deriving the asymptotic properties of a non-linear functional by means of a suitable

linearization. Under regularity conditions [see e.g. Serfling (1980)], a functional $\hat{\theta}$ possesses the (von Mises) expansion

$$\hat{\theta}(F_N) - \hat{\theta}(F_0) = N^{-1} \Sigma_{n=1}^{N} IF(z_n,\hat{\theta},F_0) + R_N(F_0) \qquad (4)$$

where $IF(\cdot,\hat{\theta},F_0)$ is a function that depends only on $\hat{\theta}$ and F_0 and $R_N(\cdot)$ is a remainder term. The function $IF(\cdot,\hat{\theta},F_0)$ is called the influence function (IF) of $\hat{\theta}$ at F_0 [Hampel (1974)] [8]. A more explicit definition of the IF is the following:

DEFINITION 2: Let $\Delta_{(z)}$ be a d.f. with mass concentrated at the point z and let $F_{\varepsilon,z} = (1 - \varepsilon) F + \varepsilon \Delta_{(z)}$. Then the IF of $\hat{\theta}$ at F is defined by

$$IF(z,\hat{\theta},F) = \lim_{\varepsilon \to 0+} [\hat{\theta}(F_{\varepsilon,z}) - \hat{\theta}(F)]/\varepsilon,$$

provided that the limit exists.

Thus, the IF is the collection of Gateaux differentials of the functional $\hat{\theta}$ in the direction of point mass distributions [9], and can therefore be interpreted as a measure of the asymptotic bias of $\hat{\theta}$, as an estimator of $\hat{\theta}(F)$, under an arbitrarily small contamination of the d.f. F by a point mass [10].

If $\hat{\theta}$ is an M-estimator defined by an equation of the form (3), the matrix $P(\hat{\theta},F)$ exists and is p.d, and the d.f. F satisfies regularity conditions sufficient to allow interchanging the order of differentiation and integration, then it follows from the Implicit Function Theorem applied to (3) that the IF of $\hat{\theta}$ at F exists and is equal to

$$IF(z,\hat{\theta},F) = P(\hat{\theta},F)^{-1}\ \eta(z,\hat{\theta}(F)), \tag{5}$$

that is, the IF of $\hat{\theta}$ at F is just a non-singular linear transformation of the score function that defines $\hat{\theta}$.

In particular, suppose that F_θ belongs to a regular parametric model with p.d. Fisher information matrix $J(\theta)$. If $\hat{\theta}$ is the ML estimator, then $P(\hat{\theta},F_\theta) = J(\theta)$ and so $IF(z,\hat{\theta},F_\theta) = J(\theta)^{-1}\ s(z,\theta)$. If $\hat{\theta}$ is a general M-estimator then

$$P(\hat{\theta},F_\theta) = E_\theta\ \eta(z,\theta)\ s(z,\theta)', \tag{6}$$

which generalizes the familiar equality between the information matrix of a correctly specified model and the expectation of the outer product of the likelihood score.

Now suppose that the domain of the functional $\hat{\theta}$ is equipped with some norm $\delta(\cdot)$. If $\hat{\theta}$ has a differential at G with respect to the given norm, then it can be shown that

$$\hat{\theta}(G) - \hat{\theta}(F) = E_G\ IF(z,\hat{\theta},F) + o(\delta(G - F)). \tag{7}$$

For a proof, see Serfling (1980). Thus, if G is some d.f. near F, the asymptotic bias of $\hat{\theta}$, as an estimator of $\hat{\theta}(F)$, can be approximated by $E_G\ IF(z,\hat{\theta},F)$ [11]. Notice that (7) implies that $E_F\ IF(z,\hat{\theta},F) = 0$.

As an illustration, consider the problem of evaluating the asymptotic bias of the Probit estimator under heteroskedasticity. Let x be a random vector with d.f. H and let y^* be a random variable that is $N(x'\beta,\ \sigma^2)$ conditionally on x_n. Denote by ϕ and Φ respectively the $N(0,1)$

density and d.f. Also denote by F_θ the joint d.f. of $z = (y, x')'$, where $\theta = \beta/\sigma$ and $y = 1(y^* > 0)$. If $\hat{\theta}$ is the Probit estimator of θ then, by (5), $IF(z,\hat{\theta},F) = J(\theta)^{-1}\ \lambda(x'\theta)\ [(y - \Phi(x'\theta)]\ x$, where $J(\theta)$ denotes the information matrix for θ and $\lambda(r) = \phi(r)/[\Phi(r)\ (1 - \Phi(r))]$. Clearly, $\hat{\theta}(F_\theta) = \theta$. Now let G be the d.f. of a single observation z when half of the observations on y^* are $N(x'\beta, \sigma^2)$ and half are $N(x'\beta, \omega^2)$. This kind of heteroskedasticity has been considered by Kiefer and Skoog (1984). By (7), the bias of the Probit estimator can be approximated by

$$\hat{\theta}(G) - \theta \simeq (1/2)\ J(\theta)^{-1}\ E_H\ [\lambda(x'\theta)\ (\Phi(x'\alpha\theta) - \Phi(x'\theta))\ x]$$

where $\alpha = \sigma/\omega$. For some distributions of the regressors this bias can be quite large. The approximation proposed by Kiefer and Skoog (1984), which is given by $\theta(\alpha - 1)/2$, is simply a linearization of the IF approximation about $\alpha = 1$.

As another example, consider the problem of evaluating the maximum asymptotic bias of $\hat{\theta}$, as an estimator of $\hat{\theta}(F)$, under the 'gross-error model' $F_{\varepsilon,z} = (1 - \varepsilon)\ F + \varepsilon\ \Delta_{(z)}$. From (6)

$$\sup_{z\in\mathcal{Z}} \|\hat{\theta}(F_{\varepsilon,z}) - \hat{\theta}(F)\| \simeq \varepsilon \sup_{z\in\mathcal{Z}} \|IF(z,\hat{\theta},F)\|$$

provided that ε is small. For a more general 'ε-contamination model' $F_{\varepsilon,G} = (1 - \varepsilon)\ F + \varepsilon\ G$, $G\in\mathcal{G}$, we have that

$$\sup_{G\in\mathcal{G}} \|\hat{\theta}(G) - \hat{\theta}(F)\| \leq \varepsilon \sup_{z\in\mathcal{Z}} \|IF(z,\hat{\theta},F)\|. \qquad (8)$$

Thus, the right hand side of (8), if finite, gives an approximate upper

bound on the asymptotic bias of $\hat{\theta}$ over a sufficiently small 'ε-contamination neighborhood' of F. This justifies the following definition:

DEFINITION 3: An estimator $\hat{\theta}$ is called B- (or bias-) robust at the d.f. F if $IF(\cdot,\hat{\theta},F)$ is a bounded function.

B-robustness is a desirable property, for it ensures protection against the effects of local failures of the model assumptions. However, many common econometric estimators, such as least squares and instrumental variable estimators do not have a bounded IF and therefore are not B-robust, because their IF is unbounded. For M-estimators with a continuous IF, B-robustness is equivalent to Hampel's qualitative robustness [see Huber (1981)].

A natural quantitative measure of robustness is given by the sup-norm of the IF, in the metric of some p.d. matrix B,

$$\gamma^*(\hat{\theta},F,B) = \sup_{z\in\mathfrak{X}} \|IF(z,\hat{\theta},F)\|_B,$$

called the estimator's sensitivity [12]. The choice of B is largely arbitrary. For example, if $B = I_p$ one obtains Hampel's (1974) unstandardized gross-error sensitivity. The self-standardized sensitivity of Krasker and Welsch (1982) corresponds to $B = AV(\hat{\theta},F)^{-1}$. The information-standardized sensitivity of Hampel et al. (1986) corresponds to $B = J(\theta)$, but is only defined when $\mathcal{F}$ is a parametric model. Unlike the unstandardized sensitivity, the last two measures are invariant to non singular reparameterizations of the model.

Now consider the asymptotic distribution of a functional $\hat{\theta}$ that possesses an expansion of the form (4). Assume that $\hat{\theta}$ satisfies regularity conditions sufficient to ensure that the remainder term $R_N(F_0)$ in (4) is $o_p(N^{-1/2})$ and the matrix $E_0\ IF(z,\hat{\theta},F_0)\ IF(z,\hat{\theta},F_0)'$ is finite and p.d. Since $N^{-1}\ \Sigma_{n=1}^{N}\ IF(z_n,\hat{\theta},F_0)$ is an average of i.i.d. random vectors with zero mean and finite, p.d. variance, it follows that

$$N^{1/2}\ [\hat{\theta}(F_N) - \hat{\theta}(F_0)] \xrightarrow{d} N(0,AV(\hat{\theta},F_0)), \tag{9}$$

where $AV(\hat{\theta},F_0) = E_0\ IF(z,\hat{\theta},F_0)\ IF(z,\hat{\theta},F_0)'$. A sufficient condition for the remainder term in (4) to be $o_p(N^{-1/2})$ is some form of differentiability of the functional $\hat{\theta}$, such as Frechét differentiability [see Serfling (1980) and Huber (1981)] or the weaker Hadamard differentiability [see Fernholz (1983) and Prakasa Rao (1987)].

In particular, when $\hat{\theta}$ is an M-estimator one obtains that $AV(\hat{\theta},F_0) = P(\hat{\theta},F_0)^{-1}\ Q(\hat{\theta},F_0)\ P(\hat{\theta},F_0)'^{-1}$, where $Q(\hat{\theta},F_0) = E_0\ \eta(z,\hat{\theta}(F_0))\ \eta(z,\hat{\theta}(F_0))'$. If $\hat{\theta}$ is the ML estimator of a regular parametric model then, provided that the model is correctly specified, $P(\hat{\theta},F_0) = Q(\hat{\theta},F_0) = J(\theta_0)$ and one obtains the familiar result that $AV(\hat{\theta},F_0) = J(\theta_0)^{-1}$. In the GMM case, if $\hat{\theta}$ is the estimator based on the weighting matrix $\tilde{Q}(F_0) = E_0\ [\psi(z,\hat{\theta}(F_0))\ \psi(z,\hat{\theta}(F_0))']$, then $P(\theta,F_0) = Q(\hat{\theta},F_0) = \tilde{P}(\hat{\theta},F_0)'\ \tilde{Q}(F_0)^{-1}\ \tilde{P}(\hat{\theta},F_0)$, and therefore $AV(\hat{\theta},F_0) = [\tilde{P}(\hat{\theta},F_0)'\ \tilde{Q}(F_0)^{-1}\ \tilde{P}(\hat{\theta},F_0)]^{-1}$. Such an estimator has minimum asymptotic variance matrix (AVM) in the class of GMM estimators based on the moment restriction $E_0\ \psi(z,\theta_0) = 0$ [see e.g. Hansen (1982)].

A consistent estimator of the AVM of $\hat{\theta}$ can be obtained by replacing F_0 with the empirical d.f. of the observations. A consistent estimator of $AV(\hat{\theta},F_0)$ is therefore given by $AV(\hat{\theta},F_N) = N^{-1}\ \Sigma_{n=1}^{N}\ IF(z_n,\hat{\theta},F_N)$

$IF(z_n,\hat{\theta},F_N)'$. When $\hat{\theta}$ is an M-estimator, $IF(z,\hat{\theta},F_N) = [-N^{-1}\sum_{n=1}^{N}(\partial/\partial\theta')\eta(z_n,\hat{\theta}_N)]^{-1}\eta(z,\hat{\theta}_N)$, and $AV(\hat{\theta},F_N)$ is just the covariance estimator proposed by White (1982).

It is clear from (5) and (9) that an estimator $\hat{\theta}$ which possesses an IF at F has the same IF and hence the same asymptotic distribution as an M-estimator based on the score function $\eta(z,\theta) = IF(z,\hat{\theta},F)$. Thus, if interest focuses only on asymptotic properties, such as the asymptotic bias and the AVM, there is no loss of generality in considering only M-estimators.

4. Optimal robust estimators

The ML estimator of a correctly specified model is asymptotically efficient, that is, the AVM of any other consistent estimator differs from the AVM of the ML estimator by a positive semi-definite matrix. The ML estimator, however, can loose its optimality properties for very small departures from the assumed parametric model. If its IF is unbounded, as typically occurs when the assumed model is Gaussian, the asymptotic bias that arises under certain types of misspecification can be quite large. This is not the case for a B-robust estimator, even though such an estimator will generally be less efficient than the ML estimator based on a correctly specified model. Hampel (1968) proposed to minimize this efficiency loss by considering estimators that are 'optimal B-robust', that is, consistent and asymptotically efficient at the assumed model among all B-robust estimators that satisfy a given sensitivity bound.

In what follows we consider a generalization of Hampel's approach to

the case of multidimensional estimators. We restrict attention to 'regular' M-estimators, namely the ones that are Fisher consistent at the assumed model, possess an IF, and have an asymptotically normal distribution with a finite, p.d. AVM. The assumed model is a regular parametric model $\{F_\theta\}$, and efficiency of estimation is defined in terms of an asymptotic mean square error (MSE) criterion of the form $MSE(\hat{\theta},F) = \text{trace}\,[Q\ AV(\hat{\theta},F)]$, where Q is some p.d. matrix. Given a bound $\gamma^*(\theta,F,B) \leq \gamma$ on an estimator's sensitivity, we consider the problem of finding an 'optimal B-robust' estimator $\tilde{\theta}$, namely one that has minimum asymptotic MSE at the assumed model among all M-estimators that are Fisher consistent at the assumed model and satisfy the given sensitivity bound [13]. Our formulation of the problem builds on earlier work of Hampel (1968), Krasker (1980), Bickel (1981, 1984) and Hampel et al. (1986), but is more general, for it allows the MSE criterion and the B-robustness constraint to be defined with respect to arbitrary and possibly different metrics. As we shall see, this has important implications for the form of the optimal solution.

In order to obtain a tractable problem we need to ensure that there is a one-to-one correspondence between the set of M-estimators and the set of score functions. We do so by imposing a 'normalization' conditions on the score function $\eta(z,\theta)$ that defines $\hat{\theta}$, namely $\eta(z,\theta)$ must be such that $E_F\ [-\ (\partial/\partial\theta')\ \eta(z,\hat{\theta}(F)] = I_p$. Under this condition, the IF of $\hat{\theta}$ is given by $\eta(z,\hat{\theta}(F))$ and its AVM by $E_F\ \eta(z,\hat{\theta}(F))\ \eta(z,\hat{\theta}(F))'$. When the assumed model $\{F_\theta\}$ is correctly specified, an optimal B-robust estimator of the unknown parameter θ_0 can then be based on the score function that solves the following problem

$$\text{Min}_{\eta(\cdot,\theta)\in\tilde{H}} \; E_\theta \; \eta(z,\theta)' \; Q \; \eta(z,\theta) \tag{10}$$

$$\text{s.t} \quad E_\theta \; \eta(z,\theta) = 0 \tag{11}$$

$$E_\theta \; \eta(z,\theta) \; s(z,\theta)' = I_p \tag{12}$$

$$\sup_{z\in\mathcal{Z}} \|\eta(z,\theta)\|_B \leq \gamma \tag{13}$$

for all θ in a neighborhood of θ_0. For a given θ, the set $\tilde{H}$ is the set of score functions that are square integrable with respect to F_θ. Constraint (11) ensures Fisher-consistency. Because of (6), constraint (12) corresponds to the normalization condition discussed earlier. Given (11) and (12), the objective functional (10) is the asymptotic MSE of the M-estimator based on the score function $\eta(z,\theta)$. Finally, constraint (13) is the bound on the estimator's sensitivity. When γ is not finite, problem (10)-(13) essentially reduces to the optimality problem of classical parametric statistics.

The set $\tilde{H}$ is clearly a Hilbert space with respect to the inner product $\langle\eta,\psi\rangle = E_\theta \; \eta(z)' \; Q \; \psi(z)$. The induced norm of a vector $\eta\in\tilde{H}$ is given by $[E_\theta \; \eta(z)' \; Q \; \eta(z)]^{1/2}$. The set of points in $\tilde{H}$ that satisfy (13) is convex. It may be viewed, heuristically, as a closed hypersphere with center at the origin and radius γ. The feasible set, being the intersection between this set of points and the hyperplane defined by constraints (11) and (12), is closed and convex but can be empty. It will be seen that the feasible set is not empty provided that γ is large enough. For a given θ, problem (10)-(13) is therefore one of finding a vector of minimum norm in a closed and convex subset of a Hilbert space. It follows by standard results [see e.g Luenberger (1969)] that, if the feasible set is not empty, an optimal solution exists, is unique and can

be characterized by a saddle point of the Lagrangean

$$E_\theta \Big[\eta(z,\theta)' Q \, \eta(z,\theta) + \lambda' \, \eta(z,\theta) + (\text{vec } \Gamma)' \{\text{vec } I_p - [s(z,\theta) \otimes I_p] \, \eta(z,\theta)\} + \mu(z) \, (\|\eta(z,\theta)\|_B - \gamma) \Big],$$

where λ is a p×1 non negative vector of Lagrange multipliers associated with the constraint (11), Γ is a p×p positive semi-definite matrix of Lagrange multipliers associated with the constraint (12), and the non negative function $\mu(\cdot)$ is the Lagrange multiplier associated with the constraint (13). Notice that all Lagrange multipliers are functions of θ, but this dependence is omitted for simplicity. The optimal solution $(\tilde{\eta}(\cdot,\theta), \tilde{\lambda}, \tilde{\Gamma}, \tilde{\mu}(\cdot))$ is characterized by (11)-(13) and

$$2 \, Q \, \tilde{\eta}(z,\theta) + \tilde{\lambda} - \tilde{\Gamma} \, s(z,\theta) + \tilde{\mu}(z) \, B \, \frac{\tilde{\eta}(z,\theta)}{\|\tilde{\eta}(z,\theta)\|_B} = 0$$

$$\tilde{\mu}(z) \, (\|\tilde{\eta}(z,\theta)\|_B - \gamma) = 0$$

for all z, except possibly a set with measure zero under F_θ. If $\|\tilde{\eta}(z,\theta)\|_B < \gamma$ for some z, then $\tilde{\mu}(z) = 0$ and $\tilde{\eta}(z,\theta) = A \, s(z,\theta) - a$, where $A = (1/2) \, Q^{-1} \, \tilde{\Gamma}$ and $a = (1/2) \, Q^{-1} \, \tilde{\lambda}$. If $\tilde{\mu}(z) > 0$ at some point z, then $\|\tilde{\eta}(z,\theta)\|_B = \gamma$ and

$$\tilde{\eta}(z,\theta) = [I_p + (\tilde{\mu}(z)/2\gamma) \, Q^{-1} \, B]^{-1} \, [A \, s(z,\theta) - a],$$

where the Lagrange multiplier $\tilde{\mu}(z)$ is given by the implicit equation

$$\gamma - \|[I_p + (\tilde{\mu}(z)/2\gamma)\ Q^{-1}\ B]^{-1}\ [A\ s(z,\theta) - a]\|_B = 0. \qquad (14)$$

The optimal score function is therefore given by

$$\tilde{\eta}(z,\theta) = W(z,\theta)\ [A\ s(z,\theta) - a] \qquad (15)$$

where the p×p p.d. matrix matrix $W(z,\theta)$ is given by

$$W(z,\theta) = \begin{cases} I_p & \text{if } \|A\ s(z,\theta) - a\|_B \le \gamma \\ [I_p + (\tilde{\mu}(z)/2\gamma)\ Q^{-1}\ B]^{-1} & \text{otherwise,} \end{cases}$$

and $\tilde{\mu}(z)$ is a non negative root of (14). The vector a and the matrix A, are determined implicitly by the constraints (11) and (12). The existence of the optimal score function depends therefore on the existence of a solution to the equations system

$$E_\theta\ W(z,\theta)\ A\ s(z,\theta) - [E_\theta\ W(z,\theta)]\ a = 0 \qquad (16)$$

$$E_\theta\ W(z,\theta)\ [A\ s(z,\theta) - a]\ s(z,\theta)' - I_p = 0, \qquad (17)$$

where $W(z,\theta)$ depends on (a, A). If a solution $(a(\theta), A(\theta))$ exists for all θ in an open neighborhood Θ_0 of θ_0 (necessary conditions are provided below), the function $\tilde{\eta}(\cdot,\theta)$: $\mathfrak{X} \longrightarrow \mathbb{R}^p$ can be extended to a function $\tilde{\eta}$: $\mathfrak{X} \times \Theta_0 \longrightarrow \mathbb{R}^p$. The resulting estimator will be denoted by $\tilde{\theta}$. We summarize this result in the following:

PROPOSITION 1: Let $\tilde{\theta}$ be the M-estimator of θ based on the score

function $\tilde{\eta}$ defined by (15), and assume that $a(\theta)$ and $A(\theta)$, implicitly defined by equations (16) and (17), exist for all θ in an open neighborhood of θ_0. Then $\tilde{\theta}$ minimizes trace $[Q\ AV(\hat{\theta},F_\theta)]$ among all regular M-estimators satisfying $\gamma^*(\hat{\theta},F_\theta,B) \leq \gamma$.

Notice that, unlike the ML estimator, $\tilde{\theta}$ is optimal only in a weak sense, namely with respect to the given MSE criterion. This implies that tests based on optimal B-robust estimators are robust but not invariant, that is, their power function is stable under small departures from the assumed model but they have optimal power properties only for certain departures from the null hypothesis [Peracchi (1987)].

Since $W(z,\theta)$ is a continuous function, the optimal B-robust estimator $\tilde{\theta}$ is qualitatively robust provided that the likelihood score function is continuous. Consistency and asymptotic normality of $\tilde{\theta}$ must be established formally. Consistency follows by standard methods [see e.g. Amemiya (1985)]. Asymptotic normality requires less standard techniques because the score function (15) is not differentiable at the points where $\|A(\theta)\ s(z,\theta) - a(\theta)\|_B = \gamma$. Usually, asymptotic normality can be established by verifying the sufficient conditions of Huber (1967). This is generally straightforward, because the boundedness of the optimal score ensures that the appropriate dominance conditions are satisfied.

The optimal B-robust estimator $\tilde{\theta}$ can be interpreted as a weighted ML estimator, obtained by applying the matrix $W(z,\theta)\ A(\theta)$ to the likelihood score, and subtracting off the vector $W(z,\theta)\ a(\theta)$ in order to correct for the bias. Geometrically, this corresponds to shrinking and twisting the likelihood score vector until it is entirely contained in a p-dimensional hypersphere with center at the origin and radius γ, at the

same time satisfying the condition for consistency at the assumed model.

For given matrices B and Q, Proposition 1 generates a whole family of estimators indexed by the value of the sensitivity bound γ. If there is no bound on an estimator's sensitivity the optimal estimator is the ML estimator, which always exists and is unique under the stated assumptions. In this case, the constant $a(\theta)$ is equal to zero because the Fisher consistency constraint is not binding. Also $A(\theta) = J(\theta)^{-1}$ and so, when $Q = I_p$, the information matrix is equal to the inverse of the matrix of Lagrange multipliers associated with the constraint (13). By a simple continuity argument [see Proposition 1 in Krasker (1980)], the optimal estimator $\tilde{\theta}$ exists and is unique, and the matrix $A(\theta)$ is p.d., when γ is finite and sufficiently large. Varying γ describes a trade-off between efficiency (relative to the Cramér-Rao bound) and protection against bias (the sup-norm of the IF). In some cases, such as linear regression, this trade-off can be characterized easily.

We now provide necessary (but not sufficient) conditions on the sensitivity bound γ for $a(\theta)$ and $A(\theta)$, and hence $\tilde{\theta}$, to exist. All proofs are given in the Appendix.

PROPOSITION 2: Let F_θ be such that $E_\theta \|s(z,\theta)\|_B$ exists. Then $a(\theta)$ and $A(\theta)$ exist only if

$$\gamma \geq \frac{\text{trace } B}{E_\theta \|s(z,\theta)\|_B}.$$

Geometrically, when $\gamma <$ trace B / $[E_\theta \|s(z,\theta)\|_B]$, the linear variety defined by the constraints (11)-(12) does not intersect the convex set

defined by (8), and so the feasible set is empty. Notice that the lower bound on γ depends on the choice of the metric B, as well as on θ. In practice, given this lower bound, the choice of γ depends on the econometrician's preferences between efficiency and protection against bias. When both the econometrician's preferences and the trade-off between efficiency and bias are known, solving the choice problem is straightforward.

The expression for the optimal score can be simplified considerably in the one-dimensional case or when the same metric is used for both the sensitivity and the asymptotic MSE criterion. The next result extends Theorem 4.3.1 of Hampel et al. (1986) to the case where the sensitivity is defined in an arbitrary metric.

PROPOSITION 3: If $Q = \alpha B$, $\alpha > 0$, the estimator that minimizes trace $[Q\ AV(\hat{\theta},F_\theta)]$ among all regular M-estimators satisfying $\gamma^*(\hat{\theta},F_\theta,B) \leq \gamma$ is based on a score function of the form (15), where $W(z,\theta) = w(z,\theta)\ I_p$ and the scalar weighting function $w(\cdot,\theta)$: $\mathfrak{X} \rightarrow [0,1]$ is defined by

$$w(z,\theta) = \min \{1,\ \gamma\ /\ \|A(\theta)\ s(z,\theta) - a(\theta)\|_B\}. \qquad (18)$$

In some cases the expression for the optimal score function simplifies even further. For example, if $Q = \alpha B$ and $a(\theta) = 0$, then $\tilde{\eta}(z,\theta) = w(z,\theta)\ A(\theta)\ s(z,\theta)$, where $w(z,\theta) = \min \{1,\ \gamma\ /\ \|A(\theta)\ s(z,\theta)\|_B\}$ and $A(\theta)$ is a symmetric p.d. matrix implicitly defined by the equation $E_\theta\ w(z,\theta)\ s(z,\theta)\ s(z,\theta)' - A(\theta)^{-1} = 0$. This case typically arises in regression models with symmetric conditional error distributions. Notice that the optimal B-robust estimator in this case is the same as the one

based on the score function $\eta^*(z,\theta) = w(z,\theta)\ s(z,\theta)$, which makes its interpretation as a weighted ML estimator even clearer.

Estimators based on Proposition 3 have been derived for a variety of models, including the linear regression model [Hampel (1978), Krasker (1980), Krasker and Welsch (1982)], the non-linear regression model [Fraiman (1984)], the SURE model [Peracchi (1988)], the Logit model [Stefanski, Carroll and Ruppert (1986)], and the censored regression model [Peracchi (1987)]. In most cases, the assumed model is Gaussian.

One problem that remains to be discussed is the choice of the matrices Q and B that determine the metrics in which the MSE criterion and the estimator's sensitivity are defined. One possibility is to choose metrics that are 'natural', for example $B = Q = I_p$. A better choice, from the point of view of invariance, is $B = Q = J(\theta)$. Unfortunately, these choices still lead to estimators that are computationally difficult, because the implicit equations (16) and (17) must be solved at each iteration. Another possibility is to choose metrics that, although less 'natural', are more convenient from the point of view of the computations [see e.g. the proposals in Peracchi (1987)].

Specification tests based on the difference between an optimal B-robust estimator $\tilde{\theta}$ and the ML estimator based on $\{F_\theta\}$ satisfy the conditions for powerful specification tests, namely a potentially large difference between estimators when the model is misspecified, and a relatively efficient alternative estimator to ML. Useful diagnostics for detecting influential observations and outliers can be based on the robust weights $W(z_n,\tilde{\theta})$, computed for each observation in the sample. The use of these weights provides an alternative to the traditional methods based on deleting a subset of observations at a time (typically a single

observation) and then comparing the resulting estimates with the ones based on the full sample [see e.g. Belsley, Kuh and Welsch (1980)]. Even single deletion methods can be quite expensive for non-linear estimators and may fail to reveal the presence of multiple outliers [see e.g. Atkinson (1986)]. The use of robust weights has several advantages over these methods. Robust weights are jointly computed with the parameter estimates and require no additional calculation. They are easy to interpret, because of the weighted ML nature of an optimal B-robust estimator, and all the information on the influence of a given observation can be summarized in a scalar number, such as the trace or the determinant of $W(z_n,\tilde{\theta})$, or simply $w(z_n,\tilde{\theta})$ in the case of estimators based on Proposition 3.

So far, applications in empirical econometrics have dealt mainly with the Gaussian linear regression model. For example, Krasker, Kuh and Welsch (1983) estimate hedonic price models for housing, Swartz and Welsch (1986) estimate and forecast energy demand, and Thomas (1987) uses a very large data set to estimate Engel curves for food. All these studies report significant differences in point estimates, inference and forecasts with respect to least squares. Peracchi (1987) estimates Engel curves using household budget data containing a significant fraction of reported zero expenditure. He compares the Tobit ML estimator with a number of semi-parametric and optimal B-robust estimators based on various choices of metrics. The Tobit estimator appears to be very sensitive to a few extreme observations and is way off in some cases. On the other hand, semi-parametric and optimal B-robust estimates are close to each other and look more reliable, but the latter appear to be more precise.

5. Some extensions

A number of additional results are easily obtained by applying the method of Proposition 1. First consider the case when the vector of observables is given by $z = (y, x')'$, with $y \in \mathcal{Y}$ and $x \in \mathcal{X}$, and the likelihood score can be written as the product of two components, $s(z,\theta) = r(y,x,\theta)\, v(x,\theta)$, with the conditional expectation of $r(y,x,\theta)$ given x equal to zero. The IF of the ML estimator can be factored in a similar way. For simplicity we assume that $r(y,x,\theta)$ is a scalar and $v(x,\theta)$ is a p-dimensional vector. Examples include the linear regression model, where $r(y,x,\theta) = (y - x'\beta)/\sigma$, the non-linear regression model, the Probit and Tobit models. Both components of the likelihood score can be unbounded, but we may be particularly concerned with the unboundedness of one of them. In this case we may consider the class of estimators based on a normalized score $\eta(z,\theta) = \vartheta(y,x,\theta)\, \xi(x,\theta)$ with the following robustness properties

$$\gamma_1^*(\hat{\theta},F_\theta) = \sup_{y,x} |\vartheta(y,x,\theta)| \leq \gamma_1 \tag{19}$$

$$\gamma_2^*(\hat{\theta},F_\theta,B) = \sup_{x} \|\xi(x,\theta)\|_B \leq \gamma_2. \tag{20}$$

where $|\cdot|$ denotes the absolute value. When both γ_1 and γ_2 are finite, each estimator in this class has an IF with a finite norm and is therefore B-robust. An optimal estimator can be based on the score function that minimizes the objective functional (10) subject to the robustness constraints (19)-(20) plus Fisher consistency and normalization conditions. A sufficient condition for Fisher consistency is that $\int \vartheta(y,x,\theta)\, dG_x(y) = 0$ for all x, where G_x denotes the conditional d.f. of y given x, and suitable normalization conditions are $\int \vartheta(y,x,\theta)$

$r(y,x,\theta)\ dG_x(y) = 1$ for all x, and $E_H\ \xi(x,\theta)\ v(x,\theta)' = I_p$, where H denotes the marginal d.f. of x (for simplicity the dependence of G_x and H on θ is omitted). Assume that the system of equations

$$E_{G_x}\ w_1(y,x,\theta)\ A_1\ r(y,x,\theta) - [E_{G_x}\ w_1(y,x,\theta)]\ a = 0$$

$$E_{G_x}\ w_1(y,x,\theta)\ [A_1\ r(y,x,\theta) - a]\ r(y,x,\theta) - 1 = 0$$

$$E_H\ w_2(x,\theta)\ A_2\ v(x,\theta)\ v(x,\theta)' - I_p = 0,$$

where

$$w_1(y,x,\theta) = \min\ \{1,\ \gamma_1\ /\ |A_1\ r(y,x,\theta) - a|\}$$

$$w_2(x,\theta) = \min\ \{1,\ \gamma_2\ /\ \|A_2\ v(x,\theta)\|_B\},$$

has a solution $(a(x,\theta),\ A_1(x,\theta),\ A_2(\theta))$ for all x and all θ in a neighborhood of θ_0. Under the above set of assumptions we obtain:

PROPOSITION 4: Let $Q = \alpha\ B$, $\alpha > 0$, and Let $\tilde{\theta}$ be the M-estimator based on the score function $\tilde{\eta}(z,\theta) = \tilde{\vartheta}(y,x,\theta)\ \tilde{\xi}(x,\theta)$, where

$$\tilde{\vartheta}(y,x,\theta) = w_1(y,x,\theta)\ [A_1(x,\theta)\ r(y,x,\theta) - a(x,\theta)]$$

$$\tilde{\xi}(x,\theta) = w_2(x,\theta)\ A_2(\theta)\ v(x,\theta).$$

Then $\tilde{\theta}$ minimizes trace $[Q\ AV(\hat{\theta},F_\theta)]$ among all regular M-estimators satisfying $\gamma_1^*(\hat{\theta},F_\theta) \le \gamma_1$ and $\gamma_2^*(\hat{\theta},F_\theta,B) \le \gamma_2$.

The generalization to the case when Q and B are both arbitrary is straightforward given the method of Proposition 1. When $\gamma_1 = \infty$, the optimal solution $\tilde{\vartheta}(y,x,,\theta)$ is a linear transformation of $r(y,x,\theta)$. Similarly, when $\gamma_2 = \infty$ the optimal solution $\tilde{\xi}(x,\theta)$ is a linear transformation of $v(x,\theta)$. When neither γ_1 nor γ_2 are finite we obtain the ML estimator. Lower bounds on γ_1 and γ_2 can easily be established.

Sometimes $r(y,x,\theta)$ is independent of x, as in the case of regression models with independence between the errors and the regressors. In this case the optimal estimator simplifies considerably, because the constants a and A_1 depend only on θ and not on x. Further simplifications occur when the errors are symmetrically distributed. Examples include the class of 'optimal Mallows-type' estimators [Mallows (1975), Hampel (1978), Maronna and Yohai (1981), Peracchi (1987)]. When $\gamma_2 = \infty$ one obtains the Huber estimator of regression [Huber (1973)].

Now consider again the case when $F_\theta(z) = G_x(y)\ H(x)$, but suppose that we only care about misspecification of the conditional d.f. G_x of y, but not of the marginal d.f. H of x. Often all G_x are of the same shape, e.g. they are all Gaussian, with mean and possibly variance depending on x. In what follows F_θ is replaced by the pair $(\mathcal{G},H)$, where $\mathcal{G}$ is the collection of all conditional d.f.'s G_x.

First we introduce the concept of partial IF of an estimator $\hat{\theta}$, which is a measure of the asymptotic bias of $\hat{\theta}$, as an estimator of $\hat{\theta}(\mathcal{G},H)$, arising from an infinitesimal amount of contamination of the collection of conditional d.f.'s $\mathcal{G}$, keeping the marginal d.f. H fixed.

DEFINITION 5: Let $G_{x,\varepsilon,y} = (1 - \varepsilon)\ G_x + \varepsilon\ \Delta_{(y)}$, and let $\mathcal{G}_{\varepsilon,y}$ be the collection of all d.f.'s $G_{x,\varepsilon,y}$, for given ε and y. Then the partial

influence function (PIF) of $\hat{\theta}$ at $(\mathcal{G},H)$ is given by

$$PIF(y,\hat{\theta},(\mathcal{G},H)) = \lim_{\varepsilon \to 0+} [\hat{\theta}(\mathcal{G}_{\varepsilon,y},H) - \hat{\theta}(\mathcal{G},H)]/\varepsilon$$

provided that the limit exists.

We now show that for an M-estimator $\hat{\theta}$ the PIF can simply be obtained by integrating the IF with respect to the d.f. of the x's.

PROPOSITION 5: Under regularity conditions sufficient to allow interchanging differentiation and integration, the PIF of an M-estimator $\hat{\theta}$ at a distribution $(\mathcal{G},H)$ is equal to

$$PIF(y,\hat{\theta},(\mathcal{G},H)) = E_H\ IF(y,x,\hat{\theta},(\mathcal{G},H)).$$

The sensitivity to gross-errors that affect only the conditional d.f. of y is defined in the usual way as

$$\gamma_p^*(\hat{\theta},(\mathcal{G},H),B) = \sup_{y \in \mathcal{Y}} \|PIF(y,\hat{\theta},(\mathcal{G},H))\|_B.$$

A 'constrained asymptotically efficient' estimator can then be based on the score function that solves problem (10)-(12), with (13) replaced by the constraint that $\gamma_p^*(\hat{\theta},(\mathcal{G},H),B) \leq \gamma$. Assume that the equations system

$$E_\theta\ w(y,\theta)\ A\ s(y,x,\theta) - [E_\theta\ w(y,\theta)]\ a = 0$$

$$E_\theta\ w(y,\theta)\ [A\ s(y,x,\theta) - a]\ s(y,x,\theta) - I_p = 0,$$

where

$$w(y,\theta) = \min \{1, \gamma / \|A E_H s(y,x,\theta) - a\|_B\},$$

has a solution $(a(\theta), A(\theta))$ for all θ in a neighborhood of θ_0. Then

PROPOSITION 6: Let $Q = \alpha B$, $\alpha > 0$, and let $\tilde{\theta}$ be the M-estimator based on the score function

$$\tilde{\eta}(y,x,\theta) = w(y,\theta) [A(\theta) s(y,x,\theta) - a(\theta)].$$

Then $\tilde{\theta}$ minimizes trace $[Q\ AV(\hat{\theta},(\mathcal{C},H))]$ among all regular M-estimators satisfying $\gamma_p^*(\hat{\theta},(\mathcal{C},H)) \leq \gamma$.

The generalization to the case when Q and B are both arbitrary is straightforward given the method of Proposition 1. A lower bound on γ can easily be established.

Finally consider the case when the parameter vector can be partitioned in two subsets of parameters, and we want to allow the corresponding subsets of estimates to have different degree of robustness. For example, in the case of the classical linear regression model, we may want to consider separately the regression parameters and the variance of the disturbances.

Thus, partition the parameter vector as $\theta = (\theta_1', \theta_2')'$, where θ_1 and θ_2 are respectively a p_1- and a p_2-vector, with $p_1 + p_2 = p$. Partition the score function and the IF accordingly. Let $\gamma^*(\hat{\theta}_j, F_\theta, B_j) = \sup_{z \in \mathcal{Z}} \|IF(z,\hat{\theta}_j,F_\theta)\|_{B_j}$ denote the gross-error sensitivity of an estimator of the

j-th subset of parameters in the metric of the $p_j \times p_j$ matrix B_j, and let $\gamma = (\gamma_1, \gamma_2)'$. Consider the class of regular M-estimators of θ for which $\gamma^*(\hat{\theta}_j, F_\theta, B_j) \leq \gamma_j$, $j = 1, 2$. If all components of γ are finite, then each estimator in this class has an IF with a finite norm and is therefore B-robust.

The fact that the sensitivity can be different for different subsets of parameters increases the flexibility of the estimation procedure. For example, if θ_1 is the parameter of interest and θ_2 is the nuisance parameter, the bound on the estimate of θ_2 can be very tight, in order to be sure that estimates are very robust, whereas the bound on θ_1 can be chosen to attain some balance between robustness and asymptotic efficiency at the assumed model.

An optimal B-robust estimator can be based on the score function that solves problem (10)-(12), with (13) replaced by the constraint

$$\gamma^*(\hat{\theta}_j, F_\theta, B_j) \leq \gamma_j, \qquad j = 1, 2. \tag{21}$$

The following result extends Proposition 2 to the partitioned parameters case. It also generalizes Theorem 4.4.1 of Hampel et al. (1986), to the case when the sensitivity of each subset of estimates is defined in an arbitrary metric. Generalizations of Propositions 4 and 6 to the partitioned parameters case are also straightforward to obtain.

PROPOSITION 7: Let Q and B be p.d. block diagonal matrices with $Q_j = \alpha_j B_j$, $\alpha_j > 0$, $j = 1, 2$, and assume that $a(\theta)$ and $A(\theta)$, implicitly defined by equations (16) and (17), exist for all θ in an open neighborhood of $F\theta_0$. Let $\tilde{\theta}$ be the M-estimator of θ based on a score function of the form

(15), with $W(z,\theta) = \text{Diag}\,[w_j(z,\theta)\, I_{p_j},\ j = 1, 2]$ and

$$w_j(z,\theta) = \min\,\{1,\ \gamma_j \,/\, \|A_j\, s(z,\theta) - a_j\|_{B_j}\},$$

where A_j is the $p_j \times p$ matrix formed by the j-th subset of rows of $A(\theta)$. Then $\tilde{\theta}$ minimizes trace $[Q\ AV(\hat{\theta},F_\theta)]$ among all regular M-estimators satisfying $\gamma^*(\hat{\theta}_j,F_\theta,B_j) \le \gamma_j$, $j = 1, 2$.

The generalization to the case when Q and B are both arbitrary is straightforward, and lower bounds on γ_1 and γ_2 can easily be derived.

The optimal weighting function of Proposition 7 simplifies somewhat when $a(\theta) = 0$ and $A(\theta) = \begin{bmatrix} A_{11} & \\ & A_{22} \end{bmatrix}$ is block-diagonal. In this case $w_j(z,\theta) = \min\,\{1,\ \gamma_j \,/\, \|A_{jj}\, s_j(z,\theta)\|_{B_j}$, where $s_j(z,\theta)$ denotes the j-th sub-vector of $s(z,\theta)$.

6. Final remarks

The robust estimators presented in this paper are based on parametric assumptions, but are explicitly designed to bound the negative effects of their violations. This makes them particularly attractive in situations where the researcher may be willing to specify a parametric model for the observations, but not to take it as literally true. This kind of situations arise quite frequently in empirical work. The reasons for the limited success of these methods in applied econometrics are therefore not very clear. Paradoxically, semi-parametric and non-parametric methods

have been used more frequently, even though their efficiency and robustness properties are often poor and their distribution sometimes unknown even for large samples.

One reason may be the fact that bounded influence methods are often viewed as ad-hoc methods for dealing with outliers. In fact, these methods simply extend classical parametric statistics by supplementing the standard requirements of consistency and efficiency with a robustness condition that refers to the behavior of a procedure in a small neighborhood of the assumed model. For this reason they are quite flexible, and can be applied to any model for which the investigator is prepared to specify a likelihood function.

Computational difficulties and the lack of readily available computer programs may be another reason. Often this problem is much less serious than it may look at first. For example, optimal B-robust estimators of regression can easily be computed by iteratively reweighted least squares.

Yet another reason may be the fact that little is known about these estimators aside from their large sample properties, and one may, quite rightly, fear that asymptotic results provide little guidance in the case of procedures that are highly non-linear. Research in this area would be very valuable.

Appendix

PROOF OF PROPOSITION 2: If $A(\theta)$ and $a(\theta)$ exist, premultiplying (17) by B and taking the trace of both sides gives

$$\begin{aligned} \text{trace } B &= E_\theta \; s(z,\theta)' \; B \; W(z,\theta) \; [A \; s(z,\theta) - a] \\ &\leq E_\theta \; \|s(z,\theta)\|_B \; \|W(z,\theta) \; [A \; s(z,\theta) - a]\|_B \end{aligned}$$

(by the Cauchy-Schwarz Inequality)

$$\leq \gamma \; E_\theta \; \|s(z,\theta)\|_B$$

(by (13)), from which the desired result follows.

PROOF OF PROPOSITION 3: When $Q = \alpha B$ equation (14) becomes

$$\gamma = \|[1 + \tilde{\mu}(z)/(2\alpha\gamma)]^{-1} \; [A \; s(z,\theta) - a]\|_B .$$

Solving for $\tilde{\mu}(z)$, by using the fact that α, γ, and $\tilde{\mu}(z)$ are all non negative, gives

$$\tilde{\mu}(z) = 2\alpha \; [\|A \; s(z,\theta) - a\|_B - \gamma]$$

and therefore

$$[1 + \tilde{\mu}(z)/(2\alpha\gamma)]^{-1} = \gamma \; / \; \|A \; s(z,\theta) - a\|_B .$$

PROOF OF PROPOSITION 4: Let $\tilde{\eta} = (\tilde{\eta}_1, \tilde{\eta}_2)$ denote the optimal score. Then the first order conditions for optimality are given by (19)-(20) and

$$E_{G_x} \tilde{\eta}_1(y,x,\theta) = 0$$

$$E_{G_x} \tilde{\eta}_1(y,x,\theta)\, r(y,x,\theta)\, dG_x(y) = 1$$

$$E_H\, \eta_2(x,\theta)\, v(x,\theta)' = I_p$$

$$\tilde{\eta}_1(y,x,\theta) + a(x) - A_1(x)\, r(y,x,\theta) + \tilde{\mu}_1(y,x)\, \text{sign}\,(\tilde{\eta}_1(y,x,\theta)) = 0$$

$$\tilde{\mu}_1(y,x)\,(|\tilde{\eta}_1(y,x,\theta)| - \gamma_1) = 0$$

$$\tilde{\eta}_2(x,\theta) + A_2\, v(x,\theta) + \tilde{\mu}_2(x)\, B\, \frac{\tilde{\eta}_2(x,\theta)}{\|\tilde{\eta}_2(x,\theta)\|_B} = 0$$

$$\tilde{\mu}_2(x)\,(\|\tilde{\eta}_2(x,\theta)\|_B - \gamma_2) = 0$$

for almost all (y,x) (i.e. except possibly a set with zero F_θ-probability), where the dependence of the Lagrange multipliers $a(x)$, $A_1(x)$, A_2, $\tilde{\mu}_1(y,x)$ and $\tilde{\mu}_2(x)$ on θ has been omitted. The result then follows by a standard argument.

PROOF OF PROPOSITION 5: Evaluating (3) at $\mathfrak{G}_{\varepsilon,y}$ gives

$$0 = \int\int \eta(u,v,\hat{\theta}(\mathfrak{G}_{\varepsilon,y},H))\, dG_{v,\varepsilon,y}(u)\, dH(v).$$

Differentiating with respect to ε gives

$$0 = \int \Big[\int \frac{\partial}{\partial\theta'} \eta(u,v,\hat{\theta}(\mathfrak{G}_{\varepsilon,y},H))\, \frac{\partial}{\partial\varepsilon} \hat{\theta}(\mathfrak{G}_{\varepsilon,y},H)\, dG_{v,\varepsilon,y}(u)$$

$$+ \int \eta(u,v,\hat{\theta}(\varsigma_{\varepsilon,y},H))\, d(-G_v(u) + \Delta_{(y)}(u))\ ]\ dH(v).$$

Evaluating at $\varepsilon = 0$ and using the definition of PIF gives

$$0 = \int \int \frac{\partial}{\partial\theta'} \eta(u,v,\hat{\theta}(\varsigma,H))\ dG_v(u)\ dH(v) \cdot PIF(y,\hat{\theta},(\varsigma,H)) + \int \eta(y,v,\hat{\theta}(\varsigma,H))\ dH(v).$$

The result then follows from the definition of IF.

PROOF OF PROPOSITION 6: The first order conditions for optimality are given by (11)-(13), and

$$\tilde{\eta}(y,x,\theta) + a - A\, s(y,x,\theta) + \tilde{\mu}(y)\, B \frac{\tilde{\eta}(y,x,\theta)}{\|E_H\, \tilde{\eta}(y,x,\theta)\|_B} = 0$$

$$\|E_H\, \tilde{\eta}(y,x,\theta)\|_B \leq \gamma$$

$$\tilde{\mu}(y)\ (\|E_H\, \tilde{\eta}(y,x,\theta)\|_B - \gamma) = 0$$

for almost all (y,x), where the dependence of the Lagrange multipliers a, A and $\tilde{\mu}(y)$ on θ has been omitted. The result then follows by a standard argument.

PROOF OF PROPOSITION 7: The first order conditions for optimality are given by (11)-(13), (21) and

$$\tilde{\eta}_j(z,\theta) + a - A\, s(z,\theta) + \tilde{\mu}_j(z)\, B_j \frac{\tilde{\eta}(z,\theta)}{\|\tilde{\eta}_j(z,\theta)\|_{B_j}} = 0,$$

$$\tilde{\mu}_j(z)\, (\|\tilde{\eta}_j(z,\theta)\|_{B_j} - \gamma_j) = 0$$

for $j = 1, 2$ and almost all z, where the dependence of the Lagrange multipliers a, A, $\tilde{\mu}_1(z)$ and $\tilde{\mu}_2(z)$ on θ has been omitted. The result then follows by a standard argument.

Footnotes

[1] An alternative, data-based definition of robustness has been proposed by Gilstein and Leamer (1983). For a given sample of observations, they consider the set of all points in the parameter space that can be ML estimators for some distribution of the observations. An estimate is not robust if this set is large. There are two problems with this approach. First, the set of possible ML estimates can be difficult to describe. Second, generalizations of this approach require strong restrictions on the distribution of the observations and/or the class of estimators.

[2] This need not always be true. For example, Beran (1977) shows that in the case of neighborhoods based on the Hellinger metric no trade- off arises. His result illustrates the importance of the type of topology, but should not be emphasized because, as pointed out by Bickel (1981), the Hellinger neighborhoods are really too small for being of practical interest .

[3] For example, in the location and regression models with symmetric errors, the Huber estimator is inadmissible because it is dominated by adaptive estimators.

[4] The assumption of identical distribution can be relaxed, but the assumption of independence is crucial. For some results in the case of dependent observations see Künsch (1984) and Martin and Yohai (1986).

[5] Bickel (1981, 1984) provides some results for the more general case when F_0 is at a distance ε from the assumed model $\mathcal{F}$ in a proper metric.

Footnotes

[1] An alternative, data-based definition of robustness has been proposed by Gilstein and Leamer (1983). For a given sample of observations, they consider the set of all points in the parameter space that can be ML estimators for some distribution of the observations. An estimate is not robust if this set is large. There are two problems with this approach. First, the set of possible ML estimates can be difficult to describe. Second, generalizations of this approach require strong restrictions on the distribution of the observations and/or the class of estimators.

[2] This need not always be true. For example, Beran (1977) shows that in the case of neighborhoods based on the Hellinger metric no trade- off arises. His result illustrates the importance of the type of topology, but should not be emphasized because, as pointed out by Bickel (1981), the Hellinger neighborhoods are really too small for being of practical interest .

[3] For example, in the location and regression models with symmetric errors, the Huber estimator is inadmissible because it is dominated by adaptive estimators.

[4] The assumption of identical distribution can be relaxed, but the assumption of independence is crucial. For some results in the case of dependent observations see Künsch (1984) and Martin and Yohai (1986).

[5] Bickel (1981, 1984) provides some results for the more general case when F_0 is at a distance ϵ from the assumed model $\mathcal{F}$ in a proper metric.

[6] Sometimes M-estimators are defined in terms of (2) rather than (1). The two definitions are really equivalent, for we can always obtain a root of (1) by minimizing the norm of the sample average of $\eta_N(z,\theta)$, which is a problem of the same form as (2).

[7] See e.g. Huber (1967) or, for simpler but stronger conditions, Amemiya (1985) and Duncan (1987).

[8] Sometimes the linear term in (4) vanishes, in which case higher-order terms have to be included and the definition of the IF is based, more generally, on the first non-vanishing term of the von Mises expansion.

[9] The same idea, applied to the asymptotic variance of a statistic, also viewed as a functional, leads to the concept of change-of-variance function [Hampel, Rousseeuw and Ronchetti (1981)].

[10] The IF is closely related to a number of other concepts. In particular, it can be shown [see e.g. Hampel et al. (1986)] that, under regularity conditions, the IF is equal to the limit, as $N \rightarrow \infty$, of Tukey's sensitivity curve [see e.g. Hoaglin, Mosteller and Tukey (1983)], which describes the normalized effect on the value of a statistic of adding one arbitrary point to a given sample. In turns, this fact relates the IF to the jackknife [see e.g. Efron (1982)] and the diagnostics for influential observations proposed by Belsely, Kuh and Welsch (1980).

[11] The IF provides only a local approximation to the asymptotic bias of an estimator. A measure of the distance from the assumed model up to which the estimator still gives some relevant information is given by the estimator's breakdown point [Hampel (1971)]. Donoho and Huber (1983) and Hampel et al. (1986) provide tractable finite-sample versions of Hampel's original asymptotic concept. The practical importance of the breakdown point has been demonstrated by Hampel (1985), who shows that the robustness properties of various location estimators in a number of simulation experiments can accurately be classified on the basis of their breakdown point alone.

[12] The concepts of V-robustness and change-of-variance sensitivity are analogously defined on the basis of the change-of-variance function. B- and V-robustness are not equivalent, because V-robustness implies B-robustness, but the converse is not always true [Hampel et al. (1986)].

[13] Results on efficient estimation under robustness constraints different from the ones discussed in this paper are very few, and are typically confined to the normal linear regression model. For example, Mallows (1975) considers the class of estimators with separate bounds on the IF and the sensitivity to rounding or grouping of the observations. Ronchetti and Rousseeuw (1985) consider the class of V-robust estimators. Yohai (1987) consider a class of high breakdown point estimators.

References

Amemiya, T. (1985), *Advanced Econometrics*, Harvard University Press, Cambridge, Mass.

Atkinson, A.C. (1986), "Masking unmasked", *Biometrika*, 73, 553-541.

Belsley, D.A., Kuh, E. and Welsch, R.E. (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley, New York.

Beran, R. (1977), "Minimum Hellinger distance estimates for parametric models", *Annals of Statistic*, 5, 431-444.

Bickel, P.J. (1981), "Quelques aspects de la statistique robuste", in *Ecole d'Ete de Probabilite de St. Flour*, Lecture Notes in Mathematics No. 876, Springer, Berlin.

Bickel, P.J. (1984), "Robust regression based on infinitesimal neighborhoods", *Annals of Statistics*, 12, 1349-1368.

Burguete, J., Gallant, A.R., and Souza, G. (1982), "On unification of asymptotic theory of non-linear econometrics", *Econometric Reviews*, 1, 151-190.

Chamberlain, G. (1982), "Multivariate regression models for panel data", *Journal of Econometrics*, 18, 5-46.

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

Donoho, D.L. and Huber, P.J. (1983), "The notion of breakdown point", in P.J. Bickel, K.A. Doksum and J.L. Hodges (eds.), *Festschrift for Erich L. Lehmann*, Wadsworth, Belmont (Calif.).

Duncan, G.M. (1987), "A simplified approach to M-estimation with application to two-stage estimators", *Journal of Econometrics*, 34, 373-389.

Efron, B. (1982), *The Jackknife, the Bootstrap and Other Resampling Plans*, SIAM, Philadelphia.

Fernholz, L.T. (1983), *Von Misesd Calculus for Statistical Functionals*, Lecture Notes in Statistics No. 19, Springer, New York.

Fraiman, R. (1983), "General M-estimators and applications to bounded influence estimation for non-linear regression", *Communications in Statistics, Series A*, 12, 2617-2631.

Gilstein, C.Z. and Leamer, E.E. (1983), "Robust sets of regression estimates", *Econometrica*, 55, 363-390.

Gourieroux, C., Monfort, A. and Trognon, A. (1984), "Pseudo maximum likelihood methods: Theory", *Econometrica*, 52, 681-700.

Hampel, F.R. (1968), Contribution to the Theory of Robust Estimation, Ph.D. Thesis, University of California, Berkeley.

Hampel, F.R. (1971), "A general qualitative definition of robustness", *Annals of Mathematical Statistics*, 42, 1887-1096.

Hampel, F.R. (1974), "The influence curve and its role in robust estimation", *Journal of the American Statistical Association*, 69, 383-393.

Hampel, F.R. (1978), "Optimally bounding the gross-error- sensitivity and the influence of position in factor space", *Proceedings of the ASA Statistical Computing Section*, ASA, Washington, D.C.

Hampel, F.R. (1985), "The breakdown points of the mean combined with some rejection rules", *Technometrics*, 27, 95-107.

Hampel, F.R., Rousseeuw, P.J. and Ronchetti, E. (1981), "The change-of-variance curve and optimal redescending M- estimators", *Journal of the American Statistical Association*, 76, 643-648.

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

Hansen, L.P. (1982), "Large sample properties of generalized method of moments estimators", *Econometrica*, 50, 1029-1054.

Hoaglin, D.C., Mosteller, F. and Tukey, J.W. (1983), *Understanding Robust and Explanatory Data Analysis*, Wiley, New York.

Huber, P.J. (1964), "Robust estimation of a location parameter", *Annals of Mathematical Statistics*, 35, 73-101.

Huber, P.J. (1967), "The behavior of maximum likelihood estimates under non-standard conditions", *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, 221-33.

Huber, P.J. (1973), "Robust regression: Asymptotics, conjectures and Monte-Carlo", *Annals of Statistics*, 1, 799-821.

Huber, P.J. (1981), *Robust Statistics*, Wiley, New York.

Kiefer, N. and Skoog, G. (1984), "Local asymptotic specification error analysis", *Econometrica*, 48, 1333-1346.

Krasker, W.S. (1980), "Estimation in linear regression models with disparate data points", *Econometrica*, 48, 1333-1346.

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

Krasker, W.S., Kuh, E., and Welsch, R.E. (1983), "Estimation for dirty data and flawed models", in Z. Griliches and M.D. Intriligator (eds.), *Handbook of Econometrics*, Vol. 1, North-Holland, Amsterdam.

Künsch, H. (1984), "Infinitesimal robustness for autoregressive processes", *Annals of Statistics*, 12, 843-863.

Luenberger, D. (1969), *Optimization by Vector Space Methods*, Wiley, New York.

Malinvaud, E. (1970), *Statistical Methods of Econometrics*, North Holland, Amsterdam.

Mallows, C.L. (1975), "On some topics in robustness", Technical Memorandum, Bell Telephone Laboratories, Murray Hill, N.J.

Maronna, R. and Yohai, V. (1981), " Asymptotic behavior of general M-estimates for regression and scale with random carriers", *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete*, 58, 7-20.

Martin, D.R. and Yohai, V.J. (1986), "Influence functionals for time series", *Annals of Statistics*, 14, 781-855.

Peracchi, F. (1987), Bounded Influence Methods in Econometrics with an Application to the Censored Regression Model", Ph.D. Thesis, Princeton

University.

Peracchi, F. (1988), "Bounded influence estimators for the SURE model", Dept. of Economics, UCLA, mimeo.

Prakasa Rao, B.L.S. (1987), *Asymptotic Theory of Statistical Inference*, Wiley, New York.

Ronchetti, E., and Rousseeuw, P.J. (1985), "Change-of-variance sensitivities in regression analysis", *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 68, 503-519.

Serfling, R.J. (1980), *Approximation Theorems of Mathematical Statistics*, Wiley, New York.

Small, K. (1986), "Effects of the 1979 gasoline shortages on Philadephia housing prices", *Journal of Urban Economics*, 19, 371-381.

Stefanski, L.A., Carrol, R.J. and Ruppert, D. (1986), "Optimally bounded score functions for generalized models with applications to logistic regression", *Biometrika*, 73, 413-424.

Swartz, S. and Welsch, R.E. (1986), "Applications of bounded-influence and diagnostic methods in energy modeling", in D.A. Belsley and E. Kuh (eds.), *Model Reliability*, MIT Press, Cambridge, Mass.

Thomas, D. (1987), "Robust regression of the food share Engel curve", Yale University, mimeo.

White, H. (1982), "Maximum likelihood estimation of misspecified models", *Econometrica*, 48, 817-838.

Yohai, V.J. (1987), "High breakdown-point and high efficiency robust estimates for regression", *Annals of Statistics*, 15, 642-656.