M-ESTIMATION FOR DISCRETE DATA:
ASYMPTOTIC DISTRIBUTION THEORY AND IMPLICATIONS

by

Douglas G. Simpson[1]
Department of Statistics
University of Illinois
Urbana, Illinois 61801

Raymond J. Carroll[2]
Department of Statistics
University of North Carolina
Chapel Hill, North Carolina 27514

David Ruppert[1]
Department of Statistics
University of North Carolina
Chapel Hill, North Carolina 27514

[1] Research supported in part by NSF Grant DMS 8400602.

[2] Research supported by the AFOSR Contract No. F49620-85-C-0144.

Abstract

The asymptotic distribution of an M-estimator is studied when the underlying distribution is discrete. Asymptotic normality is shown to hold quite generally within the assumed parametric family. When the specification of the model is inexact, however, it is demonstrated that an M-estimator with a non-smooth score function, e.g. a Huber estimator, has a non-normal limiting distribution at certain distributions, resulting in unstable inference in the neighborhood of such distributions. Consequently, smooth score functions are proposed for discrete data.

AMS 1970 Subject Classification: 62E20, 62F10, 62G35.

Keywords and Phrases: Robust estimation, M-estimator, discrete parametric model, smooth score function.

1. Introduction

M-estimation, originally proposed by Huber (1964) to estimate a location parameter robustly, has since been applied successfully to a variety of estimation problems where stability of the estimates is a concern. There is, for instance, a substantial body of literature on M-estimation for regression models; see Krasker and Welsch (1982) for a recent review. For further references on M-estimation, see Huber (1981).

Much of the popularity of M-estimators can be attributed to their flexibility. Desired properties of an M-estimator, such as relative insensitivity to or rejection of extremely outlying data points, can be specified in a direct way since the influence function of an M-estimator is proportional to its score function; see Hampel (1974) or Huber (1981) for details.

Surprisingly, M-estimation for discrete data seems to have received little attention. Discrete data are no less prone than continuous measurements to outliers or partial deviations from an otherwise reasonable model; see, for instance, data from mutation research presented in Simpson (1985). This paper investigates some aspects of M-estimation for discrete data.

A useful optimality theory has been developed by Hampel (1968, 1974) for robust M-estimation of a univariate parameter. His general prescription facilitates the construction of robust M-estimators with nearly optimum efficiency at a specified model. Proposals for robust estimation of the binomial and Poisson parameters, for instance, can be found in Hampel (1968). Hampel's univariate theory is briefly reviewed in Section 2. Extensions of this optimality theory to certain multivariate models are discussed in Krasker (1980), Krasker and Welsch (1982), Ruppert (1985), and Stefanski, Carroll, and Ruppert (1985).

The score function for Hampel's optimal M-estimator is not smooth, that is, it is not everywhere differentiable. This can lead to complications in the asymptotic theory when the data are discrete. For instance, Huber (1981, p. 51) considers the case where the underlying distribution is a mixture of a smooth distribution and a point mass. He observes that if the point mass is at a discontinuity of the derivative of the score function. then an M-estimate for location has a non-normal limiting distribution. Along the same lines, Hampel (1968, p. 97) notes that the optimal M-estimate for the Poisson parameter is asymptotically normal at the Poisson distribution, provided the truncation points of the score function are not integers. He conjectures that "under *any* Poisson distribution, it is asymptotically normal (with the usual variance); however, this remains to be seen."

This paper provides extensions to the asymptotic distribution theory of M-estimators especially relevant to discrete data, although Theorem 1 is somewhat broader in scope. The main results are given in Section 3. Among the applications of the theory are a more complete account of the asymptotics of the Huber M-estimate for location and a proof of Hampel's conjecture. Aside from providing a more complete asymptotic theory for M-estimation, the results have implications for choosing a score function when the data are discrete. These are discussed in the final sections. In particular, smooth score functions are proposed.

2. Parametric M-estimation: Definitions, optimality and examples

Suppose $X_1, X_2, \ldots$ are independent observations, each thought to have distribution function (d.f.) F , where θ belongs to a parameter set Θ; here Θ is a subset of R^d, $d \geq 1$. Define

(2.1) $$M(t;\psi,F) = \int \psi(\cdot,t)dF,$$

where F is a d.f. on R^1, $\psi(\cdot,\cdot)$ is a measurable real-valued function on $R^1 \times \Theta$, and $t \in \Theta$. Then T_n is an M-estimator for θ, based on a sample of size n, if it solves an equation of the form

(2.2) $$M(T_n;\psi,F_n) = 0,$$

where F_n is the empirical d.f.. The standard requirement

(2.3) $$M(\theta;\psi,F_\theta) = 0, \ \theta \in \Theta,$$

and additional regularity conditions ensure that T_n consistently estimates θ when the model is correct.

Suppose now that $\Theta \subset R^1$. The influence function at F_θ of an M-estimator for θ has the form

$$\Omega(x,\theta) = \frac{\psi(x,\theta)}{-\int\{\frac{d}{d\theta}\psi(\cdot,\theta)\}dF_\theta},$$

provided this exists. Assume F_θ has a density f_θ with respect to a suitable measure, and assume the parametrization is smooth. Letting $\ell(x,\theta) = \frac{d}{d\theta}\log f_\theta(x)$, the optimal score according to Hampel's criterion has the form

(2.4) $$\psi_{c(\theta)}(\ell(x,\theta) - \alpha(\theta)),$$

where

$$\psi_c(u) = \begin{cases} u, & |u| \leq c \\ c\,\text{sign}\,(u), & |u| > c, \end{cases}$$

and α is defined implicitly by (2.3). This estimator cannot be dominated by any M-estimator simultaneously with respect to the asymptotic variance and the bound on the influence function at F_θ. This is assuming, of course, that the estimator is asymptotically normal at F_θ.

The truncation point $c(\theta)$ determines the bounds on $\Omega(\cdot,\theta)$ and hence the robustness of the estimator to outlying data points. Observe that the maximum likelihood estimator has the form (2.4) with $c(\theta) \equiv \infty$ and $\alpha(\theta) \equiv 0$.

Two examples given in Hampel (1968) will be of special interest here.

Example 1. If F_θ is the normal d.f. with mean θ and unit variance then $\ell(x,\theta) = x - \theta$. By symmetry $\alpha(\theta) \equiv 0$, and constant variance suggests setting $c(\theta) \equiv c$. The resulting estimator, with score $\psi_c(x-\theta)$, is the Huber (1964) M-estimator for location.

Example 2. If F_θ is the Poisson d.f., with density $f_\theta(x) = e^{-\theta}\theta^x/x!$ on $x = 0,1,2,\dots$, then $\ell(x,\theta) = x\theta^{-1} - 1$. Hampel (1968, p. 96) suggests taking $c(\theta) = c\theta^{-1/2}$ on the grounds that $\ell(x,\theta)$ has standard deviation $\theta^{-1/2}$. For this choice (2.4) is equivalent to $\psi_c(x\theta^{-1/2} - \theta^{1/2} - \alpha(\theta))$. The version

$$(2.5) \qquad \psi_c(x\theta^{-1/2} - \beta(\theta)),$$

where $\beta(\theta) = \theta^{1/2} + \alpha(\theta)$ is defined by (2.3), is slightly more convenient.

3. Extended asymptotic distribution theory

Conditions for consistency of an M-estimator can be found in Huber (1964, 1967, 1981). Since the smoothness plays no role in the consistency proofs, consistency will usually be assumed here.

Huber (1981, theorems 3.2.4 and 6.3.1) shows under quite general conditions that if $T_n \to \theta = T(G)$ in probability as $n \to \infty$ then

$$(3.1) \qquad -n^{1/2}M(T_n;\psi,G) = n^{-1/2}\sum_{i=1}^{n}\psi(x_i,\theta) + o_p(1),$$

where M is given by (2.1). In particular, ψ need not be differentiable; monotonicity or Lipschitz continuity conditions are sufficient. That T_n is asymptotically normal follows immediately from (3.1) provided $M(t;\psi,G)$ has a nonzero derivative at θ and $0 < \int\psi^2(\cdot,\theta)dG < \infty$; see Corollary 6.3.2 of Huber (1981). For stronger almost sure representations for T_n under stronger conditions, see Carroll (1978a, 1978b).

To avoid Lipschitz conditions for score functions like (2.5) that have implicitly defined centering parameters, the following lemma is useful. The proof is contained in the proof of Theorem 2.2 of Boos and Serfling (1980). Denote by $||\cdot||_v$ the total variation norm, given by

$$||h||_v = \lim\sup\sum_{i=1}^{k}|h(x_i) - h(x_{i-1})|,$$

where the supremum is over partitions $a = x_0 < x_1 < \ldots < x_k = b$ of $[a,b]$, and the limit is as $a \to -\infty$, $b \to \infty$.

Lemma 1. Let $X_1, X_2, \ldots$ be independent, each with d.f. G, and let $\theta = T(G)$. Suppose $\theta(x,t)$ is continuous in x for $t \in \theta \subset R^d$ and

$$\lim_{t\to\theta}||\psi(\cdot,t) - \psi(\cdot,\theta)||_v = 0.$$

If $T_n \to \theta$ in probability as $n \to \infty$, then (3.1) holds.

Remark. The score functions of Examples 1 and 2 are continuous in total variation. For the former see Boos and Serfling (1980). For the latter, see Simpson (1985).

When the underlying distribution is discrete, the set of points where ψ fails to have a derivative has positive probability for certain parameter values. In light of (3.1), it is natural to ask whether M can have a derivative at such parameter values, i.e., whether T_n can be asymptotically normal.

The following theorem addresses this question. For $\theta \in \Theta \subset R^d$, F_θ is assumed to have a density $f_\theta = f(\cdot,\theta)$ with respect to a σ-finite measure μ, and $\psi_\theta = \psi(\cdot,\theta)$ is measurable for each θ. Let $||\cdot||$ denote any norm on R^d equivalent to the Euclidean norm. Some regularity conditions are needed:

(A1) There are measurable functions $\omega_t = \omega(\cdot,t)$ and $g_t = g(\cdot,t)$ for which $\int \omega_t f_t d\mu$, $\int |\psi_t| g_t d\mu$ and $\int \omega_t g_t d\mu$ are finite and, for some $\delta > 0$,

(i) $|f_s - f_t| \leq ||s - t|| g_t$, and

(ii) $|\psi_s| \leq \omega_t$

almost everywhere [μ] (a.e.) when $||s - t|| \leq \delta$;

(A2) There is a measurable R^d-valued function $\dot{f}_t = \dot{f}(\cdot,t)$ such that

$$|f_s - f_t - (s - t)^T \dot{f}_t| = o(||s - t||) \text{ a.e.};$$

(A3) $\psi_s \to \psi_t$ a.e. as $s \to t$.

<u>Theorem 1</u>. If for each $t \in \Theta$ (A1)-(A3) hold and

(3.2) $$M(t;F_t) = 0$$

then

(3.3) $$D_s M(s;F_t)|_{s=t} = -\int \psi_t \dot{f}_t d\mu,$$

where D_s denotes vector differentiation, and where the dependence of M on ψ has been suppressed.

Proof. For $s,t \in \Theta$

$$0 = M(s;F_t) - M(t;F_t) + M(s;F_s) - M(s;F_t)$$

$$= M(s;F_t) - M(t;F_t)$$

$$+ M(t;F_s) - M(t;F_t) + R_t(s), \tag{3.4}$$

where $$R_t(s) = \int(\psi_s - \psi_t)(f_s - f_t)d\mu$$

and (3.2) was used. The integrand of $R_t(s)$ is dominated in absolute value by $2\|s-t\|\omega_t g_t$ on $\|s-t\| \le \delta$ because of (A1). Hence, by (A3) and Dominated Convergence,

$$R_t(s) = o(\|s-t\|). \tag{3.5}$$

Similarly, (A2) and Dominated Convergence imply

$$|M(t;F_s) - M(t;F_t) - (s-t)^T\int\psi_t\dot{f}_t d\mu|$$

$$\le \int|\psi_t|\ |f_s - f_t - (s-t)^T\dot{f}_t|d\mu$$

$$= o(\|s-t\|) \quad \text{as } s \to t,$$

since the integrand is dominated by $2\|s-t\|\ |\psi_t|g_t$ on $\|s-t\| \le \delta$. From (3.4) to (3.6) conclude

$$|M(s;F_t) - M(t;F_t) + (s-t)^T\int\psi_t\dot{f}_t d\mu| = o(\|s-t\|).$$

Hence $D_s M(s;F_t)$ exists at t and is given by (3.3).

Remarks. 1. Note that ψ_t need not be differentiable.

2. When $\psi_t = \ell_t = \dot{f}_t/f_t$, (3.3) generalizes the usual information identity.

3. Huber (1981, p. 51) observes a special case, namely (3.3) holds when μ is Lebesgue measure, $\psi(x,t) = \psi(x - t)$, where $\psi(\cdot)$ is skew-symmetric about zero, and $f(x,t) = f(x - t)$, where $f(\cdot)$ is differentiable and symmetric about zero.

4. Equation (3.3), when it holds, also guarantees that the influence function at the model, given by

$$\{D_s M(s;F_t)|_{s=t}\}^{-1}\psi(x,t)$$

is defined for each $t \in \Theta$, provided that $\int \psi_t \dot{f}_t d\mu \neq 0$.

Example 2 (continued) Suppose $f(x,t) = e^{-t}t^x/x!$ on $\{0,1,2,\ldots\}$, $t > 0$. Recall that the optimal M-estimator has the score $\psi(x,t) = \psi_c(xt^{-1/2} - \beta)$. This estimator is known to be asymptotically normal at the Poisson distribution when t is in one of the open intervals where neither of the truncation points $t^{1/2}(\beta \pm c)$ is an integer; see Hampel (1968, p. 97).

To show that it is asymptotically normal at every Poisson distribution, as conjectured by Hampel, first use Theorem 1 with $g(x,t) = e^{2\delta}f(x - 1,t + \delta) + \delta^{-1}(e^{\delta} - 1 - \delta)f(x,t)$, $\omega(x,t) \equiv c$ and $\dot{f}(x,t) = f(x - 1,t) - f(x,t)$. Note that $c \geq 1$ is sufficient for β to be continuous, and hence for (A3); see Simpson (1985).

Since Lemma 1 applies and $0 < \int \psi_t^2 f_t d\mu \leq c^2$ for $c \geq 1$, it follows that the estimator is asymptotically normal at every Poisson distribution if it is consistent. For consistency see Hampel (1968, p. 96) and Theorem 2 of Huber (1967).

In Theorem 1, (3.2) allows smoothness of the parametrization to be substituted for smoothness of ψ within the assumed parametric model, so that the estimator is asymptotically normal under further conditions. If the specification of the model is inexact (as is often suspected), no result like (3.3) is available. In certain cases, it is still possible to obtain the limiting distribution of T_n from (3.1).

Assume for simplicity that Θ is an open subset of the real line. The score functionsused for robust estimation are generally at least piece-wise differentiable. The one-sided derivatives of $M(t;G)$ will then exist, in general, even when M fails to be differentiable. Write

$$m(t;G) = \frac{d}{dt}M(t;G)$$

when the derivative exists. By a well-known result from calculus, if $m(\theta-;G)$ and $m(\theta+;G)$ exist, they are equal to the corresponding one-sided derivatives of $M(t;G)$ at θ; see, e.g., Franklin (1940, p. 118).

Theorem 2. Suppose for some θ interior to Θ that $M(\theta;G) = 0$, and let T_n be a zero of $M(t;F_n)$, $n = 1,2,\ldots$, where F_n is the empirical d.f. Assume the following:

(B1) $M(\theta-;G)$ and $m(\theta+;G)$ exist finitely and are non-zero and of the same sign;

(B2) $0 < \sigma < \infty$, where $\sigma^2 = \int \psi_\theta^2 dG$;

(B3) $T_n \to \theta$ in probability as $n \to \infty$, and (3.1) holds.

Then

(3.7) $$\lim_{n\to\infty} \sup_{-\infty<z<\infty} |\mathrm{pr}\{n^{1/2}(T_n - \theta) \le z\} - H(z)| = 0,$$

where $$H(z) = \begin{cases} \phi(|m(\theta+;G)|z/\sigma), & z \ge 0 \\ \phi(|m(\theta-;G)|z/\sigma), & z \le 0, \end{cases}$$

and ϕ is the standard normal d.f.

Remarks. 1. Huber (1964, p. 78) alludes to a similar result for a location estimator.

2. The requirement that $m(\theta \pm; G)$ have the same sign is actually implied by the remaining conditions. If the one-sided derivatives were to have opposite signs, $M(t;G)$ would not change signs in a neighborhood of θ and (3.1) would not hold.

The proof of Theorem 2 is deferred to the Appendix.

Example 1 (continued) Recall that the Huber M-estimator for location has the score $\psi(x,t) = \psi_c(x-t)$. For any d.f. G, $M(-\infty;G) = c = -M(\infty;G)$, and $M(t;G)$ is continuous in t so it has a zero θ. Assume $\theta = 0$. This is unique if $G(c-) > G(-c+)$, in which case $T_n \to 0$ in probability by Proposition 2.2.1 of Huber (1981). Since ψ_c is continuous in total variation, (3.1) holds by Lemma 1. Letting $\dot{\psi}(x,t) = d/dt\, \psi_c(x-t) = -\psi_c'(x-t)$ if it exists, observe that $-\dot{\psi}(x,t-) = I(-c \le x-t < c)$ and $-\dot{\psi}(x,t+) = I(-c < x-t \le c)$, where $I(\cdot)$ denotes the indicator function. Bounded convergence yields $-m(0-;G) = G(c-) - G(-c-)$ and $-m(0+;G) = G(c+) - G(-c+)$. Hence, by Theorem 2, $n^{1/2}T_n$ is asymptotically normal if $G(c+) - G(c-) = G(-c+) - G(-c-)$; otherwise, it has a limiting distribution consisting of the left and right halves of two normal distributions with different variances (cf. Huber (1981, p. 51)).

4. A counterexample

It is instructive to examine the extent of the non-normality that occurs in a specific example. Consider again the optimal M-estimator for the Poisson parameter. The score function is

$$\psi(x,t) = \psi_c(xt^{-1/2} - \beta) = \begin{cases} -c, & x \le \ell(t) \\ xt^{-1/2} - \beta, & \ell(t) < x < h(t) \\ c, & h(t) \le x, \end{cases}$$

where $\ell(t) = t^{1/2}(\beta(t) - c)$ and $h(t) = t^{1/2}(\beta(t) + c)$.

Let G be the actual d.f. and let $\theta = T(G)$. The simplest situation is when θ is small. Assume henceforth that $\ell(\theta) < 0 < h(\theta) = 1$. Calculation yields $\beta(t) = c(c^t - 1)$ for $\ell(t) < 0$, $0 < h(t) \le 1$, and $\beta(t) = c\{e^t(1+t)^{-1} - 1\} + t^{1/2}(1+t)^{-1}$ for $\ell(t) < 0$, $1 \le h(t) \le 2$. Since β is continuous, equating the two expressions at θ gives

$$\theta^{1/2}e^{\theta} = c^{-1}. \tag{4.1}$$

The one-sided derivatives of β at θ are $\beta'(\theta-) = ce^{\theta}$ and $\beta'(\theta+) = \frac{1}{2}ce^{\theta}(1+\theta)^{-2}$, where (4.1) was used. Note that β is strictly increasing at θ. Since $\psi_c'(c-) = 1$ and $\psi_c'(c+) = 0$,

$$-\dot{\psi}(x,\theta-) = \begin{cases} ce^{\theta}, & x = 0 \\ 0, & x = 1,2,\ldots \end{cases} \tag{4.2}$$

and

$$-\dot{\psi}(x,\theta+) = \begin{cases} \frac{1}{2}ce^{\theta}\,(1+\theta)^{-2}, & x = 0 \\ \frac{1}{2}ce^{\theta}\,\{\theta^{-1} + (1+\theta)^{-1}\}, & x = 1 \\ 0, & x = 2,3,\ldots \end{cases}$$

Suppose G is a mixture of a Poisson distribution F_t and a point mass at an integer z, i.e., $G = (1-\varepsilon)F_t + \varepsilon\delta_z$. Assume $z > h(t)$ so $\dot{\psi}(z,\theta\underline{+}) = 0$. From (4.2) and (4.3)

$$\frac{m(\theta+;G)}{m(\theta-;G)} = \frac{1}{2}\left(\frac{t}{\theta} + \frac{1+t}{1+\theta}\right), \tag{4.4}$$

where $m(\theta-;G) = -ce^{\theta-t}(1-\varepsilon)$. The ratio (4.4) is unity only when $t = \theta$, which corresponds to $\varepsilon = 0$ or $z = t$. By Theorem 2, the limiting distribution of $n^{1/2}(T_n - \theta)$ consists of the right and left halves of two normal distributions.

The ratio of their standard deviations is (4.4).

Solving $0 = M(\theta;G) = c\{1-(1-\varepsilon)e^{\theta-t}\}$ yields $t = \theta + \log(1-\varepsilon)$. Table 1 shows the values of t and (4.4) for several values of ε when $\theta = 0.25$ and $c = \theta^{-1/2}e^{-\theta} = 1.5576\ \ldots$ (see (4.1)). In addition, the effect on a nominal .05 tail probability is shown.

For very small values of ε the effect is minimal, which accords with the robustness of T_n in the sense of weak* continuity (see Hampel (1971)), since it is asymptotically normal at the model. As ε increases, however, the effect becomes more serious, and inference based on T_n can be substantially biased.

For related work see Stigler (1973), who observes that a bias of this type can arise when the trimmed mean is used for discrete or grouped data.

Table 1 Effect of contaminating mass ε with $\theta = 0.25$ fixed

ε	t	r = (4.4)	$\Phi(-1.645r)$
0	0.25	1	.05
0.01	0.24	0.976	.054
0.05	0.199	0.877	.074
0.10	0.145	0.748	.109
0.15	0.087	0.610	.158
0.20	0.027	0.465	.222

5. Smooth score functions

In the example of the preceding section, one might argue that the parameter values where problems arise are unlikely to occur in practice, or that c can be changed slightly. It is not, however, the non-normal limiting distribution of T_n at certain distributions that is of concern, but the instability of inference based on T_n near those distributions. This phenomenon can alternatively be interpreted as a discontinuity of the asymptotic variance functional $V(T(G);G) = \{m(T(G);G)\}^{-1}\int\psi_{T(G)}^2 dG \; \{m(T(G);G)\}^{-1}$; cf. Huber (1981, p. 51). In the neighborhood of a distribution where V is discontinuous, estimates of the variance of T_n may be unstable.

Instability of tyis type can be avoided by requiring the M-estimator score function to be smooth, for example, by replacing $\psi_c(\cdot)$ in (2.4) with a smooth approximation. A natural way to construct such a function is by rescaling a smooth distribution function.

Suppose F is an absolutely continuous d.f. with density f symmetric about zero. Then

$$\psi(x) = 2c\{F(\frac{x}{2cf(0)}) - \frac{1}{2}\} \tag{5.1}$$

is monotone increasing, skew-symmetric about zero, and satisfies $\psi(\infty) = c$ and $\psi'(0) = 1$. Observe that ψ_c is obtained from (5.1) by taking F to be the uniform distribution on $[-\frac{1}{2},\frac{1}{2}]$. This can be approximated arbitrarily closely by a symmetric beta distribution with a small value for the shape parameter, i.e., $f(x) \propto \{(\frac{1}{2} + x)(\frac{1}{2} - x)\}^a$ on $[-\frac{1}{2},\frac{1}{2}]$. The resulting score function is complicated, however, and its second derivative has jump discontinuities. A more convenient choice is the logistic distribution, which leads to the smooth function

$$L_c(x) = c \tanh(x/c).$$

This has appeared previously. $L_c(x - t)$ is the maximum likelihood score for the location of a logistic distribution with scale 1. Holland and Welsch (1977) include an M-estimator using L_c in a Monte Carlo study of robust regression estimates.

For the important special case of estimating a Poisson parameter robustly, a smooth version of the optimal M-estimator solves

$$(5.2) \qquad n^{-1}\sum_{i=1}^{n} L_c(X_i t^{-1/2} - \beta(t)) = 0,$$

where β is defined in the usual manner.

Table 2 gives asymptotic variances V_θ and bounds γ_θ on influence functions for the estimator defined by (5.2), labeled L_c, and the optimal estimator, labeled ψ_c. In each case $c = 1.5$. The calculations are at the Poisson model, and V_θ and γ_θ are stabilized by dividing by θ and $\theta^{1/2}$ respectively.

Note that V_θ/θ is the asymptotic relative efficiency of the maximum likelihood estimator (sample mean) with respect to the corresponding M-estimator. The asymptotic variances for the logistic score are slightly smaller than those for the "optimal" score. This is possible because the bounds on the influence function of L_c are slightly higher for ψ_c. In terms of performance at the model, there appears to be little difference between L_c and ψ_c.

Table 2 Asymptotic variances and influence function bounds at the Poisson model

Mean	ψ_c		L_c	
θ	V_θ/θ	$\gamma_\theta/\theta^{1/2}$	V_θ/θ	$\gamma_\theta/\theta^{1/2}$
0.1	1.052	3.16	1.048	3.27
0.2	1.107	2.24	1.081	2.53
0.3	1.138	1.98	1.094	2.29
0.4	1.114	2.00	1.095	2.19
0.5	1.092	1.98	1.083	2.14
1.0	1.071	1.84	1.059	2.07
2.0	1.057	1.74	1.045	2.04
5.0	1.043	1.75	1.038	2.02
10.0	1.040	1.74	1.035	2.02
100.0	1.037	1.73	1.033	2.01

6. Further remarks

The need for smooth score functions is most clear when the data consist of counts. In this case every deviation from the model involves point masses.

An important consequence of Theorem 1 is that Hampel's optimal estimator (2.4) is indeed optimal as claimed when the model distribution is discrete. It would be disturbing if the theory were to break down at a countable number of parameter values. Moreover, the smooth versions discussed in Section 5, which provide more stable inference, are justified for every parameter value as being nearly optimal.

Although the discussion has focused on the score functions arising from Hampel's optimality theory, it is not limited to that context. For instance, a score based on Hampel's three part redescending ψ (see Huber (1981, p. 102)) will be prone to the same difficulties, and a smooth version will be more stable.

Appendix. Proof of Theorem 2.

Since the d.f. H is continuous, uniform convergence in (3.7) will follow from pointwise convergence via Polya's Theorem (Serfling (1980, p. 18)).

Write M(t) for M(t;G) and m(t) for m(t;G). denote by U(δ) the set (t: $0 < |t - \theta| < \delta$). By (B1), m is defined on U(δ) if δ is sufficiently small. Moreover, given $\varepsilon > 0$, there is a δ for which $t \in U(\delta)$ implies

$$|m(t) - m(\theta-)| < \varepsilon \quad \text{if } t < \theta$$

and

$$|m(t) - m(\theta+)| < \varepsilon \quad \text{if } t > \theta.$$

Choosing $\varepsilon < \min\{|m(\theta-)|, |m(\theta+)|\}$ then guarantees that $|m(t)|$ is bounded away from zero on U(δ). Fix such a δ.

Since $M(\theta) = 0$, $t \in U(\delta)$ implies

$$M(t) = m(\tau)(t - \theta) \tag{A.1}$$

for some τ strictly between t and θ, by the Mean Value Theorem (which only requires one-sided derivatives at the endpoints of the interval on which it is applied). Since m is bounded away from zero on $U(\delta)$, (A.1) shows

$$|t - \theta| = O(|M(t)|)$$

as $t \to \theta$. The right hand side of (A.1) equals

$$D(t)(t - \theta) + R(t), \tag{A.2}$$

where

$$D(t) = m(\theta+)I(t > \theta) + m(\theta-)I(t < \theta),$$

$$R(t) = [\{m(\tau) - m(\theta+)\}I(t > \theta) + \{m(\tau) - m(\theta-)\}I(t < \theta)](t - \theta),$$

and $I(A)$ is the indicator for the set A. Note that (A.2) also holds if $t = \theta$. Since $R(t) = o(|t - \theta|) = o(|M(t)|)$, (A.1) and (A.2) yield

$$D(T_n)n^{1/2}(T_n - \theta) = n^{1/2}M(T_n) + o(|n^{1/2}M(T_n)|). \tag{A.3}$$

Because of (B2), (B3) and the Lindeberg-Levy central limit theorem, the right hand side of (A.3) converges in distribution to a $N(0,\sigma^2)$ random variable, and, hence, so does the left hand side.

To obtain the limiting distribution of T_n, partition its range and consider cases. If $z < 0$ then

$$\text{pr}\{n^{1/2}(T_n - \theta) \le z,\ T_n > \theta\} = 0,$$

while

$$\mathrm{pr}\{n^{1/2}(T_n - \theta) \le z,\ T_n < \theta\} = \mathrm{pr}\{|D(T_n)|n^{1/2}(T_n - \theta) \le |D(T_n)|z\}.$$

Since $D(T_n) = m(\theta-)$ when $T_n < \theta$, and $D(t)$ does not change sign on $(\theta - \delta, \theta + \delta)$ by (B1), (A.3) implies that this last probability converges to $\Phi(|m(\theta-)|z/\sigma)$ as $n \to \infty$. Similar arguments establish that, for $z > 0$,

$$\mathrm{pr}\{n^{1/2}(T_n - \theta) \le z,\ T_n < \theta\} = \mathrm{pr}\{|m(\theta-)|n^{1/2}(T_n - \theta) < 0\} \to \frac{1}{2}$$

and

$$\mathrm{pr}\{n^{1/2}(T_n - \theta) \le z,\ T_n > \theta\}$$
$$= \mathrm{pr}\{0 < |m(\theta+)|n^{1/2}(T_n - \theta) \le z|m(\theta+)|\} \to \Phi(|m(\theta+)|z/\sigma) - \frac{1}{2}$$

and finally

$$\mathrm{pr}\{n^{1/2}(T_n - \theta) \le 0\} = 1 - \mathrm{pr}\{|m(\theta+)|n^{1/2}(T_n - \theta) > 0\} \to \frac{1}{2}$$

as $n \to \infty$. The result follows by collecting terms.

REFERENCES

Boos, D.D. and Serfling, R.J. (1980). A note on differentials and the CLT and LIL for statistical functions, with applications to M-estimates. Ann. Stat. 8, 618-624.

Carroll, R.J. (1978a). On almost sure expansions for M-estimates. Ann. Stat. 6, 314-318.

Carroll, R.J. (1978b). On the asymptotic distribution of multivariate M-estimates. J. Multi. Anal. 8, 361-371.

Franklin, P. (1940). A Treatise on Advanced Calculus. Wiley, New York.

Hampel, R. (1968). Contributions to the theory of robust estimation. Ph.D. thesis, University of California, Berkeley.

Hampel, F. (1971). A general qualitative definition of robustness. Ann. Math. Statist. 42, 1887-1896.

Hampel, F. (1974). The influence curve and its role in robust estimation. J. Amer. Statist. Assoc. 62, 1179-1186.

Holland, P.W. and Welsch, R.E. (1977). Robust regression using iterativity reweighted least-squares. Commun. in Statist. A6, 813-827.

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

Huber, P.J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1. University of California Press, Berkeley.

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

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

Krasker, W.S. and Welsch, R.E. (1982). Efficient bounded-influence regression estimation. J. Amer. Statist. Assoc. 77, 595-604.

Ruppert, D. (1985). On the bounded influence regression estimator of Krasker and Welsch. J. Amer. Statist. Assoc. 80, 205-208.

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

Simpson, D.G. (1985). Some contributions to robust inference for discrete probability models. Ph.D. dissertation, University of North Carolina, Chapel Hill.

Stefanski, L.A., Carroll, R.J., and Ruppert, D. (1985). Optimally bounded score functions for generalized linear models with applications to logistic regression. (Tentatively accepted by Biometrika.)

Stigler, S.M. (1973). The asymptotic distribution of the trimmed mean. Ann. Statist. 1, 472-477.