The Annals of Statistics
1984, Vol. 12, No. 4, 1349–1368

ROBUST REGRESSION BASED ON INFINITESIMAL NEIGHBOURHOODS[1]

BY P. J. BICKEL

University of California, Berkeley

We study robust estimation in the general normal regression model with random carriers permitting small departures from the model. The framework is that of Bickel (1981). We obtain solutions of Huber (1982), Krasker-Hampel (1980) and Krasker-Welsch (1982) as special cases as well as some new procedures. Our calculations indicate that the optimality properties of these estimates are more limited than suggested by Krasker and Welsch.

1. Introduction. Our aim in this paper is to compare and contrast robust regression estimates proposed by Huber (1973, 1982), Hampel (1978), Krasker (1978) and Krasker and Welsch (1982) as well as to derive and motivate other estimates using infinitesimal neighbourhood models as in Rieder (1978), Bickel (1981) for instance. Some of the results are stated in the discussion to Huber (1982) while others were presented at the 1979 Regression Special Topics Meeting in Boulder.

We consider a "stochastic" regression model. We observe (x_i, y_i), $i = 1, \cdots, n$ independent with common distribution P where the x_i are $1 \times p$, y_i scalar. We think of these observations as being obtained by contamination or some other stochastic perturbation from ideal but unobservable (x_i^*, y_i^*) which follow an ordinary Gaussian regression,

$$y_i^* = x_i^* \theta^T + u_i^*, \quad i = 1, \cdots, n$$

where the u_i^* are independent $\mathcal{N}(0, \sigma^2)$. Our aim is to estimate θ using the (x_i, y_i). For this formulation to make sense we must either:

(a) Specify P so that θ is identifiable. For instance let

$$x_i = x_i^* \quad \text{and} \quad y_i = x_i \theta^T + u_i$$

where the u_i are independent of x_i with common distribution symmetric about 0. This is the usual generalization of the linear model discussed e. g. in Huber (1973). For less drastic alternatives see Sacks and Ylvisaker (1978). This has the disadvantage of implicitly assuming that contamination conforms to the linear structure of the original model.

(b) Suppose that P is so close to the distribution P_0 of (x_i^*, y_i^*) that biases necessarily imposed by the lack of identifiability of θ are of the same order of magnitude as the standard deviations of good estimates. That is we assume P is

Received December 1982; revised January 1984.
[1] Research supported in part by Office of Naval Research Contract N00014-80-C-0163.
AMS 1980 *subject classifications.* Primary 62J05; secondary 62F35.
Key words and phrases. Regression, robustness, infinitesimal neighbourhoods, Krasker-Welsch estimates.

in "an order $1/\sqrt{n}$ neighbourhood" about P_0. By suitably choosing the metric defining the neighbourhood we can make precise our ideas about what departures we want to guard against as well as gauge the best that we can do against such departures in terms of classical decision theoretic measures such as M.S.E. For a general discussion of this point of view see Bickel (1981), hereafter [B]. This is the approach we take in this paper.

We apply this point of view to several types of neighbourhoods below and derive the optimal solutions. For regression through the origin we recapture the by now classical estimate of Hampel as well as Huber's (1982) MIA:A solution. For the general regression model we derive various natural extensions of the MIA:A procedure as well as the Hampel-Krasker and Krasker-Welsch procedures. Finally, we derive some negative results suggesting that the (1982) Krasker-Welsch conjecture is false.

Specifically, let $u_i = y_i - x_i\theta^T$, $i = 1, \cdots, n$. Suppose $\sigma^2 = 1$. Write $F = (G, H(\cdot \mid \cdot))$, $F_0 = (G_0, \Phi)$ where G, respectively G_0, is the marginal distribution of x_1, $H(\cdot \mid x)$ is the conditional distribution of u_1 given $x_1 = x$ and Φ is the standard normal distribution (of u_1^*). Since P and F determine each other we can describe neighbourhoods through conditions on $F, H(\cdot \mid \cdot)$. Such neighbourhoods, which will depend on n, will be denoted by $\mathscr{F}(t)$ (with subscripts) where $tn^{-1/2}$ is the size of the neighbourhood, $t \geq 0$.

Error-free x neighbourhoods: $G = G_0$ (or $x = x^*$).

Contamination: We suppose we can represent

$$H(\cdot \mid x) = (1 - \varepsilon(x))\Phi(\cdot) + \varepsilon(x)M(\cdot \mid x)$$

where $M(\cdot \mid x)$ is an arbitrary probability distribution. The contamination neighbourhoods $\mathscr{F}_{c0}(t)$, $\mathscr{F}_{ac0}(t)$ are completely specified by:

$$\mathscr{F}_{c0}(t)\text{: } \sup_x \varepsilon(x) \leq tn^{-1/2}, \quad \mathscr{F}_{ac0}(t)\text{: } \int \varepsilon(x)G_0(dx) \leq tn^{-1/2}.$$

That is, for both neighbourhoods the type of contamination of y for each x can be arbitrary. But under $\mathscr{F}_{c0}$ the conditional probability of contamination for each x is at most $tn^{-1/2}$ while under $\mathscr{F}_{ac0}$ only the marginal (or "average") probability of contamination is restricted. These are the types of departures considered by Huber (1982), Section 5.

Closely related are the metric neighbourhoods,

$$\mathscr{F}_{d0}(t)\text{: } \sup_x d(H(\cdot \mid x), \Phi) \leq tn^{-1/2}, \quad \mathscr{F}_{ad0}(t)\text{: } \int d(H(\cdot \mid x), \Phi)G_0(dx) \leq tn^{-1/2}$$

where d is a metric on the space of probability distributions on R. Of particular interest are the variational and Kolmogorov metrics given respectively by

$$v(P, Q) = \sup\{|P(A) - Q(A)| : A \text{ Borel}\},$$

$$k(P, Q) = \sup_x |P(-\infty, x] - Q(-\infty, x]|.$$

Recall that contamination neighbourhoods are contained in the corresponding

variational neighbourhoods which are contained in the corresponding Kolmogorov neighbourhoods. The variational neighbourhoods can be interpreted as contamination neighbourhoods where ε can be a function not only of x but also of u^* and H is the conditional distribution of u_1 given x_1 and u_1^*. The complements of Kolmogorov neighbourhoods are identifiable in the sense of [B] at least if G_0 has finite support.

Errors in variables models: We drop the requirement that $G = G_0$ and proceed naturally, defining

$$\mathscr{F}_{c1}(t)\colon F = (1 - \varepsilon)F_0 + \varepsilon M$$

where M is an arbitrary probability distribution on R^{p+1}, $\varepsilon = tn^{-1/2}$.

$$\mathscr{F}_{d1}(t)\colon d(F, F_0) \leq tn^{-1/2}$$

where d is a metric on the probability distributions on R^{p+1}. Here v extends naturally and is of particular interest.

We consider estimates T_n of θ which are regression equivariant and asymptotically linear and consistent under the normal model. That is, for all $X_{n\times p}$, y, $b_{1\times p}$, T_n which is $1 \times p$ satisfies:

$$T_n(X, y + Xb^T) = T_n(X, y) + b \quad \text{(equivariance)} \tag{1.1}$$

and there exists $\psi\colon R^{p+1} \to R^p$ square integrable under F_0 such that

$$\int \psi(x, v)\Phi(dv)G_0(dx) = 0 \tag{1.2}$$

$$\int \psi^T(x, v)xv\Phi(dv)G_0(dx) = I, \quad \text{the } p \times p \text{ identity,} \tag{1.3}$$

and if $u = (u_1, \cdots, u_n)$, $X = (x_1^T, \cdots, x_n^T)^T$,

$$T_n(X, u) = n^{-1}\sum_{i=1}^n \psi(x_i, u_i) + o_p(n^{-1/2}) \quad \text{(linearity and consistency)} \tag{1.4}$$

under F_0. Let $\Psi = \{\psi\colon \psi$ square integrable function from R^{p+1} to R^p satisfying (1.2) and (1.3)$\}$.

All the usual consistent asymptotically normal estimates have this structure. In particular, under regularity conditions, the general (M) estimate T_n, solving

$$\sum_{i=1}^n \psi(x_i, y_i - x_iT_n^T) = 0 \tag{1.5}$$

with $\psi \in \Psi$ satisfies (1.1) and (1.4). For members F of $\mathscr{F}$ leading to models contiguous to that given by F_0, (1.1)–(1.4) imply that $n^{1/2}(T_n - \theta)$ is asymptotically normal with mean

$$b(\psi, G, H) = n^{1/2}\int \psi(x, u)H(du \mid x)G(dx) \tag{1.6}$$

and variance-covariance matrix,

$$V(\psi) = \int \psi^T(x, u)\psi(x, u)\Phi(du)G_0(dx). \tag{1.7}$$

Note that b depends on n through G, H but for "regular" G, H stabilizes as $n \to \infty$.

In the univariate case, $p = 1$, we argue in [B] that we can characterize estimates which asymptotically minimize maximum (asymptotic) mean square error over $\mathscr{F}$ by minimizing $V(\psi) + \sup\{b^2(\psi, G, H): F \in \mathscr{F}\}$ over Ψ. More generally, the maximum risk of T_n as above, is for any reasonable symmetric loss function determined by $V(\psi)$ and $\sup\{|b(\psi, G, H)|: F \in \mathscr{F}\}$.

In Section 2 we study the univariate case as follows.

(1) We evaluate

$$b(\psi) = \lim \sup_n \sup\{|b(\psi, G, H)|: F \in \mathscr{F}\} \tag{1.8}$$

for the $\mathscr{F}$ we have introduced. Subscripts on b indicate which $\mathscr{F}$ we are considering.

(2) We solve the variational problem of minimizing $V(\psi)$ subject to $b(\psi) \leq m$. This is just Hampel's variational problem or a variation thereof.

The family of extremal $\{\psi_m: m \geq 0\}$ correspond formally via (1.5) to (M) estimates which are candidates for solutions to asymptotic min max problems. Checking that the (M) estimate or 1-step approximation to it actually is asymptotically minmax requires a uniformity argument such as that of Theorem 5, page 25 of [B] for the putative solution. These arguments are straightforward, requiring standard appeals to Huber (1967) or Bickel (1975) or Maronna and Yohai (1978). We therefore focus exclusively on the variational problems. No new procedures are obtained in this section. However, Theorem 2.1 formally gives some optimality properties of the Hampel and MIA:A estimates.

In Section 3 we consider the general multiple regression model and introduce WLS procedures and equivariance under change of basis in the independent variable space.

We derive various procedures on the basis of the optimality criteria we have advanced:

1) the Hampel-Krasker (nonequivariant) estimates;
2) the natural nonequivariant extension of Huber's MIA:A estimates (Theorem 3.1);
3) nonequivariant procedures which are also not WLS but are optimal for estimating one parameter at a time under $\mathscr{F}_{ac0}$;
4) an equivariant estimate which minimizes the maximum M.S.E. of prediction under $\mathscr{F}_{ac0}$ (Theorem 3.2);
5) the natural equivariant extension of Huber's MIA:A estimates which minimizes the maximum M.S.E. of prediction under $\mathscr{F}_{c0}$.

Finally we show that the optimality of the Hampel-Krasker and of the equivariant estimate minimizing the maximum M.S.E. of prediction depends on the quadratic form used in the loss function. This casts some doubt on a conjecture of Krasker and Welsch (1982). The doubt is confirmed by a recent counterexample of D. Ruppert.

2. Regression through the origin ($p = 1$). As we indicated, if $b(\psi)$ is given by (1.8), we want, for each $\mathscr{F}$, to solve the variational problem:

$$\text{(V)} \qquad \int \psi^2(x, u)\Phi(du)G_0(dx) = \min!$$

subject to (1.2), (1.3) and

$$b(\psi) \leq m.$$

For each $\mathscr{F}$ we actually have a one-parameter family of variational problems as m varies and in principle each family could generate its own family of solutions. Fortunately there are only two families of solutions which we describe below.

It will be shown in Theorem 3.1 that for $\mathscr{F}$ which are of interest to us, only ψ which are Huber functions for each fixed x need be considered. That is, we can write ψ in the form:

$$\begin{aligned} \psi(x, u) &= (a(x)/c(x))h(u, c(x)), \quad c(x) > 0 \\ &= a(x)\operatorname{sgn} u, \qquad\qquad c(x) = 0 \end{aligned} \tag{2.1}$$

for given functions a; $c \geq 0$ satisfying (1.3) and $h(u, c) = \max(-c, \min(c, u))$.

For such ψ condition (1.2) is always satisfied and (1.3) becomes

$$\int a(x)xB(c(x))G_0(dx) = 1 \tag{2.2}$$

where

$$B(c) = (2\Phi(c) - 1)/c \quad \text{with} \quad B(0) = 2\phi(0). \tag{2.3}$$

The two basic solution families of ψ which we denote $\{\psi_k\}$, $\{\tilde{\psi}_k\}$ will be defined by corresponding $\{a_k, c_k\}$, $\{\tilde{a}_k, \tilde{c}_k\}$ as follows:

For $0 < k < \infty$ let

$$c_k(x) = k/|x|, \quad a_k(x) = \operatorname{sgn} x \Big/ \int (2\Phi(c_k(x)) - 1)x^2G_0(dx). \tag{2.4}$$

We add two limiting cases

$$\psi_\infty(x, u) = xu \Big/ \int x^2G_0(dx) \tag{2.5}$$

$$\psi_0(x, u) = \operatorname{sgn}(xu)/2\phi(0) \int |x| \, G_0(dx). \tag{2.6}$$

These are just the influence functions of the Hampel-Krasker-Welsch family of estimates. The extremal cases (2.5), (2.6) correspond to least squares, $T_n = \sum x_iy_i/\sum x_i^2$ and $T_n = \text{median } (y_i/x_i)$ respectively.

For $0 < t < 2\phi(0)$ let $0 < q(t) < \infty$ be the unique solution of

$$2(\phi(q) - q\Phi(-q)) = t. \tag{2.7}$$

Let $[2\underline{k}\phi(0)]^{-1}$ be the (G_0) ess sup of $|x|$. For $\underline{k} < k < \infty$ define

$$\tilde{c}_k(x) = q(1/k|x|)$$

$$\tilde{a}_k(x) = x \Big/ \int x^2(2\Phi(\tilde{c}_k(x)) - 1)I(|x| \geq [2k\phi(0)]^{-1})G_0(dx) \tag{2.8}$$

$$\text{if} \quad |x| \geq [2k\phi(0)]^{-1}$$

$$= 0 \text{ otherwise.}$$

The limiting cases are:

$$\tilde{\psi}_\infty(x, u) = \psi_\infty(x, u) \tag{2.9}$$

$$\tilde{\psi}_k(x, u) = \frac{k \operatorname{sgn} u}{\gamma}, \quad |x| = [2k\phi(0)]^{-1} \tag{2.10}$$

$$= 0 \qquad \text{otherwise}$$

if $\gamma = G_0\{x: |x| = [2k\phi(0)]^{-1}\} > 0$.

THEOREM 2.1. *Solutions to (V) are provided by*

(i) *Family* $\{\psi_k\}$: $\mathscr{F}_{ac0}$, $\mathscr{F}_{av0}$, $\mathscr{F}_{ak0}$, $\mathscr{F}_{c1}$, $\mathscr{F}_{v1}$, $\mathscr{F}_{k1}$

(ii) *Family* $\{\tilde{\psi}\}$ $\mathscr{F}_{c0}$, $\mathscr{F}_{v0}$, $\mathscr{F}_{k0}$

where we have substituted $d = v$, k *as appropriate in our notation. For given* m, t *the optimal* k *depends on* m/t *only and*

(iii) *The solutions for* $\mathscr{F}_{av0}$, $\mathscr{F}_{ak0}$, $\mathscr{F}_{v1}$, $\mathscr{F}_{k1}$ *coincide.*

(iv) *The solutions for* $\mathscr{F}_{v0}$, $\mathscr{F}_{k0}$ *coincide.*

(v) *The solutions for* $\mathscr{F}_{c0}$ *are solutions for* $\mathscr{F}_{v0}$ *with* m/t *replaced by* $m/2t$.

The key to Theorem 2.1 is evaluation of $b(\psi)$ for the different neighbourhoods. The proof of a typical subset of the following assertions is given in the appendix.

If b is defined by (1.6), (1.8) then

$$b_{c0}(\psi) = t \int \operatorname{ess\,sup}_u |\psi(x, u)| G_0(dx) \tag{2.11}$$

$$b_{v0}(\psi) = t \int [\operatorname{ess\,sup}_u \psi(x, u) - \operatorname{ess\,inf}_u \psi(x, u)]G_0(dx) \tag{2.12}$$

$$b_{k0}(\psi) = t \int \|\psi(x, \cdot)\| G_0(dx) \tag{2.13}$$

where "ess" refers to Lebesgue measure and $\|\cdot\|$ is the variational norm of $\psi(x, \cdot)$ viewed as a distribution function.

On the other hand,

$$b_{c1}(\psi) = t \operatorname{ess\,sup}_{x,u} |\psi(x, u)| \tag{2.14}$$

$$b_{v1}(\psi) = t[\operatorname{ess\,sup}_{x,u}\psi(x, u) - \operatorname{ess\,inf}_{x,u}\psi(x, u)] \tag{2.15}$$

$$b_k(\psi) = t \text{ ess sup}_x \| \psi(x, \cdot) \|. \tag{2.16}$$

The "average" models behave like "errors in variables".

$$b_{a\cdot 0}(\psi) = b_{\cdot 1}(\psi). \tag{2.17}$$

If ψ is antisymmetric in u

$$b_{vi}(\psi) = 2b_{ci}(\psi), \quad i = 0, 1. \tag{2.18}$$

If, in addition, ψ is monotone in u, then

$$b_{ki}(\psi) = b_{vi}(\psi), \quad i = 0, 1. \tag{2.19}$$

PROOF OF THEOREM. From (2.11)–(2.19) it is clear the solutions of (V) depend on m, t through m/t only and we can take $t = 1$. We claim it is enough to show (i) for $\mathscr{F}_{c1}$, (ii) for $\mathscr{F}_{c0}$. Since all members of both familes $\{\psi_s\}$ and $(\tilde{\psi}_s\}$ are antisymmetric and monotone in u, we can apply (2.18), (2.19) and the inclusion relations between the neighbourhoods to derive (iii)–(iv). From (iii)–(iv), (i) and (ii) follow for all neighbourhoods and (v) is immediate.

Problem (V) for $\mathscr{F}_{c1}$ is just Hampel's variational problem. Existence of a solution follows from standard weak compactness arguments. For these and the derivation of the family of solutions by a standard Lagrange multiplier argument, see, for example, [B].

Problem (V) for $\mathscr{F}_{c0}$ is a little less standard. Huber (1982) essentially derives the solution indirectly from his finite minimax robust testing theory.

We will give another proof which relies on a "conditional on x" Lagrange multiplier argument for the p-variate case. See the proof of Theorem 3.1 and note (2) following it. □

Discussion.

(1) *Unknown G_0.* In practice G_0 is unknown. Strictly speaking it is not required for the calculation of any particular estimate of the families $\{\psi_k\}$, $\{\tilde{\psi}_k\}$. However, in order to pick out a member on optimality grounds, say, minimizing maximum M.S.E., and to estimate maximum M.S.E., G_0 is required. Estimating G_0 by the empirical distribution of the x_i gives the same asymptotic results.

(2) *Unknown scale.* In practice the scale σ^2 of the u_i^* is unknown. As we indicate in [B] under mild conditions, the estimate T_n solving

$$\sum_{i=1}^n \psi(x_i, (y_i - x_i T_n)/s) = 0 \tag{2.20}$$

where s is a consistent estimate of σ (over $\mathscr{F}$) and ψ is antisymmetric in u for fixed x will have influence function $\sigma\psi(x, u/\sigma)$. It follows that the optimal ψ functions derived under the assumption σ known can be modified as in (2.20) to yield estimates optimal whatever be σ. There are serious questions of computation and existence of solutions when scale is estimated simultaneously. See Maronna (1976) and Krasker and Welsch (1982).

(3) The agreement between the errors in variables and average c or v models

is interesting though, in retrospect, not surprising. As Huber (1982) reveals for the average c model, Nature can be thought of as using most of her allocated ε of contamination to create very skew conditional given x distributions of u for the largest x and this can certainly also be done for errors in variables.

(4) The qualitative behaviour for $\mathscr{F}_{c0}$ (and $\mathscr{F}_{v0}$) is surprising as noted by Huber (1982). Small x's which are relatively uninformative are cut out by the $\tilde{\psi}$ estimates and on the other hand the $\tilde{\psi}$ are not bounded. (However if G_0 is estimated as it must be by the empirical d.f. of the x_i, $\sup_{i,u} |\tilde{\psi}_k(x_i, u)| < \infty$ for each n.) In this case since Nature is required to spread her contamination evenly, it pays to take chances and use c large at the large values of x which are informative if they are not contaminated and it does not pay to take any chances at the small and uninformative values of x.

(5) Interestingly enough, the same behaviour is exhibited by the Hellinger metric neighbourhoods $\mathscr{F}_{h0}$ where $h^2(P, Q) = \int (\sqrt{dP/du} - \sqrt{dQ/du})^2\, du$. Here it may be shown

$$b_{h0}(\psi) = 2t \int \left(\int \psi^2(x, u)\Phi(du) \right)^{1/2} G_0(dx)$$

and the resulting optimal ψ are of the form

$$\psi_k^*(x, u) = a(x)u$$

where

$$\begin{aligned} a(x) &= 0, && |x| \leq k \\ &= \mu(x - k \operatorname{sgn} x), && |x| > k, \end{aligned}$$

where μ is determined by (1.3).

These solutions do not agree with the unique solution $\psi_\infty(x, u)$ (essentially least squares), appropriate for $\mathscr{F}_{ah0}$, $\mathscr{F}_{h1}$.

3. The general case. For $p > 1$ we face the usual problem of choosing adequate scalar summaries (measures of loss) of the vector $b(\psi, F)$ and the matrix $V(\psi)$ on which to optimize.

Again ψ's which are Huber functions for each x play a special role,

$$\psi(x, u) = (a(x)/c(x))h(u, c(x)) \tag{3.1}$$

where a is now a vector, $c \geq 0$. For such ψ, (1.2) is satisfied, (1.3) becomes

$$\int x^T a(x) B(c(x)) G_0(dx) = I \tag{3.2}$$

and

$$V(\psi) = \int a^T a(x) A(c(x)) G_0(dx) \tag{3.3}$$

where

$$A(c) = \frac{2\Phi(c) - 1 - 2c\phi(c)}{c^2} + 2\Phi(-c), \quad A(0) = 1. \tag{3.4}$$

Also natural are ψ corresponding to weighted least squares estimates (WLS) definable in the multivariate case by

$$T_n = \textstyle\sum_{i=1}^n w_i y_i x_i (\sum_{i=1}^n w_i x_i^T x_i)^{-1}$$

with

$$w_i = w(x_i, y_i - x_i T_n^T)$$

scalars defined up to a proportionality constant. Note that ψ corresponds to a WLS estimate $\Leftrightarrow$ the direction of ψ is that of a linear transformation of x, i.e.,

$$\psi(x, u) = w(x, u)uxR \tag{3.5}$$

with

$$R^{-1} = \int x^T x w(x, u) u^2 \Phi(du) G_0(dx).$$

We classify solutions to the p-variate problem according as they do or do not possess equivariance under changes of basis in the X-space. An estimate T_n is equivariant under change of basis if and only if

$$T_n(XB, y) = T_n(X, y)[B^T]^{-1}.$$

(a) *Nonequivariant solutions.*

(i) *The Hampel-Krasker solution.* Perhaps the most natural choice of objective function is the total M.S.E. of the components, tr $V(\psi) + bb^T(\psi, F)$. If we let $|\cdot|$ denote the Euclidean norm, this leads to the following p-variate version of (V),

$$\int |\psi|^2(x, u)\Phi(du)G_0(dx) = \min! \tag{V}$$

for $\psi \in \Psi$ and $\sup_{\mathscr{F}} |b|(\psi, F) \leq m$. Holmes (1982) has shown that for $\mathscr{F}_{ac0}$, $\mathscr{F}_{c1}$,

$$\sup_{\mathscr{F}} |b|(\psi, F) = t \operatorname{ess\,sup}_{x,u} |\psi(x, u)|$$

so that (V) is just the problem of Krasker, Hampel (1978) whose solution is of the form, for $\lambda_0 < \lambda < \infty$,

$$\psi(x, u, \lambda) = xQh(u, \lambda/|xQ|)$$

where Q is symmetric positive definite and by (3.2)

$$Q^{-1} = \int x^T x \left(2\Phi\left(\frac{\lambda}{|xQ|}\right) - 1\right) G_0(dx).$$

Here

$$\lambda = \text{ess sup}_{x,u} |\psi(x, u, \lambda)|$$

and

$$0 < \lambda_0 = \inf\{\sup_{x,u} |\psi(x, u)| : \psi \in \Psi\}.$$

The solution to (V) has $\lambda = mt$. Krasker and Hampel (see also [B]) show that whenever there exists ψ with ess $\sup_{x,u} |\psi(x, u)| = \lambda > \lambda_0$, then $\psi(\cdot, \cdot, \lambda)$ exists and is unique.

Note that $\psi(\cdot, \cdot, \lambda)$ is of the form (3.1) and also WLS with

$$a(x) = \lambda(xQ/|xQ|), \quad c(x) = \lambda/|xQ|, \quad w(x, u) \propto h(u, c(x))/u.$$

NOTES.

(1) Calculations along the lines of Maronna (1976) show that $\lambda \to Q_\lambda$ is decreasing (in the order on positive definite symmetric matrices).

(2) It may be shown that $\lambda_0 \geq p/2\phi(0) \int |x| G_0(dx)$.

(ii) *A generalization of Huber's approach.* For $\mathscr{F}_{c0}$ it seems difficult to evaluate $\sup_{\mathscr{F}} |b|(\psi, F)$ exactly. However, it is easy to show that (see appendix)

$$\sup\{|b|(\psi, F): F \in \mathscr{F}_{c0}\} \leq t \int \sup_u |\psi(x, u)| G_0(dx).$$

As in the 1-dimensional case $\int \sup_u |\psi(x, u)| G_0(dx)$ can be interpreted as an average sensitivity. The solution of the resulting problem,

$$\int |\psi(x, u)|^2 \Phi(du) G_0(dx) = \min! \tag{V'}$$

subject to (1.2), (1.3) and

$$\int \sup_u |\psi(x, u)| G_0(dx) \leq \lambda$$

for $\lambda = m/t$, yields what should be a reasonable approximation to (V).

THEOREM 3.1. *For every $\lambda > \lambda_1$ there exists a unique pair $(s(\lambda), Q(\lambda))$ such that*

$$\tilde{\psi}(\cdot, \lambda) = \rho(\cdot, Q(\lambda), s(\lambda))$$

is an influence function and

$$\int \sup_u |\tilde{\psi}(x, u, \lambda)| G_0(dx) = \lambda \tag{3.6}$$

and $\tilde{\psi}(\cdot, \lambda)$ solves (V').

The solutions to (V′) are describable as follows: Define, for $s > 0$, Q symmetric positive definite, q as in (2.7),

$$\rho(x, Q, s) = xQh(u, q([s \mid xQ \mid]^{-1})), \quad \mid xQ \mid > [2s\phi(0)]^{-1}$$
$$= 0 \qquad \text{otherwise.}$$

Let

$$\lambda_1 = \inf\left\{\int \sup_u \mid \psi(x, u) \mid G_0(dx): \psi \in \Psi\right\}.$$

$\tilde{\psi}(\cdot, \lambda)$ can be written in the form (3.1) with corresponding functions defined for $s = s(\lambda)$, $Q = Q(\lambda)$ by

$$\tilde{c}(x, \lambda) = q(\mid sxQ \mid^{-1})$$
$$\tilde{a}(x, \lambda) = xQ\tilde{c}(x, \lambda) \quad \text{for} \quad \mid xQ \mid > [2s\phi(0)]^{-1}$$
$$= 0 \qquad \text{otherwise.}$$

Preliminary calculations along the lines of Maronna (1976) and Maronna-Yohai (1981) indicate that at least if G_0 does not place mass on hyperplanes, then Q is uniquely determined by s through (3.2), i.e.

$$Q^{-1} = \int_{S(s,Q)} x^T x(2\Phi(q(\mid sxQ \mid^{-1})) - 1)G_0(dx) \tag{3.7}$$

where $S(s, Q) = \{x: \mid sxQ \mid > 2\phi(0)\}$ and then s is determined by λ through (3.6)

$$\int_{S(s,Q)} \mid xQ \mid q(\mid sxQ \mid^{-1})G_0(dx) = \lambda. \tag{3.8}$$

Moreover if we write Q_s for the solution of (3.7), $s \to Q_s$ is nondecreasing and hence $\lambda \to s(\lambda)$ is also. So we can reparametrize $\tilde{\psi}(\cdot, \lambda)$ by s for $s > \inf\{s(\lambda): \lambda > \lambda_1\}$. If, for $p = 1$, we take $k = sQ_s$, then we obtain the family $\tilde{\psi}_k$ of Theorem 2.1. Since k is an increasing function of λ we obtain the conclusions of Theorem 2.1.

PROOF. In the appendix we show by standard optimization theory arguments that a solution to (V′) exists and is also the solution to a Lagrangian problem

$$\int \left\{\mid \psi \mid^2(x, u) - 2\int u\psi(x, u)Qx^T + \frac{2}{s}\mid \psi \mid (x, u)\right\}\Phi(du)G_0(dx) = \min!$$

for $Q_{p\times p}$, $s > 0$.

If ψ_0 is the solution we can minimize

$$\int \mid \psi \mid^2(x, u)\Phi(du) - 2\int u\psi(x, u)Qx^T\Phi(du)$$

subject to $\sup_u \mid \psi(x, u) \mid \leq \sup_u \psi_0(x, u)$ and conclude that ψ_0 is of the form (3.1)

with the corresponding vector $a_0(x)$ and $c_0(x)$ minimizing

$$\int \{|a|^2(x)A(c(x)) - 2xQa^T(x)B(c(x)) + s^{-1}|a|(x)\}G_0(dx).$$

Minimizing pointwise we obtain as necessary conditions for a_0, c_0

$$a_0A(c_0) = xQB(c_0) + s^{-1}(a_0/|a_0|) = 0, \quad a_0 \neq 0 \tag{3.9}$$

$$\begin{aligned} |a_0|^2 &\leq xQa_0^Tc_0 \\ &= xQa_0^Tc_0 \quad \text{if} \quad c_0 > 0. \end{aligned} \tag{3.10}$$

From (3.10), $a_0 \neq 0 \Rightarrow c_0 > 0$. Then by (3.9)

$$a_0 = |a_0|(xQ/|xQ|) = c_0xQ$$

by (3.10). Again by (3.9)

$$c_0A(c_0) - B(c_0) + (1/s|xQ|) = 0$$

which implies $|xQ| \geq [2s\phi(0)]^{-1}$, $c_0 = q([s|xQ|]^{-1})$. Conversely, if $|x| > [2s\phi(0)]^{-1}$, $\tilde{a}(x, \lambda)$, $\tilde{c}(x, \lambda)$ yield

$$|a|^2A - 2xQa^TB(c) + s^{-1}|a| < 0$$

and hence $0 \neq a_0 = \tilde{a}$ by our previous reasoning. Since $\tilde{\psi}$ must satisfy (1.2), Q must satisfy (3.9) and be positive definite symmetric. The theorem is proved. □

(iii) *One at a time optimality.* Another nonequivariant solution of interest is obtained by minimizing the maximum M.S.E. of each component of θ separately. That is, we seek $\psi^* = (\psi_1^*, \cdots, \psi_p^*) \in \Psi$ which *simultaneously* minimizes

$$\int [\psi_j]^2(x, u)\Phi(du)G_0(dx)$$

for $\psi = (\psi_1, \cdots, \psi_p) \in \Psi$ and

$$\sup\{|b_j(\psi, F)|: F \in \mathscr{F}\} \leq m_j$$

where $b(\psi, F) = (b_1(\psi, F), \cdots, b_p(\psi))$. For neighbourhoods of the "average" or errors in variables types, the solutions ψ^*, indexed by the vector $m = (m_1, \cdots, m_p)$, are *not* of the WLS form. They are given by

$$\psi_j^*(x, u; m) = uxa_j^Th(u, m_j/|xa_j^T|), \quad j = 1, \ldots, p \tag{3.11}$$

where (1.2) and (1.3) hold. Existence of $\psi^*(\cdot, m_0)$ and their form as solutions of a Lagrange problem are guaranteed. for m_0 an interior point of $\{m: t \sup_{x,u}|\psi_j(x, u)| \leq m_j, j = 1, \cdots, p\}$. The limiting case corresponding to the median is, for $x = (x_1, \cdots, x_p)$,

$$\psi_j^*(x, u) = c_j\text{sgn}[(x_j - \textstyle\sum_{k\neq j} b_{kj}x_k)u] \tag{3.12}$$

where

$$c_j = \left[\left(\frac{2}{\pi}\right)^{1/2} \int |x_j - \textstyle\sum_{k\neq j} b_{kj}x_k| G_0(dx)\right]^{-1}$$

where $B = \| b_{ij} \|$ is determined by

$$\int \operatorname{sgn}(x_j - \textstyle\sum_{k \neq j} b_{kj} x_k) x_i G_0(dx) = 0, \quad i \neq j. \tag{3.13}$$

If $(x_{i1}, \cdots, x_{ip}, y_i)$, $i = 1, \cdots, p$ are the observations, $\hat{\theta}_1, \cdots, \hat{\theta}_p$ are the estimates, and $\hat{\varepsilon}_i = y_i - \sum_{j=1}^{p} x_{ij}\hat{\theta}_j$ are the residuals, then $\hat{\theta}_1, \cdots, \hat{\theta}_p$ are characterized by the property that

$$\operatorname{median}_i \hat{\varepsilon}_i / (x_{ij} - \textstyle\sum_{k \neq j} b_{kj} x_{ik}) \cong 0$$

for $j = 1, \cdots, p$. In view of (3.13) the b_{kj} can be interpreted as the coefficients of a least absolute residuals fit of $\sum_{k \neq j} b_k x_k$ to x_j, i.e.,

$$\int | x_j - \textstyle\sum_{k \neq j} b_{kj} x_k | \, G_0(dx) = \min \int | x_j - \textstyle\sum_{k \neq j} b_k x_k | \, G_0(dx). \tag{3.14}$$

This characterization guarantees the existence of this influence function at least if G_0 is absolutely continuous. Of course, there may be difficulties for a sample where we replace G_0 by the empirical d.f. of the X_i.

At first glance this solution appears to render the Hampel-Krasker solution inadmissible. This is, however, not the case. ψ^* here minimizes (for suitable m_j),

$$R(\psi) = \textstyle\sum_{i=1}^{p} \int \psi_i^2(x, u) \Phi(du) G_0(dx) + \sum_{i=1}^{p} \max_{\mathcal{F}} b_i^2(\psi, F)$$

while the Hampel-Krasker solution minimizes

$$S(\psi) = \textstyle\sum_{i=1}^{p} \int \psi_i^2(x, u) \Phi(du) G_0(dx) + \max_{\mathcal{F}} \sum_{i=1}^{p} b_i^2(\psi, F).$$

Of course, $S \leq R$ but the optimal solutions are not related.

(b) *Equivariant solutions.* When translated to influence functions this equivariance becomes

$$\psi(x, u, G_0) = \psi(xB, u, G_0 B^{-1}) B^T \tag{3.15}$$

where $\psi(x, u, G)$ is the influence curve if $X_1 \sim G$.

(i) *Equivariant best MSE of prediction.* Suppose that X_1 is error free so that $G = G_0$ and that $\int | x |^2 G_0(dx) < \infty$. The most natural way of obtaining invariant ψ with local optimality properties is to use as objective function the expected mean square error of prediction

$$\int \{ x V(\psi) x^T G(dx) + x b^T(\psi) b(\psi) x^T \} G_0(dx).$$

We can rewrite this as

$$\int \psi \Sigma \psi^T(x, u) \Phi(du) G_0(dx) + b(\psi, F) \Sigma b^T(\psi, F)$$

where

$$\Sigma = \int x^T x G_0(dx). \tag{3.16}$$

As in the noninvariant case we can deal easily with $\mathscr{F}_{a0c}$ since

$$\sup\{b(\psi, F)\Sigma b^T(\psi, F): F \in \mathscr{F}_{ac0}\} = \text{ess sup}_{x,u}\psi(x, u)\Sigma\psi^T(x, u). \tag{3.17}$$

Minimizing the maximum of our objective function over $\mathscr{F}_{a0c}$ is easy once we have solved

$$\int \psi\Sigma\psi^T(x, u)\Phi(du)G_0(dx) = \min! \tag{V$_I$}$$

for $\psi \in \Psi$ such that

$$\text{ess sup}_{x,u}\psi\Sigma\psi^T(x, u) \leq \lambda.$$

Let

$$\lambda_{I0} = \inf \text{ess}\{\sup_{x,u}\psi\Sigma\psi^T(x, u): \psi \in \Psi\}$$

$$d^2(x, \Sigma) = x\Sigma x^T.$$

For $\lambda > \lambda_{I0}$ let

$$\psi_I(x, u, \lambda) = xQh(u, \lambda/d(xQ, \Sigma)) \tag{3.18}$$

where Q is positive definite symmetric,

$$\int x^T x\left(2\Phi\left(\frac{\lambda}{d(xQ, \Sigma)}\right) - 1\right)G_0(dx) = Q^{-1}. \tag{3.19}$$

THEOREM 3.2. *If* $\lambda > \lambda_{I0}$, $\psi_I(\cdot, \cdot, \lambda)$ *uniquely solves* (V$_I$).

PROOF. Again by standard arguments we can establish existence of a minimizing ψ_0 which solves an equivalent Lagrangian problem

$$\int \{\psi\Sigma\psi^T(x, u) - 2\int uxQ\Sigma\psi^T(x, u)\}\Phi(du)G_0(dx) = \min!$$

subject to $|\Psi\Sigma\psi^T| \leq \lambda$. A direct minimization of $\psi\Sigma\psi^T - 2uxQ\Sigma\psi^T$ under the side condition yields (3.18) and (3.2) implies (3.19). □

Note that the uniqueness of ψ_I and (3.19) imply the equivariance property (3.15).

(ii) *An equivariant Huber solution.* As in the nonequivariant case we can bound the maximum expected squared bias of the predictor

$$\sup\left\{\int xb^T b(\psi, F)x^T G_0(dx): F \in \mathscr{F}_{c0}\right\}$$

above by

$$t \int \{\sup_u \psi(x, u)\Sigma\psi^T(x, u)\}G_0(dx).$$

The resulting variational problem

$$\int \psi\Sigma\psi^T(x, u)\Phi(du)G_0(dx) = \min!$$

subject to

$$\int \sup_u \psi(x, u)\Sigma\psi^T(x, u)G_0(dx) \leq \lambda \tag{3.20}$$

has solutions of the form

$$\tilde{\psi}(x, u, s) = \frac{\tilde{a}_I(x, s)}{\tilde{c}_I(x, s)} h(u, \tilde{c}_I(x, s)) \tag{3.21}$$

where

$$\tilde{c}_I(x, \lambda) = q(1/sd(xQ, \Sigma)), \quad \tilde{a}_I(x, s) = xQ\tilde{c}_I(x, s)$$

if

$$d(xQ, \Sigma) \geq [2s\phi(0)]^{-1}$$
$$= 0 \quad \text{otherwise.}$$

and Q, s are determined by the requirement that $\tilde{\psi}_I$ is an influence function satisfying equality in (3.20).

Reparametrizations are possible for the procedures of this section as for the Hampel-Krasker and Huber solutions.

(iii) *The Krasker-Welsch* (1982) *solution.* Based on sensitivity considerations, Krasker and Welsch proposed estimates given by

$$\psi_{KW}(x, u, \lambda) = xQh(x, \lambda/d(xQ, V^{-1})), \quad \lambda > \sqrt{p} \tag{3.22}$$

where

$$\int x^T x\left(2\Phi\left(\frac{\lambda}{d(xQ, V^{-1})}\right) - 1\right)G_0(dx) = Q^{-1}$$

and

$$V_\lambda = \int \psi^T\psi(x, u, \lambda)\Phi(du)G_0(dx). \tag{3.23}$$

Equivalently if $A^{-1} = QV^{-1}Q$, (3.23) becomes

$$A = \int x^T x[2\Phi(\lambda/d(x, A^{-1})) - 1 - 2\lambda d^{-1}(x, A^{-1})\phi(\lambda d^{-1}(x, A^{-1}))]G_0(dx)$$

and Q may be obtained directly from (3.22). Existence of the K-W solution for

$\lambda > \sqrt{p}$ is guaranteed by results of Maronna (1976). The K-W solution is also equivariant. It evidently has the property (by arguing as for Theorem 3.2) of uniquely minimizing $\int \psi V^{-1}(\psi_{KW})\psi^T$ subject to $\sup \psi V^{-1}(\psi_{KW})\psi^T \leq \lambda^2$. Krasker and Welsch conjecture a strong optimality property (see below).

(iv) *More general optimality properties.* Whatever be p, least squares estimates do not minimize only trace $V(\psi)$ but the matrix itself or equivalently $\int \psi M \psi^T$ for all M positive definite, symmetric. It is fairly easy to see (see also Stahel, 1981) that once we bound the vector influence curve as we have in this section, no such conclusion is possible. Thus $\psi M \psi^T(x, u) - 2u\psi(x, u)QMx^T$ is minimized subject to $|\psi| \leq \lambda$ by $\psi = uxQ$ if $|u| \leq \lambda/|xQ|$, but, unless $M = I$, by a boundary value other than $\lambda(xQ/|xQ|)$ if $|u| > \lambda/|xQ|$.

Krasker and Welsch seek to remedy this failing by restricting ψ to the WLS form, i.e., forcing the direction of ψ to coincide with a linear transformation of x. They conjecture that their solution minimizes $V(\psi)$ among all WLS estimates with $\sup \psi V^{-1}(\psi)\psi^T \leq \eta$. Our methods do not readily give a counterexample to their conjecture but we show below that neither the Hampel-Krasker estimate nor the equivariant estimate of section (i) possess the analogous optimality property, thus casting some doubt on the conjecture. (David Ruppert has recently discovered a counterexample to the conjecture.) Suppose G_0 is spherically symmetric, its support is bounded, has a nonempty interior, and does not contain 0. Then, by symmetry, the Hampel-Krasker, section (i) and Krasker-Welsch solutions are of the same form. For suitable λ,

$$\psi_0(x, u) = rxh(u, \lambda/r|x|)$$

where

$$r = \left[\int |x|^2\left(2\Phi\left(\frac{\lambda}{r|x|}\right) - 1\right)G_0(dx)\right]^{m-1}.$$

If ψ_0 were a universally optimal solution for the Hampel-Krasker or MSE of prediction problems among WLS estimates, it would solve, *for all S*,

$$\text{(Vs)} \qquad \int \psi S \psi^T(x, u)\Phi(du)G_0(dx) = \min!$$

subject to $|\psi| \leq \lambda$, $\psi \in \Psi$ and ψ WLS as in (3.5).

By conditioning as in the proof of Theorem 3.1 and restricting to

$$w(x, u) = \frac{\lambda}{c(x)} \frac{h(u, c(x))}{u|xR|},$$

we see that $R_0 = rI$, $c_0(x) = \lambda/r|x|$ minimizes

$$\int \lambda^2\left(\frac{d^2(xR, S)}{|xR|^2}\right)A(c(x))G_0(dx)$$

among all $c > 0$, R symmetric positive definite such that

$$\int \lambda\left(\frac{x^T xR}{|xR|}\right)B(c(x))G_0(dx) = I.$$

If we let c range over the Banach space of continuous functions vanishing at ∞ with supremum norm, it can be shown that if $p > 3$ the map

$$(c, R) \to \int \frac{x^T x R}{|xR|} B(c(x)) G_0(dx)$$

has a nonsingular differential at $c = c_0$, $R = R_0$ where r is given in the definition of ψ. Therefore by Luenberger (1969, page 243) there exists a Lagrange multiplier matrix $W_S S$ such that R_0, c_0 minimize

$$\int \frac{d^2(xR, S)}{|xR|^2} A(c(x)) G_0(dx) - 2 \int \frac{\operatorname{tr}(W_S S R x^T x)}{|xR|} B(c(x)) G_0(dx) \tag{3.24}$$

among all R symmetric positive definite, $c \geq 0$, c's vanishing at ∞. But minimization over c leads as in Theorem 3.1 to

$$c = \operatorname{tr}(RSRx^T x)/\operatorname{tr}(W_S S R x^T x)\,|xR|. \tag{3.25}$$

If we set $c = c_0$, $R = R_0$, we deduce that $W_S = R_0/\lambda$. If we now substitute (3.25) back into (3.24), find the differential of the resulting map from the set of symmetric matrices to the real line and set it equal to 0 at $R = R_0$, we obtain the equation

$$\int \alpha(c_0(x))((SR_0 + R_0 S) - 2\beta(x, S) R_0) x^T x G_0(dx) = 0 \tag{3.26}$$

where

$$\alpha(c) = 2(c\Phi(-c) - \phi(c)), \qquad \beta(x, S) = d^2(xR_0, S)/|xR_0|^2.$$

Simplifying, we get

$$S \int \alpha\left(\frac{\lambda}{r|x|}\right) x^T x G_0(dx) = \int \alpha\left(\frac{\lambda}{|x|}\right) \frac{xSx^T}{|x|^2} x^T x G_0(dx) \tag{3.27}$$

for all positive definite symmetric S. Passing to the limit, the relationship must hold for nonnegative definite S as well. Put

$$S = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & & \cdots & 0 \\ 0 & & \cdots & 0 \end{pmatrix}$$

to obtain a contradiction since by symmetry of G_0, $\int \alpha(\lambda/r|x|) x^T x G_0(dx)$ is a multiple of I and G_0 has a nonempty interior.

NOTES.

(1) For $p > 1$ as in the univariate case we would typically need to estimate G_0 and σ in order to implement adequate scale equivariant estimates. No new theoretical issues arise from optimality considerations. However the computational solution and existence of problems which arise with simultaneous estimation of scale become more serious.

(2) Our discussion in this section is essentially limited to the contamination neighbourhood since the maximum bias (as measured by different norms) in the p-variate case can only be easily calculated for these. However, these solutions are also adequate for variational and Kolmogorov neighbourhoods provided t is taken as double its value for contamination. Thus, for $\mathscr{F}_{\mathrm{a0v}}$, $\mathscr{F}_{\mathrm{1v}}$

$$\sup |b(\psi, F)| \leq 2t \sup_{x,u} |\psi(x, u)| \tag{3.28}$$

while for $\mathscr{F}_{\mathrm{0v}}$

$$\sup |b(\psi, F)| \leq 2t \int \sup_u |\psi(x, u)| G_0(dx) \tag{3.29}$$

and for $\mathscr{F}_{\mathrm{a0k}}$, $\mathscr{F}_{\mathrm{1k}}$

$$\sup_{\mathscr{F}_{\mathrm{1k}}} |b(\psi, F)| \leq t \sup_x |\, \|\psi(x, \cdot)\| \,| \tag{3.30}$$

where $\|\psi(x, \cdot)\| = (\|\psi_1(x, \cdot)\|, \cdots, \|\psi_p(x, \cdot)\|)$ and $\|\psi_i(x, \cdot)\|$ is the variational norm of $\psi_i(x, \cdot)$.

(3) The invariant estimates based on minimizing MSE of prediction are appealing and seem reasonable for the error free x models. They are seriously compromised for errors in variables, however, since the matrix $\int x^T x G_0(dx)$ is not robustly estimated by replacing G_0 by the empirical distribution. A fairly artificial way out is to down weight extreme values of x. That is, let u_2 satisfy conditions of Maronna (1976), and $\Sigma(G_0)$ be the robust covariance determined by that u_2.

$$\int u_2(d(x, \Sigma^{-1}))x^T x G_0(dx) = \Sigma. \tag{3.31}$$

Then we can easily see that the estimate which minimizes the downweighted MSE of prediction

$$\sup_{\mathscr{F}} \left\{ \int u_2(d(x, \Sigma^{-1}))\{xV(\psi)x^T + xb^T(\psi)b(\psi)x^T\} G_0(dx) \right\}$$

is given by (3.19) with Σ given by (3.31) for both $\mathscr{F}_{\mathrm{ac0}}$ and $\mathscr{F}_{\mathrm{c1}}$. The estimate is clearly equivariant. This is essentially equivalent to a proposal of Maronna, Bustos, and Yohai (1979).

APPENDIX

PROOF OF (2.11)-(2.19). For the errors in variables models these claims are proved in [B]. For the other neighbourhoods the arguments are similar. As an example here is the proof of (2.11).

Since $G = G_0$, by (1.2),

$$b(\psi, G, H) = t \int\int \psi(x, u) M(du \mid x) G_0(dx). \tag{A.1}$$

Since M is arbitrary (2.11) follows. As a second example we prove (2.17) for $\mathscr{F}_v$.

Write

$$
\begin{aligned}
b(\psi, G, H) &= \int\int \psi(x, u)[H(du \mid x) - \Phi(du)]G_0(dx) \\
&= \int\int \psi(x, u)[M^+(du \mid x) - M^-(du \mid x)]\alpha(x)G_0(dx)
\end{aligned} \tag{A.2}
$$

where $\alpha(x)$ is the common total mass of the positive and negative parts of the measure $H(\cdot \mid x) - \Phi(\cdot)$ and M^+, M^- are the probability measures obtained by normalizing these positive and negative parts. $F \in \mathscr{F}_{\text{av1}}$ means $\int \alpha(x)G_0(dx) \leq tn^{-1/2}$. Since M^+, M^- are arbitrary, (2.17) follows. □

PROOF OF (3.7). By definition

$$
\begin{aligned}
|b|(\psi, F) &= t\left\{\sum_{j=1}^p \left(\int\int \psi_j(x, u)M(du \mid x)G_0(dx)\right)^2\right\}^{1/2} \\
&\leq t\int \left\{\sum_{j=1}^p \left(\int \psi_j(x, u)M(du \mid x)\right)^2\right\}^{1/2} G_0(dx)
\end{aligned} \tag{A.3}
$$

by Jensen's inequality applied to the random vector

$$
\left(\int \psi_1(X_1, u)M(du \mid X_1), \cdots, \int \psi_p(X_1, u)M(du \mid X_1)\right).
$$

Existence of solutions in Theorem 3.1.

Sketch of argument. Consider ψ as elements of $L_2(F_0; R^p)$, square integrable p-variate functions. Define the following maps from L_2 to R or R^{p^2}

$$
\begin{aligned}
a_0&: \psi \to \int |\psi|^2(x, u)\Phi(du)G_0(dx) \\
a_1&: \psi \to \int \sup_u |\psi(x, u)| G_0(dx) \\
a_2&: \psi \to \int ux^T\psi(x, u)\Phi(du)G_0(dx) \\
a_3&: \psi \to \sup_{x,u} |\psi(x, u)|.
\end{aligned}
$$

Then a_0, a_1 are convex, a_2 is linear. Let

$$
\lambda_{1M} = \inf\{\lambda : \psi \in \Psi, a_1(\psi) \leq \lambda, a_3(\psi) \leq M\}.
$$

It is easy to see that $\lambda_{1M} \downarrow \lambda_1$ if $M \to \infty$. Suppose $\lambda > \lambda_{1M}$. Then by problem 7, page 236 of Luenberger (1969) there exist Q_M, S_M such that

$$
\begin{aligned}
&\inf\{a_0(\psi): a_1(\psi) \leq \lambda, a_2(\psi) = I, a_3(\psi) \leq M\} \\
&\quad = \inf\{a_0(\psi) - 2 \operatorname{tr} Q[a_2(\psi) - I] + (2/s)[a_0(\psi) - \lambda]\}.
\end{aligned} \tag{A.4}
$$

Moreover since $\{\psi: a_3(\psi) \leq M\}$ is weakly compact and a_0 is lower semicontinuous, the infima in (A.4) are assumed by, say, $\psi_M^* \in \Psi$. By arguing as in the proof of the theorem

$$\psi_M^*(x, u) = \rho(x, u, s_M, Q_M) \quad \text{if} \quad |\rho(x, u, s_M, Q_M)| \leq M.$$

It readily follows by considering s_M and $Q_M/\mathrm{tr}(Q_M)$ that we can extract a subsequence $\{M_r\}$ such that $\psi_{M_r}^*$ converges pointwise to a limit ψ^* as $M_r \to \infty$. Since by the optimality of $\psi_{M_r}^*$, the sequence $a_0(\psi_{M_r}^*)$ is uniformly bounded, we can conclude that $a_2(\psi_{M_r}^*) \to a_2(\psi^*)$, i.e. $\psi^* \in \Psi$ and $a_1(\psi_{M_r}^*) \to a_1(\psi^*)$. By lower semicontinuity of a_0, ψ^* is the solution to (V'). Applying (A.5) with $M = \infty$ we obtain $(s(\lambda), Q(\lambda))$ such that $\rho(x, u, Q(\lambda), s(\lambda)) = \psi^*$. Unicity of (Q, s) follows from the strict convexity of a_0. □

REFERENCES

BICKEL, P. J. (1975). One step Huber estimates in the linear model. *J. Amer. Statist. Assoc.* **70** 428–434.

BICKEL, P. J. (1981). Quelques aspects de la statistique robuste. In École d'Été de Probabilités de St. Flour. *Springer Lecture Notes in Math.* **876** 2–68.

HAMPEL, F. R. (1978). Optimally bounding the gross-error-sensitivity and the influence of position in factor space. *1978 Proceedings of the A.S.A. Statistical Computing Section.* A.S.A., Washington, D.C. 59–64.

HOLMES, R. (1981). Thesis, University of California, Berkeley.

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **1** 221–233. University of California Press.

HUBER, P. J. (1973). Robust regression: asymptotics, conjectures, and Monte Carlo. *Ann. Statist.* **1** 799–821.

HUBER, P. J. (1983). Minimax aspects of bounded influence regression. *J. Amer. Statist. Assoc.* **78** 66–80.

KRASKER, W. (1980). Estimation in linear regression models with disparate data points. *Econometrica* **48** 1333–1346.

KRASKER, W. and WELSCH, R. (1982). Efficient bounded influence regression estimation. *J. Amer. Statist. Assoc.* **77** 595–604.

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

MARONNA, R. (1976). Robust M-estimators of multivariate location and scatter. *Ann. Statist.* **4** 51–67.

MARONNA, R., BUSTOS, O., and YOHAI, V. (1979). Bias and efficiency robustness of general (M) estimates for regression with random carriers. In *Smoothing Techniques for Curve Estimation* 91–116. T. Gasser and M. Rosenblatt, Eds. Springer-Verlag, Berlin.

MARONNA, R. A. and YOHAI, V. (1981). Asymptotic behaviour of general (M) estimates for regression and scale with random carriers. *Z. Wahrsch. verw. Gebeite* **58** 7–20.

RIEDER, H. (1978). A robust asymptotic testing model. *Ann. Statist.* **6** 1080–1099.

SACKS, J. and YLVISAKER, D. (1978). Linear estimation for approximately linear models. *Ann. Statist.* **6** 1122–1137.

STAHEL, W. (1981). Thesis. E.T.H. Zurich.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 93720