The Annals of Statistics
1973, Vol. 1, No. 5, 799-821

THE 1972 WALD MEMORIAL LECTURES

ROBUST REGRESSION: ASYMPTOTICS, CONJECTURES AND MONTE CARLO[1]

BY PETER J. HUBER

Swiss Federal Institute of Technology, Zürich

Maximum likelihood type robust estimates of regression are defined and their asymptotic properties are investigated both theoretically and empirically. Perhaps the most important new feature is that the number p of parameters is allowed to increase with the number n of observations. The initial terms of a formal power series expansion (essentially in powers of p/n) show an excellent agreement with Monte Carlo results, in most cases down to 4 observations per parameter.

1. Introduction. Consider the classical least squares problem: p unknown parameters $\theta_1, \cdots, \theta_p$ are to be estimated from n observations $X_1, \cdots, X_n$ to which they are related by

$$X_i = \sum_{j=1}^{p} c_{ij}\theta_j + U_i \,. \tag{1.1}$$

The c_{ij} are known coefficients and the U_i are independent random errors with (approximately) identical distributions.

Classically, the problem is solved by minimizing the sum of squares:

$$\sum_{i=1}^{n} (X_i - \sum_{j=1}^{p} c_{ij}\theta_j)^2 = \min! \,, \tag{1.2}$$

or, equivalently, by solving the system of p equations, obtained by differentiating (1.2),

$$\sum_{i=1}^{n} (X_i - \sum_{k=1}^{p} c_{ik}\theta_k)c_{ij} = 0 \,, \qquad j = 1, \cdots, p \,. \tag{1.3}$$

The original justification for this method (due to Gauss) is somewhat circular: least squares estimates are optimal if the errors are independent identically distributed normal; on the other hand, Gauss assumed a normal error law because then the sample mean, which "is generally accepted as a good estimate," turns out to be optimal in the simplest special case, see Gauss (1821).

In the regression case, uncontrollable inhomogeneity of variance among the U_i and genuinely long-tailed error distributions have almost indistinguishable effects, both impairing the efficiency of the estimates. Just a single grossly outlying observation may spoil the least squares estimate and moreover, outliers are much harder to spot in the regression than in the simple location case. Thus robust alternatives to the method of least squares are sorely needed. In view of the Gauss–Markov theorem these alternatives cannot be linear in the observations.

Received November 1972; revised February 1973.

[1] This paper was presented as part of the Wald Lecture at the IMS Annual Meeting at Hanover, New Hampshire, August 28—September 1, 1972.

As regression calculations nowadays are done on computers, some quite mechanical procedures for deflating the influence of gross errors of any kind, wrong weights (inhomogeneous variances) or otherwise longtailed error distributions are called for, even more so than in the simple location case. I should hasten to emphasize that this does not obviate the need for a careful inspection of the pattern of the residuals.

We shall have (i) to define an estimate, (ii) to investigate its (asymptotic) properties, (iii) to estimate the covariance matrix of the estimate, (iv) to devise some numerical procedure for computing the estimate and its estimated covariance matrix, and finally, (v) to investigate (empirically) the small sample properties.

Any such estimate will in some sense generalize a robust alternative to the sample mean. There are essentially two simple ways to obtain such estimates. The first is to replace the function (1.2) to be minimized by some expression which is less sensitive to extreme values of the residuals

$$\Delta_i = X_i - \sum_j c_{ij}\theta_j \,. \tag{1.4}$$

The second is based on the remark that according to (1.3) the parameters θ_j are estimated in such a way that the residual vector $\Delta_\bullet$ and the column vectors $c_{\bullet j}$ have empirical correlations 0 (more precisely: cross-moments, since the averages are not subtracted out). So one simply replaces (1.3) by a robust alternative to the correlation coefficient (Huber (1972b)).

Of the three most obvious contenders—*M*-, *L*- and *R*-estimates (see Huber (1972a), Section 4)—the first type generalizes most straightforwardly. One simply replaces (1.2) by

$$\sum_{i=1}^n \rho(X_i - \sum_j c_{ij}\theta_j) = \min! \tag{1.5}$$

where ρ is some (usually convex) function, e.g.

$$\begin{aligned} \rho(x) &= \tfrac{1}{2}x^2 && \text{for} \quad |x| < c \\ &= c|x| - \tfrac{1}{2}c^2 && \text{for} \quad |x| \geqq c \,. \end{aligned} \tag{1.6}$$

The value of c may depend on the observations X_i, in order to obtain scale invariance.

The ρ given by (1.6) leads to estimates with well-defined asymptotic and finite sample minimax properties in the simplest special case ($p = 1$, $c_{ij} = 1$), see Huber (1964), (1968), and at least the asymptotic optimality carries over to the regression case. For another proposal for ρ, see Anscombe (1967).

If we differentiate (1.5), we obtain (with $\psi = \rho'$) the following analogue of (1.3):

$$\sum_{i=1}^n \psi(X_i - \sum_k c_{ik}\theta_k)c_{ij} = 0 \,, \qquad j = 1, \cdots, p \tag{1.7}$$

which is equivalent with (1.5) if ρ is convex. Note that this is a robustized version of the cross-moment: we are correlating modified (metrically Winsorized) residuals with the coefficient vector $c_{\bullet j}$.

Here, only the residuals Δ_i have been subjected to a modification; one wonders whether one can gain robustness with regard to errors in the coefficients c_{ij} by modifying also the other factor, the coefficient vector $c_{\cdot j}$.

We obtain R-estimates of regression if we minimize, instead of (1.2),

$$\sum_i a_n(R_i)\Delta_i = \min! \tag{1.8}$$

Here, R_i is the rank of Δ_i in $(\Delta_1, \cdots, \Delta_n)$ and $a_n(\cdot)$ is some monotone scores function satisfying $\sum_i a_n(i) = 0$ (see Jaeckel (1972)). Note, however, that these estimates are unable to estimate an additive main effect and thus do not contain estimates of location as particular cases. On the contrary, the additive main effect has to be estimated by applying an estimate of location to the residuals.

If we differentiate (1.8), which is a piecewise linear convex function of $\theta = (\theta_1, \cdots, \theta_p)$, we obtain the following approximate equalities at the minimum:

$$\sum_i a_n(R_i)c_{ij} \approx 0\,, \qquad j = 1, 2, \cdots, p\,. \tag{1.9}$$

These approximate equations in turn can be reconverted into a minimum problem, e.g.

$$\sum_j \left|\sum_i a_n(R_i)c_{ij}\right| = \min! \tag{1.10}$$

This last variant was investigated by Jurečková (1971), and the asymptotic equivalence between (1.8) and (1.10) was shown by Jaeckel (1972). Instead of correlating $a_n(R_i)$ with c_{ij}, as in (1.9), we might use ranks for the $c_{\cdot j}$ too, etc. It remains to be seen which of these approaches are sensible and fruitful.

Note that the simple straight line regression problem is basic; if we know how to treat this, we can in principle attack the general regression problem by considering one parameter at a time, keeping the others fixed at trial values. However, the existence of a fixed point is not always easy to establish, and even if there is one, the iterative search for it does not necessarily converge.

Furthermore, all of these regression estimates allow one-step versions: start with some reasonably good preliminary estimate θ^*, and then apply one step of Newton's method to (1.7) etc., just as in the location case (Andrews *et al.* (1972), Bickel (1971 a)). A kind of one-step L-estimate of regression has been investigated by Bickel (1971 b).

Our past experience with estimates of location suggests that M-estimates are easiest to cope with, as far as asymptotic theory is concerned (even though R-estimates have received a much greater coverage in the literature of the past few years). We shall therefore concentrate on the estimates defined by (1.5) or (1.7).

There is little hope to build an exact finite sample theory of robust regression, but for large n asymptotic approximations should be possible. Specifically, we have problems in mind which occur, e.g., in X-ray crystallography. With the present computing facilities, typical values for p and n there are in the range

$$\begin{aligned} p&: \quad 10 \quad \text{to} \quad 500\,, \\ n&: \quad 100 \quad \text{to} \quad 10{,}000\,. \end{aligned}$$

The unknown parameters are, essentially, coordinates of atoms; typically, one is interested in questions involving many parameters simultaneously in a non-linear fashion, e.g., whether two molecules have the same spatial configuration.

We intend to build an asymptotic theory for $n \to \infty$; but there are several possibilities for the concomitant behavior of p. In particular, with decreasing restrictiveness:

(a) $\limsup p < \infty$
(b) $\lim p^3/n = 0$
(c) $\lim p^2/n = 0$
(d) $\lim p/n = 0$
(e) $\limsup p/n < 1$
(f) $\lim n - p = \infty$.

Case (a) has been treated by Relles (1968). The generalization to case (b) is relatively straightforward. Cases (d) and (e), possibly also (f) seem to be the interesting ones for the practical applications. I may quote a crystallographer's recommendation that there should be at least 5 observations per parameter, i.e., $p/n \leqq 0.2$ (Hamilton (1970)). It will become clear in the next section why (e) and (f) are unlikely to yield to a reasonably simple asymptotic theory; then we shall attack what are, essentially, cases (b) to (d). Theoretical results and conjectures are summarized near the end of Sections 3 and 6; Monte Carlo results are summarized in Section 9.

On purpose, I have described the regression problem in terms of the classical least squares theory, where the matrix $C = (c_{ij})$ is thought to derive from a rigorous and fixed mathematical model. In statistics, it is more customary to treat the coefficients c_{ij} as "independent variables," possibly also subject to errors. Only little is known how to robustize regression procedures with respect to errors in the c_{ij}, but some of the possibilities mentioned after (1.7) and (1.10) may be attractive.

2. The classical least squares approach. In this section we sketch a simple and basic, but seemingly little known result about the limiting behavior of least squares estimates if the errors are non-normal. See also Eicker (1963), (1967).

We assume that our regression problem is imbedded in a sequence of similar problems with matrices $C^{(N)} = (c_{ij}^{(N)})$, $1 \leqq i \leqq n_N$, $1 \leqq j \leqq p_N$, $N \to \infty$, but we shall suppress the index N whenever feasible.

Assume that the errors U_i are independent and identically distributed, according to some fixed non-normal law with mean 0 and variance $\sigma^2 < \infty$. Then the least squares estimate of $\theta = (\theta_1, \cdots, \theta_p)$ is

$$\hat{\theta} = (C^T C)^{-1} C^T X . \tag{2.1}$$

The problem is to find necessary and sufficient conditions such that *all* estimates of the form $\hat{\alpha} = \sum a_j \hat{\theta}_j$ are asymptotically normal.

Evidently, the coordinate system in the parameter space is arbitrary; we can

choose it to be orthogonal:

$$C^TC = I . \tag{2.2}$$

Also, we can standardize the a_j such that $\sum a_j^2 = 1$. Then

$$\hat{\alpha} = \sum_i s_i X_i , \tag{2.3}$$

with

$$s_i = \sum_j c_{ij} a_j , \tag{2.4}$$

and

$$\sum s_i^2 = 1 , \tag{2.5}$$

hence

$$\text{Var}\,(\hat{\alpha}) = \sigma^2 . \tag{2.6}$$

LEMMA 2.1. *The estimate $\hat{\alpha}$ is asymptotically normal* iff $\max_i |s_i| \to 0$ *as* $N \to \infty$.

PROOF. If $\max_i |s_i| \not\to 0$, then the limiting law of $\hat{\alpha}$ (if it exists at all) can be decomposed into two components, one of which is not normal, hence it cannot be normal itself. On the other hand, if $\max |s_i| \to 0$, then it is easy to check that Lindeberg's condition holds, namely in our case

$$\frac{1}{\sigma^2} \sum_i E\{s_i^2 U_i^2 1_{(|s_i U_i| \geq \varepsilon\sigma)}\} \to 0 . \tag{2.7}$$

Schwarz's inequality gives

$$s_i^2 = (\sum_j c_{ij} a_j)^2 \leqq \sum_j c_{ij}^2 \sum_j a_j^2 = \sum_j c_{ij}^2$$

with equality holding if $a_j \sim c_{ij}$.

Note that

$$\gamma_{ii} = \sum_j c_{ij}^2 \tag{2.8}$$

is the ith diagonal element of the projection matrix

$$\Gamma = C(C^TC)^{-1}C^T \tag{2.9}$$

(which is independent of the particular coordinate system in the parameter space). Thus we have proved:

PROPOSITION 2.2. *A necessary and sufficient condition for all least squares estimates $\hat{\alpha}$ to be asymptotically normal is*

$$\max_i \gamma_{ii} \to 0 \qquad as \quad N \to \infty . \tag{2.10}$$

REMARK 1. Since $\sum_i \gamma_{ii} = p$, we have $\max \gamma_{ii} \geqq \text{ave}\, \gamma_{ii} = p/n$, hence (2.10) implies $p/n \to 0$.

REMARK 2. The fitted value of X_i (the least squares estimate of the expectation of X_i) is

$$T_i = \sum_j c_{ij} \hat{\theta}_j = \sum_l \gamma_{il} X_l .$$

In particular, if (2.10) fails, then some of the fitted values will not be asymptotically normal.

REMARK 3. Continuing in this vein, we note that the residuals are

$$X_i - T_i = (1 - \gamma_{ii})X_i - \sum_{l \neq i} \gamma_{il} X_l .$$

Hence, if γ_{ii} is close to 1, a gross error in X_i does not necessarily show up in the residual $X_i - T_i$, but it might show up also elsewhere, say in the residual $X_m - T_m$ if γ_{mi} is large! Thus we may say that (2.10) also ensures a kind of "*robustness of design*." This robustness of design is optimized by having $\gamma_{ii} = p/n$ for all i; then we call the design matrix C *balanced*. Note that condition (e) $\limsup p/n < 1$ is a necessary condition for the γ_{ii} to be bounded away from 1.

In view of these remarks, condition (2.10) (and hence $p/n \to 0$) appears to be indispensable for any reasonably simple general asymptotic theory of robust regression. Of course, if one is not interested in potentially *all*, but only in *some* special linear estimands α, weaker conditions might suffice.

3. *M*-estimates of regression. We shall now estimate the unknown parameters by minimizing an expression of the form (1.5). We begin with a short discussion of the regularity conditions, which separate into three parts:

(i) *Conditions on the design matrix C.* The diagonal elements of the projection matrix

$$\Gamma = C(C^T C)^{-1} C^T$$

are assumed to be uniformly small:

$$\max_{1 \leq i \leq n} \gamma_{ii} = \varepsilon \ll 1 ;$$

the precise order of smallness will be specified from case to case. Without loss of generality we may choose the coordinate system in the parameter space such that the true parameter point is $\theta^0 = 0$, and that $C^T C$ is the identity matrix.

(ii) *Conditions on the estimate.* The function ρ is assumed to be convex and not monotone, and to possess bounded derivatives of sufficiently high order. In particular, $\psi(x) = (d/dx)\rho(x)$ should be continuous and bounded.

Convexity of ρ serves to guarantee equivalence between (1.5) and (1.7), and asymptotic uniqueness of the solution; otherwise it is unimportant. Higher order derivatives are technically convenient (Taylor expansions!), but their existence is hardly essential for the results to hold.

(iii) *Conditions on the error laws.* We assume that the errors U_i are independent, identically distributed, such that

$$E(\psi(U_i)) = 0 .$$

We require this in order that the expectation of (1.5) reaches its minimum at the true value θ^0.

The assumption of independence is a serious restriction; the assumption that

the errors are identically distributed simplifies notations and calculations, but could easily be relaxed.

The cases $\varepsilon p^2 \to 0$ and $\varepsilon p \to 0$. I have a reasonably simple rigorous treatment only if $\varepsilon p^2 \to 0$ or, with somewhat less satisfactory results, if $\varepsilon p \to 0$. This implies $p^3/n \to 0$ and $p^2/n \to 0$ respectively; thus, quite moderate values of p already lead to very large and impractical values for n.

The idea is to compare the zeros of the two vector-values random functions Φ and Ψ of θ:

$$\Phi_j(\theta) = \frac{-1}{E(\psi')} \sum_i \psi(X_i - \sum_k c_{ik}\theta_k)c_{ij} , \tag{3.1}$$

$$\Psi_j(\theta) = \theta_j - \frac{1}{E(\psi')} \sum_i \psi(X_i)c_{ij} . \tag{3.2}$$

The zero $\tilde{\theta}$ of Ψ,

$$\tilde{\theta}_j = \frac{1}{E(\psi')} \sum_i \psi(X_i)c_{ij} , \tag{3.3}$$

of course is not a genuine estimate. According to the results of Section 2, all linear combinations $\tilde{\alpha} = \sum a_j \tilde{\theta}_j$ are asymptotically normal iff $\varepsilon \to 0$. The zero $\hat{\theta}$ of Φ is our estimate, and we shall have to show that the difference between $\hat{\theta}$ and $\tilde{\theta}$ is small. Let a_j be indeterminate coefficients satisfying $\sum a_j^2 = 1$ and let s_i be defined by (2.4). Write for short

$$t_i = \sum_j c_{ij}\theta_j . \tag{3.4}$$

Since $C^TC = I$, we have

$$||t||^2 = \sum t_i^2 = \sum \theta_j^2 = ||\theta||^2 . \tag{3.5}$$

Expand $\sum a_j \Phi_j(\theta)$ into a Taylor series with remainder term:

$$\sum_j a_j \Phi_j(\theta) = \frac{-1}{E(\psi')}\{\sum \psi(X_i)s_i - \sum \psi'(X_i)t_i s_i + \tfrac{1}{2}\sum \psi''(X_i - \eta t_i)t_i^2 s_i\} \tag{3.6}$$

with $0 < \eta < 1$. This can be rearranged to give

$$\sum a_j(\Phi_j(\theta) - \Psi_j(\theta)) = \sum_{jk} \Delta_{jk} a_j \theta_k - \frac{1}{2E(\psi')} \sum_i \psi''(X_i - \eta t_i)t_i^2 s_i \tag{3.7}$$

where

$$\Delta_{jk} = \frac{1}{E(\psi')} \sum_i [\psi'(X_i) - E\psi'(X_i)]c_{ij}c_{ik} . \tag{3.8}$$

We now intend to show that (3.7) is uniformly small on sets of the form

$$\{(\theta, a) \mid ||\theta||^2 \leqq Kp, ||a|| = 1\} . \tag{3.9}$$

By Schwarz's inequality,

$$(\sum \Delta_{jk} a_j \theta_k)^2 \leqq \sum_{jk} \Delta_{jk}^2 \sum_j a_j^2 \sum_k \theta_k^2 = \sum_{jk} \Delta_{jk}^2 ||\theta||^2 . \tag{3.10}$$

We have

$$E(\sum_{jk} \Delta_{jk}^2) = \sum_{jk} E(\Delta_{jk}^2) = \sum_{jki} c_{ij}^2 c_{ik}^2 \cdot \frac{\operatorname{Var}(\psi')}{(E\psi')^2}, \tag{3.11}$$

and

$$\sum_{jki} c_{ij}^2 c_{ik}^2 = \sum_i \gamma_{ii}^2 \leqq \max \gamma_{ii} \sum_i \gamma_{ii} = \varepsilon p . \tag{3.12}$$

Let $\delta > 0$; then Markov's inequality shows that there is a constant

$$K_1 = \frac{\operatorname{Var}(\psi')}{(E\psi')^2} \cdot \frac{1}{\delta} \tag{3.13}$$

such that

$$P(\sum_{jk} \Delta_{jk}^2 \geqq K_1 \varepsilon p) \leqq \delta . \tag{3.14}$$

We conclude that with probability $> 1 - \delta$,

$$(\sum_{jk} \Delta_{jk} a_j \theta_k)^2 < K K_1 \varepsilon p^2 \tag{3.15}$$

holds simultaneously for all (a, θ) in (3.9).

Assume that ψ'' is bounded, say $|\psi''(x)| \leqq 2E(\psi')M$ for some M, then

$$\left| \frac{1}{2E(\psi')} \sum_i \psi''(X_i - \eta t_i) t_i^2 s_i \right| \leqq M \max |s_i| \sum t_i^2 \leqq M \varepsilon^{\frac{1}{2}} ||\theta||^2 , \tag{3.16}$$

(see (3.5) and the remarks preceding (2.8)).

If we put things together, we obtain that with probability $> 1 - \delta$, (3.7) is bounded in absolute value by

$$r = ((KK_1)^{\frac{1}{2}} + MK)(\varepsilon p^2)^{\frac{1}{2}} \tag{3.17}$$

and this uniformly on the set (3.9). Since the result holds simultaneously for all a with $||a|| = 1$, we have in fact shown that with probability $> 1 - \delta$

$$||\Phi(\theta) - \Psi(\theta)|| \leqq r \qquad \text{for} \quad ||\theta||^2 \leqq Kp . \tag{3.18}$$

As

$$E(||\tilde{\theta}||^2) = \frac{E(\psi^2)}{(E\psi')^2} p , \tag{3.19}$$

it follows from Markov's inequality that $||\tilde{\theta}||^2 \leqq Kp/4$ with arbitrarily high probability, provided K is chosen large enough. Moreover, then

$$||\Phi(\theta) - \theta|| \leqq ||\Phi(\theta) - \Psi(\theta)|| + ||\tilde{\theta}|| \leqq r + \tfrac{1}{2}(Kp)^{\frac{1}{2}} \tag{3.20}$$

on the set $||\theta||^2 \leqq Kp$.

If $\varepsilon p \to 0$, r can be made smaller than $\frac{1}{2}(Kp)^{\frac{1}{2}}$, so that (3.20) implies

$$||\Phi(\theta) - \theta|| < ||\theta|| \tag{3.21}$$

on the sphere $||\theta||^2 = Kp$, and we conclude from Brouwer's fixed point theorem, that $\Phi(\theta)$ has a zero $\hat{\theta}$ inside the ball $||\theta||^2 < Kp$.

Moreover, if we insert $\theta = \hat{\theta}$ into (3.18), we obtain that

$$||\hat{\theta} - \tilde{\theta}|| \leqq r . \tag{3.22}$$

In particular, if $\varepsilon p^2 \to 0$, this implies

$$||\hat{\theta} - \tilde{\theta}|| \to 0 \tag{3.23}$$

in probability.

If only $\varepsilon p \to 0$, we obtain the weaker result that

$$\frac{||\hat{\theta} - \tilde{\theta}||}{p^{\frac{1}{2}}} \to 0 \tag{3.24}$$

in probability. Note that $||\tilde{\theta}|| \sim p^{\frac{1}{2}}$ in view of (3.19).

Let $\hat{\alpha} = \sum a_j \hat{\theta}_j$ and $\tilde{\alpha} = \sum a_j \tilde{\theta}_j$, with $||a|| = 1$. Recall that $\hat{\alpha}$ is the estimate to be investigated, while $\tilde{\alpha}$ is a sum of independent random variables and is asymptotically normal if $\varepsilon \to 0$.

PROPOSITION 3.1. *If* $\varepsilon p^2 \to 0$, *then*

$$\sup_{||a||=1} |\hat{\alpha} - \tilde{\alpha}| \to 0 \tag{3.25}$$

in probability.

If a *is chosen at random with respect to the invariant measure on the sphere* $||a|| = 1$, *and if* $\varepsilon p \to 0$, *then*

$$\hat{\alpha} - \tilde{\alpha} \to 0 \tag{3.26}$$

in probability. In particular, (3.26) *implies that* $\hat{\alpha}$ *is asymptotically normal.*

PROOF. (3.25) is an immediate consequence of (3.23); (3.26) follows from (3.24) and the fact that the average of $|\hat{\alpha} - \tilde{\alpha}|^2$ over the unit sphere $||a|| = 1$ is $||\hat{\theta} - \tilde{\theta}||^2/p$.

Incidentally, the assumption that the true parameter point is $\theta^0 = 0$ was used only in (3.19). For instance, if θ^* is any estimate satisfying $||\theta^* - \theta^0|| = O_P(p^{\frac{1}{2}})$, then we can show in the same way that just one step of Newtons's method (for solving $\Phi(\theta) = 0$, with trial value θ^*) leads to an estimate $\hat{\theta}^*$ satisfying

$$||\hat{\theta}^* - \tilde{\theta}|| \to 0\,, \qquad ||\hat{\theta}^* - \hat{\theta}|| \to 0$$

in probability, provided $\varepsilon p^2 \to 0$.

(I conjecture that (3.26) holds for any-fixed or random-choice of a which is independent of the observations, provided $\varepsilon p \to 0$. On the other hand, it appears that (3.26) is not true in general if $\varepsilon p \nrightarrow 0$, see Section 5.)

4. Some formal power series expansions. The proofs employed in the preceding section break down when ε converges to zero at a slower rate than $o(1/p^2)$ or $o(1/p)$. It is by no means clear whether the results break down too. In order to obtain some heuristic insights into what is going on, I resorted to (formal) asymptotic expansions, ordered according to powers of ε. Although I was not able to bound the remainder terms and thus not able to show that the formal expansions are indeed asymptotic expansions, the leading terms give an interesting picture and prepare a fertile ground for conjectures and speculations, some of

which are mentioned at the end of Section 6. In particular, the strikingly different behavior for symmetric and asymmetric error distributions is intriguing. The expansions of this and of the next two sections make sense already for $p = 1$.

Put

$$\lambda(t) = E(\psi(U + t)) . \tag{4.1}$$

As before, we denote the fitted values by

$$T_i = \sum_j c_{ij}\hat{\theta}_j \tag{4.2}$$

where $(\hat{\theta}_j)$ is the solution of (1.7).

The guiding idea is to expand (1.7) into a Taylor series around the true value $\theta = 0$, and to find asymptotic expansions of the interesting quantities, ordered according to powers of ε.

We rewrite (1.7) as

$$\hat{\theta}_j = \sum_i \frac{\psi(X_i)}{\lambda'(0)} c_{ij} + R_j(T) , \tag{4.3}$$

where the remainder term is

$$R_j(T) = \sum_i \frac{1}{\lambda'(0)} [\psi(X_i - T_i) - \psi(X_i) + \lambda'(0)T_i]c_{ij} . \tag{4.4}$$

We put for short

$$Y_{ki} = \frac{\psi^{(k)}(X_i) - \lambda^{(k)}(0)}{\lambda'(0)} , \qquad k \geqq 0 \tag{4.5}$$

$$\beta_k = \lambda^{(k)}(0)/\lambda'(0) \tag{4.6}$$

(in these two formulas the upper index k stands for k-fold differentiation) and expand the remainder term into a Taylor series

$$R_j(T) = \sum_{k\geqq 1} R_{kj}(T) \tag{4.7}$$

with

$$R_{kj}(T) = \frac{(-1)^k}{k!} \sum_i Y_{ki} T_i^k c_{ij} + \frac{(-1)^{k+1}}{(k+1)!} \beta_{k+1} \sum_i T_i^{k+1} c_{ij} . \tag{4.8}$$

In particular, (4.3) now reads

$$\hat{\theta}_j = \sum_i Y_{0i} c_{ij} + R_j(T) . \tag{4.9}$$

We start a bootstrap procedure, putting

$$\hat{\theta}_j^{(1)} = \sum_i Y_{0i} c_{ij} \tag{4.10}$$

$$T_i^{(k)} = \sum_j \hat{\theta}_j^{(k)} c_{ij} \tag{4.11}$$

$$\hat{\theta}_j^{(k+1)} = \sum_i Y_{0i} c_{ij} + R_j(T^{(k)}) . \tag{4.12}$$

We may view these quantities as formal power series in the Y_{ki}. The terms of order k in the Y's are the same in $\hat{\theta}^{(k)}$, $\hat{\theta}^{(k+1)}$, $\cdots$, so the procedure converges in

a formal way. Evidently, we obtain the same expansion if we replace the recursion formula (4.12) by

$$\hat{\theta}_j^{(k+1)} = \hat{\theta}_j^{(1)} + [R_{1j}(T^{(k)}) + R_{2j}(T^{(k-1)}) + \cdots + R_{kj}(T^{(1)})]_{k+1} \tag{4.13}$$

where $[\ ]_{k+1}$ means that only the terms of order $\leqq k+1$ in the Y are retained. Since (4.13) involves only polynomials in the Y, it is more suitable for the actual calculation of the expansion.

Let U_{kj} be the sum of all terms of order k in this expansion of $\hat{\theta}_j$, so that

$$\hat{\theta}_j^{(k)} = U_{1j} + U_{2j} + \cdots + U_{kj}\,. \tag{4.14}$$

A tedious calculation gives the following initial terms:

$$U_{1j} = \textstyle\sum_i Y_{0i} c_{ij}$$

$$U_{2j} = -\textstyle\sum_{il} Y_{0l} Y_{1i} \gamma_{il} c_{ij} + \frac{\beta_2}{2} \sum_{ilm} Y_{0l} Y_{0m} \gamma_{il} \gamma_{im} c_{ij}$$

$$\begin{aligned} U_{3j} = {} & \textstyle\sum Y_{0m} Y_{1i} Y_{1l} \gamma_{il} \gamma_{lm} c_{ij} - \frac{\beta_2}{2} \sum Y_{0m} Y_{0g} Y_{1i} \gamma_{il} \gamma_{lm} \gamma_{lg} c_{ij} \\ & - \beta_2 \textstyle\sum Y_{0m} Y_{0g} Y_{1l} \gamma_{il} \gamma_{im} \gamma_{lg} c_{ij} + \frac{\beta_2^2}{2} \sum Y_{0m} Y_{0g} Y_{0s} \gamma_{il} \gamma_{im} \gamma_{lg} \gamma_{ls} c_{ij} \\ & + \tfrac{1}{2} \textstyle\sum Y_{0m} Y_{0g} Y_{2i} \gamma_{im} \gamma_{ig} c_{ij} - \frac{\beta_3}{6} \sum Y_{0m} Y_{0g} Y_{0s} \gamma_{im} \gamma_{ig} \gamma_{is} c_{ij}\,. \end{aligned}$$

In the symmetric case (when the errors U_i are symmetrically distributed and ρ is symmetric around 0), the even numbered coefficients β_k are zero and about half of the terms disappear. For reasons of symmetry, we have then also

$$E(U_{kj}) = 0$$

and there is strong evidence (although no proof yet) that

$$E(U_{kj}^2) = O(\varepsilon^{k-1})\,.$$

In the symmetric case the situation is rather more complicated since there are very sizeable bias terms, but the U_{kj} still appear to be ordered according to powers of ε.

5. The bias terms. In the beginning of Section 3 we standardized our problem in such a way that, ideally, the estimates $\hat{\theta}_j$ should have a non-degenerate normal distribution centered at 0, but in the asymmetric case there can be sizeable shifts or "biases."

The first nonzero bias term is

$$E(\hat{\theta}_j^{(2)}) = E(U_{2j}) = K_0 \textstyle\sum_i \gamma_{ii} c_{ij}\,, \tag{5.1}$$

with

$$K_0 = -E(Y_0 Y_1) + \frac{\beta_2}{2} E(Y_0^2)\,. \tag{5.2}$$

(We are writing $Y_0, Y_1, \cdots$ instead of $Y_{0i}, Y_{1i}, \cdots$ in formulas holding for any fixed i.)

The corresponding bias term of $\hat{\alpha} = \sum a_j \hat{\theta}_j$, with $\sum a_j^2 = 1$, is

$$E(\hat{\alpha}^{(2)}) = K_0 \sum_i \gamma_{ii} s_i , \tag{5.3}$$

with $s_i = \sum_j c_{ij} a_j$. Schwarz's inequality gives

$$|\sum_i \gamma_{ii} s_i| \leqq (\sum_i \gamma_{ii}^2)^{\frac{1}{2}} \leqq (\varepsilon p)^{\frac{1}{2}} . \tag{5.4}$$

This need not be small when ε is small, and the bias can be very serious indeed. For example, take the balanced case $\gamma_{ii} = p/n$, and assume that θ_1 corresponds to a main effect equally affecting all observations, i.e., $c_{i1} \equiv 1/n^{\frac{1}{2}}$. Then

$$E(\hat{\theta}_1^{(2)}) = K_0 \frac{p}{n^{\frac{1}{2}}} \tag{5.5}$$

$$E(\hat{\theta}_j^{(2)}) = 0 , \qquad j > 1 ,$$

and the bias of $\hat{\theta}_1^{(2)}$ does not tend to 0 unless $\varepsilon p = p^2/n \to 0$. (But I should hasten to point out that this bias is still negligible in comparison to the bias $\delta n^{\frac{1}{2}}$ of $\hat{\theta}_l$ caused by a systematic error $+\delta$ in all observations.)

The leading bias term of the fitted value T_i is

$$E(T_i^{(2)}) = K_0 \sum_l \gamma_{il} \gamma_{li} . \tag{5.6}$$

The Schwarz inequality gives

$$|\sum_l \gamma_{il} \gamma_{li}| \leqq (\sum_l \gamma_{il}^2 \sum_l \gamma_{li}^2)^{\frac{1}{2}} \leqq \varepsilon p^{\frac{1}{2}} . \tag{5.7}$$

The following simple design matrix shows that this bound is asymptotically sharp:

$$C = \begin{pmatrix} e & 0 \cdots 0 & 0 \\ 0 & e \cdots 0 & 0 \\ \vdots & & \vdots \\ 0 & 0 \quad e & 0 \\ 0 & 0 \quad 0 & e \\ g & g \cdots g & g \end{pmatrix} \tag{5.8}$$

where e stands for a column vector with all r components equal to 1, and $g = r/(n - p)$.

That is, we assume that $n = pr + 1$, that each θ_j is observed r times, and that there is an additional observation X_n of $g \cdot (\theta_1 + \cdots + \theta_p)$. This C does not satisfy $C^T C = I$; but $\gamma_{ii} \equiv p/n$, and

$$\sum_l \gamma_{il} \gamma_{ln} = \left(\frac{p}{n}\right)^2 [1 + (r(n - p))^{\frac{1}{2}}] \sim \frac{p}{n} p^{\frac{1}{2}} . \tag{5.9}$$

Evidently, the small biases in the $\hat{\theta}_j$ may add up to a large bias in $T_n = g \times (\hat{\theta}_1 + \cdots + \hat{\theta}_p)$.

Thus, an outlying residual $X_n - T_n$ might have been caused not by a gross error in X_n, but by a large bias in the fitted value T_n!

By the way, in the frequent case where the diagonal vector (γ_{ii}) either belongs

to the span of the column vectors of C, or is orthogonal to it, (γ_{ii}) is an eigenvector of Γ to the eigenvalue 1 or 0 respectively, hence (5.7) improves to

$$|\sum_l \gamma_{ll}\gamma_{li}| \leqq \varepsilon .$$

Some (unchecked) calculations indicate that the higher order bias terms behave as follows. While $E(U_{1j}) = 0$, $E(U_{2j}) = O((\varepsilon p)^{\frac{1}{2}})$, it appears that $E(U_{3j})$ and $E(U_{4j})$ are of the order $O(\varepsilon(\varepsilon p)^{\frac{1}{2}})$, and $E(U_{5j})$, $E(U_{6j})$ of the order $O(\varepsilon^2(\varepsilon p)^{\frac{1}{2}})$.

6. The covariance terms. As already mentioned, calculations for $k = 1, 2, 3$ indicate that in the symmetric case

$$\text{Var}\,(U_{kj}) = O(\varepsilon^{k-1}) . \tag{6.1}$$

Thus, we can hope that

$$\text{Cov}\,(\hat\theta_j, \hat\theta_k) = \text{Cov}\,(\hat\theta_j^{(1)}, \hat\theta_k^{(1)}) + O(\varepsilon) , \tag{6.2}$$

$$\text{Cov}\,(\hat\theta_j, \hat\theta_k) = \text{Cov}\,(\hat\theta_j^{(3)}, \hat\theta_k^{(3)}) + O(\varepsilon^2) , \tag{6.3}$$

and

$$\begin{aligned}\text{Cov}\,(\hat\theta_j^{(3)}, \hat\theta_k^{(3)}) = {} & \text{Cov}\,(U_{1j}, U_{1k}) + \text{Cov}\,(U_{2j}, U_{2k}) \\ & + \text{Cov}\,(U_{1j}, U_{2k}) + \text{Cov}\,(U_{1j}, U_{3k}) \\ & + \text{Cov}\,(U_{2j}, U_{1k}) + \text{Cov}\,(U_{3j}, U_{1k}) + O(\varepsilon^2) .\end{aligned} \tag{6.4}$$

In the asymmetric case we can expect that the large bias terms will cause trouble and spoil the remainder terms. Nevertheless, we shall evaluate (6.4) in the general asymmetric case:

$$\begin{aligned}\text{Cov}\,(\hat\theta_j^{(3)}, \hat\theta_k^{(3)}) = {} & E(Y_0^2)\delta_{jk} + K_1 \sum_i \gamma_{ii} c_{ij} c_{ik} + K_2 \sum_{il} \gamma_{il}^2 c_{ij} c_{lk} \\ & + K_3 \sum_{il} \gamma_{il}\gamma_{ll} c_{ij} c_{ik} + O(\varepsilon^2)\end{aligned} \tag{6.5}$$

with

$$\begin{aligned}K_1 &= 3E(Y_0^2)E(Y_1^2) - 2E(Y_0^2 Y_1) + 3E(Y_0^2)E(Y_0 Y_2) - \beta_3(E(Y_0^2))^2 \\ K_2 &= 3(E(Y_0 Y_1))^2 + \beta_2 E(Y_0^3) - 8\beta_2 E(Y_0^2)E(Y_0 Y_1) + \frac{5\beta_2^2}{2}(E(Y_0^2))^2 \\ K_3 &= 2(E(Y_0 Y_1))^2 - 3\beta_2 E(Y_0^2)E(Y_0 Y_1) + \beta_3^2(E(Y_0^2))^2 .\end{aligned}$$

In the symmetric case one has $K_2 = K_3 = 0$.

Let $\hat\alpha = \sum a_j \hat\theta_j$ with $\sum a_j^2 = 1$ and $s_i = \sum c_{ij} a_j$ as before, then

$$\begin{aligned}\text{Var}\,(\hat\alpha) = {} & E(Y_0^2) + K_1 \sum_i \gamma_{ii} s_i^2 + K_2 \sum_{il} \gamma_{il}^2 s_i s_l \\ & + K_3 \sum_i (\sum_l \gamma_{il}\gamma_{ll}) s_i^2 + O(\varepsilon^2) .\end{aligned} \tag{6.6}$$

Evidently,

$$0 \leqq \sum_i \gamma_{ii} s_i^2 \leqq \varepsilon .$$

The next term satisfies

$$0 \leqq \sum_{il} \gamma_{il}^2 s_i s_l \leqq \varepsilon .$$

PROOF. Let (t_i) be an eigenvector belonging to the largest eigenvalue λ of (γ_{il}^2).

Then $$\lambda t_i = \sum_l \gamma_{il}^2 t_l \leqq \sum_l \gamma_{il}^2 \max_l t_l = \gamma_{ii} \max_l t_l .$$

Thus $$\lambda \max_l t_l \leqq \varepsilon \max_l t_l ,$$

hence $\lambda \leqq \varepsilon$, and thus

$$0 \leqq \sum_{il} \gamma_{il}^2 s_i s_l \leqq \varepsilon .$$

But the third term causes trouble. I do not know whether the obvious bound

$$|\sum_i \sum_l \gamma_{il} \gamma_{ll} s_i^2| \leqq \varepsilon p^{\frac{1}{2}}$$

(which follows from (5.7)) is sharp. With the matrix (5.8) and $a_j = 1/p^{\frac{1}{2}}$ one obtains

$$\sum_i \sum_l \gamma_{il} \gamma_{ll} s_i^2 \sim \left(\frac{p}{n}\right)^2 p^{\frac{1}{2}} = \varepsilon^2 p^{\frac{1}{2}} .$$

In any case, the third term in general is not of the order $O(\varepsilon)$, thus, in the asymmetric case, the conjectures (6.2) and (6.3) cannot both be true.

We summarize this section with a few conjectural conclusions based on the formal expansions.

Conjectural conclusions. Assume first that the error distribution and ρ are symmetric around 0. Then it appears that the estimates have reasonably simple asymptotic properties if and only if $\varepsilon \to 0$. In particular, $\hat{\alpha}^{(1)} = \sum a_j \hat{\theta}_j^{(1)} = \tilde{\alpha}$ then is asymptotically normal for any (a_j) with $\sum a_j^2 = 1$, and the difference between $\hat{\alpha} = \sum a_j \hat{\theta}_j$ and $\hat{\alpha}^{(1)}$ appears to tend to 0 in probability.

The asymmetric case seems to behave well only under the stronger assumption that $\varepsilon p \to 0$. In particular according to (5.5), $E(\hat{\alpha} - \hat{\alpha}^{(1)}) \sim K_0 \sum \gamma_{ii} s_i$ can be of the order $(\varepsilon p)^{\frac{1}{2}}$. And unless $\varepsilon p^{\frac{1}{2}} \to 0$, we cannot even expect that the variances of $\hat{\alpha}$ and $\hat{\alpha}^{(1)}$ are close to each other. But unless p is well above 100, these effects are hardly noticeable in practice, cf. Section 9, Table 2.

7. Estimation of the covariance matrix of $\hat{\theta}$. Analogy with the classical expression suggests to estimate the covariance matrix of $\hat{\theta}$ by a matrix of the form

$$\frac{\frac{1}{n-p} \sum \psi(X_i - T_i)^2}{\left[\frac{1}{n} \sum \psi'(X_i - T_i)\right]^2} (C^T C)^{-1} . \tag{7.1}$$

Perhaps the denominator $n - p$ is inaccurate, and perhaps $C^T C$ should be replaced by a matrix proportional to $W = (w_{jk})$:

$$w_{jk} = \sum_i \psi'(X_i - T_i) c_{ij} c_{ik} . \tag{7.2}$$

Also here, formal power series expansions may give some heuristic insights. From now on, we shall assume that $p \to \infty$, so that $1/n$ is negligible against p/n.

(i) *The term* $n^{-1} \sum \psi'(X_i - T_i)$. The expansion of $\psi'(X_i - T_i)/\lambda'(0)$ begins with

$$\frac{\psi'(X_i - T_i)}{\lambda'(0)} = 1 + Y_{1i} - \beta_2 T_i - Y_{2i} T_i + \frac{\beta_3}{2} T_i^2 + \cdots . \tag{7.3}$$

The leading bias terms are

$$E(T_i) = K_0 \sum_l \gamma_{ll}\gamma_{li} + \cdots \qquad \text{(cf. (5.6))} \tag{7.4}$$

$$E(Y_{2i}T_i) = E(Y_0Y_2)\gamma_{ii} + \cdots \tag{7.5}$$

$$E(T_i^2) = E(Y_0^2)\gamma_{ii} + \cdots . \tag{7.6}$$

Thus, after summing over i,

$$E\left\{\frac{1}{n}\sum \psi'(X_i - T_i)\right\} = \lambda'(0)\left\{1 - \frac{p}{n}\left[K_4 + \frac{\sum_{il}\gamma_{ll}\gamma_{li}}{p}K_5\right]\right\} + \cdots \tag{7.7}$$

with

$$K_4 = E(Y_0Y_2) - \frac{\beta_3}{2}E(Y_0^2),$$

$$K_5 = \beta_2 K_0 .$$

If the diagonal (γ_{ii}) belongs to the span of the columns of C, the factor in front of K_5 is 1.

The variance of $n^{-1}\sum \psi'(X_i - T_i)$ is negligible, being of the order $O(1/n)$.

(ii) *The term* $n^{-1}\sum \psi(X_i - T_i)^2$. The expansion begins with

$$\frac{\psi(X_i - T_i)^2}{\lambda'(0)^2} = Y_{0i}^2 + T_i^2(1 + Y_{1i})^2 - 2T_iY_{0i}(1 + Y_{1i}) + T_i^2Y_{0i}(\beta_2 + Y_{2i}) + \cdots \tag{7.8}$$

and the leading bias terms are

$$E(T_i^2(1 + Y_{1i})^2) = \gamma_{ii}E(Y_0^2)(1 + EY_{1i}^2) + \cdots$$

$$E(2T_iY_{0i}(1 - Y_{1i})) = \gamma_{ii}(2E(Y_0^2) + 2E(Y_0^2Y_1)) + 2(\textstyle\sum_l \gamma_{ll}\gamma_{li})K_0E(Y_0Y_1) + \cdots$$

$$E(T_i^2Y_{0i}(\beta_2 + Y_{2i})) = \gamma_{ii}E(Y_0^2)E(Y_0Y_2) + \cdots .$$

Thus we obtain

$$E\left[\frac{1}{n}\sum \psi(X_i - T_i)^2\right] = \lambda'(0)^2E(Y_0^2)\left\{1 - \frac{p}{n}\left[1 - K_6 - \frac{\sum_{il}\gamma_{ll}\gamma_{li}}{p}K_7\right]\right\} + \cdots \tag{7.9}$$

with

$$K_6 = E(Y_1^2) - 2\frac{E(Y_0^2Y_1)}{E(Y_0^2)} + E(Y_0Y_2)$$

$$K_7 = 2\frac{(E(Y_0Y_1))^2}{E(Y_0^2)} - \beta_2E(Y_0Y_1) = -2\frac{E(Y_0Y_1)}{E(Y_0^2)}K_0 .$$

The variance of $n^{-1}\sum \psi(X_i - T_i)^2$ is again negligible, being of the order $O(1/n)$, thus

$$E\left[\frac{1}{n-p}\sum \psi(X_i - T_i)^2\right] = (\lambda'(0))^2E(Y_0^2)\left\{1 + \frac{p}{n}\left[K_6 + \frac{\sum_{il}\gamma_{ll}\gamma_{li}}{p}K_7\right]\right\} + \cdots . \tag{7.10}$$

(iii) *The matrix W.* Assume $C^TC = I$ and define W as in (7.2), then we obtain

$$\frac{W}{\lambda'(0)} = I + B\,, \tag{7.11}$$

where the expansion of the matrix B begins

$$B_{jk} = \sum_i [Y_{1i} - (\beta_2 + Y_{2i})T_i + \tfrac{1}{2}\beta_3 T_i^2]c_{ij}c_{ik} + \cdots\,. \tag{7.12}$$

We are interested in $W = \lambda'(0)(I + B)$

$$W^{-1} = \frac{1}{\lambda'(0)}(I - B + B^2 - \cdots) \tag{7.13}$$

$$W^{-2} = \frac{1}{\lambda'(0)^2}(I - 2B + 3B^2 - \cdots)\,.$$

The leading bias terms are

$$E(B_{jk}) = -\sum_i [K_4\gamma_{ii} + K_5 \sum_l \gamma_{ll}\gamma_{li}]\, c_{ij}c_{ik} + \cdots$$

$$E(\textstyle\sum_m B_{jm}B_{mk}) = E(Y_1^2)\sum_i \gamma_{ii}c_{ij}c_{ik} + 2K_5 \sum_{il}\gamma_{il}^2 c_{ij}c_{lj}\,. \tag{7.14}$$

Thus, in particular,

$$E(W^{-1}_{jk}) = \frac{1}{\lambda'(0)}\{\delta_{jk} + [K_4 + E(Y_1^2)]\sum_i \gamma_{ii}c_{ij}c_{ik} + K_5 \sum_i (\sum_l \gamma_{ll}\gamma_{li})c_{ij}c_{ik} + 2K_5 \sum_{il}\gamma_{il}^2 c_{ij}c_{lk}\} + \cdots \tag{7.15}$$

$$E(W^{-2}_{jk}) = \frac{1}{\lambda'(0)^2}\{\delta_{jk} + [2K_4 + 3E(Y_1^2)]\sum_i \gamma_{ii}c_{ij}c_{ik} + 2K_5 \sum_i (\sum_l \gamma_{ll}\gamma_{li})\, c_{ij}c_{ik} + 6K_5 \sum_{il}\gamma_{il}^2 c_{ij}c_{lk}\}\cdots\,. \tag{7.16}$$

The three simplest proposals for estimating the covariance matrix of $\hat{\theta}$ seem to be

$$\frac{\frac{1}{n-p}\sum \psi(X_i - T_i)^2}{\left[\frac{1}{n}\sum \psi'(X_i - T_i)\right]^2}(C^TC)^{-1} \tag{7.17}$$

$$\frac{\frac{1}{n-p}\sum \psi(X_i - T_i)^2}{\frac{1}{n}\sum \psi'(X_i - T_i)}W^{-1} \tag{7.18}$$

$$\frac{1}{n-p}\sum \psi(X_i - T_i)^2 W^{-1}(C^TC)W^{-1}\,. \tag{7.19}$$

All three reduce to the classical expression in the classical case $\psi(x) = x$. All three agree if $p = 1$, $c_{ij} = 1/n^{\frac{1}{2}}$ (estimation of a location parameter).

We now insert the expressions derived in (i), (ii), (iii) into these estimates, in order to compare them with the covariance matrix (6.5). This leads to perfectly horrible and not easily comparable formulas. However, if we assume that the

error distribution and ρ are symmetric and that the matrix C is balanced ($\gamma_{ii} \equiv p/n$), there is a considerable simplification. It turns out that all three covariance estimates contain a bias of the order $O(p/n)$, and that it is possible to make them unbiased up to $O(p^2/n^2)$ terms, if one multiplies them by a correction factor K^2, K or K^{-1} respectively, where

$$K = 1 + \frac{p}{n} E(Y_1^2) . \tag{7.20}$$

(In actual use, one would have to replace the unknown $E(Y_1^2)$ by an estimate.)

8. Computation of such estimates. We propose to use the derivative of (1.6), i.e.,

$$\psi(x) = \max(-c, \min(c, x)) , \tag{8.1}$$

and we intend to estimate simultaneously with θ also a scale parameter σ. (Since the variance of the estimate of σ will be of the order $O(1/n) = o(p/n)$, this modification does not influence the asymptotic theory which we had derived for fixed σ.) Admittedly the above ψ does not possess the smoothness postulated in Section 3, but it is quite unlikely that this should make any difference in the results, if the distribution of the errors U is reasonably smooth.

We propose to solve the non-linear system

$$\sum_i \psi\left(\frac{X_i - \sum_k c_{ik}\theta_k}{\sigma}\right) c_{ij} = 0 , \qquad j = 1, \cdots, p \tag{8.2}$$

$$\frac{1}{n-p} \sum_i \psi\left(\frac{X_i - \sum_k c_{ik}\theta_k}{\sigma}\right)^2 = \beta . \tag{8.3}$$

with

$$\beta = E_\Phi \psi(U)^2 .$$

There are two main methods for solving (8.2), a crude and a somewhat more sophisticated one.

First the crude method. The idea behind it is simply to linearize (8.2):

$$\sum_i \sigma\psi\left(\frac{X_i}{\sigma}\right) c_{ij} - \sum_k \sum_i \psi'\left(\frac{X_i}{\sigma}\right) c_{ij} c_{ik} \theta_k = 0 , \tag{8.4}$$

to replace $\psi'(X_i/\sigma)$ by its expected value (or an estimate of it) and then to solve for θ. To improve accuracy, the procedure can be repeated with the X_i replaced by the current residuals $X_i - T_i$.

More precisely, the procedure can be described as follows

1. Compute $(C^T C)^{-1}$.
2. Take starting values θ, σ.
3. Let

$$T_i = \sum_k c_{ik}\theta_k \qquad m = \frac{1}{n} \sum \psi'\left(\frac{X_i - T_i}{\sigma}\right)$$

$$Y_j = \sum_i \sigma\psi\left(\frac{X_i - T_i}{\sigma}\right) c_{ij} \qquad q = \frac{1}{n-p} \sum \psi\left(\frac{X_i - T_i}{\sigma}\right)^2 \sigma^2 .$$

4. Put

$$\sigma_{\text{new}} = \left(\frac{q}{\beta}\right)^{\frac{1}{2}}$$

$$\Delta\theta = \frac{1}{m}(C^TC)^{-1}Y$$

$$\theta_{\text{new}} = \theta + \Delta\theta\,.$$

5. If $||\Delta\theta||$ is smaller than some predetermined threshold, stop. Otherwise begin again at 2., with the new values for θ and σ.

The idea behind the sophisticated method is as follows. If ψ is piecewise linear, and if it would be known to which linear piece of ψ each residual belongs, the solution of (8.2), (8.3) is a problem of elementary algebra. Let $\theta^{(r)}$, $\sigma^{(r)}$ be trial values, then, with (8.1) there are three classes of residuals:

the lower class	$I_-^{(r)} = \{i \mid X_i - \sum c_{ik}\theta_k^{(r)} < -c\sigma^{(r)}\}$
the middle class	$I_0^{(r)} = \{i \mid -c\sigma^{(r)} \leqq X_i - \sum c_{ik}\theta_k^{(r)} \leqq c\sigma^{(r)}\}$
the upper class	$I_+^{(r)} = \{i \mid X_i - \sum c_{ik}\theta_k^{(r)} > c\sigma^{(r)}\}$.

If the partition $I_-^{(r)}$, $I_0^{(r)}$, $I_+^{(r)}$ agrees with that induced by the solution of (8.2), (8.3) then this final solution can be reached in just one step. In order to see this, we rewrite the system as

$$\sum_0 (X_i - \sum_k c_{ik}\theta_k)c_{ij} + (\sum_+ c_{ik} - \sum_- c_{ik})c\sigma = 0 \tag{8.5}$$

$$\sum_0 (X_i - \sum_k c_{ik}\theta_k)^2 + (\sum_+ 1 + \sum_- 1)c^2\sigma^2 = (n-p)\beta\sigma^2\,, \tag{8.6}$$

where the index 0, +, − at the summation sign indicates summations over $I_0^{(r)}$, $I_+^{(r)}$, $I_-^{(r)}$ respectively.

The computational procedure now is as follows.

1. Choose some starting values $\theta^{(0)}$, $\sigma^{(0)}$. Let $r = 0$.
2. Find the sets $I_0^{(r)}$, $I_+^{(r)}$, $I_-^{(r)}$.
3. Compute the matrix

$$W_{jk} = \sum_0 c_{ij}c_{ik}\,, \tag{8.7}$$

the vectors

$$Y_k = \sum_0 X_i c_{ik} \tag{8.8}$$

$$R_k = \sum_+ c_{ik} - \sum_- c_{ik}\,,$$

and the scalars

$$Q = \sum_0 X_i^2 \tag{8.9}$$

$$M = \sum_+ 1 + \sum_- 1\,.$$

4. Solve

$$\sum W_{jk}\tau_k = Y_j\,, \qquad j = 1, \cdots, p \tag{8.10}$$

$$\sum W_{jk}\delta_k = R_j\,, \qquad j = 1, \cdots, p$$

for τ and for δ.

5. Then $\theta_k(\sigma) = \tau_k + c\sigma\delta_k$ satisfies (8.5) identically in σ; if we insert this linear function into (8.6) and solve for σ^2, we obtain

$$(\sigma^{(r+1)})^2 = \frac{\sum_0 (X_i - \sum_k c_{ik}\tau_k)^2}{(n-p)\beta - c^2[\sum_0 (\sum_k c_{ik}\delta_k)^2 + M]} \tag{8.11}$$

or equivalently (but somewhat more sensitive to rounding errors)

$$(\sigma^{(r+1)})^2 = \frac{Q - \sum_k Y_k\tau_k}{(n-p)\beta - c^2[\sum_k R_k\delta_k + M]}\,. \tag{8.12}$$

6. Put $\theta_k^{(r+1)} = \tau_k + c\sigma^{(r+1)}\delta_k$ and find the corresponding partition $I_0^{(r+1)}$, $I_+^{(r+1)}$, $I_-^{(r+1)}$. If it agrees with the preceding one $I_0^{(r)}$, $I_+^{(r)}$, $I_-^{(r)}$, then ($\theta^{(r+1)}$, $\sigma^{(r+1)}$) solves (8.2). Otherwise replace r by $r+1$ and go to 3.

In order to improve numerical accuracy, it may be preferable to replace X_i by the current value of the residual $X_i - \sum_k c_{ik}\theta_k^{(r)}$ in (8.8) ff. Then, in 6., we have

$$\theta_k^{(r+1)} = \theta_k^{(r)} + \tau_k + c\sigma^{(r+1)}\delta_k\,.$$

Since there are only finitely many possible partitions I_0, I_+, I_-, the procedure must either stop after finitely many steps, or it must repeat itself periodically.

The sophisticated procedure ordinarily shows a very fast convergence (3 to 5 iterations are typical) but it is not foolproof. For larger values of p/n (above 0.1 or so) it happens with increasing frequency that either the matrix W becomes singular or that the denominator of (8.11), (8.12) becomes negative. (This occurs in particular if the initial value of σ is chosen too small.)

The crude procedure may need 15 to 30 iterations but seems to be foolproof (especially if m is replaced by a constant; convergence has not been proved yet). Its slowness is in part counterbalanced by the fact that some time consuming matrix operations have to be performed only once; moreover, it works for arbitrary ψ.

Finally, according to Section 7, the covariance matrix of $\hat{\theta}$ might be estimated either by

$$\frac{\frac{1}{n-p}\sum \psi\left(\frac{X_i - T_i}{\sigma}\right)^2 \sigma^2}{\frac{1}{n}\sum \psi'\left(\frac{X_i - T_i}{\sigma}\right)} KW^{-1} \tag{8.13}$$

or by

$$\frac{\frac{1}{n-p}\sum \psi\left(\frac{X_i - T_i}{\sigma}\right)^2 \sigma^2}{\left[\frac{1}{n}\sum \psi'\left(\frac{X_i - T_i}{\sigma}\right)\right]^2} K^2(C^TC)^{-1}\,. \tag{8.14}$$

Incidentally, the choice of starting values for θ and σ presents a problem. In the regression case there is no easy analogue to the sample median, and despite its known inadequacy (see Andrews *et al.* (1972)) one might have to start with the least squares solution.

9. Monte Carlo results. As the power series expansions of Sections 4 ff. are rather shaky, it was doubly necessary to check them by Monte Carlo calculations. Evidently, the matrix C must be chosen such that computational shortcuts are possible; for most of the experiments the matrix (5.8) was used, with ψ as in (8.1). In most cases, there were 50 replications if $n \leqq 1025$, but only 10 or fewer for larger n.

Tables 1 and 2 summarize some results for fixed σ. For the error distribution we took the normal, a conventional contaminated normal, the Cauchy and two highly asymmetric distributions, namely, the χ^2 with 2 and 4 degrees of freedom respectively (apart from a different scale; location was adjusted such that $E\psi(U_i) = 0$).

By "Var $(\hat{\theta}_i)$" the average variance of a parameter orthogonal to the last row of (5.8) is meant; the theoretical value is that given by (6.5) or (6.6), etc. In parentheses, we give the estimated standard deviation of the Monte Carlo averages, in units of the last given digit.

ESVAR is the Monte Carlo average of the coefficient of $(C^TC)^{-1}$ in (8.14):

$$\text{ESVAR} = \text{ave}\left\{\frac{(1/(n-p))\sum \psi(X_i - T_i)^2}{m^2} K^2\right\}$$

where

$$m = \frac{1}{n}\sum \psi'(X_i - T_i)$$

and where (7.20) has been estimated by

$$K = 1 + \frac{p}{n}\frac{1-m}{m}$$

(see the end of Section 7 for the motivation). Besides ESVAR, the ratio [observed Var $(\hat{\theta}_i)$]/ESVAR is also given.

For the asymmetric distributions, the "bias" is the average of $(n/p)^{\frac{1}{2}}T_n$, as given theoretically by (5.6), (5.9). Note that the variance of $(n/p)^{\frac{1}{2}}T_n$ is asymptotically the same as Var $(\hat{\theta}_i)$.

The agreement in general is very good, for normal errors is even fantastically good, and this down to $n/p = 4$. For Cauchy errors, the agreement is very good for $n/p \geqq 16$ and tolerable for $n/p = 8$.

Some larger, but mostly explainable discrepancies show up with the density $e^{-(x+\xi)}$, $x > -\xi$. For instance, the difference between the theoretical and the observed value of $n^{-1}\sum \psi'(X_i - T_i)$ coincides almost exactly with the observed frequency of residuals $< -c$, which is wrongly calculated as 0 by the Taylor expansion.

When the nuisance parameter σ is estimated simultaneously, the results (not shown here) are essentially the same, as predicted in Section 8.

TABLE 1
Symmetric distributions, fixed scale

	$n = 1025$		$\mathrm{Var}(\hat{\theta}_i)$					$E\left[\frac{1}{n}\sum \psi'(X_i - T_i)\right]$			$E\left[\frac{1}{n-p}\sum \psi(X_i - T_i)^2\right]$		
	$(n-1)/p$	p	theor.	obs.	ratio	ESVAR	ratio	theor.	obs.	ratio	theor.	obs.	ratio
	∞		1.0371					.8664			.7785		
normal	32	32	1.0357	1.0355(17)	1.000(2)	1.022(7)	1.01(1)	.8741	.8756(14)	1.002(2)	.7838	.7760(32)	.990(4)
$N(0, 1)$	16	64	1.0343	1.0334(11)	.999(1)	1.031(7)	1.00(1)	.8818	.8824(12)	1.001(1)	.7892	.7886(36)	.999(5)
$c = 1.5$	8	128	1.0316	1.0320(11)	1.000(1)	1.031(8)	1.00(1)	.8972	.8961(13)	.999(1)	.7999	.8039(39)	1.005(5)
	4	256	1.0262	1.0249(9)	.999(1)	1.008(12)	1.02(1)	.9280	.9250(17)	.997(1)	.8213	.8274(67)	1.007(8)
	∞		1.1073					.6827			.5161		
normal	32	32	1.1051	1.1056(49)	1.000(4)	1.083(9)	1.02(1)	.6965	.7006(18)	1.006(3)	.5209	.5165(17)	.992(3)
$N(0, 1)$	16	64	1.1029	1.0974(26)	.995(2)	1.100(10)	1.00(1)	.7102	.7101(19)	1.000(3)	.5257	.5262(20)	1.001(4)
$c = 1.0$	8	128	1.0984	1.0970(25)	.999(2)	1.088(9)	1.01(1)	.7378	.7363(18)	.998(2)	.5353	.5394(20)	1.008(4)
	4	256	1.0896	1.0860(36)	.997(3)	1.030(20)	1.05(2)	.7929	.7920(37)	1.000(5)	.5545	.5682(46)	1.025(8)
	∞		1.4655					.6194			.5623		
.85 $N(0, 1)$	32	32	1.4779	1.500(29)	1.01(2)	1.478(13)	1.01(2)	.6326	.6324(18)	1.000(3)	.5694	.5693(18)	1.000(3)
+.15 $N(0, 9)$	16	64	1.4903	1.484(16)	1.00(1)	1.494(14)	.99(1)	.6457	.6445(20)	.998(3)	.5764	.5783(19)	1.003(3)
$c = 1.0$	8	128	1.5152	1.538(17)	1.02(1)	1.506(16)	1.02(2)	.6720	.6694(21)	.996(3)	.5905	.5968(22)	1.011(4)
	4	256	1.5649	1.656(30)	1.06(2)	1.487(19)	1.11(2)	.7246	.7213(24)	.995(3)	.6186	.6425(31)	1.039(5)
	∞		2.546					.5000			.6366		
Cauchy	32	32	2.701	2.789(108)	1.03(4)	2.686(38)	1.04(4)	.5072	.5093(26)	1.004(5)	.6539	.6526(24)	.998(4)
$c = 1.0$	16	64	2.856	2.886(92)	1.01(3)	2.866(41)	1.01(4)	.5145	.5141(25)	.999(5)	.6712	.6713(24)	1.000(4)
	8	128	3.165	3.853(86)	1.22(3)	3.112(30)	1.24(3)	.5289	.5311(16)	1.004(3)	.7057	.7082(16)	1.004(2)

TABLE 2
Asymmetric distributions, fixed scale

	n/p (approx.)	p	n	Var $(\hat{\theta}_i)$ theor.	obs.	ratio	ESVAR	ratio
	∞			.552				
	32	33	1025	.562	.589(28)	1.05(5)	.574	1.03(5)
	32	129	4097		.591(33)	1.05(6)	.569	1.04(6)
$f(x) = e^{-(x+\xi)}$,	16	65	1025	.572	.616(18)	1.08(3)	.586	1.05(3)
$x > -\xi$ $(c = 1)$	16	257	4097		.580(21)	1.01(4)	.592	.98(4)
	8	129	1025	.592	.666(12)	1.12(2)	.618	1.08(2)
	8	513	4097		.669(17)	1.13(3)	.616	1.09(3)
	8	1025	8193		.653	1.10	.624	1.05
	4	257	1025	.633	.750(26)	1.18(4)	.645	1.16(4)
	∞			1.519				
	32	32	1025	1.56	1.66(7)	1.06(4)	1.57(2)	1.06(4)
$f(x) = (x + \xi)e^{-(x+\xi)}$,	16	64	1025	1.60	1.60(4)	1.00(3)	1.56(1)	1.02(3)
$x > -\xi$ $(c = 1.5)$	8	8	65	1.68	1.66(7)	.99(4)	1.60(3)	1.03(4)
	8	32	257	1.68	1.71(7)	1.02(4)	1.58(3)	1.08(5)
	8	128	1025	1.68	1.66(3)	.99(2)	1.58(2)	1.04(2)
	4	256	1025	1.83	1.76(3)	.96(2)	1.61(2)	1.10(2)

$E\left[\frac{1}{n}\sum \psi'(X_i - T_i)\right]$ theor.	obs.	ratio	$E\left[\frac{1}{n-p}\sum \psi(X_i - T_i)^2\right]$ theor.	obs.	ratio	bias of $(n/p)^{\frac{1}{2}}T_n$ theor.	obs.	ratio
.841			.391					
.846	.840	.993	.400	.400(2)	1.000(5)	.175	.194(13)	1.11(7)
	.841	.994		.398(3)	.995(8)	.342	.374(27)	1.09(8)
.853	.840	.985	.408	.404(3)	.990(7)	.340	.415(16)	1.22(5)
	.839	.984		.407(2)	.997(5)	.671	.790(35)	1.18(5)
.863	.841	.975	.425	.417(2)	.982(5)	.655	.862(21)	1.32(3)
	.841	.975		.416(2)	.980(5)	1.297	1.75(7)	1.35(5)
	.839	.972		.419	.990	1.829	2.47	1.35
.885	.859	.971	.458	.439(4)	.965(10)	1.216	1.70(8)	1.40(7)
.8034			.9803					
.808	.803(2)	.993(2)	.987	.994(5)	1.007(5)	.37	.37(2)	1.01(5)
.814	.811(2)	.997(2)	.993	.997(4)	1.005(4)	.73	.68(3)	.93(4)
.824	.825(3)	1.001(4)	1.005	1.008(9)	1.002(9)	.36	.32(2)	.88(5)
.824	.825(4)	1.001(4)	1.005	1.009(10)	1.004(10)	.72	.62(4)	.86(5)
.824	.823(2)	.999(2)	1.005	1.015(5)	1.010(5)	1.42	1.29(4)	.91(3)
.844	.851(2)	1.008(2)	1.030	1.068(6)	1.037(6)	2.64	2.39(8)	.90(3)

TABLE 3
Constants of Sections 5 to 7

	$N(0, 1)$ $c = 1.5$	$N(0, 1)$ $c = 1.0$	$.85N(0, 1) + .15N(0, 9)$ $c = 1.0$	Cauchy $c = 1.0$	$e^{-(x+\xi)}$, $x > -\xi$ $c = 1.0$	$(x + \xi)e^{-(x+\xi)}$, $x > -\xi$ $c = 1.5$
$E(\psi^2)$	.77847	.51606	.56234	.63262	.39077	.98025
$E(\psi')$	.86639	.68269	.61945	.50000	.84141	.80338
K_0	0	0	0	0	.1720	.3753
K_1	−.0437	−.0709	.3975	5.948	.3230	.0821
K_2	0	0	0	0	−.0408	.8203
K_3	0	0	0	0	.0413	.3574
K_4	−.2881	−.6459	−.6788	−.4627	−.1720	−.2522
K_5	0	0	0	0	−.0324	.0499
K_6	.2196	.2982	.4002	.8684	.5522	.0690
K_7	0	0	0	0	.1396	.1356
ξ					.84141	1.81045

REFERENCES

ANDREWS, D. F., BICKEL, P. J., HAMPEL, F. R., HUBER, P. J., ROGERS, W. H., and TUKEY, J. W. (1972). *Robust Estimates of Location: Survey and Advances.* Princeton Univ. Press.

ANSCOMBE, F. J. (1967). Topics in the investigation of linear relations fitted by the method of least squares. *J. Roy. Statist. Soc. Ser. B* **29** 1-52.

BICKEL, P. J. (1971a). A note on approximate (M) estimates in the linear model. To appear in *J. Amer. Statist. Assoc.*

BICKEL, P. J. (1971b). On some analogues to linear combinations of order statistics in the linear model. Unpublished manuscript.

EICKER, F. (1963). Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *Ann. Math. Statist.* **34** 447-456.

EICKER, F. (1967). Limit theorems for regressions with unequal and dependent errors. *Proc. Fifth Berkeley Symp. Math. Statist. Prob.* **1** 59-66.

GAUSS, C. F. (1821). Göttingische gelehrte Anzeigen. 321-327 (reprinted in *Werke* Bd. **4** page 98).

HAMILTON, W. C. (1970). The revolution in crystallography. *Science* **169** 133-141.

HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73-101.

HUBER, P. J. (1968). Robust confidence limits. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **10** 269-278.

HUBER, P. J. (1972a). Robust statistics. *Ann. Math. Statist.* **43** 1041-1067.

HUBER, P. J. (1972b). Robust correlation. (In preparation.)

JAECKEL, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of the residuals. *Ann. Math. Statist.* **43** 1449-1458.

JUREČKOVÁ J. (1971). Nonparametric estimate of regression coefficients. *Ann. Math. Statist.* **42** 1328-1338.

RELLES, D. A. (1968). Robust regression by modified least squares. Ph. D. thesis, Yale Univ.

DEPARTMENT OF MATHEMATICS
E. T. H.
CLAUSIUSSTR 55
8006 ZURICH, SWITZERLAND