Chapter 2
Robust Statistics

Peter Bühlmann

2.1 Introduction to Three Papers on Robustness

2.1.1 General Introduction

This is a short introduction to three papers on robustness, published by Peter Bickel as single author in the period 1975–1984: "One-step Huber estimates in the linear model" (Bickel 1975), "Parametric robustness: small biases can be worthwhile" (Bickel 1984a), and "Robust regression based on infinitesimal neighbourhoods" (Bickel 1984b). It was the time when fundamental developments and understanding in robustness took place, and Peter Bickel has made deep contributions in this area. I am trying to place the results of the three papers in a new context of contemporary statistics.

2.1.2 One-Step Huber Estimates in the Linear Model

The paper by Bickel (1975) about the following procedure. Given a $\sqrt{n}$-consistent initial estimator $\tilde{\theta}$ for an unknown parameter θ, performing one Gauss-Newton iteration with respect to the objective function to be optimized leads to an asymptotically efficient estimator. Interestingly, this results holds even when the MLE is not efficient, and it is equivalent to the MLE if the latter is efficient. Such a result was known for the case where the loss function corresponds to the maximum likelihood estimator (Le Cam 1956). Bickel (1975) extends this result to much more general loss functions and models.

P. Bühlmann (✉)
ETH Zürich, Rämistrasse 101, HG G17 8092, Zürich, Switzerland
e-mail: buhlmann@stat.math.ethz.ch

J. Fan et al. (eds.), *Selected Works of Peter J. Bickel*, Selected Works in Probability and Statistics 13, DOI 10.1007/978-1-4614-5544-8_2,

The idea of a computational short-cut without sacrificing statistical was relevant more than 30 years ago (summary point 5 in Sect. 3 of Bickel 1975). Yet, the idea is still very important in large scale and high-dimensional applications nowadays. Two issues emerge.

In some large-scale problems, one is willing to pay a price in terms of statistical accuracy while gaining substantially with respect to computing power. Peter Bickel has recently co-authored a paper on this subject (Meinshausen et al. 2009): having some sort of guarantee on statistical accuracy is then highly desirable. Results as in Bickel (1975), probably of weaker form which do not touch on the concept of efficiency, are underdeveloped for large-scale problems.

The other issue concerns the fact that iterations in algorithms correspond to some form of (algorithmic) regularization which is often very effective for large datasets. A prominent example of this is with boosting: instead of a Gauss-Newton step, boosting proceeds with Gauss-Southwell iterations which are coordinatewise updates based on an n-dimensional approximate gradient vector (where n denotes sample size). It is known, at least for some cases, that boosting with such Gauss-Southwell iterations achieves minimax convergence rate optimality (Bissantz et al. 2007; Bühlmann and Yu 2003) while being computationally attractive. Furthermore, in view of robustness, boosting can be easily modified such that each Gauss-Southwell up-date is performed in a robust way and hence, the overall procedure has desirable robustness properties (Lutz et al. 2008). As discussed in Sect. 3 of Bickel (1975), the starting value (i.e., the initial estimator) matters also in robustified boosting.

2.1.3 Parametric Robustness: Small Biases Can Be Worthwhile

The following problem is studied in Bickel (1984a): construct an estimator that performs well for a particular parametric model $\mathscr{M}_0$ while its risk is upper-bounded for another larger parametric model $\mathscr{M}_1 \supset \mathscr{M}_0$. As an interpretation, one believes that $\mathscr{M}_0$ is adequate but one wants to guard against deviations coming from $\mathscr{M}_1$. It is shown in the paper that the corresponding optimality problem has not an explicit solution: however, approximate answers are presented and interesting connections are developed to the Efron-Morris (Efron and Morris 1971) family of translation estimates, i.e., adding a soft-thresholded additional correction term to the optimal estimator under $\mathscr{M}_0$. (The reference Efron and Morris (1971) is appearing in the text but is missing in the list of references in Bickel's paper).

The notion of parametric robustness could be interesting in high-dimensional problems. Guarding against specific deviations (which may be easier to specify in some applications than in others) can be more powerful than trying to protect nonparametrically against point-mass distributions in any direction. In this sense, this paper is a key reference for developing effective high-dimensional robust inference.

2.1.4 Robust Regression Based on Infinitesimal Neighbourhoods

Robust regression is analyzed in Bickel (1984b) using a nice mathematical framework where the perturbation is within a $1/\sqrt{n}$-neighbourhood of the uncontaminated ideal model. The presented results in Bickel (1984b) give a clear (mathematical) interpretation of various procedures and suggest new robust methods for regression.

A major issue in robust regression is to guard against contaminations in X-space. Bickel (1984b) gives nice insights for the classical case where the dimension of X is relatively small: a new challenge is to deal with robustness in high-dimensional regression problems where the dimension of X can be much larger than sample size. One attempt has been to robustify high-dimensional estimators such as the Lasso (Khan et al. 2007) or L_2Boosting (Lutz et al. 2008), in particular with respect to contaminations in X-space. An interesting and different path has been initiated by Friedman (2001) with tree-based procedures which are robust in X-space (in connection with a robust loss function for the error). There is clearly a need of a unifying theory, in the spirit of Bickel (1984b), for robust regression when the dimension of X is large.

References

Begun JM, Hall WJ, Huang W-M, Wellner JA (1983) Information and asymptotic efficiency in parametric–nonparametric models. Ann Stat 11(2):432–452

Beran R (1974) Asymptotically efficient adaptive rank estimates in location models. Ann Stat 2:63–74

Bickel P (1975) One-step Huber estimates in the linear model. J Am Stat Assoc 70:428–434

Bickel PJ (1982) On adaptive estimation. Ann Stat 10(3):647–671

Bickel P (1984a) Parametric robustness: small biases can be worthwhile. Ann Stat 12:864–879

Bickel P (1984b) Robust regression based on infinitesimal neighbourhoods. Ann Stat 12:1349–1368

Bickel PJ, Klaassen CAJ (1986) Empirical Bayes estimation in functional and structural models, and uniformly adaptive estimation of location. Adv Appl Math 7(1):55–69

Bickel PJ, Ritov Y (1987) Efficient estimation in the errors in variables model. Ann Stat 15(2):513–540

Bickel PJ, Ritov Y (1988) Estimating integrated squared density derivatives: sharp best order of convergence estimates. Sankhyā Ser A 50(3):381–393

Bickel PJ, Klaassen CAJ, Ritov Y, Wellner JA (1993) Efficient and adaptive estimation for semiparametric models. Johns Hopkins series in the mathematical sciences. Johns Hopkins University Press, Baltimore

Bickel PJ, Klaassen CAJ, Ritov Y, Wellner JA (1998) Efficient and adaptive estimation for semiparametric models. Springer, New York. Reprint of the 1993 original

Birgé L, Massart P (1993) Rates of convergence for minimum contrast estimators. Probab Theory Relat Fields 97(1–2):113–150

Birgé L, Massart P (1995) Estimation of integral functionals of a density. Ann Stat 23(1):11–29

Bissantz N, Hohage T, Munk A, Ruymgaart F (2007) Convergence rates of general regularization methods for statistical inverse problems and applications. SIAM J Numer Anal 45:2610–2636
Bühlmann P, Yu B (2003) Boosting with the L_2 loss: regression and classification. J Am Stat Assoc 98:324–339
Efron B (1977) The efficiency of Cox's likelihood function for censored data. J Am Stat Assoc 72(359):557–565
Efron B, Morris C (1971) Limiting the risk of Bayes and empirical Bayes estimators – part I: Bayes case. J Am Stat Assoc 66:807–815
Friedman J (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29:1189–1232
Hájek J (1962) Asymptotically most powerful rank-order tests. Ann Math Stat 33:1124–1147
Khan J, Van Aelst S, Zamar R (2007) Robust linear model selection based on least angle regression. J Am Stat Assoc 102:1289–1299
Kiefer J, Wolfowitz J (1956) Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. Ann Math Stat 27:887–906
Klaassen CAJ (1987) Consistent estimation of the influence function of locally asymptotically linear estimators. Ann Stat 15(4):1548–1562
Kosorok MR (2009) What's so special about semiparametric methods? Sankhyā 71(2, Ser A): 331–353
Laurent B, Massart P (2000) Adaptive estimation of a quadratic functional by model selection. Ann Stat 28(5):1302–1338
Le Cam L (1956) On the asymptotic theory of estimation and testing hypotheses. In: Proceedings of the third Berkeley symposium on mathematical statistics and probability, vol 1. University of California Press, Berkeley, pp 129–156
Lutz R, Kalisch M, Bühlmann P (2008) Robustified L2 boosting. Comput Stat Data Anal 52:3331–3341
Meinshausen N, Bickel P, Rice J (2009) Efficient blind search: optimal power of detection under computational cost constraint. Ann Appl Stat 3:38–60
Murphy SA, van der Vaart AW (1996) Likelihood inference in the errors-in-variables model. J Multivar Anal 59(1):81–108
Neyman J, Scott EL (1948) Consistent estimates based on partially consistent observations. Econometrica 16:1–32
Pfanzagl J (1990a) Estimation in semiparametric models. Lecture notes in statistics, vol 63. Springer, New York. Some recent developments
Pfanzagl J (1990b) Large deviation probabilities for certain nonparametric maximum likelihood estimators. Ann Stat 18(4):1868–1877
Pfanzagl J (1993) Incidental versus random nuisance parameters. Ann Stat 21(4):1663–1691
Reiersol O (1950) Identifiability of a linear relation between variables which are subject to error. Econometrica 18:375–389
Ritov Y, Bickel PJ (1990) Achieving information bounds in non and semiparametric models. Ann Stat 18(2):925–938
Robins J, Tchetgen Tchetgen E, Li L, van der Vaart A (2009) Semiparametric minimax rates. Electron J Stat 3:1305–1321
Schick A (1986) On asymptotically efficient estimation in semiparametric models. Ann Stat 14(3):1139–1151
Stein C (1956) Efficient nonparametric testing and estimation. In: Proceedings of the third Berkeley symposium on mathematical statistics and probability 1954–1955, vol. I. University of California Press, Berkeley/Los Angeles, pp 187–195
Stone CJ (1975) Adaptive maximum likelihood estimators of a location parameter. Ann Stat 3:267–284
Strasser H (1996) Asymptotic efficiency of estimates for models with incidental nuisance parameters. Ann Stat 24(2):879–901
Tchetgen E, Li L, Robins J, van der Vaart A (2008) Minimax estimation of the integral of a power of a density. Stat Probab Lett 78(18):3307–3311

van der Vaart AW (1988) Estimating a real parameter in a class of semiparametric models. Ann Stat 16(4):1450–1474
van der Vaart A (1991) On differentiable functionals. Ann Stat 19(1):178–204
van der Vaart A (1996) Efficient maximum likelihood estimation in semiparametric mixture models. Ann Stat 24(2):862–878
van Eeden C (1970) Efficiency-robust estimation of location. Ann Math Stat 41:172–181
Wellner JA, Klaassen CAJ, Ritov Y (2006) Semiparametric models: a review of progress since BKRW (1993). In: Frontiers in statistics. Imperial College Press, London, pp 25–44

One-Step Huber Estimates in the Linear Model

P. J. BICKEL*

Simple "one-step" versions of Huber's (M) estimates for the linear model are introduced. Some relevant Monte Carlo results obtained in the Princeton project [1] are singled out and discussed. The large sample behavior of these procedures is examined under very mild regularity conditions.

1. INTRODUCTION

In 1964 Huber [7] introduced a class of estimates (referred to as (M)) in the location problem, studied their asymptotic behavior and identified robust members of the group. These procedures are the solutions $\hat{\theta}$ of equations of the form,

$$\sum_{i=1}^{n} \psi(X_i - \hat{\theta}) = 0 \ , \tag{1.1}$$

where $X_1 = \theta + E_1, \cdots, X_n = \theta + E_n$ and $E_1, \cdots, E_n$ are unknown independent, identically distributed errors which have a distribution F which is symmetric about 0. If F has a density f which is smooth and if f is known, then maximum likelihood estimates if they exist satisfy (1.1) with $\psi = -f'/f$.

Under successively milder regularity conditions on ψ and F, Huber showed in [7] and [8] that such $\hat{\theta}$ were consistent and asymptotically normal with mean θ and variance $K(\psi, F)/n$ where

$$K(\psi, F) = \int_{-\infty}^{\infty} \psi^2(t) f(t) dt \Big/ \left[\int_{-\infty}^{\infty} f(t) d\psi(t) \right]^2 . \tag{1.2}$$

If F is unknown but close to a normal distribution with mean 0 and known variance in a suitable sense, Huber in [7] further showed that (M) estimates based on

$$\begin{aligned} \psi_K(t) &= t && \text{if} \quad |t| < K \\ &= K \operatorname{sgn} t && \text{if} \quad |t| \geq K \end{aligned} \tag{1.3}$$

have a desirable minimax robustness property. If K is finite these estimates can only be calculated iteratively. It has, however, been observed by Fisher, Neyman and others that if F is known and $\psi = (-f'/f)$, the estimate obtained by starting with a $\sqrt{n}$ consistent estimate $\tilde{\theta}$ and performing one Gauss-Newton iteration of (1.1) is asymptotically efficient even when the MLE is not and is equivalent to it when it is (cf. [13]). One purpose of this note is to show that under mild conditions this equivalence holds in the more general context of the linear model for general ψ.

Typically the estimates obtained from (1.1) are not scale equivariant.[1] To obtain acceptable procedures a scale equivariant and location invariant estimate of scale $\hat{\sigma}$ must be calculated from the data and $\hat{\theta}$ be obtained as the solution of

$$\sum_{j=1}^{n} \psi_{\hat{\sigma}}(X_j - \hat{\theta}) = 0 \ , \tag{1.4}$$

where

$$\psi_\sigma(x) = \psi(x/\sigma) \ . \tag{1.5}$$

The resulting $\hat{\theta}$ is then both location and scale equivariant. The estimate $\hat{\sigma}$ can be obtained simultaneously with $\hat{\theta}$ by solving a system of equations such as those of Huber's Proposal 2 [8, p. 96] or the "likelihood equations"

$$\begin{aligned} \sum_{j=1}^{n} \psi\left(\frac{X_j - \theta}{\sigma}\right) &= 0 \ , \\ \sum_{j=1}^{n} \chi\left(\frac{X_j - \theta}{\sigma}\right) &= 0 \ , \end{aligned} \tag{1.6}$$

where $\chi(t) = t\psi(t) - 1$. Or, we may choose $\hat{\sigma}$ independently. For instance, in this article, the normalized interquartile range,

$$\hat{\sigma}_1 = (X_{(n-[n/4]+1)} - X_{([n/4])})/2\Phi^{-1}(3/4) \ , \tag{1.7}$$

and the symmetrized interquartile range,

$$\hat{\sigma}_2 = \text{median}\ \{|X_j - m|\}/\Phi^{-1}(\tfrac{3}{4}), \tag{1.8}$$

are used where $X_{(1)} < \cdots < X_{(n)}$ are the order statistics, Φ is the standard normal cdf and m is the sample median. If $\hat{\sigma} \to \sigma(F)$ at rate $1/\sqrt{n}$ and F is symmetric as hypothesized, then the asymptotic theory for the location model continues to be valid with $K(\psi, F)$ replaced by $K(\psi(\frac{\cdot}{\sigma(F)}), F)$. (E.g., cf. [7].) We shall show (in the context of the linear model) under mild conditions that the one-step "Gauss-Newton" approximation to (1.4)—θ being the only unknown—behaves asymptotically like the root.

The estimates corresponding to ψ_K have a rather appealing form and, of course, all of these Gauss-Newton

* P.J. Bickel is professor, Department of Statistics, University of California, Berkeley, Ca. 94720. This research was performed with partial support of the O.N.R. under Contract N00014-67-A-D151-0017 with Princeton University, and N00014-67-A0114-0004 with the University of California at Berkeley, as well as that of the John Simon Guggenheim Foundation. The author would like to thank P.J. Huber, C. Kraft and C. Van Eeden and D. Relles for providing him with reprints of their work on this subject; W. Rogers III for programming the Monte Carlo computations of Section 3, which appeared in the Princeton project; and a referee who made Tables 1 and 2 reflect numerical realities.

[1] In this article location (scale) invariance refers to procedures which remain unchanged when the data are shifted (rescaled). The term "equivariant" is in accord with its usage in [2]. Thus, $\hat{\theta}$ location and scale equivariant means that $\hat{\theta}(aX_1 + b, \cdots, aX_n + b) = a\hat{\theta}(X_1, \cdots, aX_n) + b$ and $\hat{\sigma}$ scale equivariant means that $\hat{\sigma}(aX_1, \cdots, aX_n) = |a|\hat{\sigma}(X_1, \cdots, X_n)$.

© Journal of the American Statistical Association
June 1975, Volume 70, Number 350
Theory and Methods Section

procedures have the virtue of being simple and easily amenable to hand calculation for simple linear models. An analogous remark was made by Kraft and Van Eeden [11, 12] in connection with estimates based on rank tests.

Details of the model and the estimates are to be found in Section 2. Some Monte Carlo calculations are given in Section 3. Statements and proofs of the asymptotic behavior of the one-steps are given in Section 4. Finally, the proofs of some of the lemmas of Section 4 appear in an appendix.

2. THE MODEL AND ESTIMATES

The class of (M) estimates was extended to the general linear model by Relles [15] and Huber [9]. Here we observe $\mathbf{X} = (X_1, \cdots, X_n)$ where

$$X_j = \sum_{i=1}^{p} c_{ij}\beta_i + E_j \ , \quad 1 \leq j \leq n \ , \tag{2.1}$$

the E_j are as previously, the β_i unknown regression parameters and $C = \|c_{ij}\|$, the design matrix. An (M) *estimate* (scale equivariance not required) is defined quite naturally as a solution $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \cdots, \hat{\beta}_p)$ of the system of equations

$$\sum_{j=1}^{n} c_{ij}\psi(Y_j(\boldsymbol{\beta})) = 0 \ , \quad 1 \leq i \leq p \ , \tag{2.2}$$

where

$$Y_j(\mathbf{t}) = X_j - \sum_{i=1}^{p} c_{ij}t_i \quad \text{if} \quad \mathbf{t} = (t_1, \cdots, t_p) \ . \tag{2.3}$$

Again, if $\psi = -f'/f$, these are the likelihood equations, and if $\psi(t) = t$, $\hat{\boldsymbol{\beta}} = \mathbf{X}C'[CC']^{-1}$, the least squares estimate. To obtain scale equivariance, we again need a scale equivariant estimate $\hat{\sigma}$ which is "shift" invariant, i.e.,

$$\hat{\sigma}(\mathbf{x} + \mathbf{t}C) = \hat{\sigma}(\mathbf{x}) \ . \tag{2.4}$$

The (scale equivariant) (M) estimates are now defined as the solutions of the system,

$$\sum_{j=1}^{n} c_{ij}\psi_{\hat{\sigma}}(Y_j(\hat{\boldsymbol{\beta}})) = 0 \ , \quad i = 1, \cdots, p \ . \tag{2.5}$$

Under various regularity conditions Relles and Huber [15, 9] have shown that $\hat{\boldsymbol{\beta}}$ is asymptotically normal with mean $\boldsymbol{\beta}$ and covariance matrix $K(\psi, F)[CC']^{-1}$ for the nonequivariant case and $K(\psi(\frac{\cdot}{\sigma(F)}), F)[CC']^{-1}$ otherwise. The efficiencies are independent of the design matrix and Huber's robustness results carry through. Let $\boldsymbol{\beta}^*$ be a given estimate of $\boldsymbol{\beta}$ which is shift equivariant, i.e.,

$$\boldsymbol{\beta}^*(\mathbf{x} + \mathbf{t}C) = \boldsymbol{\beta}^*(\mathbf{x}) + \mathbf{t} \ . \tag{2.6}$$

We shall say $\hat{\boldsymbol{\beta}}$ is a *one-step* (M) *estimate of Type 1* if ψ is absolutely continuous with derivative ψ' and $\hat{\boldsymbol{\beta}}$ satisfies the equations

$$\sum_{j=1}^{n} c_{ij}\psi(Y_j(\boldsymbol{\beta}^*)) = \sum_{k=1}^{p} (\hat{\beta}_k - \beta_k^*)\cdot\sum_{j=1}^{n} c_{kj}c_{ij}\psi'(Y_j(\boldsymbol{\beta}^*)), \quad 1 \leq i \leq p \ . \tag{2.7}$$

This system of equations is the linear approximation to the system (2.2) if we use $\boldsymbol{\beta}^*$ as an initial estimate. In the situations we are interested in, $\sum_{j=1}^{n} c_{kj}c_{ij}\psi'(Y_j(\boldsymbol{\beta}^*))$ is well approximated by its asymptotic expectation $\sum_{j=1}^{n} c_{kj}c_{ij}A(\psi, F)$, where

$$A(\psi, F) = \int_{-\infty}^{\infty} \psi'(t)dF(t) = -\int_{-\infty}^{\infty} f(t)d\psi(t) \ . \tag{2.8}$$

(The second equality holds only under mild regularity conditions.) The term on the right makes sense even when ψ is just of bounded variation on intervals. We shall use a slightly more general definition of $A(\psi, F)$ in (4.6). If $\hat{A}(\psi, F)$ is a consistent estimate of $A(\psi, F)$, we therefore define a *one-step* (M) *estimate of Type 2 as* the solution $\hat{\boldsymbol{\beta}}$ of the equations

$$\sum_{j=1}^{n} c_{ij}\psi(Y_j(\boldsymbol{\beta}^*)) = \sum_{k=1}^{p} (\hat{\beta}_k - \beta_k^*)(\sum_{j=1}^{n} c_{kj}c_{ij})\hat{A}(\psi, F) \ , \tag{2.9}$$

or equivalently,

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + \frac{1}{\hat{A}(\psi, F)} \cdot \{\psi(Y_1(\boldsymbol{\beta}^*)), \cdots, \psi(Y_n(\boldsymbol{\beta}^*))\}C'[CC']^{-1} \tag{2.10}$$

when CC' is nonsingular. Similarly we shall speak of scale equivariant one-step (ψ) estimates defined as previously, save that ψ is replaced by $\psi_{\hat{\sigma}}$ where $\hat{\sigma}$ is "shift" invariant *throughout*.

Our principal aim in introducing the one steps was to provide a version of Huber's estimate which is readily computable by hand in the location problem and other simple models. The ψ function of (2.2) here is given by (1.3). For a given scale estimate $\hat{\sigma}$ and the location model, the Type 1 one-step estimate may be written

$$\hat{\beta} = [\{\sum X_i : i \in S_0\} + K[N_+ - N_-]]/N_0 \ , \tag{2.11}$$

where $S_0 = \{i : |X_i - \beta^*| \leq K\hat{\sigma}\}$, $S_+ = \{i : (X_i - \beta^*) > K\hat{\sigma}\}$, $S_- = \{i : (X_i - \beta^*) < -K\hat{\sigma}\}$ and N_0, N_+, N_- are the cardinalities of S_0, S_+ and S_-. If S_0 is empty the estimate is undefined. In the general case, let

$$S_0 = \{j : |X_j - \sum_{i=1}^{p} c_{ij}\beta_i^*| \leq K\hat{\sigma}\} \ ,$$

etc. Then the Type 1 estimate is obtained as follows.

Replace any residual $X_j - \sum_{i=1}^{p} c_{ij}\beta_i^*$ by $K\hat{\sigma}$ if $j \in S^+$ and by $-K\hat{\sigma}$ if $j \in S^-$. If $j \notin S_0$, replace c_{ij} by 0 for $i = 1, \cdots, p$. If we denote the resulting vector of modified residuals by $\mathbf{R}^*$ and the resulting matrix of modified c_{ij} by C^*, then

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = \mathbf{R}^*C'[C^*C^{*\prime}]^{-1} \ . \tag{2.12}$$

Alternatively, it is easy to see that if we define N_0 as before then under the conditions given in Section 4,

$$(N_0/n) \xrightarrow{p} A(\psi_\sigma, F) \tag{2.13}$$

and thus an alternative estimate (Type 2) would be

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + (n/N_0)\mathbf{R}^*C'[CC']^{-1} \tag{2.14}$$

Other possibilities are discussed in [9].

3. SOME MONTE CARLO RESULTS

As part of a larger study, [1], one-step estimates (for ψ_K) were considered as estimates for location under a variety of distributions and sample sizes. The following one-step procedures were considered. Let m denote the median, M the mean.

(1) $M15$; $K = 1.5$; $\delta = \delta_1$; $\beta = M$
(2) $D15$; $K = 1.5$; $\delta = \delta_1$; $\beta = m$
(3) $D20$; $K = 2.0$; $\delta = \delta_1$; $\beta = m$
(4) $P15$; $K = 1.5$; $\delta = \delta_2$; $\beta = m$

These were compared to the following Huber iterative estimates proposed by Hampel.

(5) $A15$; $K = 1.5$; $\delta = \delta_2$
(6) $A20$; $K = 2.0$; $\delta = \delta_2$

Note that comparison of $A15$ and $A20$ to $D15$ and $M15$ and $D20$, respectively, is reasonable since δ_2 and δ_1 are asymptotically equivalent to order $1/\sqrt{n}$ under mild regularity conditions, provided that F is symmetric.[2]

The sample sizes considered were $n = 5$, 10, 20 and 40. The distributions considered (not all being represented for each n) were:

(1) N—the normal
(2) C—the Cauchy
(3) 25 percent (NU)—a mixture of a standard normal distribution with the distribution of a standard normal variate divided by an independent variate having a uniform distribution on the interval (0, 1). The proportions were 75 percent normal, 25 percent of the latter distribution.
(4) t—the t distribution with three degrees of freedom.
(5) DE—the double exponential distribution.
(6) Pseudo-samples in which k observations were drawn from a normal distribution with variance nine (or 100) and the remaining $n - k$ were standard normal deviates. These are denoted by the notation

$$\left(\frac{k}{n}100\right)\text{ percent}\begin{cases}(3N)\\(10N)\end{cases}.$$

Tables 1 and 2 were calculated using Exhibits 5.4–5.8 of [1] as well as measures of accuracy of these exhibits.[3] We refer to [1] for details of the Monte Carlo sampling procedure, a discussion of the accuracy of the results and other material of interest. Using between 640 and 1,000 replicates for each sample and some devices discussed in [1], essentially two-figure accuracy was obtained. We use the notation x/y to denote the efficiency of x with respect to y, i.e., the ratio Var y/Var x. Entries are 0 in cases such as those involving M or $M15$ under the C or 25 percent (NU) distribution in which the variances of these estimates are known to be infinite.

The asymptotic theory of Section 4 leads us to expect that $P15$ and $D15$ will behave like $A15$ and $D20$ like $A20$ in all of these cases. On the other hand, $M15$ should behave like $A15$ in Cases (1), (4), (5) and (6) only. What actually happened can be summarized as follows.

1. The difference between the one-step $P15$ and the iterate $A15$ set to the same scale is negligible across the whole range of distributions. However, the efficiency of the starting point m to $A15$ in this case is never less than .68.

2. If the starting point is too poor for the population at hand the loss in efficiency can be substantial. An example in point is t_3 where $M/m = .6$, $M15/D15 = .8$. Unfortunately, too few shapes and starting points were considered to see if there is a reasonable relation between the efficiency of the starting point to the iterate and that of the one step to the iterate.

3. The choice of scale has quite significant effects as the $P15/A15$, $D15/A15$ comparisons indicate. Unfortunately, the iterated forms of $D15$, $D20$ were not included in the study. Of course this has no bearing on the question of whether the one step is a good substitute for the iterated estimate.

4. Figures not included in this article but available in [1] indicate that the general qualitative nature of Tables 1 and 2 is unchanged if measures of spread other than the variance are used. However, the effect of a nonrobust starting point as in $M15$ is less severe.

5. The difference in computation time between iterate and one step can be substantial. In the Princeton study the average time of computation per estimate was recorded. From these figures it can be seen that the average percent increase in time for $A15$ versus $P15$ was of the order of 25 percent to 30 percent. (This is a percentage of the time required after all constants such as δ_2 have been computed.) Preliminary computations for one steps with scale known for a standard Gaussian population ($n = 20$) indicate that the one-step starting at the median agrees with the iterate (up to two decimal places) between 80 ($K = 1.0$) and 60 ($K = 2.0$) percent of the time.

6. More extensive Monte Carlo computations need to be carried out to get a clear idea of the relationship between one-steps and iterates. This is particularly true for the smaller sample sizes for which the Princeton project figures are essentially unreliable.

1. Efficiencies of One Steps and Starting Points Versus Iterates for Sample Size 20

Efficiencies	Distributions						
	N	25% (3N)	10% (10N)	DE	t_3	25% (NU)	C
P15/A15	1.00	1.0	1.00	.99	1.0	1.0	1.0
m/A15	.70	.9	.83	1.13	.9	.9	1.5
D15/A15	1.00	1.0	.99	.96	1.0	1.0	.9
m/A20	.68	1.0	.92	1.22	1.0	1.0	1.9
D20/A20	1.00	1.0	.98	.97	1.0	1.0	.9
M/m	1.50	1.0	.16	.65	.6	0	0
M15/D15	1.00	1.0	.1–.3	1.01	.8	0	0

NOTE: For n = 20, the last significant figure is reliable at least up to ±1 for shapes other than 10 percent (10N) and up to ±2 for 10 percent (10N) unless a range is shown.

[2] It is easy to show under symmetry that if $F' = f$ is finite at $F^{-1}(\frac{1}{2})$ then $\hat{\sigma}_1$ and $\hat{\sigma}_2$ are asymptotically both Gaussian with mean $F^{-1}(\frac{1}{2})/\Phi^{-1}(\frac{1}{2})$ and variance $[4\Phi^{-1}(\frac{1}{2})f(F^{-1}(\frac{1}{2}))]^{-2}$. These assertions as well as asymptotic equivalence may be argued by replacing the quantile process by the empirical process as in Pyke-Shorack [14] or in the general linear model as in Bickel [4].

[3] Measures of accuracy of these exhibits do not appear in [1] but are available from Andrews *et al.*

2. Efficiencies of One Steps and Starting Points Versus Iterates for Sample Sizes 5, 10 and 40

Efficiencies	n = 5			n = 10				n = 40		
	N	25% NU	C	N	20% 3N	25% NU	C	N	25%NU	C
P15/A15	1.00	.8–1.2	.8–1.2	1.00	1.0	.9–1.0	.9–1.1	1.00	1.0	1.0
m/A15	.76	1.1–1.3	1.0–1.6	.77	1.0	.9–1.0	1.4–1.9	.68	.8	1.5
D15/A15	1.04	.6–.8	.5–1.1	1.02	.9	.2– .9	.1– .6	1.01	1.0	.9
m/A20	.73	1.0–1.5	1.1–1.9	.75	1.0	.9–1.0	1.8–2.3	.67	.8	1.9
D20/A20	1.02	.6–.9	.6–1.0	1.01	.9	.2– .9	.1– .6	1.06	1.0	.9
M/m	1.47	0	0	1.37	.7	0	0	1.53	0	0
M15/D15	.96	0	0	1.00	1.0	0	0	1.00	0	0

NOTE: For n = 5, 10 and shape N the last significant figure is reliable at least up to ±2. Otherwise unless a range is shown the last significant figure is reliable at least up to ±1.

4. THE LARGE SAMPLE BEHAVIOR OF ONE-STEPS

We shall prove asymptotic normality of the one-step estimates under the following simple conditions.

Condition G: The matrices CC'/n tend as $n \to \infty$ to a limit C_0 which is positive definite. Further,

$$\lim_n \max_{i,j} |c_{ij}|/\sqrt{n} = 0 \ . \tag{4.1}$$

We shall also need some smoothness conditions on ψ in addition to a consistency Condition A. The first set, labeled C, is appropriate for simple estimates while the second, S, is needed for scale equivariant estimates.

Condition A: $\int \psi(t)dF(t) = 0$.

Clearly A holds if F is symmetric and ψ is antisymmetric.

Condition C1. The function ψ is of bounded variation in every interval, i.e., it may be written as

$$\psi = \psi^+ - \psi^- \tag{4.2}$$

where $\psi^\pm$ is monotone increasing and further,

$$\int_{-\infty}^{\infty} (\psi^\pm(x+h) - \psi^\pm(x-h))^2 dF(x) = O(1) \quad \text{as} \quad h \to 0 \ . \tag{4.3}$$

$$\sup \frac{1}{|h|} \left\{ \int_{-\infty}^{\infty} (\psi^\pm(x+q+h) - \psi^\pm(x+q))dF(x) : |q| \leqq \epsilon, \ |h| \leqq \epsilon \right\} < \infty \quad \text{for some } \epsilon > 0 \ . \tag{4.4}$$

Condition C2: Suppose that there exists $A(\psi^\pm, F)$ such that

$$\int_{-\infty}^{\infty} (\psi^\pm(x+h) - \psi^\pm(x))dF(x) = hA(\psi^\pm, F) + O(h) \ . \tag{4.5}$$

In this case define

$$A(\psi, F) = A(\psi^+, F) - A(\psi^-, F) \ . \tag{4.6}$$

Condition S_1: (a) The function ψ is as in (4.2) and

$$\sup \{(1/q^2) \int (\psi^\pm((1+\lambda+q)(x+h)) - \psi^\pm((1+\lambda)(x+h)))^2 dF(x) : |h| \leqq \epsilon, \ |\lambda| \leqq \epsilon, \ |q| \leqq \epsilon\} < \infty \tag{4.7}$$

for some $\epsilon > 0$.

$$\sup \left\{ 1/|h| \left| \int (\psi^\pm((1+\lambda)x + h) - \psi^\pm((1+\lambda)x - h))dF(x) \right| : |h| \leqq \epsilon, \ |\lambda| \leqq \epsilon \right\} < \infty \tag{4.8}$$

for some $\epsilon > 0$.

Condition S_2: There exists $A(\psi^\pm, F)$ such that

$$\int [\psi^\pm((1+\lambda)x + h) - \psi^\pm(x)]dF(x) = A(\psi^\pm, F)h + o(|h|) + O(|\lambda h|) + O(\lambda^2) \ . \tag{4.9}$$

Condition S_2 is satisfied if we can formally differentiate under the integral sign, ψ is antisymmetric and F is symmetric (about zero).

Finally we require further conditions on β^* and σ.

Condition B: If $\beta = 0$,

$$\beta^* = O_p(n^{-\frac{1}{2}}) \tag{4.10}$$

which by (2.6) implies that $\beta^* - \beta = O(n^{-\frac{1}{2}})$ in probability if β is true.

Condition D: There exists a positive functional $\sigma(F)$ such that

$$\hat{\sigma} = \sigma(F) + O_p(n^{-\frac{1}{2}}) \ . \tag{4.11}$$

(Because of the invariance assumption (2.4), Assertion (4.11) holds whatever be β if it holds for $\beta = 0$.)

Moreover, writing σ for $\sigma(F)$, we shall suppose that

$$\hat{A}(\psi_{\hat{\sigma}}, F) \to A(\psi_\sigma, F) \tag{4.12}$$

in probability whatever be β.

In the definitions and arguments which follow we shall assume that all probabilities and expectations are calculated under the assumption that $\beta = 0$ unless the contrary is specifically indicated. Also let M be a generic constant. Define

$$T_n(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} c_j[\psi(Y_j(t)) - E(\psi(Y_j(t)))] \ , \tag{4.13}$$

where we write c_j for c_{1j}.

Lemma 4.1: If G and C_1 hold, then

$$\sup \{|T_n(t) - T_n(0)| : |t| \leq M/\sqrt{n}\} \xrightarrow{p} 0 \ . \tag{4.14}$$

(We use $|t|$ to denote the maximum of the absolute

values of the coordinates of $\mathbf{t}$.) Let

$$T_n(\mathbf{t}, \lambda) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} c_j[\psi((1+\lambda)Y_j(\mathbf{t})) - E(\psi((1+\lambda)Y_j(\mathbf{t})))] \quad (4.15)$$

Lemma 4.2: If G and S_1 hold, then

$$\sup\{|T_n(\mathbf{t}, \lambda) - T_n(\mathbf{0}, 0)| : |\mathbf{t}| \le M/\sqrt{n},\ |\lambda| \le \epsilon_n\} \xrightarrow{p} 0 \quad (4.16)$$

where $\epsilon_n \downarrow 0$ in any way whatever.

The proofs of these lemmas are given in the appendix.

From these lemmas we immediately get:

Proposition 4.1: (a) If G, C_1 and C_2 hold, then

$$\sup\left\{\frac{1}{\sqrt{n}}\left|\left[\sum_{j=1}^{n} c_{1j}(\psi(Y_j(\mathbf{t})) - \psi(X_j)) + \left(\sum_{i=1}^{p}\sum_{j=1}^{n} c_{1j}c_{ij}\right)A(\psi, F)\right]\right| : |\mathbf{t}| \le \frac{M}{\sqrt{n}}\right\} \xrightarrow{p} 0 \quad (4.17)$$

(b) If G, S_1 and S_2 hold, then

$$\sup\left\{\frac{1}{\sqrt{n}}\left|\left[\sum_{j=1}^{n} c_{1j}(\psi((1+\lambda)Y_j(\mathbf{t})) - \psi(X_j)) + \left(\sum_{i=1}^{p} t_i \sum_{j=1}^{n} c_{1j}c_{ij}\right)A(\psi, F)\right]\right| : |\mathbf{t}| \le \frac{M}{\sqrt{n}},\ |\lambda| \le \frac{M}{\sqrt{n}}\right\} \xrightarrow{p} 0 \quad (4.18)$$

Proof: Immediate upon expanding $E(\psi((1+\lambda)Y_j(\mathbf{t})) - \psi(X_j))$.

As an immediate consequence of this proposition we obtain

Theorem 4.1: If G, A, C_1, C_2, S_1, S_2, B and D hold and

$$\int \psi_\sigma^2(t)dF(t) < \infty$$

and $\hat{\boldsymbol{\beta}}$ is one step of Type 2, then under the model (2.1),

$$\sqrt{n}\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - (\psi_\sigma(E_1), \cdots, \psi_\sigma(E_n))C'[CC']^{-1}[A(\psi_\sigma, F)]^{-1}\} \to 0 \quad (4.19)$$

in probability. A similar assertion holds when scale is not estimated. Hence, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ has a limiting Gaussian distribution with mean zero and covariance matrix $K(\psi_\sigma, F)C_0^{-1}$ where K is defined by (1.2) with the denominator in general given by $[A(\psi, F)]^2$.

Proof: By invariance reduce to the case $\boldsymbol{\beta} = 0$. Apply (4.18) with $\psi = \psi_\sigma$. Substitute $\beta_i^* - \beta_i$ for t_i, $(\hat{\sigma} - 1)$ for λ, c_{kj} for c_{1j}. Since $\sum_{i=1}^{p}(\beta_i^* - \beta_i)\sum_{j=1}^{n} c_{kj}c_{ij}$ is bounded in probability we can replace $A(\psi_\sigma, F)$ by $\hat{A}(\psi_{\hat{\sigma}}, F)$. The final result follows by Lindeberg's form of the central limit theorem.

Estimates of Type 1 satisfy the conclusion of Theorem 4.1 iff

$$\frac{1}{n\hat{\sigma}}(\psi_{\hat{\sigma}}'(Y_1(\hat{\boldsymbol{\beta}}^*)), \cdots, \psi_{\hat{\sigma}}'(Y_n(\hat{\boldsymbol{\beta}}^*)))CC' \xrightarrow{p} A(\psi_\sigma, F)C_0 . \quad (4.20)$$

It is easy to show that this is true if, in addition to our other conditions, either

Condition E_1: ψ' is uniformly continuous, or

Condition E_2: ψ' is of bounded variation in every interval and

$$E|[\psi']^{\pm}(aX_1 + b) - [\psi']^{\pm}(X_1)| = o(1) \quad \text{as} \quad a \to 1, b \to 0 .$$

Condition E_1 applies to smooth ψ while E_2 applies to Huber's ψ_K function. These conditions are far from necessary.

Although we have for completeness indicated the theory for the general linear model, in that context our theorem is best viewed as support for the feeling that a few iterations in solving a system of equations such as (2.5) lead to estimates whose behavior is much like that of the root. The reasons are:

1) For a multilinear regression, one usually employs a computer in any case and then solving the system (2.5) is not appreciably more difficult than obtaining the least squares estimates.
2) In such a case the only candidate for $\hat{\boldsymbol{\beta}}^*$ is the least squares estimate, and as we shall see, even for moderately heavy tailed distributions the resulting one-step estimate can be poor.

However, for situations such as location, regression through the origin, and the c sample problem, where simple robust starting points such as the median or its analogues exist, the one-step estimates are easy to compute, and, as we have seen for location, quite satisfactory, at least if $\hat{\sigma}$ is chosen properly and the starting point is not too bad. Similar results hold if we replace Condition G by the more general

$$b^2(n)CC' \to C_0 \quad (4.21)$$

where $b(n) \to 0$, C_0 is positive definite,

$$b(n) \max_{i,j} |c_{ij}| \to 0 , \quad (4.22)$$

$\hat{\boldsymbol{\beta}}^*$ is $b^{-1}(n)$ consistent and all other conditions are unchanged.

If ψ is monotone rather than just of bounded variation it may be shown (see [16]) that these conditions guarantee convergence of the iterate as well as the one-step (M) estimate. If ψ is smooth and scale is known, it was shown by Huber [9] that a version of Theorem 4.1 holds for both iterates and one steps if $p \to \infty$ as well as n. The approach of this article does not extend readily to that case.

APPENDIX

Proof of Lemma 4.1: Without loss of generality take $\psi = \psi^+\nearrow$. Begin by noting that for fixed $\mathbf{t}$ with $|\mathbf{t}| \le M$,

$$T_n(\mathbf{t}/\sqrt{n}) - T_n(0) \xrightarrow{p} 0 . \quad (A.1)$$

To see this, calculate

$$E\left(T_n\left(\frac{t}{\sqrt{n}}\right) - T_n(0)\right)^2 = \frac{1}{n}\sum_{j=1}^{n} c_j^2 \operatorname{Var}\left(\psi\left(Y_j\left(\frac{t}{\sqrt{n}}\right)\right) - \psi(X_j)\right) \leq \frac{1}{n}\sum_{j=1}^{n} c_j^2 \int_{-\infty}^{\infty}\left(\psi\left(s - \sum_{i=1}^{p} c_{ij}\frac{t_i}{\sqrt{n}}\right) - \psi(s)\right)^2 f(s)ds \leq \left\{\frac{1}{n}\sum_{j=1}^{n} c_j^2\right\} \max\left\{\int_{-\infty}^{\infty} (\psi(s+h) - \psi(s))^2 f(s)ds \colon |h| \leq pM \max_{i,j}|c_{ij}|/\sqrt{n}\right\} \to 0 \quad \text{(A.2)}$$

by Condition G and (4.3). Decompose the cube $K = \{t\colon |t| \leq ([1/\delta] + 1)\delta M/\sqrt{n}\}$ as the union of cubes with vertices on the grid of points $(j_1\delta M/\sqrt{n}, \cdots, j_p\delta M/\sqrt{n})$ where the $j_i = 0, \pm 1, \cdots, \pm[1/\delta] + 1$. If $|t| \leq M/\sqrt{n}$, let $P(t)$ be (say) the lowest vertex of the cube containing t. For fixed δ, by (A.2)

$$\max\{|T_n(P(t)) - T_n(0)| \colon |t| \leq M/\sqrt{n}\} \xrightarrow{p} 0 \ . \quad \text{(A.3)}$$

On the other hand, let K_1 be any cube of the partition and let P_1 be its lowest vertex. Then, by the monotonicity of ψ,

$$\sup\{|T_n(t) - T_n(P_1)| \colon t \in K_1\} \leq \frac{1}{n}\sum_{j=1}^{n} |c_j| \left\{\left[\psi\left(Y_j(P_1) + \frac{M\delta}{\sqrt{n}}S_j\right) - \psi\left(Y_j(P_1) - \frac{M\delta}{\sqrt{n}}S_j\right)\right] + E\left[\psi\left(Y_j(P_1) + \frac{M\delta}{\sqrt{n}}S_j\right) - \psi\left(Y_j(P_1) - \frac{M\delta}{\sqrt{n}}S_j\right)\right]\right\} \quad \text{(A.4)}$$

where $S_j = \sum_{i=1}^{p} |c_{ij}|$. By arguing as for (A.2) it is easy to see that

$$\frac{1}{n}\operatorname{Var}\left\{\sum_{j=1}^{n} |c_j|\left[\psi\left(Y_j(P_1) + \frac{M\delta}{\sqrt{n}}S_j\right) - \psi\left(Y_j(P_1) - \frac{M\delta}{\sqrt{n}}S_j\right)\right]\right\} \to 0 \ . \quad \text{(A.5)}$$

It follows that to establish the lemma we need only check that

$$\max\left\{\frac{1}{\sqrt{n}}\sum_{j=1}^{n} |c_j|\left[E\left(\psi\left(Y_j(P_n(t)) + \frac{M\delta}{\sqrt{n}}S_j\right)\right) - E\left(\psi\left(Y_j(P_n(t)) - \frac{M\delta}{\sqrt{n}}S_j\right)\right)\right] \colon |t| \leq \frac{M}{\sqrt{n}}\right\} = \delta O(1) \quad \text{(A.6)}$$

uniformly in δ, as $n \to \infty$.

Again using the monotonicity of ψ, it is clear that the expression in (A.6) is bounded by

$$\frac{1}{\sqrt{n}}\left[\sum_{j=1}^{n} |c_{ij}| \max\left\{E\left[\psi\left(X_1 + q + \frac{M\delta}{\sqrt{n}}S_j\right) - \psi\left(X_1 + q - \frac{M\delta}{\sqrt{n}}S_j\right)\right] \colon |q| \leq \frac{MS_j}{\sqrt{n}}\right\}\right]$$

which by (4.4) is $O(\delta/n \sum_{i,j} |c_j c_{ij}|)$ uniformly in δ, for fixed M. The lemma follows.

Proof of Lemma 4.2: The estimate of (A.2) shows that if (4.16) holds,

$$T_n(t_n, \lambda_n) = T_n(0, 0) + 0_p(1) \quad \text{(A.7)}$$

whenever $t_n, \lambda_n \to 0$. Arguing as in Lemma 2.1 it is easy to see that it suffices to prove that

$$\sup\{|T_n(t/\sqrt{n}, \lambda) - T_n(t_0/\sqrt{n}, \lambda)| \colon |t - t_0| \leq \delta, |\lambda| \leq \epsilon_n\} = \delta O_p(1) \quad \text{(A.8)}$$

uniformly in δ, and

$$\sup\{|T_n(t_0/\sqrt{n}, \lambda) - T_n(t_0/\sqrt{n}, 0)| \colon |\lambda| \leq \epsilon_n\} = O_p(1) \quad \text{(A.9)}$$

for each t_0.

Now write

$$T_n(t/\sqrt{n}, \lambda) = T_n(t_0/\sqrt{n}, \lambda) + [T_n(t/\sqrt{n}, \lambda) - T_n(t_0/\sqrt{n}, \lambda)] \ . \quad \text{(A.10)}$$

Bound the last term, using the monotonicity of ψ as before, by

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n} |c_j|\left[\psi\left((1+\lambda)\left(X_j + \frac{\delta S_j M}{\sqrt{n}}\right)\right) - \psi\left(X_j + \frac{\delta S_j M}{\sqrt{n}}\right)\right] + \frac{1}{\sqrt{n}}\sum_{j=1}^{n} |c_j|\left[\psi\left(X_j - \frac{\delta S_j M}{\sqrt{n}}\right) - \psi\left((1+\lambda)\left(X_j - \frac{\delta S_j M}{\sqrt{n}}\right)\right)\right] + \frac{1}{\sqrt{n}}\sum_{j=1}^{n} |c_j|\left[\psi\left(X_j + \frac{\delta S_j M}{\sqrt{n}}\right) - \psi\left(X_j - \frac{\delta S_j M}{\sqrt{n}}\right)\right] + \frac{1}{\sqrt{n}}\sum_{j=1}^{n} |c_j| E\left[\psi\left((1+\lambda)\left(X_j + \frac{\delta S_j M}{\sqrt{n}}\right)\right) - \psi\left((1+\lambda)\left(X_j - \frac{\delta S_j M}{\sqrt{n}}\right)\right)\right]. \quad \text{(A.11)}$$

Let $W_n^{(i)}(\lambda, \delta)$, $1 \leq i \leq 3$, be the stochastic processes obtained by centering the preceding first three sums at their expectation. By (4.17) and Condition G,

$$E(W_n^{(1)}(\lambda_1, \delta) - W_n^{(1)}(\lambda_2, \delta))^2 \leq K_1(\lambda_1 - \lambda_2)^2 \quad \text{(A.12)}$$

where K_1 is independent of n, λ_i, and δ for all λ_i sufficiently small, n large. Hence, by [5, p. 95],

$$\max\{|W_n^{(1)}(\lambda, \delta)| \colon |\lambda| \leqq \epsilon_n\} \xrightarrow{p} 0 \ . \quad \text{(A.13)}$$

A similar argument works for $W_n^{(3)}$, while $W_n^{(2)}$ may be taken care of as in the proof of Lemma 2.1. Similarly,

$$\max\{|T_n(t_0/\sqrt{n}, \lambda) - T_n(t_0/\sqrt{n}, 0)| \colon |\lambda| \leq \epsilon_n\} \xrightarrow{p} 0 \quad \text{(A.14)}$$

uniformly in $|t_0| \leq M$. In view of (A.13) and (A.14), to prove (A.8) we need only bound

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n} |c_j|\left[E\left[\psi\left((1+\lambda)\left(X_j + \frac{\delta S_j M}{\sqrt{n}}\right)\right) - E\left(\psi\left((1+\lambda)\left(X_j - \frac{\delta S_j M}{\sqrt{n}}\right)\right)\right]\right]$$

by $|\delta| O(1)$ for $|\lambda| \leq \epsilon_n$. But this can be done using (4.18) as (4.14) was used in Lemma 2.1. Finally, (A.9) follows by using the same "tightness" argument as we employed for (A.13).

[*Received September 1971. Revised July 1974.*]

REFERENCES

[1] Andrews, D.F., *et al.*, *Robust Estimates of Location: Survey and Advances*, Princeton, N.J.: Princeton University Press, 1972.

[2] Berk, R., "A Special Structure and Equivariant Estimation," *The Annals of Mathematical Statistics*, 38 (October 1967), 1436–45.

[3] Bickel, P.J. and Wichura, M., "Convergence Criteria for Multiparameter Stochastic Processes," *The Annals of Mathematical Statistics*, 42 (October 1971), 1656–70.

[4] ———, "Analogues of Linear Combinations of Order Statistics in the General Linear Model," *Annals of Statistics*, 1 (July 1973), 597–616.

[5] Billingsley, P., *Convergence of Probability Measures*, New York: John Wiley and Sons, Inc., 1968.

[6] Hájek, J. and Sidak, Z., *Theory of Rank Tests*, New York: Academic Press, 1967.

[7] Huber, P.J., "The Behaviour of Maximum Likelihood Estimates Under Nonstandard Conditions," *Proceedings of the Fifth Berkeley Symposium*, 1 (1965), 221–33.

[8] ———, "Robust Estimation of a Location Parameter," *The Annals of Mathematical Statistics*, 35 (March 1964), 73–101.

[9] ———, "Robust Regression," *Annals of Statistics*, 1 (December 1973), 799–821.

[10] Kraft, C. and Van Eeden, C., "Efficient Linearized Estimates Based on Ranks," *Proceedings of the First International Symposium on Nonparametric Statistics*, Cambridge: Cambridge University Press, 1969, 267–73.

[11] ———, "Asymptotic Efficiencies of Methods of Computing Efficient Estimates Based on Ranks," *Journal of the American Statistical Association*, 67 (March 1969), 199–202.

[12] ———, "Linearized Rank Estimates and Signed Rank Estimates for the General Linear Hypothesis," *The Annals of Matehmatical Statistics*, 43 (January 1972), 42–57.

[13] Le Cam, L., "On the Asymptotic Theory of Estimation and Testing Hypotheses," *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, Berkeley: University of California Press, 1956, 129–56.

[14] Pyke, R. and Shorack, G., "Weak Convergence of a Two-Sample Empirical Process and a New Approach to the Chernoff-Savage Theorems," *The Annals of Mathematical Statistics*, 39 (June 1968), 755–71.

[15] Relles, D., "Robust Regression by Modified Least Squares," thesis, Yale University, 1968.

[16] Yohai, V.J., "Robust Estimation in the Linear Model." *Annals of Statistics*, 2 (May 1974), 562–67.

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 $\mathcal{F}$ by minimizing $V(\psi) + \sup\{b^2(\psi, G, H): F \in \mathcal{F}\}$ over Ψ. More generally, the maximum risk of T_n as above, is for any reasonable symmetric loss function determined by $V(\psi)$ and $\sup\{|b(\psi, G, H)|: F \in \mathcal{F}\}$.

In Section 2 we study the univariate case as follows.

(1) We evaluate

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

for the $\mathcal{F}$ we have introduced. Subscripts on b indicate which $\mathcal{F}$ 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 $\mathcal{F}_{ac0}$;
4) an equivariant estimate which minimizes the maximum M.S.E. of prediction under $\mathcal{F}_{ac0}$ (Theorem 3.2);
5) the natural equivariant extension of Huber's MIA:A estimates which minimizes the maximum M.S.E. of prediction under $\mathcal{F}_{c0}$.

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

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

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

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

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,

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

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

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

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

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

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),

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

while the Hampel-Krasker solution minimizes

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

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

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

$$\psi(x, u, G_0) = \psi(xB, u, G_0B^{-1})B^T \tag{3.15}$$

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

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

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

$$b(\psi, G, H) = \int\int \psi(x, u)[H(du \mid x) - \Phi(du)]G_0(dx)$$

$$= \int\int \psi(x, u)[M^+(du \mid x) - M^-(du \mid x)]\alpha(x)G_0(dx) \tag{A.2}$$

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

$$|b|(\psi, F) = t\left\{\sum_{j=1}^p \left(\int\int \psi_j(x, u)M(du \mid x)G_0(dx)\right)^2\right\}^{1/2}$$

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

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}

$$a_0\colon \psi \to \int |\psi|^2(x, u)\Phi(du)G_0(dx)$$

$$a_1\colon \psi \to \int \sup_u |\psi(x, u)| G_0(dx)$$

$$a_2\colon \psi \to \int ux^T\psi(x, u)\Phi(du)G_0(dx)$$

$$a_3\colon \psi \to \sup_{x,u} |\psi(x, u)|.$$

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

$$\inf\{a_0(\psi)\colon a_1(\psi) \leq \lambda, a_2(\psi) = I, a_3(\psi) \leq M\}$$

$$= \inf\{a_0(\psi) - 2 \operatorname{tr} Q[a_2(\psi) - I] + (2/s)[a_0(\psi) - \lambda]\}. \tag{A.4}$$

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

The Annals of Statistics
1984, Vol. 12, No. 3, 864-879

PARAMETRIC ROBUSTNESS: SMALL BIASES CAN BE WORTHWHILE[1]

BY P. J. BICKEL

University of California, Berkeley

We study estimation of the parameters of a Gaussian linear model $\mathcal{M}_0$ when we entertain the possibility that $\mathcal{M}_0$ is invalid and a larger model $\mathcal{M}_1$ should be assumed. Estimates are robust if their maximum risk over $\mathcal{M}_1$ is finite and the most robust estimate is the least squares estimate under $\mathcal{M}_1$. We apply notions of Hodges and Lehmann (1952) and Efron and Morris (1971) to obtain (biased) estimates which do well under $\mathcal{M}_0$ at a small price in robustness. Extensions to confidence intervals, simultaneous estimation of several parameters and large sample approximations applying to nested parametric models are also discussed.

1. Introduction. The basic aim of robust inference as developed by Huber, Hampel and others has been the production and study of statistical procedures which

(a) perform reasonably well when the parametric assumptions are perfectly satisfied; and

(b) are relatively insensitive to nonparametric departures from parametric assumptions which a given data set is believed to satisfy.

The main parametric model considered has been the Gaussian linear model and the departures, outliers and gross errors in the variables, have been modeled by assuming non-Gaussian error distributions and, where suitable, dependence between the independent and error variables.

An important aspect of this point of view is a focus on inference about parameters of interest rather than on deciding whether the parametric model provides an adequate fit. This is in contrast to the older approach of estimation and testing after a goodness of fit test or more generally rejection of outliers.

The same point of view makes sense in a purely parametric context. We have two possible parametric models in mind, $\mathcal{M}_0$, $\mathcal{M}_1$ with $\mathcal{M}_0 \subset \mathcal{M}_1$. Our primary interest is in estimating parameters which are identifiable in $\mathcal{M}_1$.

Again,

(i) we believe that $\mathcal{M}_0$ is adequate and want estimates or confidence regions based on estimates that perform well under that assumption. However

(ii) we wish to guard against the possible departures presented by $\mathcal{M}_1$.

Received April 1983; revised April 1984.

[1] Work performed with the partial support of Office of Naval Research Contract N00014-80-C-0163 and the Adolph and Mary Sprague Miller Foundation. Some of this material was presented at the 1980 Wald Lectures of the Institute of Mathematical Statistics.

AMS 1980 *classifications*. Primary 62F10; secondary 62F25.

Key words and phrases. Parametric robustness, pretesting, limited translation estimates, confidence intervals.

Here is the main situation we are thinking of with some specific examples.

Nested linear models. We observe $y_{n\times 1}$ where

$$y = \theta + e.$$

e is an n-variate normal vector with mean 0 and covariance matrix Σ. θ ranges freely over an r-dimensional linear space Θ_0 under $\mathcal{M}_0$ and over an s-dimensional linear space $\Theta_1 \supset \Theta_0$ under $\mathcal{M}_1$ where $r < s \leq n$. We suppose Σ known. Our asymptotic analysis in Section 5 will permit us as usual to substitute a consistent estimate $\hat{\Sigma}$ for Σ. We are interested in inference about $\mu(\theta)$ where μ is a linear function of θ. Special cases are:

1(a) *Pooling means* (Mosteller, 1948). We are given two samples $X_1, \cdots, X_m$ independent $\mathcal{N}(\mu, \sigma^2)$; $Y_1, \cdots, Y_n$ independent $\mathcal{N}(\mu + \Delta, \sigma^2)$. We want to estimate or set a confidence interval on μ. We believe $\Delta = 0$ ($\mathcal{M}_0$) but want to guard against arbitrary Δ ($\mathcal{M}_1$). Plausible examples, e.g. measurements in a current and previous survey, are discussed by Mosteller.

1(b) *Additive effects with possible interactions.* Suppose $\mathcal{M}_1$ is an ANOVA model in the sense of Scheffé (1959), possibly including random effects, which contains some interaction terms as well as main effects, and $\mathcal{M}_0$ is purely additive specifying all interactions to be 0. We take the variances of all random effects as well as measurement errors to be known. We want to study some or all of the main effects. An interesting special case is the crossover design discussed by B. W. Brown (1980). Here two groups of subjects I and II which for simplicity we take of equal size $n/2$ are each administered two drugs A, B in succesion and responses measured. The second drug is administered after response to the first has been measured and a time deemed sufficient for the effect of the first to wear off has elapsed. The order of administration of the drugs is AB in group 1, BA in group 2. Model $\mathcal{M}_1$ here is that the response $Y_{ijk(u)}$ of the jth subject in group i during period k who is administered drug u during that period is

$$Y_{ijk(u)} = \mu + \pi_k + \phi_u + \lambda_{uk} + \xi_{ij} + \varepsilon_{ijk}$$

where π_k, $k = 1, 2$, is the period effect, ϕ_u, $u = \text{A, B}$ is the drug effect, and λ_{uk} is the interaction of drug u and period k with $\lambda_{u1} = 0$. These are all fixed. As usual, identifiability requires further linear restrictions. On the other hand, ξ_{ij}, the effect of the jth subject in group i, is considered random $\mathcal{N}(0, \sigma_\xi)$, and ε_{ijk}, the within subject deviation for the kth period (including measurement error), is modeled as $\mathcal{N}(0, \sigma_\varepsilon^2)$. All are modeled as independent of each other. We assume σ_ξ^2, σ_ε^2 known. $\mathcal{M}_0$ specifies that, as we hope, there is no interaction, $\lambda_{uk} \equiv 0$. We are interested in estimating $\phi_b - \phi_a$, the difference in effectiveness of the drugs.

1(c) *Nested regression models.* Write $\theta = X\beta$, $\beta_{s\times 1}$, $X = (x_1, \cdots, x_s)$ an $n \times s$ matrix of rank s and think of the s columns of X as corresponding to s independent variables. Suppose β ranges freely over R^s under $\mathcal{M}_1$ but $s - r$ coordinates of β are set equal to 0 under $\mathcal{M}_0$, i.e. $s - r$ of the independent variables are irrelevant. Various linear functions $\mu(\theta)$ are of interest, for instance the

vector of expectations θ itself or one or more predicted values $x\beta$, at various values x.

From this special case we will proceed (under regularity conditions) by an asymptotic analysis to the general case of

Nested parametric models. We observe $(X_1, \cdots, X_n)$ with joint density $p_n(x, \theta)$ (with respect to some measure ν_n). Under $\mathscr{M}_1$, $\theta \in \Theta_1$, an open subset of s-dimensional space. Under $\mathscr{M}_0$, $\theta \in \Theta_0 \subset \Theta_1$, a (locally) r-dimensional subsurface of Θ_1, and μ is a smooth vector-valued function of θ. This of course covers all previous situations as well as many others including Example 1 with σ^2 unknown, nested loglinear models, etc.

Our point of view, essentially already suggested by Hodges and Lehmann (1952), page 402, is that procedures should be judged by their maximum risks under $\mathscr{M}_0$ and $\mathscr{M}_1$. So, in the context of nested parametric models, if $M(\theta, \delta)$ is the risk of a decision rule δ when θ is true we should look at

$$m(\delta) = \sup\{M(\theta, \delta): \theta \in \Theta_0\}, \quad M(\delta) = \sup\{M(\theta, \delta): \theta \in \Theta_1\}.$$

M can be thought of as a measure of robustness of δ and we should be interested in procedures which make m small subject to a bound on M.

In the basic linear model example the solutions we end up with are necessarily biased under $\mathscr{M}_1$. Robustness requires that the biases be bounded through M. The worthwhile gains are in reduction of m over the unbiased minimax estimate.

In Section 2 we apply this theory to the linear model example for quadratic loss when μ is one dimensional. The optimal procedures are difficult to compute. We motivate a family of reasonable approximately optimal solutions, compare them numerically to the optimum and other competitors and also briefly discuss the crucial question of selection within the family.

In Section 3 we discuss confidence intervals based on these estimates. In Section 4, we derive, using results of Berger (1982) and Huber (1977), some procedures for the multivariate case. In Section 5, we show how these ideas generalize to yield reasonable procedures in nested parametric models and, finally, in Section 6, give conclusions and propose open questions.

2. The nested linear models: dim(μ) = 1, quadratic loss.

a) *Optimality theory.* We specialize to estimation of μ with quadratic loss. That is, we assume that μ is real, linear, and if $\delta(x)$ is an estimate

$$M(\theta, \delta) = E_\theta(\delta(X) - \mu(\theta))^2. \tag{2.1}$$

Since we assume Σ known, we can, by taking $Y^* = Y\Sigma^{-1/2}$, $\mathscr{M}_i^* = \mathscr{M}_i\Sigma^{-1/2}$, reduce our problem to one in which the observation Y^* has covariance matrix $\sigma^2 I$, the standard linear model.

Let $\hat{\mu}_i = \mu(\hat{\theta}_i)$, $i = 0, 1$, be the least squares estimates of μ under $\mathscr{M}_0$, $\mathscr{M}_1$ respectively. Then, for $i = 0, 1$, $\hat{\mu}_i$ has constant risk and is minmax under $\mathscr{M}_i$. Let

σ_i^2 be the variance of $\hat{\mu}_i$ so that

$$\inf_\delta M(\delta) = \sigma_1^2, \quad \inf_\delta m(\delta) = \sigma_0^2.$$

Let $\hat{\mu}_c^*$ minimize $m(\delta)$ subject to $M(\delta)/\sigma_1^2 \leq 1/c$ so that $\hat{\mu}_i^* = \hat{\mu}_i$, $i = 0, 1$. Note that $M(\hat{\mu}_0) = \infty$ and $\hat{\mu}_0$ is certainly not robust. Let

$$\rho = \operatorname{corr}(\hat{\mu}_0, \hat{\mu}_1) = \sigma_0/\sigma_1 \tag{2.2}$$

which is independent of the error variance σ^2,

$$\hat{\Delta} = \hat{\mu}_1 - \hat{\mu}_0 \tag{2.3}$$

and

$$\sigma_\Delta^2 = \sigma_1^2(1 - \rho^2), \tag{2.4}$$

its variance.

PROPOSITION 1. *The estimate $\hat{\mu}_c^*$ may be written*

$$\hat{\mu}_c^* = \hat{\mu}_0 + \sigma_{\hat{\Delta}} w_q^*(\hat{\Delta}/\sigma_{\hat{\Delta}}) \tag{2.5}$$

where

$$q^2 = (1 - c)/c(1 - \rho^2) \tag{2.6}$$

$$\begin{aligned} &w_q^* \text{ is odd and obtained by minimizing } Ew^2(Z) \text{ subject to} \\ &\sup_\Delta E(w(Z + \Delta) - \Delta)^2 \leq 1 + q^2 \text{ for } Z \sim \mathcal{N}(0, 1). \end{aligned} \tag{2.7}$$

NOTE. Evidently w_q^* is the solution of the special case $\mu = \theta$, $r = 0$, $s = 1$, $\sigma^2 = 1$. We call this problem (P).

PROOF. By sufficiency reduce to $\hat{\theta}_1$ and without loss of generality choose a canonical basis so that $\hat{\theta}_0$ consists of the first r components of $\hat{\theta}_1$ and all components of $\hat{\theta}_1$ are independent normal variables with variance σ^2. Moreover we can arrange that $\hat{\mu}_0/\sigma_0$ is the first component of $\hat{\theta}_1$ and $\hat{\Delta}/\sigma_1(1 - \rho^2)^{1/2}$ is the $(r+1)$st component. Note by Hodges and Lehmann (1952) that $\hat{\mu}_c^*$ is unrestrictedly minimax for the "mixed" model: for suitable $\lambda(c)$ and $\theta = (\theta^{(1)}, \cdots, \theta^{(s)})$, $\hat{\theta}_1$ has density $(1 - \lambda)p_1 + \lambda p_0$ where p_1 is the density of $\hat{\theta}_1$ under $\mathcal{M}_1$ and θ, while p_0 is the density of $\hat{\theta}_1$ under $(\theta^{(1)}, \cdots, \theta^{(r)}, 0, \cdots, 0)$, i.e. under $\mathcal{M}_0$. We can reduce this unrestricted problem by invariance, using for instance Kiefer's (1957) general results. Since we want to estimate

$$\sigma_0\theta^{(1)} + (1 - \rho^2)^{1/2}\sigma_1\theta^{(r+1)},$$

the problem is invariant under arbitrary translations of $\theta^{(i)}$, $i \neq 1, r + 1$, and we can reduce to $\hat{\mu}_0$, $\hat{\Delta}$. The problem is also invariant under translations of $\hat{\mu}_0$, keeping $\hat{\Delta}$ fixed. Since $\hat{\mu}_c^*$ is unique it therefore must be of the form $\mu_0 + w(\hat{\Delta})$. Claims (2.7) and (2.6) follow by calculation. □

Unfortunately calculation of w_q^* is difficult. See Bickel (1983) for its rather unpleasant qualitative features.

In view of these unpleasant features, it is natural to seek other families of robust estimates with more satisfactory behaviour. By invariance it seems reasonable to look for $\hat{\mu}$ of the form

$$\hat{\mu}_0 + \sigma_{\hat{\Delta}} w(\hat{\Delta}/\sigma_{\hat{\Delta}}). \tag{2.8}$$

For any such estimate

$$M(\hat{\mu}) = \sigma_1^2(\rho^2 + (1 - \rho^2)\sup_{\Delta} E(w(Z + \Delta) - \Delta)^2) \tag{2.9}$$

$$m(\hat{\mu}) = \sigma_1^2(\rho^2 + (1 - \rho^2)Ew^2(Z)). \tag{2.10}$$

Abusing notation, let us call the coefficients of $(1 - \rho^2)$ inside parentheses in these expressions $M_0(w)$, $m_0(w)$. They correspond to M and m in problem (P).

b) *"Approximate" optimality in problem* (P). From (2.9) and (2.10) reasonable w in problem (P) correspond to reasonable $\hat{\mu}$. In problem (P) we observe $X = Z + \Delta$, $Z \sim \mathcal{N}(0, 1)$ and we want to minimize $m_0(w)$ subject to a bound on $M_0(w)$. Three approximate optimality principles lead to the same family, the limited translation estimates of Efron and Morris (1971) defined by

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

which leads to $M_0(e_q) = 1 + q^2$.

I. *Optimality in a related problem* (Bickel, 1983, Marazzi, 1980). Suppose π is a prior distribution, $r(\pi)$ the Bayes risk, w_π the Bayes estimate, and $G_\pi = \pi * \Phi$, where $*$ denotes convolution, is the marginal distribution of X. Then,

$$r(\pi) = 1 - I(G_\pi) \tag{2.11}$$

$$w_\pi(x) = x + (g_\pi'/g_\pi)(x) \tag{2.12}$$

where g_π is the density of G_π, $I(G)$ is the Fisher information where

$$\begin{aligned} I(G) &= \int \frac{[g']^2}{g}(x)\, dx, && \text{if the integral is defined} \\ &= \infty && \text{otherwise.} \end{aligned}$$

By Hodges and Lehmann (1952) and (2.11), the optimal w_q^* corresponds to G_q^* which for some $\lambda(q)$ minimizes $I(G)$ over $\mathcal{G}_0 = \{G = (1 - \lambda)\Phi + \lambda\Phi * H, H \text{ arbitrary}\}$. If we "approximate" $\mathcal{G}_0$ by $\mathcal{G}_1 = \{G = (1 - \lambda)\Phi + \lambda H, H \text{ arbitrary}\}$ we arrive at Huber's (1964) problem with solution G_1 where

$$\begin{aligned} (g_1'/g_1)(x) &= -x, && |x| \leq q \\ &= -q \operatorname{sgn} x, && |x| > q. \end{aligned}$$

Substituting into (2.12), we get the Efron-Morris family.

II. *Bounding unbiased estimate of risk* (Berger, 1982). If

$$\psi(x) = x - w(x) \tag{2.13}$$

under mild conditions

$$M(\Delta, w) = 1 + E_\Delta(\psi^2(x) - 2\psi'(x))$$

so that $1 + \psi^2(x) - 2\psi'(x)$ is the UMVU estimate of $M(\eta, w)$. Berger (in a more general context) proposes minimizing $m_0(w)$ subject to $\psi^2(x) - 2\psi'(x) \leq q^2$. The solution is easily seen to be e_q.

In fact Berger's approach must yield the same results as approach I both in our context and his more general restricted Bayes models. To see this in our model, note that

$$\begin{aligned}
&\inf_w\{(1 - \lambda)m_0(w) + \lambda \sup_x(1 + \psi^2(x) - 2\psi'(x))\} \\
&\quad = 1 + \inf_\psi \sup\left\{\int (\psi^2(x) - 2\psi'(x))G(dx): G \in \mathscr{G}_1\right\} \\
&\quad = 1 - \min\{I(G): G \in \mathscr{G}_1\}
\end{aligned}$$

by a minmax argument.

III. *Bounding unbiased estimate of bias.* Note that $\psi(X)$ is the UMVU estimate of the bias of $w(X)$. Thus it seems reasonable to minimize $m_0(w)$ subject to $\sup_x |\psi(x)| \leq q$. This is the exact analogue of Hampel's robustness formulation. The solution is again e_q.

For further optimality properties of Efron-Morris estimates, see Bickel (1983).

c) *Performance of Efron-Morris* (E-M) *estimates and competitors.* We measure the relative performance of estimates $\hat{\mu}$ by their relative savings and losses in risk with respect to $\hat{\mu}_1$

$$S(\hat{\mu}) = 1 - m(\hat{\mu})/m(\hat{\mu}_1), \quad L(\hat{\mu}) = M(\hat{\mu})/M(\hat{\mu}_1) - 1.$$

For estimates of the form (2.8),

$$S(\hat{\mu}) = (1 - \rho^2)(1 - m_0(w)), \quad L(\hat{\mu}) = (1 - \rho^2)(M_0(w) - 1).$$

Table 1 gives $1 - m_0(w)$ as a function of $q^2 = M_0(w) - 1$ for the E-M estimates, for w_q^* (calculated by Dr. A. Marazzi) and for some competitors which we now discuss.

Pretesting estimates. A type of procedure long advocated by Bancroft and others (see Bancroft and Han, 1977, for a review) are estimates

$$\begin{aligned}
\hat{\mu} &= \hat{\mu}_0, \quad |\hat{\theta}_1 - \hat{\theta}_0| \leq c\sigma \\
&= \hat{\mu}_1, \quad \text{otherwise}
\end{aligned}$$

with c chosen to produce an appropriate level for the test of H: $\mathscr{M}_0$ vs. $\mathscr{M}_1$ based

TABLE 1
Gain at 0, $g = 1 - m_0(\omega)$, as a function of the increase in maximum risk $q^2 = M_0(w) - 1$.

q^2	g_e	g_b	g_s	g_j	q	$d(q)$
.1	.413	.085	—	—	.316	.715
.2	.538	.155	—	.330	.447	.903
.3	.619	.225	—	.438	.548	1.053
.4	.676	.290	—	.523	.632	1.175
.5	.721	.350	.711	.592	.707	1.281
.6	.758	.405	.753	.648	.775	1.370
.7	.786	.455	.788	.695	.837	1.461
.8	.811	.500	.816	.735	.894	1.538
.9	.832	.540	.840	.768	.949	1.608
1.0	.850	.58	.859	.796	1.000	1.679

Note: g_e is the increase for the E-M estimate, g_b for the pretest, g_s for the Sacks family, g_j for Jeffreys' type of generalized Bayes estimate. q and $d(q)$ are the critical values for the E-M and pretest estimates.

on $(|\hat{\theta}_1 - \hat{\theta}_0|)/\sigma$. If $|\hat{\mu}_1 - \hat{\mu}_0| \neq |\hat{\theta}_1 - \hat{\theta}_0|$, this estimate is not of the form (2.8). A version of that form can be based on testing $H: E\hat{\Delta} = 0$ vs. $E\hat{\Delta} \neq 0$ and is given by

$$\hat{\mu}_c^B = \hat{\mu}_0 + \sigma_{\hat{\Delta}} b_q\left(\frac{\hat{\Delta}}{\sigma_{\hat{\Delta}}}\right)$$

with

$$\begin{aligned} b_q(x) &= 0, \quad |x| \leq d(q) \\ &= x, \quad |x| > d(q) \end{aligned} \tag{2.14}$$

and d chosen so that

$$M_0(b_q) = 1 + q^2.$$

The ψ function corresponding to b_q via (2.13) corresponds to hard rejection which is known not to work well. This seems true here too. The Bancroft-Han estimate is even worse. (See also Sclove et al. (1972).

Another interesting and desirable feature of the E-M family is monotonicity of $M(\Delta, e_q)$ as a function of $|\Delta|$, i.e. $M_0(e_q)$ is assumed at $|\Delta| = \infty$. This is not true of the pretest estimates and more generally estimates which correspond to redescending ψ functions. Nevertheless we can expect smooth versions of such estimates to perform reasonably well. Motivated by Sacks and Ylvisaker (1978), J. Sacks has proposed a family of such ψ,

$$\psi_\gamma(x) = 2(2 + (|x| - \gamma)_+^2)^{-1}x.$$

Another natural family consists of the Jeffreys' type estimates which are generalized Bayes with respect to a prior distribution placing mass p at 0 and corresponding to Lebesgue measure otherwise.

$$\delta_p(x) = x((1/p - 1)\varphi(x) + 1)^{-1}.$$

Table 1 shows very substantial gains in m_0 for small payments in M_0. Small

biases can be very worthwhile. The pretest estimates are clearly poor and the Jeffreys type estimates are inferior to both the E-M and Sacks estimates.

There is, of course, a serious question as to which E-M estimate to use. The natural way is to calibrate by the maximum $L(\hat{\mu})$ we are willing to tolerate. This of course depends both on ρ^2 and $M_0(w)$. For instance, if $n_1 = n_2$ in the pooling example $\rho^2 = \frac{1}{2}$. If we are willing to accept a 10% loss we would take $q = .2$ and obtain a gain of (.5) (.538) = 26.9%.

Another idea is to bound the maximum squared bias of $\hat{\mu}$ standardized by the variance of $\hat{\mu}_1$. For the E-M estimates this equals $L(\hat{\mu})$. The remaining approach of choosing d according to a reasonable level for the test of H: $\Delta = 0$ based on $\hat{\Delta}$ yields unreasonably high values of $L(q)$ and is not recommended.

The performance of E-M is markedly better than that of the "Jeffreys" or pretest procedures for small q^2. This is in accordance with the asymptotic results of Bickel (1983). Since the Sacks' procedures which are on the whole comparable with E-M cannot be extended over the whole q^2 range, we are left with E-M as the candidate of choice.

The best we can do in terms of $m_0(w)$ for given $M_0(w)$ cannot be calculated exactly. However effective numerical procedures have been derived in Marazzi (1980, 1982). Here is a table of the optimal g based on results he has supplied.

q	.06	.12	.19	.29	.44	.70
g_0	.39	.49	.57	.66	.74	.82

3. Nested linear models: μ univariate.

Confidence intervals and other loss functions. In univariate estimation problems, we usually want confidence intervals as well as point estimates. Since, given our assumed knowledge of σ, we can form fixed width confidence intervals based on $\hat{\mu}_1$, it seems reasonable to ask how intervals of the same width based on estimates $\hat{\mu}$ perform. This boils down to fixing a width $2z\sigma_1$ and using the loss function

$$\begin{aligned} \ell(\theta, d) &= 1 \text{ if } |d - \mu(\theta)| \geq z\sigma_1 \\ &= 0 \text{ otherwise} \end{aligned} \tag{3.1}$$

$$M(\theta, \hat{\mu}) = P[|\hat{\mu} - \mu(\theta)| \geq z\sigma_1] = 1 - P_\theta[\mu(\theta) \in \hat{\mu} + z\sigma_1]. \tag{3.2}$$

From the argument of Proposition 1 it is easy to see that for any loss function of the form $\ell(|\mu(\theta) - d|)$, equivariant estimates are of the form (2.8). Calculation of the optimal procedures is even more hopeless for this loss function. However, it is easy to see that approximate optimality approach III continues to yield the E-M estimate. More generally

PROPOSITION 2. *Suppose $\ell(\theta, d) = \ell(|\mu(\theta) - d|)$ and ℓ is nondecreasing. Then $m(\hat{\mu})$ is minimized among all equivariant $\hat{\mu}$ of the form* (2.8) *with $|\psi(x)| \leq q$ by an* E-M *estimate*

$$\hat{\mu}_c^e = \hat{\mu}_0 + \sigma_{\hat{\Delta}} e_q(\hat{\Delta}/\sigma_{\hat{\Delta}}). \tag{3.3}$$

PROOF. Without loss of generality, suppose $\sigma_{\hat{\Delta}} = 1$. If $\theta \in \Theta_0$ and $\hat{\mu}$ is given by (2.8)

$$m(\hat{\mu}) = E\ell(|U + w(V)|)$$

where U, V are independent normal with mean 0. By Anderson's theorem (Anderson, 1955) $E(\ell(|U + w(V)|) \mid V)$ is monotone increasing in $|w(V)|$. The proposition follows. □

The risk of an E-M estimate (3.3) for a loss function $\ell(|\theta - d|)$ is given by

$$M(\theta, \hat{\mu}_c^e) = \int_{-\infty}^{\infty} \Big\{ \ell(\sigma_0 u - \Delta)[\Phi(d - \tilde{\Delta}) - \Phi(-q - \tilde{\Delta})]$$

$$+ \int_{q-\tilde{\Delta}}^{\infty} \ell(\sigma_0 u + \sigma_1(1 - \rho^2)^{1/2}(w - q))\phi(w)\, dw \tag{3.4}$$

$$+ \int_{-\infty}^{-q-\tilde{\Delta}} \ell(\sigma_0 u + \sigma_1(1 - \rho^2)^{1/2}(w + q))\phi(w)\, dw \Big\} \phi(u)\, du$$

where $\Delta = \mu(\theta) - \mu(\theta_0)$, $\tilde{\Delta} = \Delta/\sigma_1(1 - \rho^2)^{1/2}$. Evidently M depends on θ through Δ only, as it must, and moreover,

PROPOSITION 3. *If ℓ is as in Proposition 2, then M is a nondecreasing function of $|\Delta|$ for the estimator $\hat{\mu}_c^e$.*

PROOF. It is enough to consider ℓ such that ℓ' exists and is bounded since we can then obtain the general case by approximation. Differentiate M with respect to Δ and interchange limits to get

$$\frac{\partial M}{\partial \tilde{\Delta}}(\theta, \hat{\mu}_c^e)$$

$$= \sigma_1(1 - \rho^2)^{1/2}[\Phi(q - \tilde{\Delta}) - \Phi(-q - \tilde{\Delta})] \int_{-\infty}^{\infty} \tilde{\ell}'(\sigma_0 u - \Delta)\phi(u)\, du \geq 0. \quad \square$$

NOTE. This establishes monotonicity of risk for an arbitrary monotone loss function in the original problem considered by Efron and Morris. Thus

$$m(\hat{\mu}_c^e) = \left(\int_{-\infty}^{\infty} \ell(\sigma_0 u)\phi(u)\, du \right)(2\Phi(q) - 1)$$

$$+ 2 \int_{-\infty}^{\infty} \int_{d}^{\infty} \ell(\sigma_0 u + \sigma_1(1 - \rho^2)^{1/2}(v - q))\phi(v)\phi(u)\, du\, dv \tag{3.5}$$

$$M(\hat{\mu}_c^e) = \int_{-\infty}^{\infty} \ell(\sigma_1(u - (1 - \rho^2)^{1/2}q))\phi(u)\, du. \tag{3.6}$$

TABLE 2
Minimum probabilities of coverage of fixed length intervals centered at E-M estimates: z = 1.960.

q^2	.2	.4	.6	.8
.2	.982	.978	.972	.962
	.932	.936	.941	.945
.4	.988	.985	.977	.965
	.912	.922	.932	.941
.6	.992	.989	.980	.966
	.894	.908	.922	.936
.8	.994	.991	.982	.968
	.874	.894	.913	.932

Note: For each table, the first entry in each box is the minimum probability of coverage on $\mathscr{M}_0$ given by (3.7), the second the minimum on $\mathscr{M}_1$ given by (3.8).

If we specialize to confidence intervals as in (3.1), we obtained minimum probabilities of coverage,

$$1 - m(\hat{\mu}_c^e) = (2\Phi(z/\rho) - 1)(2\Phi(q) - 1) + 2P[-z - (1 - \rho^2)^{1/2}q \le A \le z - (1 - \rho^2)^{1/2}d,\ B \ge q] \tag{3.7}$$

where (A, B) are bivariate standard normal with correlation $(1 - \rho^2)^{1/2}$.

$$1 - M(\hat{\mu}_c^e) = \Phi(z - (1 - \rho^2)^{1/2}q) + \Phi(z + (1 - \rho^2)^{1/2}q) - 1. \tag{3.7a}$$

We give these probabilities for $z = 1.96$ (corresponding to a 95% confidence level) and selected q in Table 2. The results are similar for the 90% and 99% levels. Again the cost benefit structure seems attractive.

Brown (1980) essentially uses pretest estimate based confidence intervals on a data set to illustrate the dangers of the crossover method. If we treat σ_ξ^2, σ_ϵ^2 as equal to their estimated values so that $\rho^2 = .48$ for these data and say select $q = .2$ in Table 1 so that $L(\hat{\mu}_q^e) \cong .10$ we obtain significant results for all ($\mathscr{M}_1$) confidence levels tabled and a fortiori all corresponding ($\mathscr{M}_0$) levels, which is consistent with an analysis of the data based on first period results only.

4. Nested linear models: Quadratic loss in the multivariate case. Suppose $\dim(\mu) = p$. Then $\hat{\mu}_1 \sim \mathscr{N}_p(\mu(\theta), \Sigma_1)$, $\hat{\mu}_0 \sim \mathscr{N}_p(\mu(\theta_0), \Sigma_0)$ where θ_0 is the projection of θ on Θ_0. If $\ell(\theta, d)$ is a function of $\mu(\theta) - d$, invariance considerations lead as before to estimates

$$\hat{\mu} = \hat{\mu}_0 + w(\hat{\Delta}) \tag{4.1}$$

where $\hat{\Delta} = \hat{\mu}_1 - \hat{\mu}_0$ is independent of $\hat{\mu}_0$ with an $\mathscr{N}_p(\Delta, \Sigma_1 - \Sigma_0)$ distribution, $\Delta = \mu(\theta) - \mu(\theta_0)$. Specialize further to,

$$\ell(\mu(\theta) - d) = (\mu(\theta) - d)A(\mu(\theta) - d)^T, \quad A \text{ positive definite.}$$

Then,

$$m(\hat{\mu}) = \operatorname{tr}(A\Sigma_0) + \operatorname{tr}(AE_0(w^T w(\hat{\Delta})))$$

$$M(\hat{\mu}) = \operatorname{tr}(A\Sigma_0) + \sup_\Delta \operatorname{tr}(AE_\Delta((w(\hat{\Delta}) - \Delta)^T(w(\hat{\Delta}) - \Delta)))$$

and in minimizing $m(\hat{\mu})$ subject to a bound on M we need only consider the second terms above. That is, it is enough to consider the special case $r = 0$, $s = p$. Exact solution is impossible. However we can attempt approximations. We can always reduce to the case $A = \| a_i^2 \delta_{ij} \|$ diagonal, $\Sigma_1 - \Sigma_0$ the identity. That is, we observe $X = \Delta + Z$, $Z \sim \mathcal{N}_p(0, I)$, $\Delta = (\Delta_1, \cdots, \Delta_p)$. The risk of an estimate $w = (w_1, \cdots, w_p) = x - \Psi(x)$ is

$$M(\Delta, w) = \sum_{i=1}^p a_i^2 E(w_i(X) - \Delta_i)^2$$

$$= \sum_{i=1}^p a_i^2 + E\left\{\sum_{i=1}^p a_i^2(\psi_i^2(X) - 2\frac{\partial \psi_i}{\partial x_i}(X))\right\}$$

under mild conditions. If π is a Bayes prior distribution with Bayes risk $r(\pi)$, Bayes estimate w_π, and marginal density g_π then

$$w_\pi(x) = x + \nabla \log g_\pi(x)$$

$$r(\pi) = \sum_{i=1}^p a_i^2 - I(G_\pi) \tag{4.2}$$

where ∇ is the gradient $((\partial/\partial x_1), \cdots, (\partial/\partial x_p))$

$$I(G) = \sum a_i^2 \int \left(\frac{\partial g}{\partial x_i}(x)\right)^2 g^{-1}(x)\, dx \tag{4.4}$$

(and $= \infty$ if the quantity on the right is undefined). Again the original problem is to minimize $I(G)$ over $\mathcal{G}_0$ and approximation (I) is to minimize over $\mathcal{G}_1$ (with Φ now the p-variate standard normal). By the argument given for one dimension, this yields the same solution as does approximation (II) which minimizes $M(0, w)$ subject to a bound on $[\sum a_i^2(\psi_i^2(x) - 2(\partial \psi_i/\partial x_i)(x))] \leq q^2$, for suitable q^2. Unfortunately this approximation is also difficult to compute (but see Chen, 1983), unless all the a_i^2 are equal, say to $1/p$. In this case the solution is given for $p = 3$ by Huber (1977) and for general p by Berger (1981), Theorem 3. Here

$$w(x) = 0 \qquad |x| \leq q$$

$$= \rho(|x|^2)x, \quad |x| > q \tag{4.5}$$

with ρ a ratio of Bessel functions with parameters depending on p and scale depending on q^2 and $\rho(|q|^2) = 0$. For $p \geq 3$ we can take $q = 0$, i.e., find the minimax estimate in this class which minimizes $M(0, w)$. The answer is the Stein positive part estimate, $q^2 = 2(p - 2)$,

$$\rho(r) = \left(1 - \frac{2(p-2)}{r}\right).$$

As Berger points out, $M(0, w)$ for this estimate drops very sharply from .296 when $p = 3$ to .07 for $p = 5$. Although ths solution is appealing we face the usual ambiguities of the multivariate case. For $p \geq 3$ we could, for instance, also reduce $M(\theta, \hat{\mu})$ for $|\mu(\theta_0)|$ small by applying Steinian shrinking to $\hat{\mu}_0$. Moreover, the effect of the choice of loss function on the suitability of the estimate is difficult to make precise.

For $a_i^2 = 1/p$, it seems reasonable to consider average squared bias and,

$$\text{minimize } E\{\textstyle\sum_{i=1}^p w_i^2(X)\} \text{ subject to } p^{-1} \textstyle\sum_{i=1}^p \psi_i^2 \leq q^2.$$

The solution is as in the one-dimensional case,

$$\begin{aligned} \tilde{w}(x) &= 0, && |x|^2 \leq q^2 \\ &= (1 - (q/|x|)x, && |x|^2 > q^2. \end{aligned} \tag{4.6}$$

If we define M as in the introduction then for fixed $M(w) = 1 + q^2$, estimate (4.5) improves (4.6) at $\Delta = 0$. This follows since the estimates (4.6) also have, if $\tilde{\psi}$ corresponds to $\tilde{w}$,

$$M(\tilde{w}) = 1 + p^{-1} \sup_x \sum \left[\tilde{\psi}_i^2(x) - 2 \frac{\partial \tilde{\psi}_i}{\partial x_i}(x)\right] = 1 + q^2. \tag{4.7}$$

The difference is substantial and despite its attractive feature of computability for more general loss functions, this analogue to Hampel robustness seems unsatisfactory for this application.

5. Nested parametric models: Asymptotics. We extend the approaches of Sections 3 and 4 to general nested parametric models by using large sample approximations. Related results are given by Sen (1979) for pretesting estimates. For simplicity we consider estimation of $\mu(\theta)$ where μ is a smooth real-valued function of θ.

Suppose Θ_1, Θ_0 are as we described previously, respectively an open subset of R^s and a (locally) r-dimensional submanifold of Θ_1. Suppose that the models are approximable locally in the sense of Le Cam, to scale $n^{-1/2}$, by nested Gaussian linear models and admit estimates $\hat{\theta}_{0n}$, $\hat{\theta}_{1n}$ (typically M.L.E.'s under $\mathcal{M}_0$, $\mathcal{M}_1$) which are efficient and locally sufficient uniformly on compact subsets of Θ_0, Θ_1 respectively. See Le Cam (1969), Chapters 3, 4 for a detailed description of these concepts and suitable conditions.

Fix $\theta_0 \in \Theta_0$ and reparametrize Θ by $\theta_0 + an^{-1/2}$ in Pitman form. Locally Θ permits arbitrary a while Θ_0 specifies $a \in V(\theta_0)$ an r-dimensional subspace of R^s. Also $\mu(\theta_0 + an^{-1/2}) = \mu(\theta_0) + a\dot{\mu}(\theta_0) + O(n^{-1/2})$ where $\dot{\mu}$ is the differential of μ. Finally, $n^{1/2}\{(\hat{\theta}_{0n} - \theta_0), (\hat{\theta}_{1n} - \theta_0)\}$ is asymptotically normal uniformly on compact sets of (θ_0, a) with means $(a\Pi(\theta_0), a)$ and covariance matrix $\Sigma(\theta_0)$ where $\Pi(\theta_0)$ is the projection matrix of $V(\theta_0)$.

These approximations suggest that in order to minimize maximum M.S.E. of estimates of $\mu(\theta)$ over large Pitman neighbourhoods of θ_0 in Θ_0, subject to a bound on the maximum M.S.E. over large Pitman neighbourhoods of θ_0 in Θ, we

use asymptotically equivariant estimates as follows. Let

$$\hat{\Delta}_n = \mu(\hat{\theta}_{1n}) - \mu(\hat{\theta}_{0n}), \quad \sigma^2_{\Delta}(\theta_0) = \dot{\mu}^T(\theta_0)\begin{pmatrix}-1\\1\end{pmatrix}\Sigma(\theta_0)\begin{pmatrix}-1\\1\end{pmatrix}^T\dot{\mu}(\theta_0)$$

denote the asymptotic variance of $n^{1/2}\hat{\Delta}_n$ under $\theta_0 + an^{-1/2}$,

$$\Delta = a(I - \Pi(\theta_0))\dot{\mu}(\theta_0)$$

denote its asymptotic mean, and $\hat{\sigma}_{\Delta,n}$ be a consistent estimate of σ_{Δ}, e.g.

$$\hat{\sigma}_{\Delta_n} = \sigma_{\Delta}(\hat{\theta}_{1n}).$$

Then, an asymptotically equivariant estimate is one of the form

$$\mu(\hat{\theta}_{0n}) + \hat{\sigma}_{\Delta_n} w(\hat{\Delta}_n/\hat{\sigma}_{\Delta_n}) \tag{5.1}$$

and n times the M.S.E. at $\theta_0 + an^{-1/2}$ of such an estimate is (under mild conditions) approximated by

$$M(\theta_0, a, w) = \sigma_1^2(\theta_0)(\rho^2(\theta_0) + (1 - \rho^2(\theta_0))E(w(Z + \Delta) - \Delta)^2) \tag{5.2}$$

where $\sigma_i^2(\theta_0)$ is the asymptotic variance of $n^{1/2}\mu(\hat{\theta}_{in})$ and,

$$\rho^2(\theta_0) = \sigma_0^2(\theta_0)/\sigma_1^2(\theta_0).$$

From (5.2), given a bound $1/c$ on $\sup_a M(\theta_0, a, w)/\sigma_1^2(\theta_0)$, we minimize $\sup_{a \in V(\theta_0)} M(\theta_0, a, w)$ by taking $w = w_q^*$. As in Section 2, we obtain reasonable results by taking $w = e_q$, with q related to c via (2.6) and $\rho = \rho(\sigma_0)$. The asymptotic sufficiency and efficiency properties of $\hat{\theta}_{in}$, $i = 0, 1$, enable us to formulate asymptotic optimality and near optimality properties of these estimates in the class of all estimates. For simplicity, we omit these.

We give a simple illustration of this approach by applying it to the case of nested linear models with $\Sigma = \sigma^2 I$, σ^2 unknown, and μ a linear function of the mean θ. Then our prescription is merely to replace $\sigma^2_{\hat{\Delta}}$ in (2.8) by

$$\hat{\sigma}^2_{\hat{\Delta}} = \tau^2[\sigma^{-2}(\sigma_1^2 - \sigma_0^2)] \tag{5.2a}$$

where $\tau^2 = \| Y - \hat{\theta}_1 \|^2/(n - 2)$, the usual estimate of σ^2. The ratio in parentheses in (5.2) depends on the models only. For general Σ, given a consistent estimate $\hat{\Sigma}$ of Σ, we can calculate $\hat{\theta}_0$, $\hat{\theta}_1$ by generalized least squares using $\hat{\Sigma}$ and then plug $\hat{\Sigma}$ into $\sigma^2_{\hat{\Delta}}$ appropriately calculated.

As a second illustration, consider pooling two binomial samples. Let $\hat{p}_i = N_i/n_i$, $i = 1, 2$, where N_i is $\text{bin}(n_i, p_i)$, $0 < p_i < 1$, $n_1/n_2 = \lambda$, $0 < \lambda < 1$. We want to estimate p_1. $\mathscr{M}_0$ prescribes $p_1 = p_2$. So, if we use $n = n_1 + n_2$ as an index,

$$\hat{\theta}_{1n} = (\hat{p}_1, \hat{p}_2), \quad \hat{\theta}_{0n} = (\hat{p}, \hat{p})$$

where

$$\hat{p} = (N_1 + N_2)/n = (\lambda\hat{p}_1 + \hat{p}_2)/(1 + \lambda).$$

If $\theta = (p, p)$,

$$\sigma_0^2(\theta) = p(1 - p), \quad \sigma_1^2(\theta) = p(1 - p)\frac{(1 + \lambda)}{\lambda}, \quad \rho^2(\theta) = \frac{\lambda}{1 + \lambda}.$$

Then if $\hat{r}_i = 1 - \hat{p}_i$, $i = 1, 2$, putting $w = e_q$ in (5.1),

$$\hat{\mu}_c^e = \hat{p} + \left(\frac{\hat{p}_1\hat{r}_1}{\lambda n}\right)^{1/2} e_q\left(\frac{(\lambda n)^{1/2}(\hat{p}_1 - \hat{p}_2)}{(1 + \lambda)(\hat{p}_1\hat{r}_1)^{1/2}}\right)$$

or

$$\begin{aligned} \hat{\mu}_c^e &= \hat{p} \text{ if } |(\lambda n)^{1/2}(\hat{p}_1 - \hat{p}_2)/(\hat{p}_1\hat{r}_1)^{1/2}(1 + \lambda)| \leq q \\ &= \hat{p}_1 - q \operatorname{sgn}(\hat{p}_1 - \hat{p}_2)(\lambda n)^{-1/2}(\hat{p}_1\hat{r}_1)^{1/2} \text{ otherwise.} \end{aligned} \tag{5.3}$$

This yields, by (5.1), for quadratic loss, a relative loss in risk of

$$\sigma_1^{-2}(\theta) \sup_a M(\theta, a, w) - 1 = q^2/(1 + \lambda) \tag{5.4}$$

while the relative savings in risk are

$$1 - \sigma_1^{-2}(\theta) \sup_{V(\theta)} M(\theta, a, w) = (1 - m_0(e_q))/(1 + \lambda). \tag{5.5}$$

Clearly we can extend this approach to confidence intervals and the p-variate case. What we are doing should be clear from the examples. We essentially interpolate between the M.L.E.'s of $\mu(\theta)$ under $\mathscr{M}_0$ and $\mathscr{M}_1$ using weights which are functions of Wald's form of the test statistic for H: $\mu(\theta) \in \mu(\Theta_0)$ vs. K: $\mu(\theta) \in \mu(\Theta_1)$.

When we consider the limit of ordinary risks $M(\theta, \{\delta_n\})$ we find that procedures (5.1) generally exhibit a discontinuity at points of Θ_0, i.e. convergence of the risk is not uniform. This is reminiscent of Hodges' example of a super efficient estimate which is essentially a pretest estimate corresponding to a sequence of levels tending to 0. However the Hodges procedure has infinite relative loss in risk whereas we propose to pay a small price in the relative loss in exchange for improved behaviour on Θ_0.

6. Conclusions: Open questions.

(1) We have applied robustness ideas to derive what we judge are useful biased estimates in the estimation of single parameters under a simple model $\mathscr{M}_0$ when we want to guard against deviations towards a larger model $\mathscr{M}_1$. The solutions involve both an approximation to the optimality principle and in general a large sample approximation. Tables 1 and 2 show that the first approximation is not serious for quadratic loss and the solutions give reasonable confidence intervals. The adequacy of the large sample approximation remains to be assessed in different models by obtaining approximate solutions of the Berger-Bickel type to the exact model, where possible.

(2) In the p-variate case, even approximate solutions can only be calculated in special cases and their structure depends on the loss function. It may be appropriate to apply Steinian "pulling in" within the simple model towards a yet simpler model as well as further "pulling in" towards the simple model itself. Alternatively, if we do not believe that losses from errors made in estimation of different components of μ should be combined it may still make sense to apply pulling in towards $\mathscr{M}_0$ on each component individually.

(3) This approach is applicable, in principle, to large sample problems when $\mathscr{M}_1$ is nonparametric. For example, suppose we want to estimate features of distributions such as medians, means, or even the whole distribution or its density. Our approach suggests reasonable ways of interpolating between estimates based on parametric assumptions and nonparametric estimates.

(4) Typically we have more than one simple candidate model $\mathscr{M}_0$. It would be very interesting to obtain reasonable estimates of $\mu(\theta)$ which do well at each member of a set of simple models while still performing adequately at a super model $\mathscr{M}_1$.

(5) This work is closely connected with the recent studies of Marazzi (1980) and Berger (1982) on robust Bayesian inference. See also the thesis of Y. Ritov (1982) and Masreliez and Martin (1977). Problem (P) is precisely of that form, minimize the Bayes risk for a prior degenerate at $\{0\}$ subject to a bound on the maximum risk—interpreted as the worst that misspecification of the prior can do. On the other hand, if in our original problem we replace the maximum risk over $\mathscr{M}_0$ by an average, we are again in the robust Bayesian framework. We prefer not to try to specify prior distributions. Our point is just that a possibly naive belief in a simpler model can be catered to with reasonable safety.

Acknowledgement. I am grateful to B. Efron and P. Huber for helpful conversations and A. Marazzi for the calculation of the lower bounds.

REFERENCES

ANDERSON, T. W. (1955). The integral of a symmetric unimodal function over a symmetric convex set. *Proc. Amer. Math. Soc.* **6** 170–176.

BANCROFT, T. A. and HAN, C. P. (1977). Inference based on conditional specification: a note and a bibliography. *Internat. Statist. Rev.* **45** 117–128.

BERGER, J. (1982). Estimation in continuous exponential families: Bayesian estimation subject to risk restrictions and inadmissibility results. *Statistical Decision Theory and Related Topics III.* S. S. Gupta and J. Berger, eds. Academic, New York.

BICKEL, P. J. (1983). Minimax estimation of the mean of a normal distribution subject to doing well at a point. *Recent Advances in Statistics, Festschrift for H. Chernoff* 511–528. H. Rizvi and D. Siegmund, eds. Academic, New York.

BROWN, B. W. (1980). The crossover experiment for clinical trials. *Biometrics* **36** 69–80.

CHEN, S. Y. (1983). Restricted risk Bayes estimation for the mean of a multivariate normal distribution. Tech. Report 83-33, Purdue University.

HODGES, J. L. JR. and LEHMANN, E. L. (1952). The use of previous experience in reaching statistical decisions. *Ann. Math. Statist.* **23** 396–407.

HUBER, P. J. (1977). Robust covariances. *Statistical Decision Theory and Related Topics III.* S. S. Gupta and J. Berger, Eds. Academic, New York.

KIEFER, J. (1957). Invariance, minimax sequential estimation and continuous time processes. *Ann. Math. Statist.* **28** 573–601.

LECAM, L. (1969). Théorie asymptotique de la décision statistique. Presses de l'Université de Montréal.

MARAZZI, A. (1980). Robust Bayesian estimation for the linear model. Technical Report No. 27. E.T.H. Zurich.

MARAZZI, A. (1982). On constrained minimization of the Bayes risk for the linear model. Technical Report No. 34. E.T.H. Zurich.

MASRELIEZ, C. J. and MARTIN, R. D. (1977). Robust Bayesian estimation for the linear model and robustifying the Kalman filter. *I.E.E.E. Trans. Automat. Control* **AC-22** June 1977. 361–371.

MORRIS, C., RADHAKRISHNAN, R. and SCLOVE, S. L. (1972). Nonoptimality of preliminary test estimators for the mean of a multivariate normal distribution. *Ann. Math. Statist.* **43** 1481–1490.

MOSTELLER, F. (1948). On pooling data. *J. Amer. Statist. Assoc.* **43** 231–242.

RITOV, Y. (1982). Robust quasi Bayesian inference. Thesis, Hebrew University, Jerusalem.

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

SCHEFFÉ, H. (1959). *The Analysis of Variance*. Wiley, New York.

SEN, P. K. (1979). Asymptotic properties of maximum likelihood estimators based on conditional specification. *Ann. Statist.* **7** 1019–1033.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720