Statistics 240 Lecture Notes

P.B. Stark www.stat.berkeley.edu/∼stark/index.html

November 23, 2010

1 Part 9: Robustness and related topics. ROUGH DRAFT

References:

Hampel, F.R., Rousseeuw, P.J., Ronchetti, E.M., and Strahel, W.A., 1986. *Robust Statistics: The approach based on influence functions*, Wiley, NY.

Huber, P.J., 1981. *Robust Statistics*, Wiley, N.Y.

1.1 Heuristics

So far, we have been concerned with nonparametric tests and procedures: tests and procedures with minimal assumptions about the probability distribution that gives rise to the data. Examples of the kinds of assumptions we have made in various places are that the distribution is continuous, that the observations are iid, that subjects are randomized into treatment and control, *etc.*

We now consider assumptions that are more restrictive, but still less restrictive than the parametric assumptions common to much of statistical theory. In a loose sense, an estimator or test is *robust* if its performance is good over a neighborhood of distributions including the family in which the truth is modeled to lie.

Virtually every real data set has some "gross errors," which are in some sense corrupted measurements. Data can be transcribed incorrectly, bits can be corrupted in transmission or storage, cosmic rays can hit satellite-borne instruments, power supplies can have surges, operators can miss their morning cups of coffee, *etc.* Hampel *et al.* (1986) quote studies that find gross errors comprise

up to 20% of many data sets in physical, medical, agricultural and social sciences, with typical rates of 6% or more.

Even without gross errors, there is always a limit on the precision with which data are recorded, leading to truncation or rounding errors that make continuous models for data error distributions only approximate. Furthermore, parametric error models rarely are dictated by direct physical arguments; rather, the central limit theorem is invoked, or some historical assumption becomes standard in the field.

One early attempt to deal with gross outliers is due to Sir Harold Jeffreys, who (in the 1930s) modeled data errors as a mixture of two Gaussian distributions, one with variance tending to infinity. The idea is that the high-variance Gaussian represents a fraction of gross outliers possibly present in the data; one wants to minimize the influence of such observations on the resulting estimate or test. Considerations in favor of fitting by minimizing mean absolute deviation instead of least squares go back much further.

We will be looking at ways of quantifying robustness and of constructing procedures whose performance when the model is true is not much worse than the optimal procedure, but whose performance when the model is wrong (by a little bit) is not too bad, and is often much better than the performance of the optimal procedure when the model for which it is optimal is wrong by a little bit.

The kinds of questions typically asked in the robustness literature are:

1. Is the procedure sensitive to small departures from the model?

2. To first order, what is the sensitivity?

3. How wrong can the model be before the procedure produces garbage?

The first issue is that of qualitative robustness; the second is quantitative robustness; the third is the "breakdown point."

1.2 Resistance and Breakdown Point

Resistance has to do with changes to the observed data, rather than to the theoretical distribution underlying the data. A statistic is *resistant* if arbitrary changes to a few data (such as might be

caused by *gross outliers*, or small changes to all the data (such as might be caused by rounding or truncation), result in only small changes to the value of the statistic.

Suppose we are allowed to change the values of the observations in the sample. What is the smallest fraction would we need to change to make the estimator take an arbitrary value? The answer is the *breakdown point* of the estimator.

For example, consider the sample mean $\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$ as an estimator of the mean of F. (The mean can be written as a functional of the distribution: $T(F) = \int x dF$; $\theta = T(F_\theta)$.) Corrupting a single observation can make the sample mean take any real value: the breakdown point is $\frac{1}{n}$.

In contrast, consider the "α-trimmed mean," defined as follows: Let $k_\ell = \lfloor \alpha n \rfloor$ and $k_h = \lceil (1-\alpha)n \rceil$. Let $X_{(j)}$ be the jth order statistic of the data. Define

$$\bar{X}_\alpha = \frac{1}{k_h - k_\ell + 1} \sum_{j=k_\ell}^{k_h} X_{(j)}. \tag{1}$$

This measure of location is less sensitive to outliers than the sample mean is: Its breakdown point is $n^{-1}\min(k_\ell, n - k_h + 1)$. An alternative, not necessarily equivalent, definition of the α-trimmed mean is through the functional

$$T(F) = \frac{1}{1-2\alpha} \int_\alpha^{1-\alpha} F^{-1}(t) dt. \tag{2}$$

This version has breakdown point α.

We shall make the notion of breakdown point more precise presently; a few definitions are required.

Definition 1 *The Lèvy distance between two distribution functions F and G on $\mathbb{R}$ is*

$$\lambda(F,G) = \inf\{\epsilon : F(x-\epsilon) - \epsilon \le G(x) \le F(x+\epsilon) + \epsilon, \ \ \forall x \in \mathbb{R}\}. \tag{3}$$

The Lévy distance makes sense for probability distributions on the real line, where we can define the cdf. We want to extend this kind of notion to probability distributions on more complicated sample spaces, including $\mathbb{R}^k$. Ultimately, we want a metric on probability distributions on Polish spaces, which are defined below. To do that, we need a bit more abstract structure.

The following definitions are given to refresh your memory. For more detail, see Rudin, W. (1974). *Real and Complex Analysis*, 2nd edition, McGraw-Hill, N.Y., pages 636–650 of Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*, Springer-Verlag, NY, or Halmos, P.R. (1974). *Measure Theory*, Springer-Verlag, N.Y.

Definition 2 *A* topology $\mathcal{V}$ *on a space* $\mathcal{X}$ *is a collection of subsets* V *of* $\mathcal{X}$ *such that*

1. $\emptyset \in \mathcal{V}$.
2. *If* $V_j \in \mathcal{V}$, $j = 1, \ldots, n$, *then* $\cap_{j=1}^{n} V_j \in \mathcal{V}$.
3. *For any collection* $\{V_\alpha\}_{\alpha \in A}$ *of elements of* $\mathcal{V}$, $\cup_{\alpha \in A} V_\alpha \in \mathcal{V}$.

The elements of $\mathcal{V}$ *are called* open sets*, and we call* $\mathcal{X}$ *a* topological space*. Complements of elements of* $\mathcal{V}$ *are called* closed sets*. Every open set* V *that contains* $x \in \mathcal{X}$ *is called a* neighborhood *of* x*. The collection of sets* $\mathcal{B} \subset \mathcal{V}$ *is a* base *if for every* $x \in \mathcal{X}$ *and every neighborhood* V *of* x, $x \in B \subset V$ *for some* $B \in \mathcal{B}$*. The space* $\mathcal{X}$ *is* separable *if it has a countable base. Let* A *be a subset of* $\mathcal{X}$*. The largest open set contained in* A *is called the* interior *of* A*, denoted* A°*. The smallest closed set that contains* A *is called the* closure *of* A*, denoted* $\bar{A}$.

Definition 3 *Let* $\mathcal{X}$ *be a topological space with topology* $\mathcal{V}$ *and let* $\mathcal{Y}$ *be a topological space with topology* $\mathcal{W}$*. A mapping* $f : \mathcal{X} \to \mathcal{Y}$ *is* continuous *if* $f^{-1}(W) \in \mathcal{V}$ *for every* $W \in \mathcal{W}$.

That is, a function from one topological space into another is continuous if the preimage of every open set is an open set. (The preimage of a set $A \subset \mathcal{Y}$ under the mapping $f : \mathcal{X} \to \mathcal{Y}$ is $f^{-1}(A) \equiv \{x \in \mathcal{X} : f(x) \in A\}$. Note that the preimage of a set that contains a single point can be a set containing more than one point.)

Definition 4 *A collection* $\mathcal{M}$ *of subsets of a set* $\mathcal{X}$ *is a* σ-algebra *on* $\mathcal{X}$ *if*

1. $\mathcal{X} \in \mathcal{M}$
2. $A \in \mathcal{M}$ *implies* $A^c \in \mathcal{M}$
3. *if* $A_j \in \mathcal{M}$, $j = 1, 2, 3, \ldots$, *then* $\cup_{j=1}^{\infty} A_j \in \mathcal{M}$

If $\mathcal{M}$ *is a* σ*-algebra on* $\mathcal{X}$*, then we call* $\mathcal{X}$ *a* measurable space *and we call the elements of* $\mathcal{M}$ *the* measurable sets *of* $\mathcal{X}$*. If* $\mathcal{X}$ *is a measurable space with* σ*-algebra* $\mathcal{M}$ *and* $\mathcal{Y}$ *is a topological space with topology* $\mathcal{W}$ *then* $f : \mathcal{X} \to \mathcal{Y}$ *is a* measurable function *if* $f^{-1}(W) \in \mathcal{M}$ *for every* $W \in \mathcal{W}$.

For any collection $\mathcal{F}$ of subsets of a set $\mathcal{X}$, there is a smallest σ-algebra containing $\mathcal{F}$.

Theorem 1 *If each member of a sequence* $\{f_j\}_{j=1}^{\infty}$ *is a measurable function from a measurable space* $\mathcal{X}$ *into* $[-\infty, \infty]$*, then so is* $\sup_j f_j$ *and so is* $\limsup_j f_j$*.*

Definition 5 *A real-valued function* $s : \mathcal{X} \to \Re^+$ *on a measurable space* $\mathcal{X}$ *is* simple *if its range consists of finitely many points in* $\Re^+$*. Such a function can be written in the form*

$$s(x) = \sum_{j=1}^{n} a_j 1_{x \in A_j} \tag{4}$$

for a set $\{a_j\}_{j=1}^n \subset \Re^+$ *and a collection of subsets* $\{A_j\}_{j=1}^n$ *of* $\mathcal{X}$*. If* $\{A_j\}_{j=1}^n$ *are measurable, then* s *is a measurable simple function.*

Definition 6 *A* positive measure μ *on a measurable space* $\mathcal{X}$ *with* σ*-algebra* $\mathcal{M}$ *is a mapping from* $\mathcal{M} \to \Re^+$ *such that*

1. $\mu(A) \geq 0 \ \forall A \in \mathcal{M}$
2. *if* $\{A_j\}_{j=1}^{\infty}$ *are pairwise disjoint, then*

$$\mu(\cup_{j=1}^{\infty} A_j) = \sum_{j=1}^{\infty} \mu(A_j). \tag{5}$$

A positive measure μ *for which* $\mu(\mathcal{X}) = 1$ *is called a* probability measure*. A measurable space on which a measure is defined is called a* measure space*.*

Definition 7 *The integral of a measurable simple function* s *of the form 4 with respect to the measure* μ *is*

$$\int_{\mathcal{X}} s d\mu \equiv \sum_{j=1}^{n} a_j \mu(A_j). \tag{6}$$

Integrals of more general measurable functions are defined as limits of integrals of simple functions.

Theorem 2 *Every nonnegative real-valued measurable function* f *is the limit of an increasing sequence* $\{s_n\}$ *of measurable simple functions.*

Definition 8 *The integral of a nonnegative measurable function* f *with respect to the measure* μ *is*

$$\int_{\mathcal{X}} f d\mu \equiv \sup_{s \leq f, s \ simple} \int_{\mathcal{X}} s d\mu. \tag{7}$$

The *positive part* f^+ of a real-valued function f is $f^+(x) \equiv \max(f(x), 0) \equiv f \vee 0$. The *negative part* f^- of a real-valued function f is $f^-(x) \equiv -\min(f(x), 0) \equiv f \wedge 0$. If f is measurable, so are $|f|$, f^+ and f^-.

Definition 9 *The integral of a measurable function f with respect to the measure μ is*

$$\int_{\mathcal{X}} f d\mu \equiv \int_{\mathcal{X}} f^+ d\mu - \int_{\mathcal{X}} f^- d\mu \tag{8}$$

if $\int_{\mathcal{X}} |f| d\mu < \infty$.

The following three theorems are the workhorses of the theory:

Theorem 3 (Lebesgue's Monotone Convergence Theorem.) *Let $\{f_j\}_{j=1}^{\infty}$ be a nonnegative sequence of functions such that for every $x \in \mathcal{X}$, $f_1(x) \leq f = 2(x) \leq \cdots \leq \infty$ and $f_j(x) \to f(x)$ as $j \to \infty$. Then f is measurable and $\int_{\mathcal{X}} f_j d\mu \to \int_{\mathcal{X}} f d\mu$.*

Theorem 4 (Fatou's Lemma.) *If $\{f_j\}$ is a sequence of nonnegative measurable functions, then*

$$\int_{\mathcal{X}} (\liminf_{j\to\infty} f_j) d\mu \leq \liminf_{j\to\infty} \int_{\mathcal{X}} f_j d\mu. \tag{9}$$

Theorem 5 (Lebesgue's Dominated Convergence Theorem.) *Let $\{f_j\}_{j=1}^{\infty}$ be a sequence of measurable functions such that for every $x \in \mathcal{X}$, $f(x) = \lim_j f_j(x)$ exists. If there is a function g such that $\int_{\mathcal{X}} |g| d\mu < \infty$ and $|f_j(x)| \leq g(x)$ for all $x \in \mathcal{X}$ and all $j \geq 1$, then $\int_{\mathcal{X}} |f| d\mu < \infty$, $\lim_j \int_{\mathcal{X}} |f_j - f| d\mu = 0$, and $\lim_j \int_{\mathcal{X}} f_j d\mu = \int_{\mathcal{X}} f d\mu$.*

Definition 10 *If $\mathcal{X}$ is a topological space with topology $\mathcal{V}$, the* Borel σ-algebra *$\mathcal{B}$ on $\mathcal{X}$ is the smallest σ-algebra containing $\mathcal{V}$. The elements of $\mathcal{B}$ are called the* Borel sets *of $\mathcal{X}$.*

Definition 11 *A* metric space *is a set $\mathcal{X}$ and a* metric *function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ such that*

1. $d(x, x) = 0$
2. $d(x, y) = 0$ *implies* $x = y$
3. $d(x, y) = d(y, x)$, *for all* $x, y \in \mathcal{X}$ *(symmetry)*
4. $d(x, y) \leq d(x, z) + d(y, z)$ *for all* $x, y, z \in \mathcal{X}$ *(triangle inequality)*

If d satisfies all these axioms but (2), it is called a pseudometric.

An *open ball* with radius $\epsilon \geq 0$ centered at the point x in the metric space $\mathcal{X}$ is the set $B_\epsilon(x) = \{y \in \mathcal{X} : d(x,y) < \epsilon\}$. If $\mathcal{X}$ is a metric space and $\mathcal{V}$ is the collection of all sets in $\mathcal{X}$ that are unions of open balls, then $\mathcal{V}$ is a topology on $\mathcal{X}$; it is called the *topology induced by d.*

If $\mathcal{V}$ is a topology on $\mathcal{X}$ such that there exists a metric d on $\mathcal{X}$ for which $\mathcal{V}$ is the topology induced by d, then $\mathcal{V}$ is *metrizable*.

Definition 12 *A sequence $\{x_j\}_{j=1}^\infty$ of elements of a metric space $\mathcal{X}$ is a* Cauchy sequence *if for every $\epsilon > 0$ there exists an integer $N = N(\epsilon)$ such that $d(x_n, x_m) < \epsilon$ if $\min(n,m) > N$.*

Definition 13 *A sequence $\{x_j\}_{j=1}^\infty$ of elements of a metric space $\mathcal{X}$* converges to *$x \in \mathcal{X}$ if*

$$\lim_{j\to\infty} d(x_j, x) = 0. \tag{10}$$

Definition 14 *A metric space $\mathcal{X}$ is* complete *if every Cauchy sequence in $\mathcal{X}$ converges to an element of $\mathcal{X}$.*

Definition 15 *A subset D of a metric space $\mathcal{X}$ is* dense *if for every element x of $\mathcal{X}$ and every $\epsilon > 0$, there exists $y \in D$ such that $d(y,x) < \epsilon$.*

Definition 16 *A metric space $\mathcal{X}$ is* separable *if it contains a countable dense subset.*

Definition 17 *A* vector space over the real numbers *is an Abelian group $\mathcal{X}$ of* vectors *with the group operation $(+ : \mathcal{X} \times \mathcal{X} \to \mathcal{X},\ (x,y) \mapsto x+y)$, together with the operation of scalar multiplication $(: \Re \times \mathcal{X} \to \mathcal{X},\ \ (\lambda, x) \mapsto \lambda x)$, such that for all $\lambda, \mu \in \Re$ and all $x, y \in \mathcal{X}$,*

1. $\lambda(x+y) = \lambda x + \lambda y$
2. $(\lambda+\mu)x = \lambda x + \mu x$
3. $\lambda(\mu x) = (\lambda\mu)x$
4. $1x = x$
5. $0x = 0$

Definition 18 *A function $\|\cdot\| : \mathcal{X} \to \mathbb{R}^+$ on a vector space $\mathcal{X}$ is a* norm *if for all $x, y \in \mathcal{X}$ and all $\lambda \in \mathbb{R}$,*

1. $\|x\| \geq 0$ *(positive semidefinite)*
2. $\|x\| = 0$ *implies* $x = 0$ *(with (1), this means the norm is positive definite)*
3. $\|\lambda x\| = |\lambda|\|x\|$ *(positive scalar homogeneity)*
4. $\|x + y\| \leq \|x\| + \|y\|$ *(triangle inequality)*

If $\|\cdot\|$ satisfies all but (2), it is a seminorm. *If $\|\cdot\|$ is a norm on a vector space $\mathcal{X}$, we call $\mathcal{X}$ a normed (linear) vector space.*

A normed linear vector space is a metric space under the metric $d(x, y) \equiv \|x - y\|$.

Definition 19 *A* Banach space *is a complete normed linear vector space.*

Definition 20 *A* Polish space $\mathcal{X}$ *is a separable topological space whose topology is metrizable by a metric d with respect to which $\mathcal{X}$ is complete.*

Equivalently, a Polish space is a topological space that is homeomorphic to a separable Banach space.

Examples of Polish spaces include $\mathbb{R}^k$. Let $\mathcal{M}$ denote the space of probability measures on the Borel σ-algebra $\mathcal{B}$ of subsets of $\mathcal{X}$. (The Borel σ-algebra is the smallest σ-algebra containing all the open sets in $\mathcal{X}$.) Let $\mathcal{M}'$ denote the set of finite signed measures on $(\mathcal{X}, \mathcal{B})$; this is the linear space of measures generated by $\mathcal{M}$. The measures in $\mathcal{M}'$ are *regular* in the sense that for every $F \in \mathcal{M}'$,

$$\sup_{C \subset A;\, C \text{ compact}} F(C) = F(A) = \inf_{G \supset A;\, G \text{ open}} F(G). \tag{11}$$

The weak-star topology in $\mathcal{M}'$ is the weakest topology for which the functional

$$\int \psi dF \tag{12}$$

is continuous for every continuous, bounded function $\psi : \mathcal{X} \to \mathbb{R}$.

In this section, we assume that the space $\mathcal{X}$ of possible data is a Polish space, and that all measures are defined on $\mathcal{B}$. An overbar (*e.g.*, $\bar{A}$) denotes topological closure, the superscript $\circ$ (*e.g.*, A°) denotes the topological interior, and the superscript c will denote complementation ($A^c = \{x \in \mathcal{X} : x \notin A\}$).

Definition 21 *For any subset A of a metric space $\mathcal{X}$ with metric $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$, the* closed ϵ-neighborhood of A *is*

$$A^\epsilon = \{x \in \mathcal{X} : \inf_{a \in A} d(x, a) \leq \epsilon\}. \tag{13}$$

It will be important presently that

$$A^\epsilon = \bar{A}^\epsilon = \overline{A^\epsilon} = \overline{\bar{A}^\epsilon}. \tag{14}$$

Definition 22 *The Prohorov distance between two measures F and G defined on a common algebra $\mathcal{A}$ of subsets of a metric space $\mathcal{X}$ is*

$$\pi(F, G) = \inf\{\epsilon \geq 0 : F(A) \leq G(A^\epsilon) + \epsilon, \ \ \forall A \in \mathcal{A}\} \tag{15}$$

Expanding the events by ϵ to form A^ϵ corresponds to the measure G being "shifted" slightly from F, for example, by rounding. The addition of ϵ corresponds to a fraction ϵ of the observations being from a completely different distribution.

We shall verify that the Prohorov distance really is a metric if the sample space $\mathcal{X}$ is a Polish space. Clearly, it is nonnegative, and $\pi(F, F) = 0$. We need to show symmetry, the triangle inequality, and that $\{\pi(F, G) = 0\} \Rightarrow \{F = G\}$. The following proof follows that in Huber (1981).
Symmetry. This will follow immediately if we can show that if $F(A) \leq G(A^\epsilon) + \epsilon$ for all $A \in \mathcal{A}$, then $G(A) \leq F(A^\epsilon) + \epsilon$ for all $A \in \mathcal{A}$. Recall that because $\mathcal{A}$ is an algebra, if $A \in \mathcal{A}$, then $A^c \in \mathcal{A}$ as well. Take any $\delta > \epsilon$, and consider $A = (B^\delta)^c = B^{\delta c}$ for any $B \in \mathcal{A}$. Note that $A \in \mathcal{A}$, so by the premise,

$$F(B^{\delta c}) \leq G(B^{\delta c \epsilon}) + \epsilon, \tag{16}$$

or

$$\begin{aligned} 1 - F(B^\delta) &\leq 1 - G(B^{\delta c \epsilon c}) + \epsilon & (17) \\ G(B^{\delta c \epsilon c}) &\leq F(B^\delta) + \epsilon. & (18) \end{aligned}$$

However, $B \subset B^{\delta c \epsilon c}$, as we shall see. This statement is equivalent to $B^{\delta c \epsilon} \subset B^c$. This is essentially immediate from $\delta > \epsilon$: if $x \in B^{\delta c \epsilon}$, then $\exists y \notin B^\delta$ s.t. $d(x, y) < \epsilon$ (typo in Huber here). Thus $x \in B^c$, because otherwise $d(x, y) > \delta > \epsilon$. Thus

$$G(B) \leq G(B^{\delta c \epsilon c}) \leq F(B^\delta) + \epsilon. \tag{19}$$

But $B^\epsilon = \cap_{\delta>\epsilon} B^\delta$, so the result follows.

Positive definite. To show that $(\pi(F,G) = 0) \Rightarrow (F = G)$, note that the closure of A is $\bar{A} = \cap_{\epsilon>0} A^\epsilon$. Thus $\pi(F,G) = 0$ implies $F(A) \leq G(A)$ and $G(A) \leq F(A)$ for all closed sets $A \in \mathcal{A}$.

Triangle inequality. If $\pi(F,G) \leq \epsilon$ and $\pi(G,H) \leq \delta$, then for every $A \in \mathcal{A}$,

$$F(A) \leq G(A^\epsilon) + \epsilon \leq H((A^\epsilon)^\delta) + \epsilon + \delta. \tag{20}$$

But by the triangle inequality for the metric d on $\mathcal{X}$, $(A^\epsilon)^\delta \subset A^{\epsilon+\delta}$, so we are done.

Note that the Prohorov distance between $\hat{F}_n$ and the "contaminated" empirical distribution one gets by changing k of the data by an arbitrary amount is $\frac{k}{n}$.

Theorem 6 *(Strassen, 1965; see Huber Thm 2.3.7.) The following are equivalent:*

1. *$F(A) \leq G(A^\delta) + \epsilon$ for all $A \in \mathcal{B}$*
2. *There are (possibly) dependent $\mathcal{X}$-valued random variables X, Y such that $\mathcal{L}(X) = F$, $\mathcal{L}(Y) = G$, and $\mathbb{P}\{d(X,Y) \leq \delta\} \geq 1 - \epsilon$. (Here, $\mathcal{L}(X)$ denotes the probability law of X,* etc.*)*

Definition 23 *Suppose that the distance function d on $\mathcal{X} \times \mathcal{X}$ is bounded by one (one can replace d by $d(x,y)/(1+d(x,y))$ to make this so). The* bounded Lipschitz metric *on $\mathcal{M}$ is*

$$d_{BL}(F,G) = \sup_{\psi:|\psi(x)-\psi(y)|\leq d(x,y)} \left| \int \psi dF - \int \psi dG \right|. \tag{21}$$

The bounded Lipschitz metric is truly a metric on $\mathcal{M}'$.

Theorem 7 *The set of regular Borel measures $\mathcal{M}'$ on a Polish space $\mathcal{X}$ is itself a Polish space with respect to the weak topology, which is metrizable by the Prohorov metric and by the bounded Lipschitz metric.*

Consider a collection of probability distributions indexed by ϵ, such as the Prohorov neighborhood

$$\mathcal{P}_\pi(\epsilon; F) = \{G \in \mathcal{M} : \pi(F,G) \leq \epsilon\} \tag{22}$$

or the "gross error contamination neighborhood" (not truly a neighborhood in the weak topology)

$$\mathcal{P}_{\text{gross error}}(\epsilon; F) = \{G \in \mathcal{M} : G = (1-\epsilon)F + \epsilon H, H \in \mathcal{M}\}. \tag{23}$$

Let $M(G, T_n)$ denote the median of the distribution of $T_n(G) - T(F)$. Let $A(G, T_n)$ denote some fixed percentile of the distribution of $|T_n(G) - T(F)|$. If $T(F_\theta) = \theta, \ \ \forall \theta \in \Theta$, we say that T is *Fisher consistent* for θ.

Definition 24 *Consider a sequence $\{T_n\}$ of estimators that is Fisher consistent and converges in probability to a functional statistic T. The* maximum bias of $\{T_n\}$ at F over the collection $\mathcal{P}(\epsilon)$ *is*

$$b_1(\epsilon) = b_1(\epsilon, \mathcal{P}, F) = \sup_{G \in \mathcal{P}(\epsilon)} |T(G) - T(F)|. \tag{24}$$

The maximum asymptotic bias of $\{T_n\}$ at F over the collection $\mathcal{P}(\epsilon)$ *is*

$$b(\epsilon) = b(\epsilon, \mathcal{P}, F) = \lim_{n \to \infty} \sup_{G \in \mathcal{P}(\epsilon)} |M(G, T_n)|. \tag{25}$$

If b_1 is well defined, $b(\epsilon) \geq b_1(\epsilon)$. Note that for the gross-error model and for the Lèvy and Prohorov distances, $b(\epsilon) \leq b(1)$, because the set $\mathcal{P}(1) = \mathcal{M}$.

Definition 25 *(Following Huber, 1981.) For a given collection $\mathcal{P}(\epsilon)$ of distributions indexed by $\epsilon \geq 0$, the* asymptotic breakdown point *of T at F is*

$$\epsilon^* \equiv \epsilon^*(F, T, \mathcal{P}(\cdot)) = \sup\{\epsilon : b(\epsilon, \mathcal{P}(\epsilon), F) < b(1)\}. \tag{26}$$

Definition 26 *(Following Hampel* et al.*, 1986.) The* breakdown point *of a sequence of estimators $\{T_n\}$ of a parameter $\theta \in \Theta$ at the distribution F is*

$$\epsilon^*(T_n, F) \equiv \sup\{\epsilon \leq 1 : \exists K_\epsilon \subseteq \Theta, \ K_\epsilon \text{ compact, s.t. } \pi(F, G) < \epsilon \Rightarrow G(\{T_n \in K_\epsilon\}) \to 1 \ \text{ as } n \to \infty\}. \tag{27}$$

That is, the breakdown point is the largest Prohorov distance from F a distribution can be, and still have the estimators almost surely take values in some compact set as $n \to \infty$.

Definition 27 *The* finite-sample breakdown point of the estimator T_n at $x = (x_j)_{j=1}^n$ *is*

$$\epsilon^*(T_n, x) \equiv \frac{1}{n} \max \left\{ m : \max_{i_1, \ldots, i_m} \sup_{y_1, \ldots, y_m \in \mathcal{X}} |T_n(z_1, \ldots, z_n)| < \infty \right\}, \tag{28}$$

where

$$z_j = \begin{cases} x_j, & j \notin \{i_k\}_{k=1}^m \\ y_k, & j = i_k \text{ for some } k. \end{cases} \tag{29}$$

This definition makes precise the notion that corrupting some fraction of the measurements can corrupt the value of the statistic arbitrarily. Note that the finite-sample breakdown point is a function not of a distribution, but of the sample and the estimator. Typically, its value does not depend on the sample. It is this "breakdown point" that we saw was zero for the sample mean; it is a measure of *resistance*, not *robustness*.

Definition 28 *A sequence of estimators* $\{T_n\}$ *is* qualitatively robust at F *if for every* $\epsilon > 0$, $\exists \delta > 0$ *such that for all* $G \in \mathcal{M}$ *and all* n,

$$\pi(F, G) < \delta \Rightarrow \pi(\mathcal{L}_F(T_n), \mathcal{L}_G(T_n)) < \epsilon. \tag{30}$$

That is, $\{T_n\}$ *is qualitatively robust at* F *if the distributions of* T_n *are equicontinuous w.r.t.* n.

1.3 The Influence Function

The next few sections follow Hampel *et al.* (Ch2) fairly closely. Consider an estimator of a real parameter $\theta \in \Theta$, where Θ is an open convex subset of $\mathbb{R}$, based on a sample X_n of size n. The sample space for each observation is $\mathcal{X}$. We consider a family $\mathcal{F}$ of distributions $\{F_\theta : \theta \in \Theta\}$, which are assumed to have densities $\{f_\theta : \theta \in \Theta\}$ with respect to a common dominating measure.

In our treatment of the bootstrap, we considered problems in which it suffices to know the empirical distribution of the data: the estimator could be assumed to be a function of the empirical distribution—at least asymptotically. We do the same here. Consider estimators $\hat{\theta} = T_n(\hat{F}_n)$ that asymptotically can be replaced by functional estimators. That is, either $T_n(\hat{F}_n) = T(\hat{F}_n)$ for all n, or there is a functional $T : \text{dom}(T) \to \mathbb{R}$ such that if the components of X_n are iid G, with $G \in \text{dom}(T)$, then

$$T_n(\hat{G}_n) \to T(G) \tag{31}$$

in probability as $n \to \infty$. $T(G)$ is the asymptotic value of $\{T_n\}$ at G. We assume that $T(F_\theta) = \theta$, $\forall \theta \in \Theta$ (this is Fisher consistency of the estimator T).

Definition 29 *A functional T defined on probability measures is Gâteaux differentiable at the measure F in* $\mathrm{dom}(T)$ *if there exists a function $a : \mathcal{X} \to \mathbb{R}$ s.t. $\forall G \in \mathrm{dom}(T)$,*

$$\lim_{t\to 0} \frac{T((1-t)F + tG) - T(F)}{t} = \int a(x) dG(x). \tag{32}$$

That is,

$$\frac{\partial}{\partial t} T((1-t)F + tG)|_{t=0} = \int a(x) dG(x). \tag{33}$$

Gâteaux differentiability is weaker than Fréchet differentiability. Essentially, Gâteaux differentiability at F ensures that the directional derivatives of T exist in all directions that (at least infinitesimally) stay in the domain of T.

Let δ_x be a point mass at x.

Definition 30 *The* influence function *of T at F is*

$$\mathrm{IF}(x; T, F) \equiv \lim_{t\to 0} \frac{T((1-t)F + t\delta_x) - T(F)}{t} \tag{34}$$

at the points $x \in \mathcal{X}$ where the limit exists.

The influence function is similar to the Gâteaux derivative, but it can exist even when the Gâteaux derivative does not, because the set of directions it involves is not as rich. The influence function gives the effect on T of an infinitesimal perturbation to the data at the point x. This leads to a "Taylor series" expansion of T at F:

$$T(G) = T(F) + \int \mathrm{IF}(x; T, F) d(G - F)(x) + \text{ remainder}. \tag{35}$$

(Note that $\int \mathrm{IF}(x; T, F) dF(x) = 0$.)

1.3.1 Heuristics using $\mathrm{IF}(x; T, F)$

Consider what happens for large sample sizes. The empirical distribution $\hat{F}_n$ tends to the theoretical distribution F, and $T_n(\hat{F}_n)$ tends to $T(\hat{F}_n)$. We will use $\hat{F}_n$ to denote both the empirical cdf and the empirical measure, so $\hat{F}_n = \frac{1}{n}\sum_{j=1}^n \delta_{X_j}$. Thus

$$(T_n(\hat{F}_n) - T(F)) \approx \frac{1}{n}\sum_{j=1}^n \mathrm{IF}(X_j; T, F) + \text{ remainder}, \tag{36}$$

or

$$\sqrt{n}(T_n(\hat{F}_n) - T(F)) \approx \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \mathrm{IF}(X_j; T, F) + \text{ remainder.} \tag{37}$$

Now $\{\mathrm{IF}(X_j; T, F)\}_{j=1}^n$ are n iid random variables with mean zero and finite variance, so their sum is asymptotically normal. If the remainder vanishes asymptotically (not that easy to verify, typically, but often true), then $\sqrt{n}(T_n(\hat{F}_n) - T(F))$ is also asymptotically normal, with asymptotic variance

$$V(T, F) = \int \mathrm{IF}(x; T, F)^2 dF(x). \tag{38}$$

When the relationship holds, an asymptotic form of the Cramér-Rao bound relates asymptotic efficiency to the influence function. Suppose T is Fisher consistent. Recall that the Fisher information at F_{θ_0} is

$$I(F_{\theta_0}) = \int \left(\frac{\partial}{\partial \theta} \ln f_\theta(x) \Big|_{\theta_0} \right)^2 dF_{\theta_0}. \tag{39}$$

The Cramér-Rao information inequality says that (under regularity conditions on $\{F_\theta\}_{\theta \in \Theta}$, including being dominated by a common measure μ, sharing a common support, and having densities $f_\theta(x)$ w.r.t. μ that are differentiable in θ) for any statistic T with $\mathbb{E}_\theta T^2 < \infty$ for which

$$\frac{d}{d\theta} \mathbb{E}_\theta T = \int \frac{\partial}{\partial \theta} T f_\theta \mu(dx), \tag{40}$$

$$V(T, F_\theta) \geq \frac{\left[\frac{\partial}{\partial \theta} \mathbb{E}_\theta T \right]^2}{I(F_\theta)}. \tag{41}$$

Thus T is asymptotically efficient only if

$$\mathrm{IF}(x; T, F) = I(F_{\theta_0})^{-1} \frac{\partial}{\partial \theta} (\ln f_\theta(x))|_{\theta_0}. \tag{42}$$

1.3.2 Relation to the Jackknife

There is an asymptotic connection between the jackknife estimate of variance and the influence function as well. Recall that for a functional statistic $T_n = T(\hat{F}_n)$, we define the jth pseudovalue by

$$T_{nj}^* = nT(\hat{F}_n) - (n-1)T(\hat{F}_{(j)}), \tag{43}$$

where $\hat{F}_{(j)}$ is the cdf of the data with the jth datum omitted. The jackknife estimate is

$$T_n^* = \frac{1}{n} \sum_{j=1}^{n} T_{nj}^*. \tag{44}$$

The (sample) variance of the pseudovalues is

$$V_n = \frac{1}{n-1}\sum_{j=1}^{n}(T_{nj}^* - T_n^*)^2. \tag{45}$$

The jackknife estimate of the variance of T_n is $\frac{1}{n}V_n$. Plugging into the definition 34 and taking $t = \frac{-1}{n-1}$ gives

$$\begin{aligned}\frac{n-1}{-1}\left(T\left((1-\frac{-1}{n-1})\hat{F}_n + \frac{-1}{n-1}\delta_{x_j}\right) - T(\hat{F}_n)\right) &= (n-1)[T(\hat{F}_n) - T(\hat{F}_{(j)})] \\ &= T_{nj}^* - T(\hat{F}_n). \end{aligned}\tag{46}$$

(Note that $(1-\frac{-1}{n-1})\hat{F}_n + \frac{-1}{n-1}\delta_{x_j} = \hat{F}_{(j)}$.) The jackknife pseudovalues thus give an approximation to the influence function.

1.3.3 Robustness measures defined from $\mathrm{IF}(x;T,F)$

The *gross error sensitivity of T at F* is

$$\gamma^* = \gamma^*(T,F) = \sup_{\{x:\mathrm{IF}(x;T,F) \text{ exists}\}} |\mathrm{IF}(x;T,F)|. \tag{47}$$

This measures the maximum change to T a small perturbation to F at a point can induce, which is a bound on the asymptotic bias of T in a neighborhood of F. If $\gamma^*(T,F) < \infty$, T is *B-robust* at F (B is for bias). For Fisher-consistent estimators, there is typically a minimum possible value of $\gamma^*(T,F)$, leading to the notion of *most B-robust* estimators. There is usually a tradeoff between efficiency and B-robustness: for a given upper bound on γ^*, there is a most efficient estimator.

The gross error sensitivity measures what can happen when the difference between what is observed and F can be anywhere (perturbing part of an observation by an arbitrary amount). There is a different notion of robustness related to changing the observed values slightly. The (infinitesimal) effect of moving an observation from x to y is $\mathrm{IF}(y;T,F) - \mathrm{IF}(x;T,F)$. This can be standardized by the distance from y to x to give the *local shift sensitivity*

$$\lambda^* = \lambda^*(T,F) = \sup_{\{x\neq y:\mathrm{IF}(x;T,F) \text{ and } \mathrm{IF}(y;T,F) \text{ both exist}\}} \frac{|\mathrm{IF}(y;T,F) - \mathrm{IF}(x;T,F)|}{|y-x|}. \tag{48}$$

This is the Lipschitz constant of the influence function. (A real function f with domain $\mathrm{dom}(f)$ is *Lipschitz continuous at x* if there exists a constant $C > 0$ such that $|f(x) - f(y)| \leq C|x-y|$

for all $y \in \mathrm{dom}(f)$. A real function f is Lipschitz continuous if it is Lipschitz continuous at every $x \in \mathrm{dom}(f)$. The *Lipschitz constant* of a Lipschitz-continuous function f is the smallest C such that $|f(x) - f(y)| \leq C|x - y|$ for all $x, y \in \mathrm{dom}(f)$. These definitions extend to functions defined on metric spaces, and to bounding $|f(x) - f(y)|$ by a multiple of $|x - y|^\nu$ with $\nu \neq 1$.)

Another measure of robustness involving the influence function is the *rejection point*

$$\rho^* = \rho^*(T, F) = \inf\{r > 0 : \mathrm{IF}(x; T, F) = 0, \ \ \forall |x| > r\}. \tag{49}$$

This measures how large an observation must be before the estimator ignores it completely. If very large observations are almost certainly gross errors, it is good for ρ^* to be finite.

1.3.4 Examples

Suppose $X \sim N(\theta, 1)$; $\theta \in \Theta = \mathbb{R}$; $\theta_0 = 0$; $F = \Phi$; $T_n(X_n) = \bar{X} = \frac{1}{n}\sum_{j=1}^n X_j$;

$$T(G) \equiv \int x dG(x). \tag{50}$$

The functional T is Fisher-consistent. Calculating IF gives

$$\begin{aligned} \mathrm{IF}(x; T, F) &= \lim_{t \to 0} \frac{\int u d((1-t)\Phi + t\delta_x)(u) - \int u d\Phi(u)}{t} \\ &= \lim_{t \to 0} \frac{(1-t)\int u d\Phi(u) + t \int u d\delta_x(u) - \int u d\Phi(u)}{t} \\ &= \lim_{t \to 0} \frac{tx}{t} \\ &= x. \end{aligned} \tag{51}$$

The Fisher information of the standard normal with unknown mean is $I(\Phi) = 1$, so

$$\int \mathrm{IF}^2(x; T, \Phi) d\Phi = 1 = I^{-1}(\Phi), \tag{52}$$

and IF $\propto (\partial/\partial\theta)(\ln f_\theta)|_0$. The arithmetic mean is the MLE and is asymptotically efficient. The gross-error sensitivity is $\gamma^* = \infty$, the local shift sensitivity is $\lambda^* = 1$, and the rejection point is $\rho^* = \infty$.

Suppose $\mathcal{X} = \{0, 1, 2, \ldots\}$, $\lambda(A) = \#A$, $\{X_j\}_{j=1}^n$ iid Poisson(θ), $\theta \in \Theta = (0, \infty)$. Then $f_\theta(k) = e^{-\theta}\theta^k/k!$. The MLE of θ is

$$\hat{\theta} \ = \ \arg \max_{\eta \in (0,\infty)} \prod_{j=1}^n f_\eta(X_j)$$

$$
\begin{aligned}
&= \arg\max_{\eta} e^{-n\eta} \prod_j \eta^{X_j}/X_j! \\
&= \arg\max_{\eta} e^{-n\eta}\eta^{n\sum_j X_j} \prod_j 1/X_j! \\
&\equiv \max_{\eta} L(\eta|\{X_j\}_{j=1}^n). \qquad (53)
\end{aligned}
$$

Stationarity requires

$$
\begin{aligned}
0 &= \partial/\partial\eta L(\eta|\{X_j\}_{j=1}^n)|_{\hat\theta} \\
&\propto e^{-n\hat\theta}\sum_j X_j\hat\theta^{\sum_j X_j - 1} - ne^{-n\hat\theta}\hat\theta^{\sum_j X_j}. \qquad (54)
\end{aligned}
$$

Which reduces to

$$
\sum_j X_j\hat\theta^{-1} - n = 0; \text{ i.e., } \hat\theta = \frac{1}{n}\sum_j X_j = \bar X. \qquad (55)
$$

This can be written as the functional estimator $T(\hat F_n)$ where T is defined by

$$
T(F) \equiv \int_{\mathcal{X}} xf(x)\lambda(dx) \qquad (56)
$$

This estimator is clearly Fisher consistent because

$$
T(F_\theta) = \sum_{k=0}^{\infty} ke^{-\theta}\theta^k/k! = \theta, \quad \forall\theta\in\Theta. \qquad (57)
$$

The influence function is defined only at points $x \in \mathcal{X}$. The functional T is linear, so

$$
\begin{aligned}
\mathrm{IF}(x;T,F_\theta) &\equiv \lim_{t\to 0}\frac{T((1-t)F_\theta + t\delta_x)) - T(F_\theta)}{t} \\
&= \lim_{t\to 0}\frac{(1-t)T(F_\theta) + tT(\delta_x) - T(F_\theta)}{t} \\
&= x - \theta, \quad x\in\mathcal{X}. \qquad (58)
\end{aligned}
$$

1.4 M-estimators

We shall consider in more detail what is needed for an M-estimator to be robust. An M-estimator is one of the form

$$
T_n(X) = \arg\min_{U_n}\sum_{j=1}^n \rho(X_j, U_n), \qquad (59)
$$

where ρ is a function on $\mathcal{X}\times\Theta$. M-estimators generalize maximum likelihood estimators. Maximum likelihood estimators minimize the negative log-likelihood, which for independent data would be of the form just given, with $\rho(X_j, U_n) = -\log f(x;\theta)$. If ρ is differentiable in θ, with

$$
\psi(x,\theta) = \frac{\partial}{\partial\theta}\rho(x,\theta), \qquad (60)
$$

then T_n is a stationary point, so it satisfies

$$\sum_{j=1}^{n} \psi(X_j, T_n) = 0. \tag{61}$$

An estimator that can be characterized as the solution of an equation in the form of either 59 or 61 is an *M-estimator*. The two are not quite equivalent: If ρ is not differentiable, there is no corresponding ψ. Conversely, ψ need not be the partial derivative of some ρ. An M-estimator is characterized by (and identified with) ρ or ψ.

Consider the second form, based on ψ. Define T implicitly as the functional for which

$$\int \psi(x, T(G))dG = 0 \tag{62}$$

for all G for which the integral exists; then the estimator is $T(\hat{F}_n)$. We calculate the IF at F of an M-estimator of this form. Define $F_{t,x} \equiv (1-t)F + t\delta_x$. We want

$$T' \equiv \lim_{t\to 0} \frac{T(F_{t,x}) - T(F)}{t}. \tag{63}$$

Assume we can interchange integration and differentiation; then we can differentiate 62 to get

$$\int \psi(x, T(F))d(\delta_x - F) + \int \left.\frac{\partial}{\partial\theta}\psi(x,\theta)\right|_{T(F)} dF \cdot \left.\frac{\partial}{\partial t}T(F_{t,x})\right|_{t=0} = 0. \tag{64}$$

This yields

$$\mathrm{IF}(x; \psi, F) = \frac{\psi(x, T(F))}{-\int \left.\frac{\partial}{\partial\theta}\psi(y,\theta)\right|_{T(F)} dF(y)}, \tag{65}$$

if the denominator is not zero. It turns out that even when ψ is not smooth, equation 65 holds.

The estimator given by ψ is B-robust iff $\psi(\cdot, T(F))$ is bounded. The asymptotic variance of the M-estimate defined by ψ is

$$V(T, F) = \frac{\int \psi^2(x, T(F))dF(x)}{\left(\int \left.\partial\psi(y,\theta)/\partial\theta\right|_{T(F)} dF(y)\right)^2}. \tag{66}$$

The maximum likelihood estimate (MLE) is a special case of an M-estimate, taking

$$\rho(x, \theta) = -\ln f_\theta(x). \tag{67}$$

The IF of the MLE is

$$\mathrm{IF}(x; T, F_\theta) = \frac{\partial \ln f_\gamma/\partial\gamma|_\theta}{V(T, F_\theta)}. \tag{68}$$

The asymptotic variance is in this case the reciprocal of the Fisher information $J(F_\theta)$, which shows the asymptotic efficiency of the MLE.

The *maximum likelihood scores function* $s(x, \theta)$ is

$$s(x, \theta) \equiv \left. \frac{\partial \ln f_\gamma(x)}{\partial \gamma} \right|_\theta ; \tag{69}$$

the MLE solves

$$\sum_{j=1}^{n} s(X_j, \mathrm{MLE}_n) = 0. \tag{70}$$

For estimates of location parameters, ψ typically depends on x and θ only through their difference: $\psi(x; \theta) = \psi(x - \theta)$. For the estimate to be Fisher-consistent, we need $\int \psi dF = 0$; then

$$\mathrm{IF}(x; F, T) = \frac{\psi(x - T(F))}{\int \psi'(x - T(F)) dF}. \tag{71}$$

The influence function is proportional to ψ. (We might as well take ψ to be the IF.)

Therefore, M-estimates have finite gross error sensitivity only if ψ is bounded, and have a finite rejection point only if ψ *redescends* to zero for large values of its argument. For the mean, this does not occur.

The asymptotic variance of such an M-estimate of a location parameter is

$$V(\psi, F) = \frac{\int \psi^2 dF}{(\int \psi' dF)^2}. \tag{72}$$

1.4.1 Robustness of M-estimates

See Huber, Ch3 for more details; this is drawn from there.

Let's calculate $b_1(\epsilon)$ for the Lévy metric for an M-estimate of location, with $\psi(x; t) = \psi(x - t)$, with ψ monotonically increasing. We will use λ to denote something new in this section, so let $d_L(F, G)$ denote the Lévy distance between distributions (or cdfs) F and G. Accordingly, we take $\mathcal{P}(\epsilon) \equiv \{G : d_L(F, G) \leq \epsilon\}$. Assume that $T(F) = 0$. Define

$$b_+(\epsilon) = \sup_{G \in \mathcal{P}(\epsilon)} T(G) \tag{73}$$

and

$$b_-(\epsilon) = \inf_{G \in \mathcal{P}(\epsilon)} T(G). \tag{74}$$

Then $b_1(\epsilon) = \max\{-b_-(\epsilon), b_+(\epsilon)\}$ (because $T(F) = 0$). Define

$$\lambda(t; G) = \mathbb{E}_G \psi(X - t) = \int \psi(x - t) dG(x). \tag{75}$$

Because ψ is monotonic, λ is decreasing in t, but not necessarily strictly decreasing, so the solution of $\lambda(t; G) = 0$ is not necessarily unique. Define

$$T^*(G) = \sup\{t : \lambda(t; G) > 0\} \tag{76}$$

and

$$T^{**}(G) = \inf\{t : \lambda(T; G) < 0\}. \tag{77}$$

Then $T^*(G) \leq T(G) \leq T^{**}(G)$. Note that $\lambda(t; G)$ increases if G is made stochastically larger. The stochastically largest element of $\{G : d_L(F, G) \leq \epsilon\}$ is

$$F_1(x) = (F(x - \epsilon) - \epsilon)_+ = \begin{cases} 0, & x \leq x_0 + \epsilon \\ F(x - \epsilon) - \epsilon & x > x_0 + \epsilon, \end{cases} \tag{78}$$

where x_0 solves $F(x_0) = \epsilon$. (Assume that x_0 exists; the discontinuous case introduces some additional bookkeeping.) Note that this distribution puts mass ϵ at $x = \infty$. For $G \in \mathcal{P}(\epsilon)$,

$$\lambda(t; G) \leq \lambda(t; F_1) = \int_{x_0}^{\infty} \psi(x - t + \epsilon) dF(x) + \epsilon\psi(\infty). \tag{79}$$

It follows that

$$\begin{aligned} b_+(\epsilon) &= \sup_{G \in \mathcal{P}(\epsilon)} T(G) \\ &= T^{**}(F_1) \\ &= \inf\{t : \lambda(t; F_1) < 0\}. \end{aligned} \tag{80}$$

Note that

$$\ell_+ = \lim_{t \to \infty} \lambda(t; F_1) = (1 - \epsilon)\psi(-\infty) + \epsilon\psi(\infty). \tag{81}$$

Provided $\ell_+ < 0$ and $\psi(\infty) < \infty$, $b_+(\epsilon) < b_+(1) = \infty$. Thus to avoid breakdown from above we need

$$\frac{\epsilon}{1 - \epsilon} < -\frac{\psi(-\infty)}{\psi(\infty)}. \tag{82}$$

We can calculate $b_-(\epsilon)$ in the same way: the stochastically smallest element of $\mathcal{P}_(\epsilon)$ is

$$F_{-1} = (F(x+\epsilon)+\epsilon) \wedge 1 = \begin{cases} F(x+\epsilon)+\epsilon, & x \le x_1 - \epsilon \\ 1, & x > x_1 - \epsilon, \end{cases} \tag{83}$$

where x_1 solves $F(x_1) = 1-\epsilon$. This distribution assigns mass ϵ to $x = -\infty$. We have

$$\lambda(t;G) \ge \lambda(t;F_{-1}) = \epsilon\psi(-\infty) + \int_{-\infty}^{x_1} \psi(x-t-\epsilon)dF(x). \tag{84}$$

Thus

$$\begin{aligned} b_-(\epsilon) &= \inf_{G\in\mathcal{P}(\epsilon)} T(G) \\ &= T^*(F_{-1}) \\ &= \sup\{t : \lambda(t;F_{-1}) > 0\}. \end{aligned} \tag{85}$$

Note that

$$\ell_- = \lim_{t\to-\infty} \lambda(t;F_{-1}) = \epsilon\psi(-\infty) + (1-\epsilon)\psi(\infty). \tag{86}$$

To avoid breakdown from below, we need $\psi(-\infty) > -\infty$ and $\ell_- > 0$, which leads to

$$\frac{\epsilon}{1-\epsilon} > -\frac{\psi(\infty)}{\psi(-\infty)}. \tag{87}$$

Combining this with 82 gives

$$-\frac{\psi(\infty)}{\psi(-\infty)} < \frac{\epsilon}{1-\epsilon} < -\frac{\psi(-\infty)}{\psi(\infty)}. \tag{88}$$

Define

$$\eta \equiv \min\left\{-\frac{\psi(-\infty)}{\psi(\infty)}, -\frac{\psi(\infty)}{\psi(-\infty)}\right\}. \tag{89}$$

The breakdown point is then

$$\epsilon^* = \frac{\eta}{1+\eta}. \tag{90}$$

The maximum possible value $\epsilon^* = 1/2$ is attained if $\psi(\infty) = -\psi(-\infty)$. The breakdown point is $\epsilon^* = 0$ if ψ is unbounded.

This calculation also shows that if ψ is bounded and $\lambda(t;F)$ has a unique zero at $t = T(F)$, then T is continuous at F; otherwise, T is not continuous at F.

Things are much more complicated for non-monotone functions ψ, including "redescending influence functions," which we shall examine presently.

1.4.2 Minimax Properties for location estimates

It is straightforward to find the minimax bias location estimate for symmetric unimodal distributions; the solution is the sample median (see Huber, §4.2). Minimizing the maximum variance is somewhat more difficult. Define

$$v_1(\epsilon) = \sup_{G \in \mathcal{P}(\epsilon)} A(G, T), \tag{91}$$

where $A(G, T)$ is the asymptotic variance of T at G. Assume the observations are iid $G(\cdot - \theta)$. The shape varies over the family $\mathcal{P}(\epsilon)$; the parameter varies over the reals. Such families are not typically compact in the weak topology. Huber uses the *vague* topology to surmount this problem. The *vague* topology is the weakest topology on the set $\mathcal{M}_+$ of sub-probability measures for which $F \to \int \psi dF$ is continuous for all continuous functions ψ with compact support. (A subprobability measure can have total mass less than one, but is otherwise the same as a probability measure.) Because $\mathbb{R}$ is locally compact, $\mathcal{M}_+$ is compact.

Define F_0 to be the distribution in $\mathcal{P}(\epsilon)$ with smallest Fisher information

$$I(G) = \sup_{\psi \in \mathcal{C}_K^1} \frac{(\int \psi' dG)^2}{\int \psi^2 dG}, \tag{92}$$

where $\mathcal{C}_K^1$ is the set of all compactly supported, continuously differentiable functions ψ s.t. $\int \psi^2 dF > 0$. This extends the definition of the Fisher information beyond measures that have densities; in fact, $I(F) < \infty$ iff F has an absolutely continuous density w.r.t. Lebesgue measure, and $\int (f'/f)^2 f dx < \infty$.

Proof. (Following Huber, pp78ff.) By assumption, $\int \psi^2 dx < \infty$. If $\int (f'/f)^2 f dx < \infty$,

$$\begin{aligned}
\left(\int \psi' f dx \right)^2 &= \left(\psi f|_{-\infty}^{\infty} - \int \psi \frac{f'}{f} f dx \right)^2 \\
&= \left(\int \psi \frac{f'}{f} f dx \right)^2 \\
&\leq \left(\int \psi^2 f dx \right) \left(\int \left(\frac{f'}{f} \right)^2 f dx \right),
\end{aligned} \tag{93}$$

by the weighted Cauchy-Schwarz inequality. Thus $I(F) \leq \int (f'/f)^2 f dx < \infty$. Now suppose $I(F) < \infty$. Then $L(\psi) = -\int \psi' dF$ is a bounded linear functional on the (dense) subset $\mathcal{C}_K^1$ of $L_2(F)$, the Hilbert space of square-integrable functions w.r.t. F. By continuity, L can be extended to a

continuous linear functional on all of $L_2(F)$. By the Riesz Representation Theorem, there then exists a function $g \in L_2(F)$ such that for all $\psi \in L_2(F)$,

$$L\psi = \int \psi g dF. \tag{94}$$

Clearly, $L1 = \int g dF = 0$. Define

$$f(x) \equiv \int_{y<x} g(y)F(dy) = \int 1_{y<x} g(y)F(dy). \tag{95}$$

By the Cauchy-Schwarz inequality,

$$|f(x)|^2 \leq (\int 1^2_{y<x} F(dy))(\int g(y)^2 F(dy)) = F((-\infty, x)) \int g^2 F(dy), \tag{96}$$

which tends to zero as $x \to -\infty$; $|f(x)|$ also tends to zero as $x \to \infty$. For $\psi \in \mathcal{C}^1_K$,

$$-\int \psi'(x)f(x)dx = -\int_{y<x} \int \psi'(x)g(y)F(dy)dx = \int \psi(y)g(y)F(dy) = L\psi \tag{97}$$

by Fubini's theorem. Thus the measure $f(x)dx$ and the measure $F(dx)$ give the same linear functional on derivatives of functions in $\mathcal{C}^1_K$. This set is dense in $L_2(F)$, so they define the same measure, and so f is a density of F. We can now integrate the definition of the functional L by parts to show that

$$L\psi = -\int \psi' f dx = \int \psi \frac{f'}{f} f dx. \tag{98}$$

Thus

$$I(F) = \|L\|^2 = \int g^2 dF = \int \left(\frac{f'}{f}\right)^2 f dx. \tag{99}$$

The functional $I(G)$ is lower-semicontinuous with respect to the vague topology, so $I(G)$ attains its infimum on any vaguely compact set. Furthermore, $I(G)$ is a convex function of G.

Theorem 8 *(Huber, Proposition 4.5) Let $\mathcal{P}$ be a set of measures on $\mathbb{R}$. Suppose $\mathcal{P}$ is convex, $F_0 \in \mathcal{P}$ minimizes $I(G)$ over $\mathcal{P}$, $0 < I(F_0) < \infty$, and the set where the density f_0 of F_0 is strictly positive is (a) convex and (b) contains the support of every distribution in $\mathcal{P}$.*

Then F_0 is the unique minimizer of $I(G)$ over $\mathcal{P}$.

The reciprocal of $I(F_0)$ lower-bounds the (worst) asymptotic variance of any estimator over all $G \in \mathcal{P}$, so if one can find an estimator whose asymptotic variance is $1/I(F_0)$, it is minimax (for asymptotic variance).

Finding F_0 can be cast as a variational problem; see Huber, §4.5.

The least-informative distributions in neighborhoods of the normal tend to have thinner tails than the normal. If one believes that outliers might be a problem, it makes sense to abandon minimaxity in favor of estimators that do somewhat better when the truth has thicker tails than the normal. That leads to considering *redescending influence functions*, for which $\psi = 0$ for x sufficiently large.

One can develop "minimax" estimators in this restricted class. For example, we could seek to minimize the asymptotic variance subject to $\psi(x) = 0$, $|x| > c$. For the ϵ-contamination neighborhood of a normal, the minimax ψ in this class is

$$\psi(x) = -\psi(-x) = \begin{cases} x, & 0 \le x \le a \\ b \tanh\left(\frac{b(c-x)}{2}\right) & a \le x \le c \\ 0, & x \ge c. \end{cases} \tag{100}$$

The values of a and b depend on ϵ.

Other popular redescending influence functions include Hampel's piecewise linear influence functions:

$$\psi(x) = -\psi(-x) = \begin{cases} x, & 0 \le x \le a \\ a & a \le x \le b \\ a\frac{c-x}{c-b} & b \le x \le c \\ 0, & x \ge c, \end{cases} \tag{101}$$

and Tukey's "biweight"

$$\psi(x) = \begin{cases} x(1-x^2)^2, & |x| \le 1 \\ 0, & |x| > 1. \end{cases} \tag{102}$$

A complication in using redescending influence functions is that scaling (some form of Studentizing) is much more important for them to be efficient than it is for monotone influence functions. The slope of the influence function in the descending regions also can inflate the asymptotic variance (recall that $(\int \psi' dF)^2$ is in the denominator of $A(F,T)$).

1.5 Estimates of Scale

We require that a scale estimate S_n be equivariant under changes of scale, so that

$$S_n(aX) = aS_n(X), \quad \forall a > 0. \tag{103}$$

It is common also to require that a scale estimate be invariant under sign changes and translations, so that

$$S_n(-X) = S_n(X) = S_n(X + b\mathbf{1}), \tag{104}$$

where $b \in \mathbb{R}$ and $\mathbf{1}$ is an n-vector of ones. The most common need for a scale estimate is to remove scale as a nuisance parameter in a location estimate, by Studentizing.

It turns out that the bias properties of a scale estimate are more important for studentizing than the variance properties. That leads to considering something involving the median deviation. The most widely used robust scale estimator is the median absolute deviation (MAD). Let $M_n(x)$ be the median of the list of the elements of x: $\text{med}\{x_j\}_{j=1}^n$. Then

$$\text{MAD}_n(x) = \text{med}\{|x_j - M_n(x)|\}_{j=1}^n. \tag{105}$$

The breakdown point of the MAD is $\epsilon^* = 1/2$. Typically, the MAD is multiplied by 1.4826 $(1/\Phi^{-1}(3/4))$ to make its expected value unity for a standard normal.

1.6 Robust Regression

This section follows Hampel *et al.* (1986), Chapter 6, rather closely. Suppose we have a linear statistical model: we observe $\{(X_j, Y_j)\}_{j=1}^n$ iid with $Y_j \in \mathbb{R}$, $X_j \in \mathbb{R}^p$ $(p \geq 1)$, and

$$Y = X\theta + \epsilon, \tag{106}$$

where $Y = (Y_j)_{j=1}^n$, X is an n by p matrix with entries $X_{jk} = (X_j)_k$, $1 \leq j \leq n$, $1 \leq k \leq p$, $\theta \in \Theta \subset \mathbb{R}^p$ is unknown, Θ is an open convex subset of $\mathbb{R}^p$, and $\epsilon \in \mathbb{R}^n$ is a vector of iid noise with mean zero. We assume that ϵ_j is independent of X_j, that the common distribution of the noise components is symmetric about zero and continuous. Let K be the common distribution of $\{X_j\}$, and let k be the density of K with respect to Lebesgue measure on $\mathbb{R}$. The joint density of (X_j, Y_j) with respect to Lebesgue measure on $\mathbb{R}^2$ is

$$f_\theta(x, y) = \sigma^{-1} g\left((y - x\theta)/\sigma\right) k(x). \tag{107}$$

Let $F_\theta(x, y)$ be the corresponding measure. For now, we take X and Y to be random (we don't condition on X).

The least-squares (LS) estimate of θ is

$$\hat{\theta}_{LS} \equiv \arg\min_{\gamma \in \Theta} \sum_{j=1}^{n} \left((y_j - x_j\gamma)/\sigma \right)^2 . \tag{108}$$

The least squares estimate minimizes the 2-norm of the residuals. The Gauss-Markov theorem says that if $\mathbb{E}\epsilon_j = 0$, $1 \le j \le n$, and $cov(\epsilon) = \sigma^2 I$, then $b\hat{\theta}_{LS}$ is the unique minimum variance unbiased linear estimator of any estimable linear functional $b\theta$ of θ. The restriction to linear estimators is severe—for example, the MLE is nonlinear for many distributions. When the error distribution is normal, the minimum variance estimate is in fact linear, but otherwise, there can be a large loss of efficiency in restricting to linear estimators.

Moreover, a single outlier can influence the least-squares estimate arbitrarily—the least-squares estimate is neither robust nor resistant. (See `http://www.stat.berkeley.edu/users/stark/Java/Correlatio` for an interactive demonstration.)

Huber (1973, Robust regression: asymptotics, conjectures, and Monte Carlo. *Ann. Stat.*, *1*, 799–821.) proposed using least squares with weights on the residuals computed iteratively to reduce the influence of extreme observations. The weights are

$$w_j = \min(1, c/|r_j|), \tag{109}$$

where c is a positive constant and r_j is the jth residual. This is a special case of the estimator

$$\hat{\theta}_\rho \equiv \arg\min_{\gamma \in \Theta} \sum_{j=1}^{n} \rho\left((y_j - x_j\gamma)/\sigma \right), \tag{110}$$

where ρ is a nonnegative function. The case $\rho(r) = r^2/2$ is least squares. The case $\rho(r) = \rho_c(r) = wr^2$ yields the MLE when the errors have density proportional to $\exp(-\rho_c(r))$.

The case $\rho(r) = |r|$ is minimum L_1 estimation, which is resistant. Minimum L_1 regression can also be performed using linear programming. We shall use linear programming in our discussion of density estimation later, so I'll take a bit of space here to set up the L_1 regression problem as a linear program.

Linear programming solves the problem

$$\min_{z: z \ge 0, Az = b} \ell z, \tag{111}$$

where $z, \ell \in \mathbb{R}^k$, A is an m by k matrix, and $b \in \mathbb{R}^m$. (There are other canonical forms for linear programs, such as $\min_{z: Az \le b} \ell z$.) Linear programming and its infinite-dimensional extensions play a central role in game theory and minimax statistical estimation.

Linear programming involves minimizing a convex functional over a convex set, so any local minimum attains the global minimal value. However, neither the constraint set nor the objective functional is strictly convex, so in general more than one parameter vector attains the global minimal value of the objective functional. The nonnegativity constraint and the linear equality constraints restrict the solution to lie within a convex *polytope*, which is analogous to a convex polygon in $\mathbb{R}^2$: it is the intersection of a collection of closed half-spaces. The fundamental theorem of linear programming says that if the linear program is consistent (if there is some vector z that satisfies all the constraints), then some vertex of the polytope attains the global minimal value of ℓ.

We seek to cast the problem

$$\min_{\gamma \in \mathbb{R}^p} \sum_{j=1}^{n} |X_j \gamma - y_j| \tag{112}$$

as a linear program. The minimizing γ is the L_1 regression estimate of θ. Define

$$A = [X \quad -X \quad I \quad -I], \tag{113}$$

$$b = y, \tag{114}$$

$$z = [f \quad g \quad d \quad e]^T, \tag{115}$$

(with $f, g \in \mathbb{R}^p$ and $d, e \in \mathbb{R}^n$), and

$$\ell = [\mathbf{0} \quad \mathbf{1}] \tag{116}$$

(with $\mathbf{0} \in \mathbb{R}^{\mathbf{2p}}$ and $\mathbf{1} \in \mathbb{R}^{\mathbf{2n}}$). Then

$$Az = b \Longleftrightarrow X(f - g) + d - e = y, \tag{117}$$

and minimal value of ℓz minimizes

$$\sum_{j=1}^{n} (d_j + e_j). \tag{118}$$

Note that if, for some j, both d_j and e_j are greater than zero, then replacing the smaller of the two with 0 and the larger of the two with $|d_j - e_j|$ reduces ℓz but does not change the value of Az. Thus for the optimal z, either $d_j = 0$ or $e_j = 0$ (or both) for each j. Therefore, at the optimum, $|d_j - e_j| = d_j + e_j$; i.e.,

$$|X_j(f - g) - y_j| = d_j + e_j, \tag{119}$$

and

$$\sum_{j=1}^{n} |X_j(f - g) - y_j| = \ell z. \tag{120}$$

The L_1 regression estimate of θ is clearly $\hat{\theta}_{L_1} = f - g$, where f and g are the first p and second p components of z^*, the solution to the linear programming problem above.

Linear programming also can solve L_∞ regression problems:

$$\min_{\gamma} \max_{j} |X_j\gamma - y_j|. \tag{121}$$

L_∞ regression is extremely non-resistant.

The most popular algorithm for solving linear programs is the simplex algorithm. The simplex algorithm starts with a feasible point, then moves from vertex to vertex of the polytope defined by the linear constraints, continually decreasing the value of ℓz. (The fundamental theorem of linear programming guarantees that there is a vertex that attains the optimal value.) At each vertex, the algorithm checks the Kuhn-Tucker conditions, which say that at an optimal solution, every direction of descent points out of the constraint set.

We continue with the discussion of robust regression for a more general weight function ρ. If ρ is differentiable with derivative $\partial\rho(r)/\partial r = \psi(r)$, then $\hat{\theta}_\rho$ can be characterized as a stationary point

$$\sum_{j=1}^{n} \psi\left((y_j - x_j\hat{\theta}_\rho)/\sigma\right) x_j = 0. \tag{122}$$

(Note that for L_1 regression, ρ is not differentiable at 0.)

Typically, σ is not known. It could be estimated together with θ; I think it is more common to use the MAD of the residuals, estimated iteratively for a few iterations, then "frozen" for the rest of the iterations in the estimation of θ.

For the Huber estimate using the weight function w, the influence of the residual is bounded, but the influence of the x-coordinate of a datum is not bounded. (See Fig. 1, p. 314, in *Hampel et al.*) If we want to use the errors-in-variables model, that should be considered as well. The influence function can be factored into the influence of the residual (IR) and the influence of position in factor space (IP)–analogous to the leverage of the point.

Hampel et al. (1986, ch. 6.3, p. 315) define an M-estimator T_n for linear models as the solution to the equation

$$\sum_{j=1}^{n} \eta\left(X_j, (y_j - X_jT_n)/\sigma\right) x_j = 0, \tag{123}$$

where $\eta : \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}$ is suitably well behaved:

1. For every $x \in \mathbb{R}^p$, $\eta(x, \cdot)$ is continuous except possibly at a finite set, where it has finite left and right limits.

2. For every $x \in \mathbb{R}^p$, $\eta(x, r)$ is an odd function of r, and is nonnegative for nonnegative r.

3. For every $x \in \mathbb{R}^p$, the set of points at which $\eta'(x, r) \equiv \partial\eta(x, r)/\partial r$ fails to exist or fails to be continuous, is a finite set.

4. $M \equiv \mathbb{E}\eta'(x, r)xx^T$ exists and is a nonsingular matrix.

5. $Q \equiv \mathbb{E}\eta^2(x, r)xx^T$ exists and is a nonsingular matrix.

This is more restrictive than the general M-estimation set up: the function $\psi(X_j, y_j, \theta)$ that this implicitly defines has the same direction as X_j, and depends on y_j only through the residual $y_j - X_j\theta$.

According to *Hampel et al.*, the functions η that have been proposed in the literature all factor into a weight that depends on x and function of the residual weighted by a different function of x:

$$\eta(x, r) = w(x)\psi(rv(x)), \tag{124}$$

where $w : \mathbb{R}^p \to \mathbb{R}^+$, $v : \mathbb{R}^p \to \mathbb{R}^+$, and $\psi : \mathbb{R} \to \mathbb{R}$. Three different types of η functions are sketched on p. 322 of *Hampel et al.*:

1. The Huber estimator discussed above takes $w(x) = v(x) = 1$:

$$\sum_{j=1}^{n} \psi_c(r_j)X_j = 0. \tag{125}$$

   This estimator does not limit the influence of position.

2. The Mallows estimator takes $v(x) = 1$:

$$\sum_{j=1}^{n} \psi_c(r_j)w_jX_j = 0. \tag{126}$$

   This estimator limits the leverage of a datum regardless of the size of the residual at that design point.

3. The Hampel-Krasker estimator takes $w(x) = 1/v(x)$:

$$\sum_{j=1}^{n} \psi_c(r_j/u_j)u_jX_j = \sum_{j=1}^{n} \psi_{cu_j}(r_j)X_j = 0. \tag{127}$$

Figure 2 on p. 323 of *Hampel et al.* illustrates the difference in the fitted regression function for these three estimators for the bivariate case with one outlier with large leverage.

Below is a collection of Matlab routines that solve iteratively reweighted least squares using robust estimates of scale. The routines do not bound the influence of position; they just bound the influence of the residuals. (The extension to weight by position is straightforward.) Included are Matlab routines to calculate Tukey's Biweight and the weight function for minimum L_1 regression by reweighting. At the end is a routine for minimum L_1 regression using linear programming. I wrote this code in 1997 and would write it differently now ...

```
function [ b, res, wssr, it ] = irwls(y,X,wfunc,param,start)
% irwls    iteratively re-weighted least squares
%          Can solve "robust" regression problems using
%          robust weights.
% Usage:
%        [b, res, wssr, it ] = irwls(y,X,wfunc,param,start)
%
% Arguments:
%       y  n-vector of data
%       X      nxp design matrix
%       wfunc  the name of the weight function subroutine, passed as a string
%                W = wfunc(res, scale, param) must take three arguments:
%           res     vector of residuals
%           scale   resistant estimate of scale (MAD/.6745)
%           param   array of particular parameters passed by caller
%           wfunc must return a vector W of the same dimension as res
%       param  a vector of parameters to pass to wfunc.
%           DEFAULT = 1.
%       start  a flag for what to use as a starting model to get
%       the estimate of scale that is passed to the weight
%       function. To get the estimate of scale, the algorithm
```

```
%       uses the median absolute deviation (MAD) of the residuals,
%       standardized to that of the normal distribution.
%       If start = 'ols', the median absolute deviation of
%       the residuals from the ordinary least-squares regression
%       function are used.
%       If start = 'median', the median absolute deviation
%       of the residuals from the median are used, completely
%       ignoring the possible variation of the "true" function.
%       If start = 'one-step', the first estimate of scale
%       is the same as for 'median', but the scale estimate is
%       updated based on the residuals from the first step
%       of the least squares fit using that scale (the scale estimate
%       is only updated once, after the first wls).
%       DEFAULT = 'ols'
%
% Output:
%       b      p-vector of fitted coefficients
%       res    n-vector of final residuals
%       wssr   final sum of squared weighted residuals
%       it     number of iterations taken

% P.B. Stark    stark@stat.berkeley.edu
% 9 July 1997

chgTol = 0.001;                 % tolerance for convergence
                                % of coefficients
maxit = 500;                    % maximum number of iterations

% check dimensions
    [n,p] = size(X);
```

```
[ny, py] = size(y);

if (n ~= ny),
    error('number of rows of y ~= number of rows of X')
end

if (py ~= 1),
    error('y has more than one column')
end

if (nargin < 3),
    param = 1;
    start = 'ols';
elseif (nargin < 4),
    start = 'ols';
end

if (~(strcmp(start,'ols') | strcmp(start,'median') | ...
    strcmp(start,'one-step'))),
    error(['start = ', start, ' not supported'])
end

%
if (strcmp(start,'ols')),
    W = ones(n,1);              % initial weight matrix
    [b, res, wssr] = wls(y,X,W);        % initial regression
    scale = 1.483*median(abs(res-median(res)));
elseif (strcmp(start,'median')),
    res = y - median(y);
    scale = 1.4843*median(abs(res-median(res)));
```

```
        W = diag(sqrt(feval(wfunc,res,scale,param)));
        [b, res, wssr] = wls(y, X, W);
    elseif (strcmp(start,'one-step')),
        res = y - median(y);
        scale = 1.4843*median(abs(res-median(res)));
        W = diag(sqrt(feval(wfunc,res,scale,param)));
        [b, res, wssr] = wls(y, X, W);
        scale = 1.4843*median(abs(res-median(res)));
    end

% loop
    relchg = 1;                         % relative change in coefficients
    it = 0;
    while (relchg > chgTol & it < maxit),       % stop when the relative change in
                                        % regression coefficients between
                                        % iterations is small enough
        lastB = b;
        it = it + 1;
            W = diag(sqrt(feval(wfunc,res,scale,param)));
        [b, res, wssr] = wls(y, X, W);
        relchg = max( abs( b - lastB ) ./ abs(b));
    end
    if (it == maxit ),
        error(['maximum number of iterations (' num2str(maxit) ...
             ') exceeded'])
    end
return


function [b, res, wssr] = wls(y,X,W)
```

```
%  wls  Multiple linear regression using weighted least squares.
%
%  Usage:
%   [b, res, wssr] = wls(y,X,W) returns the p-vector b of
%       regression coefficients such that
%           (Xb - y)' W'W (Xb - y)    (**)
%       is minimal, along with the residual n-vector
%           res = y - Xb,
%       and the weighted sum of squared residuals
%           wssr = (Xb - y)'  W  (Xb - y).
%  Arguments:
%   y   n-vector of observations
%   X   nxp design matrix
%   W   nxn positive semi-definite square matrix of weights,
%       or a vector of nonnegative weights.
%       Note that W is the square-root of what one would think
%       of as the data error covariance matrix.
%
%
%  Algorithm:   Uses the QR decomposition as follows.
%       Define Z = WX and d = Wy. Then (**) can
%       be written as an OLS (ordinary least squares)
%       problem: minimize
%           || Zb - d ||^2.
%       If [Q, R] is the QR decomposition of Z,
%       stationarity yields
%           2Z'(Zb - d) = 0,
%       or  Z'Zb = Z'd.
%       The QR decomposition gives
%           R'Rb = R'Q'd; for nonsingular R,
```

```
%           Rb = Q'd.
%       Matlab can solve this triangular system stably using
%           b = R\(Q'd).
%

    if  nargin < 3,
        error('wls requires three input arguments.');
    end

% Check that X, y, and W have compatible dimensions
    [n,p] = size(X);
    [ny,py] = size(y);
    [nW,pW] = size(W);
    dimW = max(nW, pW);
    minDimW = min(nW, pW);
    %
    if (n ~= ny),
        error( 'The number of rows in y must equal the number of rows in X' );
    elsif ( py ~= 1 ),
        error( 'y must be a vector, not a matrix' );
    elsif ( minDimW > 1 & nW ~= pW ),
        error( 'W must be a square matrix or a vector' );
    elsif ( dimW ~= n ),
        error( 'W must have the same number of rows as X and y' );
    elsif ( minDimW == 1 & min(W) < 0 ),
        error( 'W must be nonnegative' );
    end

% build the weighted design matrix and data
```

```
    if ( minDimW == 1),
        W = diag(W);        % in case W is a vector
    end

    Z = W*X;
    d = W*y;

    [Q, R] = qr(Z, 0);
    b = R\(Q'*d);

    yHat = X*b;             % predicted data values
    res = y - yHat;         % residuals
    wssr = norm(W*res);     % weighted sum of squared residuals

return;


function W = l1wght(res, scale, param)
% function W = l1wght(res, scale, param)
%
% computes the weight function for minimum l_1 regression
% for iteratively reweighted least squares: scale/|res|
%
% arguments:
%       res:    vector of residuals
%       scale:  robust estimate of scale such as MAD
%       param:  not used--present only for consistent
%           calling signature with other weights
%
% returns:
```

```
%       W:  the vector of l_1 weights
%
%
% P.B. Stark  stark@stat.berkeley.edu
% 11 July 1997.
    thresh = eps^(1/3);

    unit = ones(size(res))*thresh;
    res(abs(res) <= thresh ) = unit(abs(res) <= thresh);
    W = (abs(res)).^(-1);
return;

function W = biwght(res, scale, param)
% function W = biwght(res, scale, param)
%
% computes Tukey's biweight function for robust regression
%
% biwght(x) =   (1-x^2/param^2)^2 , |x| < param; 0, |x| >= param.
%
% arguments:
%       res:    vector of residuals
%       scale:  robust estimate of scale, such as MAD
%       param:  parameter of the biweight function.
%           8 is a reasonable number, and is the
%           default.
%
% returns:
%       W:  the vector of biweight weights
%
%
```

```
% P.B. Stark  stark@stat.berkeley.edu
% 11 July 1997.

    p = 8;
    if (nargin == 3),
        p = param;
    end

    res = res/scale;
    W = (ones(size(res)) - res.^2/p^2).^2 .* (abs(res) < p);

return;
```

Here is an implementation of L_1 regression using Matlab's linear programming routine. Note that this requires the Optimization Toolbox.

```
function [ b, res, msft ] = l1rgres(y, X, verb, alg)
% l1rgres   minimum l_1 misfit regression using linear programming
%
%   WARNING: Matlab's lp implementation is unstable!
%
% Usage:
%        [ b, res, msft ] = l1rgres(y, X, verb, alg)
%
% Input:
%        y  n-vector of data
%        X      nxp design matrix
%        verb   if verb = -1, LP runs silently.
%        alg    string for the linear programming algorithm to use.
```

```
%            If alg == 'lp', uses the MATLAB Optimization toolbox
%            LP.M routine.
%            Defaults to 'lp'.
%
%
% Output:
%        b      p-vector of fitted coefficients
%        res    n-vector of final residuals
%        msft   l_1 misfit of solution
%
% P.B. Stark    stark@stat.berkeley.edu
% 10 July 1997



%
% check dimensions
    [n,p] = size(X);
    [ny, py] = size(y);

    if (n ~= ny),
        error('number of rows of y ~= number of rows of X')
    end

    if (py ~= 1),
        error('y has more than one column')
    end

% output level?
    talk = 0;
    if (nargin >= 3),
```

```
        talk = verb;
    end;

%
% set up slack variables and their bounds
%
% want  to minimize
%   sum_{j=1}^n ( pos_j + neg_j )
% subject to
%   y = Ab + pos - neg
%   0 <= pos <= infty
%   0 <= neg <= infty
%
% what to do depends on the linear programming algorithm used.
%
    lpalg = 'lp';
    if (nargin == 4),
        lpalg = alg;
        if (~strcmp(lpalg,'lp') ),
            error(['1: the algorithm ' lpalg ' is not supported'])
        end;
    end;

%
% use ols to get a cheap feasible starting model
    [xols, olsres, rssols] = wls(y, X, ones(n,1));

% the residuals give the values of the slacks
    pos = zeros(size(olsres));
    neg = pos;
```

```
    pos(olsres > 0) = olsres(olsres > 0);
    neg(olsres < 0) = -olsres(olsres < 0);
    clear olsres;               % free the memory

% now the lp phase
    if (talk >= 0),
        disp('starting LP phase')
    end

% cases, by LP algorithm invoked
%
    if (strcmp(lpalg,'lp')),
        lpn = p + 2*n;          % number of variables, including slacks
        lpm = n;            % total number of constraints
        neq = n;            % number of equality constraints
        f = [zeros(1,p) ones(1,2*n)];   % objective vector (a row vector)
        vlb = [-inf * ones(p,1); ...
                zeros(2*n,1) ];     % lower bounds on variables
        vub = inf * ones(lpn,1);    % upper bounds
        A = [X ones(n) -ones(n)];   % add the slacks to the constraints
                        % this gets HUGE quickly!
        x0 = [xols; pos; neg];
        lpx = lp(f, A, y, vlb, vub, x0, neq, talk);
        msft = f * lpx;
        b = lpx(1:p);           % extract solution
        res = y - X*b;
    else
        error(['impossible error: algorithm ', alg, ' not supported'])
    end;
```

```
return
```