WI

9209

UNIVERSITY
OF
WISCONSIN-
MADISON

SSRI

Identification and Robustness in the Presence of Errors in Data

9209

Joel L. Horowitz
Charles F. Manski

SOCIAL SYSTEMS RESEARCH INSTITUTE

IDENTIFICATION AND ROBUSTNESS IN THE PRESENCE OF ERRORS IN DATA

by

Joel L. Horowitz
Department of Economics
University of Iowa
Iowa City, IA 52242

and

Charles F. Manski
Department of Economics
University of Wisconsin
Madison, WI 53706

May 1992

ABSTRACT

Robust estimation aims at developing point estimators that are not highly sensitive to errors in the data. However, the population parameters of interest are not identified under the assumptions of robust estimation, so the rationale for point estimation is not apparent. This paper shows that under the assumptions of robust estimation, population parameters can be bounded, even though they are not identified. Several features of the bounds are related to the breakdown point and gross-error sensitivity of robust estimation. A method for estimating the bounds is given and illustrated with an application to data on the distribution of household incomes in the U.S. It is argued that in the presence of errors in data, it is more natural to estimate the bounds than to attempt point estimation of unidentified parameters.

We thank Marianthi Markatou for comments on a previous draft. The research of Joel L. Horowitz was supported in part by NSF grant no. SES-8922460. The research of Charles F. Manski was supported in part by NSF grant no. SES-8808276.

IDENTIFICATION AND ROBUSTNESS IN THE PRESENCE OF ERRORS IN DATA

1. Introduction

Inference in the presence of errors in the data is problematic because the sampling process does not identify the probability distribution of interest. As with other identification problems, it is natural to analyze the errors-in-data problem in two stages: first determine which features of the relevant population are identified given the available information, and then develop methods for estimating the identified features.

One of the main approaches to the errors-in-data problem, robust estimation, follows a different analytical strategy. Studies of robustness aim at characterizing how point estimators of population parameters behave in the presence of errors in the data. The main objective is to find point estimators that are not greatly affected by errors. There is no attempt to undertake the separate though related task of determining what information about the parameters is available from a sampling process that produces data errors and how that information can be extracted.

In this paper, we address the identification problem directly. Our aim is to find out what can be learned about parameters of interest when the sampling process generates erroneous observations. We show that under the error-generation models used in robust estimation, it is often possible to obtain informative bounds on the values of unidentified population parameters. These bounds exhaust the information about the parameters that is available from the sampling process. We also show that estimating the bounds consistently is often very easy.

Our main assumption is that the analyst has *a priori* knowledge of an upper bound on the probability of data errors. This assumption is implicit in all robust estimation. Bounded-influence estimation assumes implicitly that the probability of data errors is "small," and high-breakdown estimation assumes implicitly that the probability of data errors is less than the breakdown point.

In bounded-influence estimation it is further assumed that the occurrence of data errors is independent of the sample realizations from the population of interest (the "contaminated sampling" model). This assumption is usually not made in high-breakdown estimation, thereby permitting more general forms of data corruption (the "corrupted sampling" model). We analyze identification under both the contaminated and corrupted sampling models.

Section 2 presents these models formally and develops basic identification results obtainable when no structure is imposed on the sample space or the parameter of interest. We introduce the concept of "identification breakdown" and relate it to the robustness concept of breakdown.

Section 3 gives further results that apply when the sample space is the real line. We obtain tight bounds on quantiles (Section 3.1) and on the more general class of parameters that respect stochastic dominance (Section 3.2).

Section 4 develops an infinitesimal identification analysis that applies when the error probability is small and the parameter of interest is a smooth functional on the space of probability distributions. This work yields an identification version of the gross-error sensitivity.

Section 5 outlines how the bounds obtained in Sections 2 and 3 can be estimated consistently. The estimation method is illustrated with an application to data on the distribution of household incomes in the U.S.

The findings reported in Sections 2-5 lead us in Section 6 to question the focus on point estimation found in robustness studies. If, given the available information, a population parameter can only be bounded, then the rationale for point estimation is not apparent. It seems to us more natural to estimate the bounds, a task that is often easily accomplished.

Our work also suggests that the perspective of robust analysis is more conservative than necessary. Robust analysis views the inference problem before the data are collected. The objective is to guard against the worst outcomes that errors in the data could conceivably produce. But some outcomes that are possible *ex ante* can be ruled out *ex post*, after the data have been collected. Identification analysis characterizes the inferences that can be made given knowledge of the empirical distribution of the data.

The proofs of propositions are in the appendix.

2. Basic Identification Analysis

2.1 STATEMENT OF THE PROBLEM

To pose the errors-in-data problem that we investigate, let (Y,Ω) be a measurable space, let $(y_0,y_1,z) \in Y \times Y \times \{0,1\}$ be a random triple distributed P, and let a random sample be drawn from P. The objective is to make an inference about the marginal distribution of y_1. One does not, however, observe the sample realizations of any of the components of (y_0,y_1,z). One observes only the realizations of $y = y_0(1 - z) + y_1z$. That is, one observes y_0 when $z = 0$ and y_1 when $z = 1$, and one never observes z. Since interest centers on y_1, this sampling process generates data with errors, namely the realizations of y that are observed when $z = 0$. Realizations of y corresponding to $z = 1$ are error free.

Let Ψ denote the space of all probability distributions on (Y,Ω). Let $Q \equiv Q(y)$ denote the distribution of the observable y. Let $P_i \equiv P_i(y_i)$ denote the marginal distribution of y_i $(i = 0, 1)$. Let $P_{ij} \equiv P_{ij}(y_i|z=j)$ denote the distribution of y_i conditional on the event $z = j$ for $i = 0, 1$ and $j = 0, 1$. Finally, let $p \equiv P(z=0)$ be the marginal probability of a data error. The object of interest is P_1. In particular, one may wish to infer a parameter $\tau(P_1)$, where $\tau(\cdot)$ maps Ψ into $\Re^1$.

The inferential problem is that the sampling process does not identify P_1 but only Q. These two distributions may be decomposed as follows:

$$(1) \qquad P_1 = (1 - p)P_{11} + pP_{10}$$

and

$$(2) \qquad Q = (1 - p)P_{11} + pP_{00}.$$

In robust analysis, the unknown P_1 is held fixed and Q is allowed to range over all distributions consistent with (1) and (2). The objective is to characterize the maximum possible difference between $\tau(Q)$, which can be estimated consistently, and $\tau(P_1)$, which cannot be estimated consistently because it is not identified. In identification analysis, which is the approach developed in this paper, Q is held fixed because it is identified by the data, and P_1 is allowed to range over all distributions consistent with (1) and (2). The objective is to set bounds on the unknown quantity $\tau(P_1)$.

It is easily seen that in the absence of prior information, identification of Q implies no restrictions on P_1. Simply observe that $Q = P_{00}$ if $p = 1$, in which case Q provides no information about P_1. On the other hand, restrictions on P_1 may arise if suitable prior information is available.

2.2 THE CONTAMINATED AND CORRUPTED SAMPLING MODELS

A piece of prior information that is frequently assumed to be available in robust estimation is that the occurrence of data errors is independent of the sample realizations from the population of interest. That is,

$$(3) \qquad P_1 = P_{11},$$

in which case inferences about P_1 are equivalent to inferences about P_{11}. This assumption underlies the influence function and bounded-influence estimation. It will be seen below that tighter bounds on P_1 can be established when (3) holds than when it does not. Accordingly, this paper provides parallel treatments of the problems in which (3) does and does not hold. Results when (3) holds are formulated in terms of bounds on P_{11} and functionals $\tau(P_{11})$ since, under (3), P_{11} and P_1 are the same. Results when (3) does not hold are formulated in terms of bounds on P_1 and functionals $\tau(P_1)$.

Following common terminology, we refer to the case where (3) holds as the "contaminated sampling" model (see, e.g., Huber 1981, p. 11). We refer to the case in which (3) does not hold as the "corrupted sampling" model. This case, which permits arbitrary corruption of an arbitrarily selected fraction of the data, underlies much of the literature on high-breakdown estimation. The process of data corruption used by Hampel *et al.* (1986), referred to hereinafter as HRRS, in their definition of the finite-sample breakdown point is a finite-sample version of our corrupted sampling model. The same is true of the process of data corruption used by Donoho and Huber (1983), referred to hereinafter as DH, in their definition of the breakdown point under ϵ-replacement. The

corrupted sampling model is also prominent in the far-removed literature on switching regressions without sample separation (see, e.g., Maddala 1983).

2.3 IMPLICATIONS OF AN UPPER BOUND ON THE ERROR PROBABILITY

Another useful piece of prior information is an upper bound, λ, on the probability, p, that a data point is erroneous. The analysis of this paper assumes that one knows a λ such that $p \leq \lambda < 1$. Huber (1964) makes this assumption explicitly in his development of minimax estimators of location in the presence of gross errors. Some more recent literature on robust estimation takes an ambiguous position. HRRS suggest a particular value for λ when they state that "altogether, 1-10% gross errors in routine data seem to be more the rule rather than the exception" (p. 28). On the other hand, these authors later seem to argue that robust estimation requires at most a vague idea of the error probability, or perhaps none at all (p. 399).

As was discussed in Section 1 of this paper, we would argue that robust estimation always implicitly assumes a bound on the error probability. Using the influence function and gross error sensitivity to guide empirical practice makes sense only if the error probability is "close to zero." Knowing the breakdown point of an estimator is of interest only if one knows whether p lies above or below this value. So the basic concepts of robust estimation seem irrelevant in the absence of a known upper bound on p.

Knowledge that $p \leq \lambda$ combined with the fact that Q is identified implies that P_{11} belongs to a set of distributions $\Psi_{11}(\lambda)$, defined below, and that P_1 belongs to a larger set $\Psi_1(\lambda)$. [Both $\Psi_{11}(\lambda)$ and $\Psi_1(\lambda)$ depend on Q as well as λ, but we leave this fact implicit as we shall not be varying Q.] As λ increases, the sets $\Psi_{11}(\lambda)$ and $\Psi_1(\lambda)$ expand but remain informative in the sense that they

are proper subsets of Ψ for all $\lambda < 1$. Proposition 1 presents the resulting restrictions on P_{11} and P_1:

Proposition 1:

A. Let the error probability p be known with $p < 1$. Then

$$(4) \qquad P_{11} \in \Psi_{11}(p) = \Psi \cap \{(Q - p\psi_{00})/(1 - p): \psi_{00} \in \Psi\}$$

and

$$(5) \qquad P_1 \in \Psi_1(p) = \{(1 - p)\psi_{11} + p\psi_{10}: (\psi_{11},\psi_{10}) \in \Psi_{11}(p)\times\Psi\}.$$

In the absence of further information, these restrictions on P_{11} and P_1 are tight.

B. $\Psi_{11}(p) \subset \Psi_1(p)$.

C. Let $\delta > 0$ and $p + \delta < 1$. Then $\Psi_{11}(p) \subset \Psi_{11}(p + \delta)$ and $\Psi_1(p) \subset \Psi_1(p + \delta)$.

D. Let it be known only that $p \leq \lambda < 1$. Then $P_{11} \in \Psi_{11}(\lambda)$ and $P_1 \in \Psi_1(\lambda)$. In the absence of further information, these restrictions on P_{11} and P_1 are tight. ■

Although Proposition 1 is formulated in terms of P_{11} and P_1, it implicitly characterizes the identifiability of P_{00} and P_0 as well. The proposition shows that $P_{00} \in \Psi_{00}(\lambda) = \Psi_{11}(1 - \lambda)$ and $P_0 \in \Psi_0(\lambda) = \Psi_1(1 - \lambda)$. We use these facts later in Section 4.

In Section 2.2 it was stated that tighter bounds on P_1 can be established under the contaminated sampling model than under the corrupted sampling model. This fact, which is implied by (3) and part B of Proposition 1, is important enough to warrant statement as a corollary to Proposition 1:

Corollary 1.1: Let it be known that $p \leq \lambda < 1$ and that $P_1 = P_{11}$. Then $P_1 \in \Psi_{11}(\lambda)$. In the absence of further information, this restriction on P_1 is tight. ■

2.4 IDENTIFICATION BREAKDOWN

Given any real-valued functional $\tau(\cdot)$ on Ψ, let $T = \{\tau(\psi): \psi \in \Psi\}$ denote the range of τ. Let T_L and T_U, respectively, denote the lower and upper bounds of T. Under the assumptions of Proposition 1D,

$$(6a) \qquad \tau(P_{11}) \in T_{11}(\lambda) = \{\tau(\psi): \psi \in \Psi_{11}(\lambda)\}$$

and

$$(6b) \qquad \tau(P_1) \in T_1(\lambda) = \{\tau(\psi): \psi \in \Psi_1(\lambda)\}.$$

Let $T_{11L}(\lambda)$ and $T_{11U}(\lambda)$, respectively, denote the lower and upper bounds of $T_{11}(\lambda)$, and let $T_{1L}(\lambda)$ and $T_{1U}(\lambda)$ denote the lower and upper bounds of $T_1(\lambda)$.

In robust estimation, the breakdown point of a functional $\tau(\cdot)$ can be defined as the largest fraction of erroneous data that $\tau(\cdot)$ can tolerate without being driven to either boundary of its range (HRRS, p. 98). Following this convention, define

$$(7a) \qquad \lambda_{11} = \sup\{\lambda: T_L < T_{11L}(\lambda) \leq T_{11U}(\lambda) < T_U\}$$

and

$$(7b) \qquad \lambda_1 = \sup\{\lambda: T_L < T_{1L}(\lambda) \leq T_{1U}(\lambda) < T_U\}$$

We call λ_{11} and λ_1 the "identification breakdown" points of $\tau(P_{11})$ and $\tau(P_1)$. In general, λ_{11} and λ_1 depend on Q. [DH define the breakdown point as the smallest fraction of erroneous data that can drive $\tau(\cdot)$ to a boundary of its

range. As noted by HRRS, their definition and that of DH differ by 1/n in a sample of size n. Our analysis could be based on either definition. We adopt that of HRRS to avoid ambiguity.]

When Y is a finite-dimensional real space, the identification breakdown point under the corrupted sampling model, λ_1, is a limiting form of the HRRS version of the finite-sample breakdown point of robust analysis. (The ϵ-replacement breakdown point of DH also could be used.) To show this, let λ_{1n} denote the HRRS finite-sample breakdown point of $\tau(\cdot)$ at a random sample of size n drawn from Q. Then we have the following:

Proposition 2: Assume that Y is a finite-dimensional real space and that

$$(8) \qquad \lim_{n \to \infty} |\tau(F_n) - \tau(G_n)| = 0$$

for any sequences of distribution functions $\{F_n\}$ and $\{G_n\}$ such that $(F_n - G_n) \to 0$ pointwise as $n \to \infty$. Then

$$(9) \qquad \lim_{n \to \infty} \lambda_{1n} = \lambda_1$$

almost surely. ■

Note that λ_1 is evaluated at the observed distribution Q, whereas the breakdown point in standard robust analysis is evaluated at the distribution of interest, P_1. This difference between the identification and standard robust breakdown points reflects the focus of identification analysis on *ex post* inference, whereas robust estimation is concerned with *ex ante* analysis of the behavior of estimators.

2.5 TIGHT BOUNDS ON PROBABILITIES

We now develop the implications of the general, but abstract, Proposition 1 for identification of P_{11} and P_1. Corollary 1.2 of Proposition 1 begins this process by obtaining tight bounds on $P_{11}(A)$ and $P_1(A)$ for all measurable sets A.

Corollary 1.2: Let it be known that $p \leq \lambda < 1$. Let $A \in \Omega$. Then

$$(10) \qquad P_{11}(A) \in \Psi_{11}(A;\lambda) = [0,1] \cap [\{Q(A) - \lambda\}/(1 - \lambda),\ Q(A)/(1 - \lambda)],$$

and

$$(11) \qquad P_1(A) \in \Psi_1(A;\lambda) = [0,1] \cap [Q(A) - \lambda,\ Q(A) + \lambda].$$

In the absence of further information, these bounds on $P_{11}(A)$ and $P_1(A)$ are tight. ■

An equivalent representation of the intervals $\Psi_{11}(A;\lambda)$ and $\Psi_1(A;\lambda)$ can be obtained by dividing the range of possible values of Q(A) into four regimes:

$$(12a) \qquad 1 - \lambda \leq Q(A) \leq \lambda \Rightarrow \Psi_{11}(A;\lambda) = \Psi_1(A;\lambda) = [0,1]$$

$$(12b) \qquad Q(A) \leq \min(1 - \lambda,\ \lambda) \Rightarrow \Psi_{11}(A;\lambda) = [0,\ Q(A)/(1 - \lambda)]$$

$$\Psi_1(A;\lambda) = [0,\ Q(A) + \lambda]$$

$$(12c) \qquad Q(A) \geq \max(1 - \lambda,\ \lambda) \Rightarrow \Psi_{11}(A;\lambda) = [\{Q(A) - \lambda\}/(1 - \lambda),\ 1]$$

$$\Psi_1(A;\lambda) = [Q(A) - \lambda,\ 1]$$

$$(12d) \quad \lambda \leq Q(A) \leq 1 - \lambda \Rightarrow \Psi_{11}(A;\lambda) = [\{Q(A) - \lambda\}/(1 - \lambda),\ Q(A)/(1 - \lambda)]$$

$$\Psi_1(A;\lambda) = [Q(A) - \lambda,\ Q(A) + \lambda].$$

It can be seen from (12) that the bounds on $P_{11}(A)$ and $P_1(A)$ are trivial if $1 - \lambda \leq Q(A) \leq \lambda$ but are informative otherwise. Moreover, the identification breakdown point of both $P_{11}(A)$ and $P_1(A)$ is $\min[Q(A), 1 - Q(A)]$.

Corollary 1.2 shows that $Q(A) \in \Psi_{11}(A;\lambda) \subset \Psi_1(A;\lambda)$ for all A. Hence, Q belongs to both $\Psi_{11}(\lambda)$ and $\Psi_1(\lambda)$. This means that if the only available information is a bound on p, one cannot reject the hypothesis that $P_1 = P_{11} = Q$. Moreover, P_{11} and P_1 cannot differ from Q by too much. Equation (10) implies that P_{11} is dominated by Q (i.e., sets of zero Q measure also have zero P_{11} measure), and (11) implies that

$$(13) \qquad \sup_{A \in \Omega} |P_1(A) - Q(A)| \leq \lambda.$$

It is important to understand the relation between the tight restriction $P_{11} \in \Psi_{11}(\lambda)$ reported in Proposition 1 and the collection of tight restrictions $P_{11}(A) \in \Psi_{11}(A;\lambda)$, $A \in \Omega$, found in Corollary 1.2. Every distribution $\psi_{11} \in \Psi_{11}(\lambda)$ satisfies $\psi_{11}(A) \in \Psi_{11}(A;\lambda)$, $A \in \Omega$. However, not every function $\phi:\Omega \to \Re^1$ satisfying $\phi(A) \in \Psi_{11}(A;\lambda)$, $A \in \Omega$, is a probability distribution. The same relation applies, of course, to the restrictions on P_1.

3. Identification when Y Is the Real Line

In Section 2, apart from Proposition 2, (Y,Ω) was an arbitrary measurable space. It does not seem possible to go much further than the results of Section 2 without imposing additional structure on (Y,Ω). In this section, we assume that Y is the extended real line and that Ω consists of the Lebesgue measurable sets. Note that distributions on the extended real line may place probability mass at $-\infty$ or ∞. Section 3.1 obtains tight bounds on quantiles of P_{11} and P_1. Section 3.2 does the same for parameters that respect stochastic dominance.

3.1 TIGHT BOUNDS ON QUANTILES

For $\alpha \in (0,1]$, the α-quantile of P_{11} is

$$q_{11}(\alpha) = \inf\{t: P_{11}[-\infty,t] \geq \alpha\}. \tag{14}$$

The α-quantile of P_1 is

$$q_1(\alpha) = \inf\{t: P_1[-\infty,t] \geq \alpha\}. \tag{15}$$

Proposition 3 shows that $q_{11}(\alpha)$ lies between two quantiles of Q and that $q_1(\alpha)$ lies between two more widely spaced quantiles of Q.

Proposition 3: Let Y be the extended real line and Ω the Lebesgue measurable sets. Let it be known that $p \leq \lambda < 1$. For $\gamma \in \mathfrak{R}^1$, let

$$r(\gamma) = \begin{cases} \gamma\text{-quantile of Q if } 0 < \gamma \leq 1 \\ -\infty \text{ if } \gamma \leq 0 \\ \infty \text{ if } \gamma > 1. \end{cases}$$

Then

$$q_{11}(\alpha) \in [r\{\alpha(1 - \lambda)\}, r\{\alpha(1 - \lambda) + \lambda\}], \tag{16}$$

and

$$q_1(\alpha) \in [r(\alpha - \lambda), r(\alpha + \lambda)]. \tag{17}$$

In the absence of further information, these bounds on $q_{11}(\alpha)$ and $q_1(\alpha)$ are tight. ■

If λ is fixed, $r[\alpha(1 - \lambda)]$ and $r[\alpha(1 - \lambda) + \lambda]$ increase as α increases. Hence, the bounds on $q_{11}(\alpha)$ shift to the right as α increases. If α is fixed and λ increases from 0 to 1, the set of possible values of $q_{11}(\alpha)$ widens from the α-quantile of Q to the smallest interval enclosing the support of Q. The

bound is informative both above and below for all $\alpha \in (0,1)$, all Q, and all $\lambda < 1$. Therefore, the identification breakdown point of $q_{11}(\alpha)$ is always 1.

The bounds on $q_1(\alpha)$ also shift to the right as α increases for fixed λ. If α remains fixed and λ increases from 0 to 1, the set of possible values of $q_1(\alpha)$ widens from the α-quantile of Q to $[-\infty,\infty]$. The lower bound is informative if $\lambda < \alpha$, and the upper bound is informative if $\lambda < 1 - \alpha$. Therefore, the identification breakdown point of $q_1(\alpha)$ is $\min(\alpha, 1 - \alpha)$.

3.2 TIGHT BOUNDS ON PARAMETERS THAT RESPECT STOCHASTIC DOMINANCE

If F and G are distributions on the extended real line Y, F is said to stochastically dominate G if $F[-\infty,t] \leq G[-\infty,t]$ for all $t \in Y$. We say that a parameter $\tau(\cdot)$ respects stochastic dominance if $\tau(F) \geq \tau(G)$ whenever F stochastically dominates G. Familiar examples include quantiles and the means of monotone functions of the random variable of interest. Proposition 4 provides tight bounds on parameters that respect stochastic dominance. We give applications following the statement of the proposition.

Proposition 4: Let Y be the extended real line and Ω the Lebesgue measurable sets. Let it be known that $p \leq \lambda < 1$. Let $\tau: \Psi \to \Re^1$ respect stochastic dominance. Define the following distributions on (Y,Ω):

$$L_\lambda[-\infty,t] \equiv \begin{cases} Q[-\infty,t]/(1 - \lambda) \text{ if } t < r(1 - \lambda) \\ 1 \text{ if } t \geq r(1 - \lambda) \end{cases}$$

$$U_\lambda[-\infty,t] \equiv \begin{cases} 0 \text{ if } t < r(\lambda) \\ (Q[-\infty,t] - \lambda)/(1 - \lambda) \text{ if } t \geq r(\lambda). \end{cases}$$

Then

(18) $\tau(P_{11}) \in [\tau(L_\lambda), \tau(U_\lambda)]$.

Let $\delta_{-\infty}$ and δ_∞ be the probability measures on Y that place all their mass at $-\infty$ and ∞, respectively. Then

(19) $\tau(P_1) \in [\tau\{(1 - \lambda)L_\lambda + \lambda\delta_{-\infty}\}, \tau\{(1 - \lambda)U_\lambda + \lambda\delta_\infty\}]$.

In the absence of further information, these bounds on $\tau(P_{11})$ and $\tau(P_1)$ are tight. ■

Proposition 4 can be applied to obtain an alternative proof of Proposition 3. More importantly, Proposition 4 can be used to obtain tight bounds on the means of bounded, increasing functions on Y. Corollary 4.1 gives the result.

Corollary 4.1: Let $g: Y \to \mathbb{R}^1$ be a bounded, increasing function with $K_0 = \lim_{t \to -\infty} g(t)$ and $K_1 = \lim_{t \to \infty} g(t)$ being the finite lower and upper bounds. For $\psi \in \Psi$, let $\tau(\psi) = \int g(y)d\psi$ be the mean of g(y) when y is distributed ψ. Then tight bounds on $\tau(P_{11})$ and $\tau(P_1)$ are

(20) $\tau(P_{11}) \in [\int g(y)dL_\lambda, \int g(y)dU_\lambda]$

and

(21) $\tau(P_1) \in [(1 - \lambda)\int g(y)dL_\lambda + \lambda K_0, (1 - \lambda)\int g(y)dU_\lambda + \lambda K_1]$. ■

Observe that if $\int g(y)dL_\lambda$ and $\int g(y)dU_\lambda$ are held fixed, the range $[K_0, K_1]$ of $g(\cdot)$ does not affect the bounds on $\tau(P_{11})$. Therefore, (20) provides tight bounds on $\tau(P_{11})$ even if $g(\cdot)$ is unbounded. In particular, letting $g(y) = y$, we find that for the contamination model

(22) $E(y_1|z=1) \in [\int ydL_\lambda, \int ydU_\lambda]$.

This interval is informative whenever the mean of Q exists because $\int y dQ > -\infty$ implies $\int y dL_\lambda > -\infty$, and $\int y dQ < \infty$ implies $\int y dU_\lambda < \infty$. Thus, we obtain finite bounds on the mean under contaminated sampling.

This finding does not contradict the well-known result in the literature on robust estimation that the mean is not robust under contaminated sampling. Identification analysis and the theory of robust estimation analyze different quantities and make different assumptions about the information that is available to the analyst. Identification analysis gives the range of possible values of the mean of P_{11} subject to the information on Q that is revealed by the sampling process. As is shown by (22), the resulting range of values is finite if the mean of Q is finite. In contrast, robust estimation supposes that Q is not yet known and, holding P_{11} fixed, obtains the feasible values of the mean of Q for $Q \in \{(1 - p)P_{11} + p\psi: \psi \in \Psi, p \leq \lambda\}$. The range of possible values of the mean is unbounded under this setup.

4. Infinitesimal Identification Analysis for Smooth Functionals

The identification findings obtained thus far hold for all $\lambda < 1$. Simplifications that facilitate further analysis occur when λ is close to 0 and $\tau(\cdot)$ is a suitably smooth functional. These simplifications are central to the literature on bounded-influence estimation. Here we exploit them to develop an infinitesimal identification analysis.

Let (Y,Ω) be an arbitrary measurable space, as in Section 2. Observe that the sets $\Psi_{11}(\lambda)$ and $\Psi_1(\lambda)$ of possible values of P_{11} and P_1, originally defined in Proposition 1, can alternatively be expressed as follows:

$$(23) \qquad \Psi_{11}(\lambda) = \{Q - [\lambda/(1 - \lambda)](\psi - Q): \psi \in \Psi_{00}(\lambda)\}$$

and

$$(24) \qquad \Psi_1(\lambda) = \{Q - \lambda(\psi - \omega): \psi \in \Psi_{00}(\lambda),\ \omega \in \Psi\},$$

where $\Psi_{00}(\lambda)$ is the set of possible distributions P_{00}.

Let $\psi \in \Psi_{00}(\lambda)$, $\omega \in \Psi$, and $0 \leq \beta \leq \lambda/(1 - \lambda)$. When it exists, define

$$(25) \qquad \tau'(Q,\psi,\omega) = \lim_{\beta \downarrow 0} \frac{\tau[Q - \beta(\psi - \omega)] - \tau(Q)}{\beta}$$

to be the derivative of $\tau(\cdot)$ at Q in the direction $-(\psi - \omega)$. Recall that $T_{11}(\lambda)$ and $T_1(\lambda)$, defined in (6a) and (6b), denote the sets of possible values for $\tau(P_{11})$ and $\tau(P_1)$, respectively. Then we have the following:

Proposition 5: Let it be known that $p \leq \lambda < 1$. Assume that $\tau'(Q,\psi,Q)$ exists and is bounded uniformly over $\psi \in \Psi_{00}(\lambda)$ and that

$$(26) \qquad \lim_{\beta \downarrow 0} \sup_{\psi \in \Psi_{00}(\lambda)} \left|\tau'(Q,\psi,Q) - \{\tau[Q - \beta(\psi - Q)] - \tau(Q)]\}/\beta\right| = 0.$$

Then

$$(27) \qquad T_{11}(\lambda) = \{\tau(Q) + \lambda\tau'(Q,\psi,Q) + o(\lambda;Q): \psi \in \Psi_{00}(\lambda)\},$$

where $o(\lambda;Q)$ denotes a term that, for fixed Q, is $o(\lambda)$ uniformly over $\psi \in \Psi_{00}(\lambda)$. Tight bounds on $\tau(P_{11})$ are

$$(28) \qquad \tau(Q) + \lambda \inf_{\psi \in \Psi_{00}(\lambda)} \tau'(Q,\psi,Q) + o(\lambda;Q) \leq \tau(P_{11})$$

$$\leq \tau(Q) + \lambda \sup_{\psi \in \Psi_{00}(\lambda)} \tau'(Q,\psi,Q) + o(\lambda;Q).$$

Assume that $\tau'(Q,\psi,\omega)$ exists and is bounded uniformly over $(\psi,\omega) \in \Psi_{00} \times \Psi$ and that

$$(29)\qquad \lim_{\beta \downarrow 0} \sup_{\substack{\psi \in \Psi_{00}(\lambda) \\ \omega \in \Psi}} \left| \tau'(Q,\psi,\omega) - \{\tau[Q - \beta(\psi - \omega)] - \tau(Q)\}/\beta \right| = 0.$$

Then

$$(30)\qquad T_1(\lambda) = \{\tau(Q) + \lambda\tau'(Q,\psi,\omega) + o(\lambda;Q): \psi \in \Psi_{00}(\lambda),\ \Psi \in \Psi\},$$

where $o(\lambda;Q)$ here denotes a term that, for fixed Q, is $o(\lambda)$ uniformly over $\psi \in \Psi_{00}(\lambda)$ and $\omega \in \Psi$. Tight bounds on $\tau(P_1)$ are

$$(31)\qquad \tau(Q) + \lambda \inf_{\substack{\psi \in \Psi_{00}(\lambda) \\ \omega \in \Psi}} \tau'(Q,\psi,\omega) + o(\lambda;Q) \leq \tau(P_1)$$

$$\leq \tau(Q) + \lambda \sup_{\substack{\psi \in \Psi_{00}(\lambda) \\ \omega \in \Psi}} \tau'(Q,\psi,\omega) + o(\lambda;Q). \quad \blacksquare$$

Proposition 5 is an abstract result whose implications can be investigated most easily by imposing additional structure on τ'. Corollary 5.1 assumes that τ' is a bounded (equivalently, continuous) linear functional and obtains results that may be compared with ones appearing in the literature on bounded-influence estimation.

Corollary 5.1: Let the assumptions of Proposition 5 hold. Let $\tau'(Q,\psi,\omega)$ be a bounded linear functional of $(\psi - \omega)$ with the integral representation

$$(32)\qquad \tau'(Q,\psi,\omega) = \int f_Q(y)d(\psi - \omega).$$

Assume without loss of generality that

$$(33)\qquad \int f_Q(y)dQ = 0.$$

(If $\int f_Q dQ = \mu_Q \neq 0$, replace f_Q with $f_Q - \mu_Q$.) Define

(34) $$B_U = \sup_{y \in Y} f_Q(y)$$

and

(35) $$B_L = \inf_{y \in Y} f_Q(y).$$

Bounds on $\tau(P_{11})$ are

(36) $$\tau(Q) + \lambda B_L + o(\lambda;Q) \leq \tau(P_{11}) \leq \tau(Q) + \lambda B_U + o(\lambda;Q).$$

Define

(37) $$B_U^* = \sup_{y \in Y} f_Q(y) - \inf_{y \in Y} f_Q(y)$$

and

(38) $$B_L^* = - B_U^*.$$

Bounds on $\tau(P_1)$ are

(39) $$\tau(Q) + \lambda B_L^* + o(\lambda;Q) \leq \tau(P_{11}) \leq \tau(Q) + \lambda B_U^* + o(\lambda;Q). \quad \blacksquare$$

In robust estimation under contaminated sampling, the unknown distribution P_{11} is held fixed. Suppose that the derivative of $\tau(\cdot)$ at P_{11} in the direction $(\psi - P_{11})$ has the representation $\int f_{P_{11}} d\psi$. Then the quantity

(40) $$\max\left[\left|\inf_{y \in Y} f_{P_{11}}(y)\right|, \sup_{y \in Y} f_{P_{11}}(y)\right]$$

is called the gross-error sensitivity of the functional $\tau(\cdot)$ at P_{11}. By comparing (34) and (35) with (40), it can be seen that $\max(|B_L|, B_U)$ is also a form of gross error sensitivity of $\tau(\cdot)$, except the derivative is evaluated at Q in the direction of $-(\psi - Q)$ instead of at P_{11} in the direction of $(\psi - P_{11})$.

Thus, in identification analysis as in robust estimation, the gross-error sensitivity governs the maximum possible value of $|\tau(Q) - \tau(P_{11})|$ under the contamination model with infinitesimal contamination probability. However, in identification analysis, the gross error sensitivity is evaluated at the observed distribution Q, not the unknown "correct" distribution P_{11}. Like the difference between the identification and robust-estimation breakdown points, this reflects the focus of identification analysis on *ex post* inference.

5. Estimation of Identified Features with an Application to the Income Distribution in the U.S.

We stated at the outset that it is natural to analyze the errors-in-data problem in two stages: first determine what is identified, and then consider estimation of the identified quantities. This paper has focussed on identification. We now provide a brief discussion of estimation and give an illustrative application.

All of the restrictions on P_{11} and P_1 reported in Sections 2-4 are functionals of the distribution Q. So an obvious estimation approach is to estimate Q by its empirical distribution Q_n and compute the restrictions on P_{11} and P_1 under Q_n. For example, the bounds $\Psi_{11}(A;\lambda)$ and $\Psi(A;\lambda)$ on $P_{11}(A)$ and $P_1(A)$ found in Corollary 1.2 may be estimated consistently by

$$(41) \qquad \Psi_{n11}(A;\lambda) \equiv [0,1]\cap[\{Q_n(A) - \lambda\}/(1 - \lambda),\ Q_n(A)/(1 - \lambda)]$$

and

$$(42) \qquad \Psi_{n1}(A;\lambda) \equiv [0,1]\cap[Q_n(A) - \lambda,\ Q_n(A) + \lambda].$$

Similarly, the bounds $[r\{\alpha(1 - \lambda)\},\ r\{\alpha(1 - \lambda) + \lambda\}]$ and $[r(\alpha - \lambda),\ r(\alpha + \lambda)]$ on $q_{11}(\alpha)$ and $q_1(\alpha)$, found in Proposition 3, may be estimated consistently by

$[r_n\{\alpha(1 - \lambda)\}, r_n\{\alpha(1 - \lambda) + \lambda\}]$ and $[r_n(\alpha - \lambda), r_n(\alpha + \lambda)]$, respectively, where $r_n(\gamma)$ is the γ-quantile of Q_n for $\gamma \in (0,1]$, $r_n(\gamma) = -\infty$ for $\gamma \leq 0$, and $r_n(\gamma) = \infty$ for $\gamma > 1$. These estimates are the best possible in the absence of prior information on Q. [The problem of developing confidence intervals for the bounds is not treated here since our main focus is on identification.]

To illustrate estimation of the bounds, we consider data on the income distribution in the U.S. The data are based on household interviews obtained in the Current Population Survey (CPS) and are published by the U.S. Bureau of the Census in series P-60 of Current Population Reports. Two sampling problems identified by the Bureau of the Census are "interview nonresponse," wherein some households in the CPS sampling frame are not interviewed, and "item nonresponse," wherein some of those interviewed do not provide complete income responses. U.S. Bureau of the Census (1991, pp. 387-388) states that in the March 1990 CPS, which provides data on incomes during 1989, approximately 4.5% of the 60,000 households in the sampling frame were not interviewed and that incomplete income data were obtained from approximately 8% of the persons in interviewed households. Faced with these nonresponse problems, the Bureau of the Census uses available information to impute missing income data. The Bureau of the Census mixes actual and imputed data to produce the household income statistics reported in its Series P-60 publications.

From the perspective of this paper, y_1 is the income a CPS household would report if it were to complete the survey, y_0 is the income the Bureau of the Census would impute to the household if the household were not to complete the survey, and $z = 1$ if a household completes the survey. The distribution of income reported by those CPS households who do complete the survey is P_{11}. The distribution of household income found in the Series P-60 publications is $Q =$

$(1-p)P_{11} + pP_{00}$, where p is the probability that a CPS household does not complete the survey and P_{00} is the distribution of incomes imputed by the Bureau of the Census to those households who do not complete the survey. The distribution of interest is $P_1 = (1-p)P_{11} + pP_{10}$, where P_{10} is the (latent) distribution of incomes that would have been reported by CPS households who did not complete the survey, had they done so. That is, P_1 is the distribution of reported incomes if all households in the CPS sampling frame were to report their incomes. [Our reference to P_1 as the "distribution of interest" does not imply that P_1 is necessarily the distribution an applied researcher would wish to learn. Applied researchers are typically interested in the distribution of actual income, not the distribution of reported income. The distribution of actual income is P_1 if households report their incomes accurately but not if some households misreport their incomes. Misreporting, although an important sampling problem, will not be addressed here.]

The Bureau of the Census imputation practice is valid if $P_{10} = P_{00}$, implying that $P_1 = Q$. Our concern, however, is with the worst-case situation in which one has no prior information about the relation between P_{00} and P_{10}. We can compute the bound estimates given at the beginning of this section if we have a consistent estimate of Q and can place an upper bound on p. Bureau of the Census publications provide both quantities. We focus on the distribution of household income in 1989.

As has been noted, U.S. Bureau of the Census (1991) reports that 4.5% of the CPS households were not interviewed and 8% of the persons in interviewed households did not provide complete income responses. The Bureau of the Census publication does not report how the latter group are spread across households

but we can be sure that no more than 7.6% (i.e., 8% x .955) of the households have item nonresponse. So $p \leq 12.1\%$.

Now consider Q. U.S. Bureau of the Census (1991, Table 5, p. 17) provides estimates for each of twenty-one income intervals (in thousands of dollars):

$Q_n[0,5) = .053$	$Q_n[35,40) = 0.066$	$Q_n[70,75) = .018$
$Q_n[5,10) = .103$	$Q_n[40,45) = 0.060$	$Q_n[75,80) = .015$
$Q_n[10,15) = .097$	$Q_n[45,50) = 0.048$	$Q_n[80,85) = .013$
$Q_n[15,20) = .092$	$Q_n[50,55) = 0.043$	$Q_n[85,90) = .009$
$Q_n[20,25) = .087$	$Q_n[55,60) = 0.032$	$Q_n[90,95) = .008$
$Q_n[25,30) = .083$	$Q_n[60,65) = 0.028$	$Q_n[95,100) = .006$
$Q_n[30,35) = .076$	$Q_n[65,70) = 0.023$	$Q_n[100,+) = .039$

Let us "fill out" this estimate of Q by imposing the auxiliary assumption that income is distributed uniformly within each interval except the last. We may now estimate bounds on features of P_{11} and P_1.

For example, consider the probability that household income is below \$30,000. We have $Q_n[0,30) = 0.515$ and $\lambda = 0.121$. Hence, the estimated bounds on $P_{11}[0,30)$ are $[0.448,0.586]$ and the estimated bounds on $P_1[0,30)$ are $[0.394,0.636]$. Now consider median household income. The median of P_{11} must lie between the .440 and .561 quantiles of Q, while the median of P_1 must lie between the .379 and .621 quantiles of Q. Replacing Q by Q_n and invoking the auxiliary assumption that Q is uniform within \$5000 intervals, the estimated bounds on $q_{11}(.5)$ are $[25.482,33.026]$, and the estimated bounds on $q_1(.5)$ are $[21.954,37.273]$.

6. Discussion

The literature on robust estimation aims at characterizing the behavior of point estimators of population parameters in the presence of errors in the data and at developing point estimators that are not highly sensitive to such

errors. However, the parameters of interest in robust estimation are not identified under the assumptions that are made, so the rationale for concentrating on point estimation is not apparent. This paper has shown that the parameters can be bounded under the assumptions of robust estimation, and it has shown how the bounds can be estimated. It seems to us more natural to estimate the bounds, which are identified, than to attempt point estimation of parameters that are not identified.

Point estimation is especially pernicious if the probability limit of the point estimator need not be in the space of feasible values of $\tau(P_1)$. The estimators most commonly considered in the robustness literature do not have this failing. In robust estimation, the functional $\tau(\cdot)$ is weakly continuous on Ψ, and the estimator of $\tau(P_1)$ typically is $\tau(Q_n)$. In this situation, $\text{plim}_{n \to \infty} \tau(Q_n) = \tau(Q)$. We observed in Section 2 that Q is in the space $\Psi_1(\lambda)$ of feasible values for P_1. Therefore, $\tau(Q)$ is in the space of feasible values for $\tau(P_1)$.

It might be argued that use of $\tau(Q_n)$ as a point estimator of $\tau(P_1)$ is preferable to estimation of the identification bounds given in our propositions. A researcher reporting $\tau(Q_n)$ need not take an explicit stand on the size of the error probability. A researcher estimating identification bounds must specify at least an upper bound on p. We do not find this argument compelling. In the absence of an upper bound on p, one cannot assess the size of the asymptotic bias $|\tau(Q) - \tau(P_1)|$ of the estimate $\tau(Q_n)$ or even be sure it is finite. The usefulness of a point estimator with unknown and, possibly, unbounded asymptotic bias is not obvious. Moreover, given an upper bound on p, estimation of $\tau(Q)$ yields no information on $\tau(P_1)$ beyond that contained in our tight bounds.

Appendix

PROOF OF PROPOSITION 1:

A. By (2), the feasible values for P_{11} and P_{00} are

$$(P_{11},P_{00}) \in \{(\psi_{11},\psi_{00}) \in \Psi\times\Psi: \ Q = (1 - p)\psi_{11} + p\psi_{00}\}.$$

Hence the feasible values for P_{11} are given by (4) and the ones for P_{00} are the ψ_{00} such that $(Q - p\psi_{00})/(1-p) \in \Psi$. Knowledge of p and identification of Q convey no information about P_{10}. Hence, by (1), the feasible values for P_1 are given by (5).

B. If $\psi_{11} \in \Psi_{11}(p)$, then $(\psi_{11},\psi_{10}) \in \Psi_{11}(p)\times\Psi$. So $\psi_{11} \in \Psi_1(p)$, by (5).

C. Let $\psi_{11} \in \Psi_{11}(p)$. Now let the error probability increase to $p + \delta$. Define $\psi_{00\delta} = (\psi_{00}p + \psi_{11}\delta)/(p + \delta)$. Then $\psi_{00\delta}$ is a probability measure and $(\psi_{11},\psi_{00\delta}) \in \Psi\times\Psi$ solves the equation $Q = \psi_{11}(1 - p - \delta) + \psi_{00\delta}(p + \delta)$. Hence $\psi_{11} \in \Psi_{11}(p + \delta)$. That $\Psi_1(p) \subset \Psi_1(p + \delta)$ follows directly from the above and from (5).

D. It follows from Part A that the feasible values for P_{11} and P_1 are $P_{11} \in \cup_{p \le \lambda} \Psi_{11}(p)$ and $P_1 \in \cup_{p \le \lambda} \Psi_1(p)$. Part C showed that $\cup_{p \le \lambda} \Psi_{11}(p) = \Psi_{11}(\lambda)$ and $\cup_{p \le \lambda} \Psi_1(p) = \Psi_1(\lambda)$. Q.E.D.

PROOF OF COROLLARY 1.2: Consider first the situation in which p is known. Part A of Proposition 1 implies that

$$P_{11}(A) \in [0,1]\cap\{[Q(A) - pa]/(1 - p): \ a \in [0,1]\}$$

$$= [0,1]\cap[\{Q(A) - p\}/(1 - p), \ Q(A)/(1 - p)] \equiv \Psi_{11}(A;p).$$

This shows that $P_{11}(A)$ is a member of $\Psi_{11}(A;p)$ but does not show that the bound is tight. To show tightness, we need to prove that, for each $a_{11} \in \Psi_{11}(A;p)$,

there exist distributions (ψ_{11},ψ_{00}) such that $\psi_{11}(A) = a_{11}$ and $Q = (1 - p)\psi_{11} + p\psi_{00}$.

To prove that such distributions exist, fix $a_{11} \in \Psi_{11}(A;p)$ and let a_{00} solve the equation $Q(A) = (1 - p)a_{11} + pa_{00}$. Because $a_{11} \in \Psi_{11}(A;p)$, it follows that $a_{00} \in [0,1]$. Now choose (ψ_{11},ψ_{00}) as follows:

For $Q(A) > 0$ and measurable $B \subset A$,

$\psi_{11}(B;p) = [Q(B)/Q(A)]a_{11}$; $\psi_{00}(B;p) = [Q(B)/Q(A)]a_{00}$.

For $Q(A) = 0$ and measurable $B \subset A$,

$\psi_{11}(B;p) = \psi_{00}(B;p) = 0$.

For $Q(Y - A) > 0$ and measurable $B \subset Y - A$,

$\psi_{11}(B;p) = [Q(B)/Q(Y - A)](1 - a_{11})$; $\psi_{00}(B;p) = [Q(B)/Q(Y - A)](1 - a_{00})$.

For $Q(Y - A) = 0$ and measurable $B \subset Y - A$,

$\psi_{11}(B;p) = \psi_{00}(B;p) = 0$.

Then $\psi_{11}(A;p) = a_{11}$ and $Q(B;p) = (1 - p)\psi_{11}(B) + p\psi_{00}(B)$ for all measurable B.

The above concerns $P_{11}(A)$. The sampling process implies no restrictions on $P_{10}(A)$. Hence the tight bound on $P_1(A)$ is

$$P_1(A) \in \{(1 - p)a_{11} + pa_{10}: \ a_{11} \in \Psi_{11}(A;p), \ a_{10} \in [0,1]\}$$

$$= \{[0, 1 - p] \cap [Q(A) - p, Q(A)] + p[0,1]\}$$

$$= [0,1] \cap [Q(A) - p, Q(A) + p] \equiv \Psi_1(A;p).$$

Now suppose only that $p \leq \lambda$. Then the feasible values for $P_{11}(A)$ and $P_1(A)$ are $P_{11}(A) \in \cup_{p \leq \lambda} \Psi_{11}(A;p)$ and $P_1(A) \in \cup_{p \leq \lambda} \Psi_1(A;p)$. But $\cup_{p \leq \lambda} \Psi_{11}(A;p) = \Psi_{11}(A;\lambda)$ and $\cup_{p \leq \lambda} \Psi_1(A;p) = \Psi_1(A;\lambda)$. Q.E.D.

PROOF OF PROPOSITION 2

Let Q and Q_n (n = 1,2.,...) be given. In the proof of Proposition 2, Q and Q_n are cumulative distribution functions (CDFs). Define equations (1') and (2') to be (1) and (2) with probability measures replaced by the corresponding CDFs. Let $p \in [0,1]$. Let Ψ^* denote the space of CDFs, and define

$$\Psi_{11}^*(p) \equiv \Psi^* \cap \{(Q - pP_{00})/(1 - p): \ P_{00} \in \Psi^*\}.$$

Also define

$$\text{(A1)} \qquad \gamma_L(p) \equiv \inf_{\substack{P_{11} \in \Psi_{11}^*(p) \\ P_{10} \in \Psi^*}} \tau(P_1),$$

where P_1 is given by (1'). Define $\gamma_U(p)$ by replacing "inf" with "sup" (A1).

Let m be the smallest integer such that $p \leq m/n$. Define $p_n \equiv m/n$. For any positive integer j, let Ψ_j^* denote the set of CDFs corresponding to discrete distributions that have at most j mass points. Let $\Psi_{11,n-m}^*(p)$ denote the set of CDFs $P_{11,n-m} \in \Psi_{n-m}^*$ such that

$$\text{(A2)} \qquad Q_n = (1 - p_n)P_{11,n-m} + p_n P_{00,m}$$

for some $P_{00,m} \in \Psi_m^*$. Define

$$\text{(A3)} \qquad \gamma_{Ln}(p) \equiv \inf_{\substack{P_{11,n-m} \in \Psi_{11,n-m}^*(p) \\ P_{10,m} \in \Psi_m^*}} \tau(P_{1n})$$

where

$$\text{(A4)} \qquad P_{1n} = (1 - p_n)P_{11,n-m} + p_n P_{10,m}.$$

Define $\gamma_{Un}(p)$ by replacing "inf" with "sup" in (A3).

The proof of Proposition 2 requires two preliminary lemmas.

Lemma 1: Suppose that (A2) holds for all n and that as $n \to \infty$, (a) $Q_n \to Q$ and (b) $P_{00,m} \to P_{00}^*$ for some P_{00}^*. Then there is a P_{11}^* such that $P_{11,n-m} \to P_{11}^*$, and $(P_{11}^*, P_{00}^*) \in \Psi_{11}^* \times \Psi^*$.

Proof: It follows from (2') and (A2) that $P_{11}^* = (1-p)^{-1}(Q - pP_{00}^*)$ and $Q = (1-p)P_{11}^* + pP_{00}^*$. P_{11}^* and P_{00}^* are nondecreasing because they are limits of sequences of nondecreasing functions. It remains to prove that P_{11}^* and P_{00}^* are continuous from below. Given any y, let $P_{11}^*(y^-) = \lim_{\nu \downarrow 0} P_{11}^*(y - \nu)$. Let $P_{00}^*(y^-)$ be defined analogously. Q is continuous from below because it is a distribution function, so

$$(A5) \qquad Q(y) = (1-p)P_{11}^*(y^-) + pP_{00}^*(y^-).$$

Moreover, since P_{11}^* and P_{00}^* are non-decreasing,

$$(A6) \qquad P_{ii}^*(y^-) \leq P_{ii}^*(y)$$

for i = 0 or 1. Therefore, by (A5) and (A6)

$$(A7) \qquad Q(y) =$$

$$(1-p)P_{11}^*(y^-) + pP_{00}^*(y^-) \leq (1-p)P_{11}^*(y) + pP_{00}^*(y) = Q(y).$$

(A6) and (A7) imply that $P_{11}^*(y^-) = P_{11}^*(y)$ and $P_{00}^*(y^-) = P_{00}^*(y)$. Q.E.D.

Lemma 2: Let (8) hold and $Q_n \to Q$ as $n \to \infty$. Then $\gamma_{Ln}(p) \to \gamma_L(p)$ and $\gamma_{Un}(p) \to \gamma_U(p)$.

Proof: Let $\epsilon > 0$ be given. Let $\{P_{11,n-m}, P_{10,m}, P_{00,m}\} \in \Psi_{11,n-m}^* \times \Psi_m^* \times \Psi_m^*$ be a sequence satisfying (A2) and such that $\tau(P_{1n}) < \gamma_{Ln}(p) + \epsilon$ for each n, where P_{1n} is given by (A4). By Helly's selection principle there is a subsequence

$\{n(i): i = 1,2,\ldots\}$ such that $P_{00,m(i)}$ converges, where m(i) is the value of m corresponding to n(i). Call the limit function P_{00}^*. By Lemma 1, the corresponding subsequence of $\{P_{11,n-m}\}$ converges to some P_{11}^*, P_{11}^* and P_{00}^* are distribution functions, and $(P_{11}^*, P_{00}^*) \in \Psi_{11}^* \times \Psi^*$. Recall that m is a function of n, and define

$$P_{1n}^* = (1 - p)P_{11}^* + pP_{10,m}.$$

$\gamma_L(p) \leq \tau(P_{1n}^*)$ for every n by definition of $\gamma_L(p)$, so (8) implies that

$$\gamma_L(p) - \epsilon < [\tau(P_{1n(i)}) - \tau(P_{1n(i)}^*)] + \tau(P_{1n(i)}^*) < \gamma_{Ln(i)}(p) + \epsilon$$

for all sufficiently large i. Therefore, since ϵ is arbitrary

$$\text{(A8)} \qquad \liminf_{i \to \infty} \gamma_{Ln(i)}(p) \geq \gamma_L(p).$$

Since (A8) holds for any convergent subsequence of $\{P_{11,n-m}, P_{00,m}\}$ and every infinite subsequence has a convergent subsequence, there can be at most finitely many values of n for which $\tau(P_{1n}) \leq \gamma_L(p) - \epsilon$. It follows that

$$\text{(A9)} \qquad \liminf_{n \to \infty} \gamma_{Ln}(p) \geq \gamma_L(p).$$

Now let P_1 be a CDF such that (1') and (2') hold and $\tau(P_1) < \gamma_L(p) + \epsilon$. Let $\{P_{1n}\}$ be a sequence of CDFs satisfying (A4) such that $P_{1n} \to P_1$ as $n \to \infty$. By (8)

$$\gamma_{Ln}(p) - \epsilon < [\tau(P_1) - \tau(P_{1n})] + \tau(P_{1n}) < \gamma_L(p) + \epsilon$$

for all sufficiently large n. Therefore,

$$\text{(A10)} \qquad \limsup_{n \to \infty} \gamma_{Ln}(p) \leq \gamma_L(p).$$

$\gamma_{Ln}(p) \to \gamma_L(p)$ follows by combining (A9) and (A10). Similar arguments apply to $\gamma_{Un}(p)$ and $\gamma_U(p)$. Q.E.D.

Proof of Proposition 2: Assume that breakdown occurs due to $\tau(P_1) = T_H$. A similar proof applies if breakdown occurs due to $\tau(P_1) = T_L$. Suppose that $Q_n \to Q$. Let $p < \lambda_1$. Define $\delta = T_H - \gamma_U(p)$. Note that $\delta > 0$. Let ϵ be such that $0 < \epsilon < \delta$. By (8) and Lemma 2

$$\gamma_{Un}(p) \leq \gamma_U(p) + \epsilon = T_H - \delta + \epsilon < T_H$$

for all sufficiently large n. It follows that for p_n defined as in (A2), $p \leq p_n < \lambda_{1n}$. Therefore, $p < \lambda_1$ implies $p < \lambda_{1n}$ for all sufficiently large n, and

$$\text{(A11)} \qquad \lambda_1 \leq \liminf_{n \to \infty} \lambda_{1n}$$

Now let $\lambda_{1n(i)}$ $(i = 1,2,\ldots)$ be a convergent subsequence of $\{\lambda_{1n}\}$. Denote the limit point by λ_1^*, and let $p < \lambda_1^*$. There is a $\delta > 0$ such that for all sufficiently large i, $\gamma_{Un(i)}(p) \leq T_H - \delta$. Let ϵ be such that $0 < \epsilon < \delta$. By (8) and Lemma 2

$$\gamma_U(p) < \gamma_{Un(i)}(p) + \epsilon \leq T_H - \delta + \epsilon < T_H$$

for all sufficiently large i. Therefore, $\gamma_U(p) < T_H$, and $p < \lambda_1^*$ implies that $p < \lambda_1$. It follows that $\lambda_1^* \leq \lambda_1$. Since this is true for every limit point of $\{\lambda_{1n}\}$,

$$\text{(A12)} \qquad \limsup_{n \to \infty} \lambda_{1n} \leq \lambda_1.$$

The theorem follows from (A11), (A12) and the fact that $Q_n \to Q$ almost surely. Q.E.D.

PROOF OF PROPOSITION 3: Corollary 1.2 shows that, for each $t \in \mathfrak{R}^1$,

(A13) $\quad P_{11}[-\infty,t] \in [0,1]\cap[\{Q[-\infty,t] - \lambda\}/(1 - \lambda),\ Q[-\infty,t]/(1 - \lambda)].$

Hence,

$$Q[-\infty,t] < (1 - \lambda)\alpha \Rightarrow P_{11}[-\infty,t] < \alpha$$

and

$$Q[-\infty,t] \geq (1 - \lambda)\alpha + \lambda \Rightarrow P_{11}[-\infty,t] \geq \alpha.$$

It follows that $q_{11}(\alpha) \in [r\{\alpha(1 - \lambda)\}, r\{\alpha(1 - \lambda)+\lambda\}]$. This bound on $q_{11}(\alpha)$ is tight because the bound (A13) on $P_{11}[-\infty,t]$ is tight. For $t \geq r[\alpha(1 - \lambda)]$, $Q[-\infty,t] \geq \alpha(1 - \lambda)$; hence, the upper bound in (A13) is no less than α. For $t < r[\alpha(1 - \lambda) + \lambda]$, $Q[-\infty,t] < \alpha(1 - \lambda) + \lambda$; hence, the lower bound in (A13) is below α. It follows that all $t \in [r\{\alpha(1 - \lambda)\}, r\{\alpha(1 - \lambda) + \lambda\}]$ are feasible values for $q_{11}(\alpha)$.

Now consider $q_1(\alpha)$. Corollary 1.2 shows that, for each $t \in \mathbb{R}^1$,

$$P_1[-\infty,t] \in [0,1]\cap\{Q[-\infty,t] - \lambda,\ Q[-\infty,t] + \lambda\}.$$

Hence,

$$Q[-\infty,t] < \alpha - \lambda \Rightarrow P_1[-\infty,t] < \alpha$$

and

$$Q[-\infty,t] \geq \alpha + \lambda \Rightarrow P_1[-\infty,t] \geq \alpha.$$

It follows that $q_1(\alpha) \in [r(\alpha - \lambda),\ r(\alpha + \lambda)]$ and, by the same reasoning as above, that this bound is tight in the absence of further information. Q.E.D.

PROOF OF PROPOSITION 4: To show that $r(L_\lambda)$ is the tight lower bound on $r(P_{11})$, let

$$\psi_{00}[-\infty,t] \equiv \begin{cases} 0 \text{ if } t < r(1-\lambda) \\ \{Q[-\infty,t] - (1-\lambda)\}/\lambda \text{ if } r \geq r(1-\lambda) \end{cases}$$

and observe that

$$Q[-\infty,t] = (1-\lambda)L_\lambda[-\infty,t] + \lambda\psi_{00}[-\infty,t], \quad \forall\, t \in \mathfrak{R}^1.$$

This proves that $L_\lambda \in \Psi_{11}(\lambda)$; hence, $\tau(L_\lambda)$ is a feasible value for $\tau(P_{11})$.

$\tau(L_\lambda)$ is the smallest feasible value for $\tau(P_{11})$ because L_λ is stochastically dominated by every member of $\Psi_{11}(\lambda)$. We need to show that $L_\lambda[-\infty,t] \geq \psi_{11}[-\infty,t]$ for all $\psi_{11} \in \Psi_{11}(\lambda)$ and $t \in \mathfrak{R}^1$. If $t \geq r(1-\lambda)$, then

$$L_\lambda[-\infty,t] - \psi_{11}[-\infty,t] = 1 - \psi_{11}[-\infty,t] \geq 0.$$

If $t < r(1-\lambda)$, then

$$\psi_{11}[-\infty,t] > L_\lambda[-\infty,t] \Rightarrow (1-\lambda)\psi_{11}[-\infty,t] > Q[-\infty,t].$$

Hence $(1-\lambda)\psi_{11}[-\infty,t] + \lambda\psi_{00}[-\infty,t] > Q[-\infty,t]$ for all $\psi_{00} \in \Psi$. This contradicts the assumption that $\psi_{11} \in \Psi_{11}(\lambda)$.

The proof that $\tau(U_\lambda)$ is the tight upper bound on $\tau(P_{11})$ is similar. Let

$$\psi_{00}[-\infty,t] \equiv \begin{cases} Q[-\infty,t]/\lambda \text{ if } t < r(\lambda) \\ 1 \text{ if } t \geq r(\lambda) \end{cases}$$

and observe that

$$Q[-\infty,t] = (1-\lambda)U_\lambda[-\infty,t] + \lambda\psi_{00}[-\infty,t], \quad \forall\, t \in \mathfrak{R}^1.$$

Hence $U_\lambda \in \Psi_{11}(\lambda)$. Moreover, U_λ stochastically dominates every $\psi_{11} \in \Psi_{11}(\lambda)$. If $t < r(\lambda)$, then

$$U_\lambda[-\infty,t] - \psi_{11}[-\infty,t] = 0 - \psi_{11}[-\infty,t] \leq 0.$$

If $t \geq r(\lambda)$, then

$$\psi_{11}[-\infty,t] \;<\; U_{\lambda}[-\infty,t] \;\Rightarrow\; (1 - \lambda)\psi_{11}[-\infty,t] \;<\; Q[-\infty,t] - \lambda.$$

Hence $(1 - \lambda)\psi_{11}[-\infty,t] + \lambda\psi_{00}[-\infty,t] < Q[-\infty,t]$ for all $\psi_{00} \in \Psi$. This contradicts the assumption that $\psi_{11} \in \Psi_{11}(\lambda)$.

Now consider the bounds on $\tau(P_1)$. By Proposition 1, P_{11} lies in the set $\Psi_1(\lambda) = \{(1 - \lambda)\psi_{11} + \lambda\psi_{10}: (\psi_{11},\psi_{10}) \in \Psi_{11}(\lambda)\times\Psi\}$. We found above that $L_{\lambda} \in \Psi_{11}(\lambda)$ and that L_{λ} is stochastically dominated by all the members of $\Psi_{11}(\lambda)$. The distribution $\delta_{-\infty}$ belongs to Ψ and is stochastically dominated by all the members of Ψ. Hence, $(1 - \lambda)L_{\lambda} + \lambda\delta_{-\infty} \in \Psi_1(\lambda)$ and $(1 - \lambda)L_{\lambda} + \lambda\delta_{-\infty}$ is stochastically dominated by all the members of $\Psi_1(\lambda)$. It follows that $\tau\{(1 - \lambda)L_{\lambda} + \lambda\delta_{-\infty}\}$ is the smallest feasible value for $\tau(P_1)$. The proof for the upper bound is analogous. Q.E.D.

PROOF OF PROPOSITION 5: Conditions (26) and (29) imply the following uniform Taylor's expansions of $\tau\{Q - [\lambda/(1 - \lambda)](\psi - Q)\}$ and $\tau[Q - \lambda(\psi - \omega)]$ around $\tau(Q)$:

(A14) $$\tau\{Q - [\lambda/(1 - \lambda)](\psi - Q)\} = \tau(Q) + \lambda\tau'(Q,\psi,Q) + o(\lambda;Q)$$

and

(A15) $$\tau[Q - \lambda(\psi - \omega)] = \tau(Q) + \lambda\tau'(Q,\psi,\omega) + o(\lambda;Q).$$

Applying (A14) to (23) yields (27) and applying (A15) to (24) yields (30). The bounds (28) and (31) follow immediately. Q.E.D.

PROOF OF COROLLARY 5.1: $\Psi_{00}(\lambda)$ is a subset of Ψ. Hence the bounds in (28) and (31) lie within the bounds that result when Ψ replaces $\Psi_{00}(\lambda)$. By (32) and (33),

$$\inf_{\psi \in \Psi} \tau'(Q,\psi,Q) = \inf_{\eta \in Y} f_Q(\eta)$$

$$\sup_{\psi \in \Psi} \tau'(Q,\psi,Q) = \sup_{\eta \in Y} f_Q(\eta)$$

$$\inf_{\substack{\psi \in \Psi \\ \omega \in \Psi}} \tau'(Q,\psi,\omega) = \inf_{\eta \in Y} f_Q(\eta) - \sup_{\eta \in Y} f_Q(\eta)$$

$$\sup_{\substack{\psi \in \Psi \\ \omega \in \Psi}} \tau'(Q,\psi,\omega) = \sup_{\eta \in Y} f_Q(\eta) - \inf_{\eta \in Y} f_Q(\eta).$$

Q.E.D.

References

Donoho, D.L. and Huber, P.J. (1983), The Notion of a Breakdown Point, in Festschrift for Erich L. Lehmann, P.J. Bickel, K.A. Doksum, and J.L. Hodges, Jr. (eds.), Wadsworth, Belmont, CA, pp. 157-184.

Hampel, F., Ronchetti, E., Rousseeuw, P., and Stahel, W. (1986), Robust Statistics, New York: Wiley.

Huber, P.(1964), "Robust Estimation of a Location Parameter," Annals of Mathematical Statistics, 35, 73-101.

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

Maddala, G.(1983), Qualitative and Limited Dependent Variables in Econometrics, Cambridge: Cambridge University Press.

U.S. Bureau of the Census (1991), "Money Income of Households, Families, and Persons in the United States: 1988 and 1989", Current Population Reports, Series P-60, No. 172, Washington, D.C.: U.S. Government Printing Office.

RECENT SSRI WORKING PAPERS

9121
Lin, Wen-Ling, Robert F. Engle, and Takatoshi Ito
DO BULLS AND BEARS MOVE ACROSS BORDERS? INTERNATIONAL TRANSMISSION OF STOCK RETURNS AND VOLATILITY AS THE WORLD TURNS

9122
Samuelson, Larry
HOW TO TREMBLE IF YOU MUST

9123
Che, Yeon-Koo
DESIGN COMPETITION THROUGH MULTIDIMENSIONAL AUCTIONS

9124
Holmes, Thomas J. and James A. Schmitz, Jr.
MEASURING SMALL BUSINESS DYNAMICS WHEN BUSINESS AGE AND MANAGERIAL TENURE CAN BE SEPARATELY IDENTIFIED

9125
Phelan, Christopher
RECURSIVE OPTIMAL MECHANISMS WITH HISTORY-DEPENDENT, PRIVATELY-OBSERVED SHOCKS

9126
LeBaron, Blake
TRANSACTIONS COSTS AND CORRELATIONS IN A LARGE FIRM INDEX

9127
Manski, Charles F.
IDENTIFICATION OF ENDOGENOUS SOCIAL EFFECTS: THE REFLECTION PROBLEM

9201
LeBaron, Blake
PERSISTENCE OF THE DOW JONES INDEX ON RISING VOLUME

9202
Berkowitz, Daniel and Beth Mitchneck
FISCAL DECENTRALIZATION IN THE SOVIET ECONOMY

9203
Streufert, Peter A.
A GENERAL THEORY OF SEPARABILITY FOR PREFERENCES DEFINED ON A COUNTABLY INFINITE PRODUCT SPACE

9204
Baek, Ehung G. and William A. Brock
A NONPARAMETRIC TEST FOR INDEPENDENCE OF A MULTIVARIATE TIME SERIES

9205
Mailath, George J., Larry Samuelson and Jeroen M. Swinkels
NORMAL FORM STRUCTURES IN EXTENSIVE FORM GAMES

9206
Mirman, Leonard J., Larry Samuelson and Amparo Urbano
DUOPOLY SIGNAL JAMMING

9207
Andreoni, James and Ted Bergstrom
DO GOVERNMENT SUBSIDIES INCREASE THE PRIVATE SUPPLY OF PUBLIC GOODS?

9208
Loretan, Mico and Peter C.B. Phillips
TESTING THE COVARIANCE STATIONARITY OF HEAVY-TAILED TIME SERIES: AN OVERVIEW OF THE THEORY WITH APPLICATIONS TO SEVERAL FINANCIAL DATASETS

9209
Horowitz, Joel L. and Charles F. Manski
IDENTIFICATION AND ROBUSTNESS IN THE PRESENCE OF ERRORS IN DATA