# IDENTIFICATION AND ESTIMATION WITH CONTAMINATED DATA
# WHEN DOES COVARIATE DATA SHARPEN INFERENCE?

by

Charles H. Mullin

**Working Paper No. 01-W09**

May 2001

DEPARTMENT OF ECONOMICS
VANDERBILT UNIVERSITY
NASHVILLE, TN 37235

www.vanderbilt.edu/econ

# Identification and Estimation with Contaminated Data: When Does Covariate Data Sharpen Inference?

By

Charles H. Mullin*
Vanderbilt University

May 2001

## Abstract

When data contain errors, parameters of interest typically are not identified without imposing strong assumptions. However, in many cases, bounds on these parameters can be constructed under relatively weak assumptions. This paper addresses under what conditions variables in addition to the one of interest, covariate data, tighten these bounds and how to optimally incorporate that information. In particular, covariate data are unable to sharpen inference without imposing some exogenous knowledge about the distribution of errors conditional on the covariates. For example, knowing that the probability of erroneous data is either orthogonal to a covariate or monotonically increasing in a covariate is typically sufficient to sharpen inference. The identification region for the distribution of the variable of interest is constructed and used to develop bounds both on probabilities and on parameters of this distribution that respect stochastic dominance. For the case of bounding parameters that respect stochastic dominance, the necessary and sufficient conditions for covariate data to sharpen inference are derived.

KEY WORDS: Robust Estimation; Contaminated Sampling; Covariate Data; Bounds; Identification.

JEL Classification Numbers: C13 and C14

---

* Department of Economics, Vanderbilt University, Nashville, TN 37235. Email comments to: charles.mullin@vanderbilt.edu.

## 1. Introduction

Suppose the marginal distribution of $y_1$, or a function of that distribution, is the object of interest. However, instead of observing $y_1$, one observes a random variable $y$ whose value can be that of either $y_1$ or another random variable $y_0$, whose distribution is unknown. Horowitz and Manski (1995, hereafter HM) analyzed what could be learned about the parameter of interest from observations on $y$ when an upper bound on the probability of data error is known. This paper extends that research by characterizing when and in what manner the presence of covariate data, $x$, enables sharper inferences about $y_1$.

Specifically, suppose that each member of a population is characterized by a triple $(y, z, x)$, where $z$ is a binary variable for the presence of erroneous data. Our first result is that covariate data are unable to sharpen inference without imposing some exogenous knowledge (or assumptions) about the distribution of errors conditional on the covariates. In other words, the mere presence of covariate data, regardless of its relationship to $y$ and $z$, is insufficient to improve inference.

Although this result may be seen as discouraging, it highlights the additional assumption in the present work which enables sharper inference on the parameter of interest. In particular, some knowledge of the distribution of errors as a function of the covariates must be known. We stress that only some knowledge is necessary; the actual distribution of errors conditional on the covariates may remain unknown. For example, knowing that the probability of erroneous data is either orthogonal to a covariate or monotonically increasing in a covariate is typically sufficient to sharpen inference. Whether or not a particular assumption imposes sufficient restrictions on the relationship between the covariates and the occurrence of errors to improve inference is a function of the distributions of $y_1$ conditional on $x$. Although we characterize this function, since these conditional distributions are unknown, the value of covariate data in any given application remains an empirical question.

Dominitz and Sherman (1999) consider a special case of this problem, utilizing the Love Canal data considered by Lambert and Tierney (1997). These data consist of measurements of the concentrations of pollutants in the water, which are contaminated because the testing procedures occasionally measure the level of a compound other than the pollutant of interest. However, each observation is subjected to a validation test. If it passes the test, then the researcher knows the observation is drawn from the population of interest. However, failure to pass the test does not guarantee that the observation is spurious. The authors point out that throwing out unverified observations induces an upward bias in the estimates of mean concentrations since lower concentrations are harder to verify. In this case, the result of the verification test corresponds to the covariate $x$ and is correlated with the probability of having erroneous data (as well as the level of contamination). Dominitz and Sherman derive sharp bounds on the level of pollutants based on this additional information.

In another example, Hotz, Mullin and Sanders (1997) use miscarriages as an instrumental variable for teenage births. However, they note that not all miscarriages are random; some are behaviorally induced by activities such as smoking and drinking. First, they bound the fraction of non-random miscarriages. Then, they use the techniques of HM to construct bounds of the effects of teenage childbearing. The applied example at the end of this paper takes these data and conditions on smoking and drinking behavior. Unlike the verification test used with the Love Canal data, neither smoking nor drinking guarantee a particular outcome, but the odds of observing a non-random miscarriage increases when a woman engages in these activities. This additional information turns out to be sufficient to tighten the bounds by 50 percent.

Cross and Manski (2001) consider a similar problem to the one addressed here. They are concerned with making inference about the long regression $E\{y|x,z\}$ when the data only identify the short conditional distributions $P(y|x)$ and $P(z|x)$. As part of their work, Cross and Manski characterize the identification region of the family of conditional distributions of *y* given *x* and a particular value of *z*, treating the remaining values of *z* as erroneous data. They demonstrate that the Total Law of Probability restricts the identification region to a bounded convex set whose extreme points are the expectations of *J*-vectors of stacked distributions.

Interpreting *z* as an indicator variable for erroneous data and switching the object of interest to $P(y|z)$, Cross and Manski analyzes the identical problem as this paper. Given the assumption that $P(z|x)$ is known, for each element in the identification region we could aggregate the conditional distributions derived by Cross and Manski across the covariates *x* to attain the distribution of *y* conditional solely on *z*. In fact, this solution is a special case of our results and a useful baseline for demonstrating our contributions.

The major contribution of this paper is to relax the assumption that $P(z|x)$ is known. Without this assumption, the identification region for the family of conditional distributions of *y* given *x* and *z* expands. We construct the expanded identification region and demonstrate how to recover the identification region for the distribution of $y_1$, the distribution of *y* conditional on error-free data. Then, we use this identification region for the distribution of $y_1$ to develop bounds both on probabilities and on parameters of this distribution that respect stochastic dominance. Finally, we establish that the presence of a binding constraint beyond the Total Law of Probability on the distribution of error-ridden data conditional on the covariates is a necessary condition to sharpen inference, but it is not a sufficient condition. For the case of bounding parameters that respect stochastic dominance, we derive the sufficient conditions.

We illustrate key results of our analysis with numerical examples and conclude with an applied example, which demonstrates the ability of covariate data to substantial sharpen inference. All proofs are in the Appendix.

## 2. Basic Identification Analysis

### 2.1 Statement of the Problem

Suppose the marginal distribution of $y_1$ is the object of interest. However, instead of observing $y_1$, one observes a random variable $y$ whose value can be that of either $y_1$ or another random variable $y_0$, whose distribution is unknown. Since interest focuses on the distribution of $y_1$, realizations of $y$ corresponding to $y_0$ are erroneous. Let $z$ be an indicator variable for the presence of erroneous data ($z=1$ when the realization of $y$ is erroneous and zero otherwise). Then, $y \equiv y_0(1-z)+y_1 z$. Finally, let $\pi$ be the probability of observing erroneous data.

HM addressed the problem of what can be learned about the distribution of $y_1$ and parameters $\tau(P_1)$ (where $\tau(\cdot)$ maps $\Psi$, the space of probability distributions, into $\Re$), given observations on $y$ and an upper bound on $\pi$. We extend that research to quantify the value of covariate data that is correlated with $y$ and/or $z$. Specifically, let $X$ be a set of covariates with $d$ distinct values;[1] let $(Y,\Omega)$ be a measurable space; let $(y_0, y_1, z, x) \in Y \times Y \times \{0,1\} \times X$ be a random quadruplet distributed $P$; and let a random sample be drawn from $P$.

Denote the space of all probability distributions on $(Y,\Omega)$ by $\Psi$. Let $Q_k \equiv Q(y|x_k)$ represent the conditional distribution of $y$ given $x_k$ and $p_k \equiv P(x_k)$ denote the marginal distribution of the covariates. Both of these functions are identified by the observable data. Let $\pi_k \equiv P(z=1|x_k)$ represent the probability of observing erroneous data conditional on the covariates. Then, $w_k \equiv p_k(1-\pi_k)/(1-\pi) = \Pr(x_k|z=0)$ is the marginal distribution of the

[1] Although it is not theoretically necessary for $X$ to have a finite number of distinct values, it will be treated in this manner in empirical applications. In order to avoid additional notation, the discrete case is assumed from the start.

covariates given error-free data. Let $P_i \equiv P\left(y_i\right)$ denote the marginal distribution of $y_i$ and $P_{ij} \equiv P\left(y_i \middle| z = j\right)$ be the conditional distribution of $y_i$ conditional on the event $z = j$ for $i = 0,1$ and $j = 0,1$. Finally, let $P_{ijk} \equiv P\left(y_i \middle| z = j, x_k\right)$ represent the distribution of $y_i$ conditional on the event $z$ and the covariates $x_k$.

The inferential problem is that the sampling process does not identify $P_1$, but only $Q_k$. These distributions may be decomposed as

$$P_1 = \sum_{k=1}^{d} p_k \left[ \left(1-\pi_k\right) P_{11k} + \pi_k P_{10k} \right] \tag{2.1}$$

and

$$Q_k = \left(1-\pi_k\right) P_{11k} + \pi_k P_{00k} . \tag{2.2}$$

In robust estimation, the unknown $P_1$ is held fixed and the family of conditional distributions, $\left\{Q_k\right\}$, is allowed to range over all possible distributions consistent with (2.1) and (2.2). In identification analysis, these conditional distributions are held fixed since they are identified by the data and $P_1$ is allowed to range over all possible distributions consistent with (2.1) and (2.2). The objective is to set bounds on the unknown quantity $\tau\left(P_1\right)$.

## 2.2 The Contaminated and Corrupted Sampling Models

A frequently imposed assumption is that the occurrence of data errors is independent of the sample realizations from the population of interest, i.e.

$$P_1 = P_{11} . \tag{2.3}$$

When this assumption holds, the data are labeled contaminated and inferences about $P_{11}$ are equivalent to inferences about $P_1$. When this assumption fails to hold, the data are called corrupted.

For the sake of brevity, we focus on the contaminated case. The changes to the propositions in HM for the case of corrupted data closely parallel those that are covered for contaminated data and are available from the authors upon request.

### 2.3 Implications of an Upper Bound on the Error Probability

Assume that an upper bound, $\lambda$, on the probability of observing erroneous data is either known or can be consistently estimated. Furthermore, assume that this upper bound is non-trivial, i.e. $\pi \le \lambda < 1$. The Total Law of Probability requires

$$\pi = \sum_{k=1}^{d} p_k \pi_k \ , \tag{2.4}$$

so

$$\lambda \ge \sum_{k=1}^{d} p_k \pi_k \ . \tag{2.5}$$

Define Θ as the set of all *d*-dimensional vectors, $\boldsymbol{\pi} \equiv (\pi_1, \ldots, \pi_d)$, that satisfy both equation (2.5) with equality and any other known constraints on the conditional probability of observing erroneous data. For example, if *x* is known to be orthogonal to *z*, then $\pi_i = \pi_j$ for all *i* and *j*. We assume throughout the discussion that any additional restrictions on Θ leave it a compact set.[2]

We are now prepared to generalize HM's first proposition.

**PROPOSITION 1:** A. *Let the set* Θ *be known, then*

$$P_{11} \in \mathrm{Y}_{11}(\Theta) \equiv \mathrm{Y} \cap \left\{ \sum_{k=1}^{d} w_k P_{11k} : P_{11k} \in \mathrm{Y}_{11k}(\pi_k), \boldsymbol{\pi} \in \Theta \right\} \tag{2.6}$$

*where* $\Psi_{11k}(\pi_k) \equiv \Psi \cap \left\{ (Q_k - \pi_k \psi_{00k}) / (1 - \pi_k), \psi_{00k} \in \Psi \right\}$. *This restriction on* $P_{11}$ *is sharp.*

B. *Let* $\Theta_1 \subset \Theta_2$, *then* $\Psi_{11}(\Theta_1) \subset \Psi_{11}(\Theta_2)$.

[2] This assumption guarantees the existence of minimum and maximum values. Relaxing the assumption requires the notation to switch from min and max to inf and sup.

For a given conditional level of erroneous data, $\pi_k$ , HM demonstrates that $\Psi_{11k}(\pi_k)$ places sharp restrictions on the conditional distribution of $P_{11k}$. Proposition 1 states that sharp restriction on $P_{11}$ can be attained in a two-step process. First, for a fixed distribution of the erroneous data across the covariates, compute the set of all possible combinations of the conditional distributions where each conditional distribution is feasible ( $P_{11k} \in \Psi_{11k}(\pi_k)$ ) and is given a weight proportional to the probability of observing its associated covariates. Second, take the union of these sets across all feasible distributions of the erroneous data.

When the decomposition of erroneous error across the covariates is known, Θ reduces to a single point and the second step above is no longer necessary. In essence, this case reduces the restrictions on $P_{11}$ to a conditional distribution by conditional distribution application of the HM results.

A corollary of Proposition 1 is that the Total Law of Probability alone places insufficient constraints on the set Θ to provide additional information about $P_{11}$. In other words, the mere presence of covariate data does not provide additional information in the absence of knowledge about how the erroneous data is distributed across the covariates.

**COROLLARY 1.1:** *The Total Law of Probability places insufficient constraints on the conditional distribution of erroneous data with respect to the covariates to sharpen the restrictions on $P_{11}$ relative to the absence of covariate data.*

## 2.4 Sharp Bounds on Probabilities

We now develop the implications of Proposition 1 for the identification of probabilities. Specifically, Corollary 1.2 states sharp bounds on $P_{11}(A)$ for all measurable sets *A*.

**COROLLARY 1.2:** *Let it be known that* $\boldsymbol{\pi} \in \Theta$ *and* $A \in \Omega$. *Then*

$$P_{11}(A) \in \Upsilon_{11}(A;\Theta) \equiv \begin{bmatrix} \min\limits_{\boldsymbol{\pi}\in\Theta} \sum\limits_{k=1}^{d} w_k \max\left\{0, \left(Q_k(A) - \pi_k\right)/\left(1-\pi_k\right)\right\}, \\ \max\limits_{\boldsymbol{\pi}\in\Theta} \sum\limits_{k=1}^{d} w_k \min\left\{Q_k(A)/\left(1-\pi_k\right), 1\right\} \end{bmatrix}. \tag{2.7}$$

*These bounds on* $P_{11}(A)$ *are sharp.*

The bounds given by (2.7) are uninformative if and only if there exists a $\boldsymbol{\pi}$ such that

$$1-\pi_k \le Q_k(A) \le \pi_k \tag{2.8}$$

for each *k*. This condition is stronger than what is necessary in the absence of covariates. In particular, the necessary and sufficient condition for the bounds to be trivial in the absence of covariate data may be expressed as

$$1-\lambda = \sum_{k=1}^{d} p_k(1-\pi_k) \le \sum_{k=1}^{d} w_k Q_k(A) \le \sum_{k=1}^{d} p_k \pi_k = \lambda\,. \tag{2.9}$$

Equation (2.9) requires the weighted average conditional probability of the event *A* to fall between the weighted average level of error-free and erroneous data. If equation (2.8) is satisfied for all possible values of the covariate data then equation (2.9) must be satisfied. The converse is not true, so equation (2.8) is a stronger condition. In other words, covariate data can only reduce the range of events for which the data are uninformative.

An implication of Corollary 1.2 is that covariate data not only weakly reduces the range of events over which the data are uninformative, but also weakly tightens the bounds on the probability of all other events. Proposition 2 formalizes the conditions under which this weak improvement in the bounds becomes strict.

**PROPOSITION 2:** *Let it be known that* $\boldsymbol{\pi} \in \Theta$ *and* $A \in \Omega$. *Let* $\Pi_L$ *be the set of all* $\boldsymbol{\pi} \in \Theta$ *that minimize the lower bound on* $P_{11}(A)$. *Then, covariate data strictly increases the lower bound on* $P_{11}(A)$ *if and only if* $\exists \boldsymbol{\pi} \in \Pi_L$ *such that* (*i*) $Q_k(A) > \pi_k$ *for at least one k and* (*ii*) $Q_k(A) < \pi_k$ *for at least one other k.*

*Similarly, let* $\Pi_U$ *be the set of all* $\boldsymbol{\pi} \in \Theta$ *that maximize the upper bound on* $P_{11}(A)$. *Then, covariate data strictly decreases the upper bound on* $P_{11}(A)$ *if and only if* $\exists \boldsymbol{\pi} \in \Pi_U$ such that (*i*) $Q_k(A) > 1-\pi_k$ *for at least one k and* (*ii*) $Q_k(A) < 1-\pi_k$ *for at least one other k.*

The first requirement ensures that the lower (upper) bound on the probability is non-trivial. The second requirement, conditional on the first, ensures that the bounds using covariates data are tighter. To see why the second requirement is sufficient, consider the lower bound. When the lower bound is non-trivial in the absence of covariate data, it is constructed by assuming that all of the erroneous data is in the set *A*. However, when covariate data is used, the second requirement guarantees that for at least one covariate some (at a minimum, the fraction $\pi_k - Q_k(A)$) of the erroneous data is not in the set *A*. This ability to exclude erroneous data from the set *A* is what raises the lower bound on $P_{11}(A)$.

The following example illustrates both Corollary 2.1 and Proposition 2.

**EXAMPLE 1:** Let each observed conditional distribution be normal with mean *x* and variance one. Let $\lambda = 0.1$, *z* be independent of *x* and *x* be a binary variable taking the values plus and minus one with equal probability. Letting $\Phi$ denote the cumulative standard normal distribution, applying equation (2.7) yields that the cumulative distribution function (CDF) of $y_1$ is bounded below by

$$\frac{1}{2}\left[\max\left\{0, \frac{\Phi(t+1)-0.1}{0.9}\right\}+\max\left\{0, \frac{\Phi(t-1)-0.1}{0.9}\right\}\right]$$

and above by

$$\frac{1}{2}\left[\min\left\{1, \frac{\Phi(t+1)}{0.9}\right\}+\min\left\{1, \frac{\Phi(t-1)}{0.9}\right\}\right].$$

Figure 1 plots the observed distribution of *y*, as well as the lower and upper bounds on $P_{11}[-\infty, t]$ both utilizing the covariate data and ignoring the covariate data. Notice that the bounds using the covariate information <u>weakly</u> sharpen inference in all regions of the CDF. In particular, in both tails of the distribution and in the middle of the distribution there is no change in the bounds when covariate data is utilized. In the tails, the first requirement of Proposition 2 is not satisfied (the bounds remain uninformative). In the middle of the distribution, the second requirement is not met (all of the erroneous data can be assigned to the set of interest for each covariate).

## 3. Identification When *Y* Is the Real Line

In this section, we restrict *Y* to the extended real line and $\Omega$ to Lebesgue measurable sets. We also introduce some additional notation. First, let $r_k(\gamma)$ equal the $\gamma$-quantile of $Q_k$ for $\gamma \in (0,1]$. Second, it is useful to characterize the distribution function of *Y* conditional on *X* that is stochastically dominated by all other feasible conditional distributions. To attain this distribution, place all of the erroneous data as far out as possible in the right-hand tail of the observed distribution. This approach yields

$$L_k[-\infty,t] \equiv \begin{cases} Q_k[-\infty,t]/(1-\pi_k) & \text{if } t < r_k(1-\pi_k) \\ 1 & \text{if } t \ge r_k(1-\pi_k) \end{cases}. \tag{3.1}$$

Similarly, to attain the conditional distribution of $Y$ that stochastically dominates all other feasible distribution, place all of the erroneous data as far out as possible in the left-hand tail. This allocation produces

$$U_k[-\infty,t] \equiv \begin{cases} 0 & \text{if } t < r_k(\pi_k) \\ (Q_k[-\infty,t]-\pi_k)/(1-\pi_k) & \text{if } t \ge r_k(\pi_k) \end{cases}. \tag{3.2}$$

Note that both of these distributions are function of $\pi_k$, the conditional rate of erroneous data.

### 3.1 Sharp Bounds on Quantiles

For $\alpha \in (0,1]$, the $\alpha$-quantile of $P_{11}$ is $q_{11}(\alpha) \equiv \inf\{t : P_{11}[-\infty,t] \ge \alpha\}$. Proposition 2 gives sharp bounds on all quantiles of $P_{11}$.

**PROPOSITION 3:** *Let Y be the extended real line and* $\Omega$ *be the Lebesgue measurable sets. Let it be known that* $\boldsymbol{\pi} \in \Theta$. *Then*

$$q_{11}(\alpha) \in [q_L(\alpha), q_U(\alpha)] \tag{3.3}$$

*where*

$$q_L(\alpha) = \min_{\boldsymbol{\pi}\in\Theta} \inf\left\{q : \sum_{k=1}^{d} w_k L_k[-\infty,q] \ge \alpha\right\} \tag{3.4}$$

*and*

$$q_U(\alpha) = \max_{\boldsymbol{\pi}\in\Theta} \sup\left\{q : \sum_{k=1}^{d} w_k U_k[-\infty,q-\varepsilon] < \alpha \quad \forall \varepsilon > 0\right\}. \tag{3.5}$$

*These bounds on* $q_{11}(\alpha)$ *are sharp.*

For a fixed Θ, $q_L(\alpha)$ and $q_U(\alpha)$ are increasing functions of $\alpha$. So, the bounds on the quantile shift to the right as $\alpha$ increases. Also, for a fixed $\alpha$, the bounds on any quantile are weakly expanding in both directions as the set Θ expands. However, as long as Θ can rule out the entire sample being erroneous, the bounds on all quantiles remain informative.

The following example illustrates Proposition 3:

**EXAMPLE 2:** Let each observed conditional distribution be normal with mean *x* and variance one. Let $\lambda = 0.1$, *z* be independent of *x* and *x* be a binary variable taking the values plus and minus one with equal probability. Figure 2 plots the observed quantile function, as well as the upper and lower bounds on the quantile function for the population of interest both utilizing the covariate data and ignoring the covariate data.

As in Example 1, covariate data fails to tighten the bounds in the middle of the distribution. Since the quantile function is the inverse of the CDF, this result was predictable given the results in Example 1 (or Proposition 2). However, unlike Example 1, covariate data tightens the bounds on the quantile function in the tails. This change follows from the fact that the quantile function remains informative in the tails.

### 3.2 Sharp Bounds on Parameters that Respect Stochastic Dominance

If *F* and *G* are distributions on the extended real line *Y*, *F* stochastically dominates *G* if $F[-\infty,t] \le G[-\infty,t]$ for all $t \in Y$. A parameter $\tau(\cdot)$ respects stochastic dominance if $\tau(F) \ge \tau(G)$ whenever *F* stochastically dominates *G*. Common examples includes quantiles and means of monotone functions of random variables. Proposition 3 provides sharp bounds on parameters that respect stochastic dominance.

**PROPOSITION 4:** *Let Y be the extended real line and* Ω *be the Lebesgue measurable sets. Let it be known that* $\boldsymbol{\pi} \in \Theta$. *Let* $\tau : \Psi \to \Re$ *respect stochastic dominance. Then*

$$\tau(P_{11}) \in \left[ \min_{\boldsymbol{\pi}\in\Theta} \tau\left( \sum_{k=1}^{d} w_k L_k \right), \max_{\boldsymbol{\pi}\in\Theta} \tau\left( \sum_{k=1}^{d} w_k U_k \right) \right]. \tag{3.6}$$

*These bounds on* $\tau(P_{11})$ *are sharp.*

When the distribution of the erroneous data across the covariates is known (Θ is a singleton), the bounds on $\tau(P_{11})$ are the weighted average of the bounds derived in HM for each of the conditional distribution. On the other hand, when the distribution of the erroneous data by covariates is uncertain, the Total Law of Probability creates a negative relationship between the bounds for the separate covariates. Consider the lower bound. If $\pi_k$ increases to $\pi'_k$, then $L_k$ stochastically dominates $L'_k$, which implies that $\tau(L_k) \geq \tau(L'_k)$, i.e. the lower bound weakly decreases. However, whenever the amount of erroneous data attributed to one covariate increase, the Total Law of Probability guarantees that the amount of erroneous data attributable to at least one other covariate must decrease. So, the lower bound for another covariate will weakly increase.

To determine when Θ places sufficient constraints on $\boldsymbol{\pi}$ for the covariate data to tighten the bounds, consider the problem in the absence of covariate data. In the absence of covariate data, labeling all the data above the (1-$\lambda$)-quantile of the *Q* distribution erroneous attains the lower bound. Similarly, claiming all the data below the $\lambda$-quantile of the *Q* distribution is erroneous attains the upper bound. Let $\eta_{Lk}$ be the proportion of $Q_k$ that falls below the $\lambda$-quantile of the *Q* distribution. Similarly, let $\eta_{Uk}$ be the proportion of $Q_k$ that falls above the (1-$\lambda$)-quantile of the *Q* distribution. Proposition 5 shows that if $\boldsymbol{\eta}_L \equiv (\eta_{L1} \dots \eta_{Ld})$ and $\boldsymbol{\eta}_U$ are both elements of Θ, then the covariates provide no additional restrictions on $\tau(P_{11})$.

**PROPOSITION 5:** *Let Y be the extended real line and Ω be the Lebesgue measurable sets. Let it be known that $\boldsymbol{\pi} \in \Theta$. Let $\tau : \Psi \to \Re$ respect stochastic dominance. If $\boldsymbol{\eta}_L \in \Theta$, then the covariates and the maintained assumptions about the set Θ fail to increase the lower bound on $\tau(P_{11})$. Similarly, if $\boldsymbol{\eta}_U \in \Theta$, then the covariates and the maintained assumptions about the set Θ fail to decrease the upper bound on $\tau(P_{11})$. These conditions become necessary if $\tau(F) > \tau(G)$ whenever F stochastically dominates G and y is a continuous random variable.*

In essence, covariate data tighten the bounds whenever assigning the erroneous data to the most extreme realizations in either tail of the observed distribution is inconsistent with the restrictions on the distribution of erroneous data across the covariates. Proposition 5 formalizes when such inconsistencies occur. However, although such a proposition can be stated, whether or not $\boldsymbol{\eta}_L, \boldsymbol{\eta}_U \in \Theta$ remains an empirical question.

The following example illustrates Proposition 4:

**EXAMPLE 3:** Let each observed conditional distribution be normal with mean $x$ and variance one. Let $\lambda = 0.1$, $z$ be independent of $x$ and $x$ be a binary variable taking the values plus and minus one with equal probability. The tenth and ninetieth percentiles of $Q$ are ±1.85, respectively. So, ignoring covariate data

$$\frac{1}{2}\left(\frac{1}{0.9}\right)\int_{-\infty}^{1.85} u\left[\phi\left(u+1\right)+\phi\left(u-1\right)\right]du \le E\left\{y_1\right\} \le \frac{1}{2}\left(\frac{1}{0.9}\right)\int_{-1.85}^{\infty} u\left[\phi\left(u+1\right)+\phi\left(u-1\right)\right]du .$$

Therefore, $-0.448 \le E\left\{y_1\right\} \le 0.448$.

For each conditional distribution, the tenth and ninetieth percentiles of $Q_x$ are ±1.282 standard deviations from the mean, respectively. Thus, Proposition 4 yields

$$\begin{aligned}\frac{1}{2}\left(\frac{1}{0.9}\right)\left[\int_{-\infty}^{0.282} u\phi\left(u+1\right)du + \int_{-\infty}^{2.282} u\phi\left(u-1\right)du\right] &\le E\left\{y_1\right\} \\ &\le \frac{1}{2}\left(\frac{1}{0.9}\right)\left[\int_{-2.282}^{\infty} u\phi\left(u+1\right)du + \int_{0.282}^{\infty} u\phi\left(u-1\right)du\right].\end{aligned}$$

Therefore, $-0.329 \le E\left\{y_1\right\} \le 0.329$, which represents a 27 percent tightening of the bounds.

## 4. An Application to Teenage Childbearing

To illustrate estimation of the bounds, we consider data on maternal outcomes for teenage mothers. Hotz, Mullin and Sanders (1997) used these same data from the NLSY and the results in HM to construct bounds of the effects of teenage childbearing on future earnings. The basic idea of their work was to treat women who miscarried as teenagers as a control group for those who gave birth. However, not all miscarriages are random; some are behaviorally induced by activities such as smoking and drinking. Furthermore, some miscarriages occur to women intending to have an abortion. In other words, the researcher observes a population of women intending to have a birth who experience random miscarriages contaminated both with women who experience non-random miscarriages and with women intending to have abortions. The goal of this section is to determine how much tighter the bounds on future earning become when covariate data is incorporated into the analysis.

Hotz, Mullin and Sanders estimate an upper bound on the probability of erroneous data based on the relative frequency of births to abortions in the population and the probability of smoking and/or drinking during pregnancy.[3] Based on their techniques, the contamination in the miscarriage population does not exceed 24 percent for black women and 27 percent for non-black women.[4] The first row of Table 1 presents the bounds on the effect of teenage births on women's annual labor market earnings at age 27. As seen in the table, the width of the bounds for both racial groups is between $4,000 and $5,000.

The second row of Table 1 displays the bounds of the same variable after conditioning on smoking and drinking behavior.[5] An upper bound on the level of contamination for each cell is estimated under the identical assumptions used in the unconditional bounds (no additional assumptions have been invoked). Utilizing these covariate data reduces the width of the bounds by approximately 50 percent.

The third and final row of Table 1 shows the bounds conditional on the quartile of a woman's AFQT (Armed Forces Qualifying Test) score. Hotz, Mullin and Sanders presented bounds conditional on this variable under the strong assumption that quartile of AFQT is orthogonal to erroneous data. Although not shown here, the data can reject that assumption (although this rejection does not affect the qualitative nature of their findings). Maintaining the same assumptions as above, the correct upper bound on contamination for each quartile was constructed and used in the estimation of these bounds. As evidenced in the table, incorporating

[3] The estimates presented here differ slightly from those in Hotz, Mullin and Sanders (1997). The primary cause for the difference is that all observations with missing data in any of the covariates utilized (smoking, drinking or AFQT) have been dropped in the current analysis. Additionally, the estimates presented here employ none of the kernel-smoothing techniques implemented by Hotz, Mullin and Sanders to estimate means when conditioning on quartile of AFQT.

[4] Black women are more likely to smoke or drink, but less likely to have an abortion than non-black women.

[5] The conditional bounds are based on the behavior in the miscarriage sample, but the weights for aggregating the conditional distributions, $\{w_k\}$, are estimated by the covariate distribution in the sample of teenage mothers.

AFQT tightens the bounds, but by less than smoking and drinking status. Additionally, the impact differs substantially by racial group. The width of the bounds reduces 36 percent for non-black women, but only 10 percent for black women.

## 5. Conclusion

Identifying bounds on parameters of interest under relative weak and, hence, more plausible, sets of assumptions has the potential to clarify numerous outstanding questions in economics and the social sciences in general. For example, bounding the returns to schooling or the benefits of prenatal care under relatively weak assumptions that most researchers would find believable could help focus policy debates and the allocation of government resources. The current paper has shown how incorporating covariate data into the construction of those bounds has the potential to increase their precision.

Although the focus of the paper has been identification, more work is needed on estimation. Empirical work will encounter the "curse of dimensionality." The estimation of the conditional distributions is similar to non-parametric estimation, since each unique combination of covariate values is treated in isolation. Therefore, the necessary sample size for meaningful estimation grows rapidly with the number of covariates considered. Kernel smoothing methods can be used to address this problem, but when samples sizes are "small," the bounds conditional on covariates are no longer guaranteed to fall within the bounds ignoring the covariate data. Additionally, it is unclear how to optimally smooth across cells when the degree of contamination in neighboring cells differs.

## Appendix

PROOF OF PROPOSITION 1:

A. Start with the case in which $\Theta$ is a singleton, so $\boldsymbol{\pi}$ is known. HM Proposition 1 shows that $\Psi_{11k}(\pi_k)$ places sharp restrictions on $P_{11k}$ for a fixed value of $\pi_k$. Hence, the feasible values of $P_{11}$ are given by equation (2.6), $\Psi_{11}(\boldsymbol{\pi})$. When $\Theta$ is not a singleton, $P_{11} \in \bigcup_{\boldsymbol{\pi}\in\Theta} \Psi_{11}(\boldsymbol{\pi}) = \Psi_{11}(\Theta)$.

B. If $\Theta_1 \subset \Theta_2$, then $\Psi_{11}(\Theta_1) \equiv \bigcup_{\boldsymbol{\pi}\in\Theta_1} \Psi_{11}(\boldsymbol{\pi}) \subset \bigcup_{\boldsymbol{\pi}\in\Theta_2} \Psi_{11}(\boldsymbol{\pi}) \equiv \Psi_{11}(\Theta_2)$.

(Restricting the set $\Theta$ to be the boundary of the set of all vectors that satisfy equation (2.5) and the exogenous restrictions on the distribution of the erroneous data (i.e. requiring equation (2.5) to hold with equality in the definition of the set $\Theta$) does not affect the set $\Psi_{11}(\Theta)$. HM Proposition 1 (C) demonstrates that $\Psi_{11k}(\pi_k) \subset \Psi_{11k}(\pi'_k)$ for all $\pi_k < \pi'_k$. Therefore, any point on the interior generates a set of distributions that are a subset of the set of distributions generated by a boundary point.)

PROOF OF COROLLARY 1.1:

HM Proposition 1 shows $\Psi_{11}(\pi) \equiv \Psi \cap \{(Q-\pi\psi_{00})/(1-\pi), \psi_{00} \in \Psi\}$ places sharp restrictions on $P_{11}$ in the absence of covariate data, so $\Psi_{11}(\Theta) \subseteq \Psi_{11}(\pi)$. Furthermore, for each $P_{11} \in \Psi_{11}(\pi)$ there exists a $\psi_{00}$ such that $(Q-\pi\psi_{00})/(1-\pi) \in \Psi$. By the Total Law of Probability $\exists \boldsymbol{\gamma} \in \Re^d$ such that $0 \le \gamma_k \le p_k/\pi$, $\sum_{k=1}^{d} \gamma_k = 1$ and $\psi_{00} = \sum_{k=1}^{d} \gamma_k \psi_{00k}$. Let $\boldsymbol{\pi}^* \in \Re^d$ be a vector whose $k^{\text{th}}$ component is $\pi_k = \pi\gamma_k / p_k$. Therefore, $P_{11} \in \Psi_{11}(\boldsymbol{\pi}^*)$. Since $\sum_{k=1}^{d} p_k \pi_k = \sum_{k=1}^{d} p_k (\pi\gamma_k / p_k) = \pi$, $\boldsymbol{\pi}^* \in \Theta$ and $\Psi_{11}(\boldsymbol{\pi}^*) \subset \Psi_{11}(\Theta)$. Thus, $\Psi_{11}(\pi) \subseteq \Psi_{11}(\Theta)$. Hence, $\Psi_{11}(\pi) = \Psi_{11}(\Theta)$.

PROOF OF COROLLARY 1.2:

Start with the case in which $\Theta$ is a singleton, so $\boldsymbol{\pi}$ is known. HM Corollary 1.2 demonstrates that $P_{11k}(A) \in \Psi_{11k}(A, \pi_k) \equiv [0,1] \cap \left[ (Q_k(A) - \pi_k)/(1-\pi_k), Q_k(A)/(1-\pi_k) \right]$. Thus, $P_{11}(A)$ is the weighted sum of the lower and upper bounds on the conditional distributions as given in equation (2.7). When $\Theta$ is not a singleton, it immediately follows that the bounds are given by the union of the bounds over all possible $\boldsymbol{\pi} \in \Theta$.

PROOF OF PROPOSITION 2:

HM Corollary 1.2 demonstrates that the bounds in the absence of covariate data are $P_{11}(A) \in \Psi_{11}(A, \pi) \equiv [0,1] \cap \left[ (Q(A) - \pi)/(1-\pi), Q(A)/(1-\pi) \right]$. Suppose $\exists \boldsymbol{\pi} \in \Pi_L$ such that $Q_{k_1}(A) > \pi_{k_1}$ and $Q_{k_2}(A) < \pi_{k_2}$. Let $\boldsymbol{\pi}' = \boldsymbol{\pi}$ except that $\pi'_{k_1} = \pi_{k_1} + \left[ \pi_{k_2} - Q_{k_2}(A) \right]$ and

$\pi'_{k_2} = Q_{k_2}(A)$. Then, the lower bound associated with $\boldsymbol{\pi}'$ is smaller than the lower bound associated with $\boldsymbol{\pi}$. Since $\boldsymbol{\pi}$ and $\boldsymbol{\pi}'$ are both feasible in the absence of covariate data, the lower bound without covariate data is no greater than the minimum of these two. Thus, these conditions are sufficient for covariate data to increase the lower bound.

Suppose $Q_k(A) < \pi_k$ for all *k*. Then, the lower bound is zero and the covariate data fails to increase the lower bound. Instead, suppose $Q_k(A) > \pi_k$ for all *k*. Then, the lower bound is $\sum_{k=1}^{d} w_k \left(Q_k(A) - \pi_k\right) / (1-\pi_k)$. Substitute in $w_k \equiv p_k(1-\pi_k)/(1-\pi)$ and simplify to get $(1-\pi)^{-1} \sum_{k=1}^{d} p_k \left(Q_k(A) - \pi_k\right) = \left(Q(A) - \pi\right) \big/ (1-\pi)$. Again, the covariate data fails to decrease the lower bound. Thus, these conditions are necessary for covariate data to increase the lower bound.

The same arguments can be applied to the upper bound.

PROOF OF PROPOSITION 3:

Since the quantile function respects stochastic dominance, this result is an application of Proposition 4.

PROOF OF PROPOSITION 4:

Start with the case in which $\Theta$ is a singleton, so $\boldsymbol{\pi}$ is known. In the proof of HM Proposition 4, they demonstrate that $L_k \in \boldsymbol{\Psi}_{11k}(\pi_k)$. Thus, by construction, $\sum_{k=1}^{d} w_k L_k \in \boldsymbol{\Psi}_{11}(\boldsymbol{\pi})$; hence, $\tau\left(\sum_{k=1}^{d} w_k L_k\right)$ is a feasible value for $\tau(P_{11})$. Furthermore, HM establish that $L_k$ is stochastically dominated by every member of $\boldsymbol{\Psi}_{11k}(\pi_k)$. Therefore, $\sum_{k=1}^{d} w_k L_k$ is stochastically dominated by every member of $\boldsymbol{\Psi}_{11}(\boldsymbol{\pi})$, which implies that $\tau\left(\sum_{k=1}^{d} w_k L_k\right)$ is the smallest feasible value of $\tau(P_{11})$ for this fixed value of $\boldsymbol{\pi}$. When $\Theta$ is not a singleton, take the minimum lower bound over all possible $\boldsymbol{\pi} \in \Theta$.

An analogous argument establishes the sharpness of the upper bound.

PROOF OF PROPOSITION 5:

HM Proposition 4 guarantees that the lower bound without covariate data is no greater than the lower bound with covariate data. If $\boldsymbol{\eta}_L \in \Theta$, then the lower bound with covariate data no greater than the lower bound without covariate data. Thus, the two lower bounds must be equal.

Similarly, HM Proposition 4 guarantees that the upper bound without covariate data is no smaller than the upper bound with covariate data. If $\boldsymbol{\eta}_U \in \Theta$, then the upper bound with covariate data no smaller than the upper bound without covariate data. Thus, the two upper bounds must be equal.

To establish necessity in the last statement of the proposition, note that $\tau\left(\sum_{k=1}^{d} w_k L_k\left(\eta_L\right)\right)$ is equal to the lower bound in the absence of covariate data. If *y* is a continuous random variable, then $\sum_{k=1}^{d} w_k L_k\left(\eta_L\right)$ is the only member of $\Psi_{11}(\pi)$ stochastically dominated by all members of $\Psi_{11}(\pi)$. Therefore, if $\boldsymbol{\eta}_L \notin \Theta$, then $P_{11}$ must stochastically dominate $\sum_{k=1}^{d} w_k L_k\left(\eta_L\right)$. Thus, $\tau\left(P_{11}\right) > \tau\left(\sum_{k=1}^{d} w_k L_k\left(\eta_L\right)\right)$ because $\tau(F) > \tau(G)$ whenever *F* stochastically dominates *G*.

Similarly, $\tau\left(\sum_{k=1}^{d} w_k U_k\left(\eta_U\right)\right)$ is equal to the upper bound in the absence of covariate data. If *y* is a continuous random variable, then $\sum_{k=1}^{d} w_k U_k\left(\eta_U\right)$ is the only member of $\Psi_{11}(\pi)$ stochastically dominated by all members of $\Psi_{11}(\pi)$. Therefore, if $\boldsymbol{\eta}_U \notin \Theta$, then $\sum_{k=1}^{d} w_k U_k\left(\eta_U\right)$ must stochastically dominate $P_{11}$. Thus, $\tau\left(P_{11}\right) < \tau\left(\sum_{k=1}^{d} w_k U_k\left(\eta_U\right)\right)$ because $\tau(F) > \tau(G)$ whenever *F* stochastically dominates *G*.

Cross, Phillip J. and Charles F. Manski. “Regressions, Short and Long,” *Econometrica* (2001).

Dominitz, Jeff and Robert P. Sherman. “Identification and Estimation with Contaminated and partially Verified Data,” working paper (November, 1999).

Horowitz, Joel L. and Charles F. Manski. “Identification and Robustness with Contaminated and Corrupted Data,” *Econometrica* Volume 63, Number 2 (March, 1995): 281 - 302.

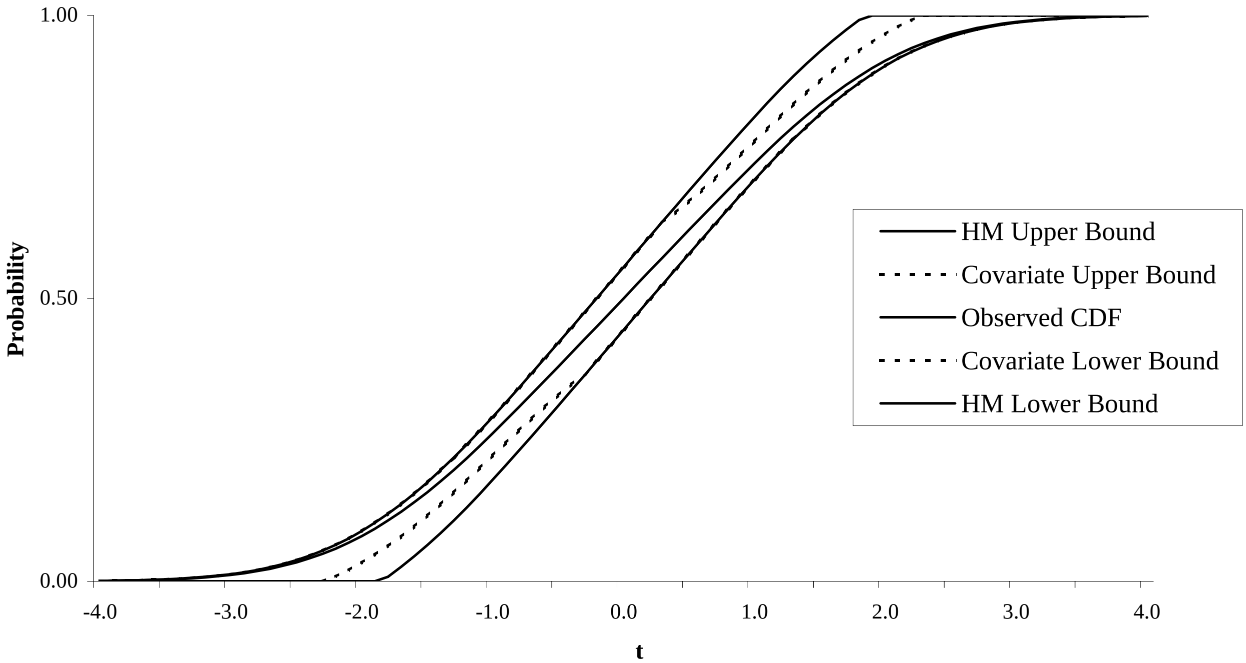
Horowitz, Joel L. and Charles F. Manski. “What Can Be Learned about Population Parameters when the Data Are Contaminated,” Maddala, G. S. and C. R. Rao (eds.) *Handbook of Statistics*: Robust Inference, Volume 15, Amsterdam, North Holland (1997): 439 - 466.

Hotz, V. Joseph, Charles H. Mullin, and Seth Sanders. “Bounding Causal Effects Using Data from a Contaminated Natural Experiment: Analyzing the Effects of Teenage Childbearing,” *Review of Economic Studies* Volume 64, Issue 4 (October 1997): 575 - 603.

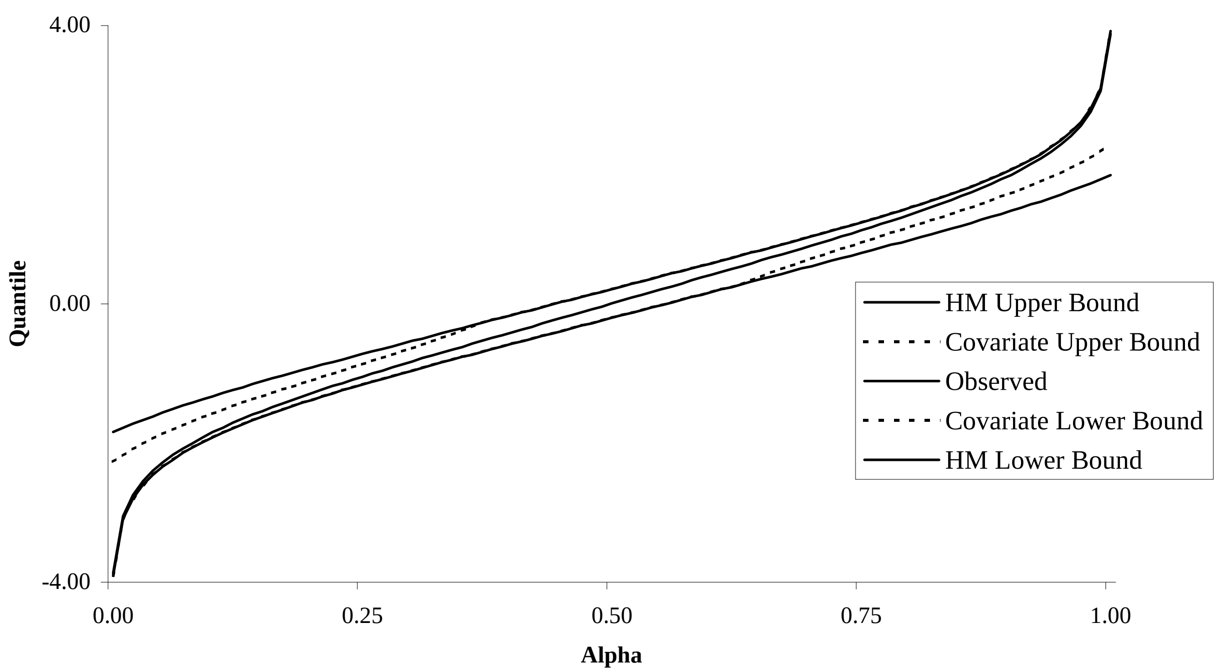
Lambert, Diane and Luke Tierney. “Nonparametric maximum Likelihood Estimation from Samples with Irrelevant Data and Verification Bias,” *Journal of the American Statistical Association* Volume 92 (September, 1997): 937 – 944.

Manski, Charles F. “Non-Parametric Bounds on Treatment Effects,” *American Economic Review* Volume 80, Issue 2 (May, 1990): 319 - 323.

## Figure I - Bounds on Probabilities




## Figure 2 - Quantiles



# Table 1

**Bounds on the Effects of Teenage Births on Women's Annual Labor Market Earnings at Age 27**

**Construction of Bounds Conditional on No Covariates, Smoking and Drinking, or AFQT**

| | Black Women | | | Non-Black Women | | |
|---|---|---|---|---|---|---|
| **Covariates** | **Lower Bound** | **Upper Bound** | **Percent Improvement** | **Lower Bound** | **Upper Bound** | **Percent Improvement** |
| **None - Baseline** | -1260 | 3200 | | 641 | 5411 | |
| **Smoking and Drinking** | -477 | 1705 | 0.51 | 1623 | 3981 | 0.51 |
| **AFQT** | -927 | 3087 | 0.10 | 1509 | 4547 | 0.36 |