UNIVERSITÄT
BERN

Faculty of Business, Economics
and Social Sciences

Department of Economics

Set Identification of Generalized Linear Predictors in the Presence of Non-Classical Measurement Errors

Kaspar Wüthrich

13-04

August 2013

DISCUSSION PAPERS

Schanzeneckstrasse 1
Postfach 8573
CH-3001 Bern, Switzerland
http://www.vwi.unibe.ch

Set Identification of Generalized Linear Predictors in the Presence of Non-Classical Measurement Errors

Kaspar Wüthrich*
University of Bern

August 19, 2013

Abstract

This paper studies the identification of coefficients in generalized linear predictors where the outcome variable suffers from non-classical measurement errors. Combining a mixture model of data errors with the bounding procedure proposed by Stoye (2007), I derive bounds on the coefficient vector under different non-parametric assumptions about the structure of the measurement error. The method is illustrated by analyzing a simple earnings equation.

JEL Classification: C2, C21, J24

Keywords: Generalized linear predictor; Non-classical measurement error; Contaminated sampling; Corrupt sampling; Multiplicative mean independence; Stochastic dominance; Nonparametric bounds.

*University of Bern, Department of Economics, Schanzeneckstrasse 1, CH-3001 Bern, Switzerland, phone: +41 31 631 4089, email: kaspar.wuethrich@vwi.unibe.ch.

1 Introduction

Validation studies suggest that many important economic variables suffer from non-classical measurement error (see e.g. Bound et al., 2001, for a survey).[1] Point identification in the presence of non-classical measurement error typically relies on strong assumptions regarding the structure of the data and the distribution of the measurement errors that may be inappropriate in many applications (Chen et al., 2011). This motivates a recent line of research to focus on set identification of the parameters of interest under weaker and more credible assumptions. Examples include the estimation of features of marginal distributions in the presence of data errors (e.g. Horowitz and Manski, 1995; Dominitz and Sherman, 2006; Kreider and Pepper, 2011), direct misclassification in discrete distributions (Molinari, 2008), estimation of treatment effects in the presence of non-standard data errors (e.g. Gundersen et al., 2012; Kreider et al., 2012), and the identification of coefficients in linear regressions where the covariates suffer from misclassification (Bollinger, 1996; Kreider, 2010).

In this paper, I propose a flexible two-step approach to construct bounds for the coefficients in generalized linear predictors in the presence of non-classical measurement error in the dependent variable. In the first step, I derive identified sets for the conditional cumulative density function (CDF) and the conditional mean of the outcome variable that are translated into bounds for the coefficient vector in generalized linear predictors in the second step. Following a growing body of literature data errors are conceptualized in a mixture model.[2] This framework models the observed conditional distribution of the outcome variable as a mixture of the true distribution that is of interest and an unknown and unrestricted erroneous distribution. I explore the identifying power of different non-parametric assumptions about the underlying structure of the measurement error. In particular, the data corruption and the data contamination assumption (Horowitz and Manski, 1995) and the multiplicative mean independence assumption (Kreider and Pepper, 2011) are discussed. Moreover, I show how to incorporate the presumption of under- and overreporting in surveys for example due to social desirability by imposing monotonicity assumptions in the spirit of Dominitz and Sherman (2006). These assumptions are related to the restrictions considered by Molinari (2008) who addresses the problem of data errors in discrete variables.

[1] I refer to classical measurement error if the errors in the dependent variable and the covariates are independent of the true variables and the error term in the true model.

[2] Examples include Horowitz and Manski (1995); Dominitz and Sherman (2004, 2006); Kreider and Pepper (2011).

The second step uses the results in Stoye (2007) to translate the identified sets for the conditional CDF and the conditional mean into identified sets for the parameter vector in generalized linear predictors. The two-step procedure naturally separates the specification of the data error process from the computation of the bounds on the coefficient vector which increases transparency and illustrates the identifying power of the underlying restrictions on the data error process.

Importantly for the interpretation and applicability of the results in this paper, Ponomareva and Tamer (2011) show that the identified set for the best linear predictor in Stoye (2007) coincides with what they call the "least squares set". The least squares set consists of the set of parameters that are best linear approximations to the conditional mean function. Consequently, the best linear predictor considered in this paper provides an easy to interpret quantity of interest under misspecification.

The method is applied by analyzing the problem of drawing inferences in a simple earnings regression where the dependent variable suffers from non-classical measurement error. The results based on data from the Swiss Household Panel[3] suggest that imposing additional non-parametric assumptions substantially narrows the data corruption bounds. Moreover, the empirical illustration indicates that allowing for non-classical and unrestricted measurement error substantially impedes identification.

The remainder of this paper is structured as follows: section 2 presents general framework and introduces the mixture model of data errors, section 3 discusses the two-step approach to identification. The empirical illustration is presented in section 4 and section 5 concludes. All proofs are collected in the appendix.

2 Notation and Framework

I am interested in the characterization of the identified sets for the coefficients in generalized linear predictors where the dependent variable Y is continuous and suffers from non-classical measurement error. The generalized linear predictor $\hat{Y}$ of Y from X is given by (Stoye, 2007),

$$\hat{Y} = G(x\theta) \tag{1}$$

[3]This study has been realized using the data collected by the Swiss Household Panel (SHP), which is based at the Swiss Centre of Expertise in the Social Sciences FORS. The project is financed by the Swiss National Science Foundation.

with

$$\begin{aligned}\theta &\equiv \left(\int x'xF_{yx}\right)^{-1}\int x'G(y)^{-1}dF_{yx} \\ &= (EX'X)^{-1}EX'G^{-1}(Y)\end{aligned}$$

where $F_{yx}(y,x)$ is the cumulative density function (CDF), $Y \in [K_0, K_1] \subseteq \mathbb{R}$ is the dependent variable and $X \in \mathbb{R}^K$ denotes a row vector of covariates. $G : \mathbb{R} \to \mathbb{R}$ is a strictly increasing and preassigned link function and θ is the coefficient vector of interest. An important special case of this framework is the best linear predictor under square loss or Ordinary Least Squares regression.

I decompose the identification analysis into two steps. In the first step, I analyze set identification of $F_{y|x}(y,x)$ and the conditional expectation $E(Y|X=x)$ of the distribution of $(Y|X)$ when Y suffers from non-classical measurement error.

Measurement errors are conceptualized in the mixture model of data errors (Horowitz and Manski, 1995). In this model it is assumed that the researcher is interested in the conditional distribution of Y. But instead she observes the conditional distribution of the random variable O that is generated by the following probability mixture,

$$O = YZ + V(1-Z) \tag{2}$$

where the conditioning on $X = x$ is kept implicit. Z is a binary variable indicating whether a realization of O is a draw from the distribution of the random variable Y that is of interest, or from the erroneous distribution of the random variable V that is unknown and left completely unrestricted. In order to gain intuition and to compare the mixture model with other data error models, I rewrite Equation 2 as follows

$$\begin{aligned}O &= YZ + V(1-Z) \\ &= Y + \xi\end{aligned} \tag{3}$$

where $\xi \equiv (1-Z)(V-Y)$. Hence, the mixture model of data errors can be viewed as a special case of continuous variables with non-classical error ξ that is not independent of the outcome variable Y (Chen et al., 2011). It is worth noting that whilst most of the statistics literature is concerned with errors that affect every observation, the mixture model assumes

that some realizations of the observed conditional distribution are error-free while others are subject to arbitrary error patterns (Bound et al., 2001).

Whenever $P(Z = 1|X = x) < 1$, i.e. measurement error occurs with a positive probability, $F_{y|x}(y, x)$ and $E(Y|X = x)$ are generally not point identified in the mixture model. However, it is possible to derive identified sets for $F_{y|x}(y, x)$ and $E(Y|X = x)$ under non-parametric assumptions about the measurement error.[4] For later reference, let $(\underline{F}_{y|x}(y, x), \overline{F}_{y|x}(y, x))$ denote the bounding functions for $F_{y|x}(y, x)$ such that $\underline{F}_{y|x}(y, x) \leq F_{y|x}(y, x) \leq \overline{F}_{y|x}(y, x)$ for all (y, x) and let $(\underline{E}(x), \overline{E}(x))$ denote the bounds on the conditional mean, $E(Y|X = x)$ such that $\underline{E}(x) \leq E(Y|X = x) \leq \overline{E}(x)$.

In the second step of the identification analysis, I translate the identified sets for $F_{y|x}(y, x)$ and $E(Y|X = x)$ into bounds for the coefficient vector θ. This is achieved by applying Proposition 2.1 of Stoye (2007). This proposition derives tight bounds for the coefficient vector of interest when the distribution of the dependent variable is only set identified. For a general link function $G(\cdot)$ the analysis requires the characterization of $\underline{F}_{y|x}(y, x)$ and $\overline{F}_{y|x}(y, x)$. In the important special case where $G(\cdot)$ is linear, knowledge of $\underline{E}(x)$ and $\overline{E}(x)$ is sufficient.

Throughout this paper, I assume perfect observability of the covariates and that the data are missing at random. The framework can be extended to the case of missing and erroneous outcome data by incorporating the results of Manski (2003, Chapter 1). The analysis of erroneous covariate and outcome data is beyond the scope of this paper.[5]

3 Identification Analysis

This section is concerned with the identification analysis. Finite sample issues are discussed in Section 4.

[4] Examples include Horowitz and Manski (1995); Dominitz and Sherman (2004); Kreider and Pepper (2011) among others.

[5] For example Horowitz et al. (2003) analyze the identification of general statistical functionals when covariate and outcomes are missing arbitrarily.

3.1 Step 1: Bounding the Outcome Distribution

To illustrate the identification problem, decompose conditional CDFs of O and Y using the Law of Total Probability.

$$F_{o|x}(o,x) = p(x)F_{y|z,x}(y,1,x) + (1-p(x))F_{v|z,x}(v,0,x) \tag{4}$$

$$F_{y|x}(y,x) = p(x)F_{y|z,x}(y,1,x) + (1-p(x))F_{y|z,x}(y,0,x) \tag{5}$$

where $p(x) \equiv P(Z=1|X=x)$. I write $F_{o|x}(o,x)$ and $F_{v|x}(v,x)$ for the CDFs of O and V given $X=x$ and $F_{y|z,x}(y,i,x)$ and $F_{v|z,x}(v,i,x)$ for the CDFs of Y and V given $X=x$ and $Z=i$ for $i=0,1$.

The identification problem arises from the fact that the sampling process reveals only $F_{o|x}(o,x)$ but not $F_{y|x}(y,x)$, the object of interest. In particular, knowledge of $F_{o|x}(o,x)$ does not imply restrictions on $F_{y|x}(y,x)$ (Horowitz and Manski, 1995). To obtain bounds that are tighter than the logical unit range additional assumptions are needed. To start out, assume that the probability of misreporting is known,

Assumption 1. *$p(x)$ is known $\forall x \in \mathbb{R}^K$*

I will maintain this assumption for deriving the main results in this section. At the end of the section, I then show how to relax assumption 1 to the case of a known lower bound on $p(x)$.

Data Contamination and Data Corruption

Under assumption 1 only, Horowitz and Manski (1995) show that the following bounds on the conditional CDF and the conditional mean apply.

$$F_{y|x}(y,x) \in [0,1] \cap [F_{o|x}(o,x) - (1-p(x)), F_{o|x}(o,x) + (1-p(x))] \tag{6}$$

and

$$\begin{aligned} E(Y|X=x) \in [&p(x)E(O|O \leq \tau_{o|x}(p(x), X=x) + (1-p(x))K_0, \\ &p(x)E(O|O > \tau_{o|x}(1-p(x)), X=x) + (1-p(x))K_1] \end{aligned} \tag{7}$$

where $\tau_{o|x}(\cdot)$ is the quantile function of $F_{o|x}(o,x)$. Following Huber (1981) and Horowitz and Manski (1995), I will refer to these bounds as data corruption bounds. These bounds can

be considered as worst case bounds under assumption 1 since they only impose that $p(x)$ is known. Unfortunately, they are often wide and uninformative in practice. Thus, it might be desirable to impose additional structure on the measurement error process in order to narrow the bounds. One possibility is consider to a situation where the mixing process Z is independent of the outcome variable.

Assumption 2. *Z and Y are independent given $X = x$, $\forall x \in \mathbb{R}^K$*

Assumption 2 implies that $F_{y|z,x}(y, 1, x) = F_{y|x}(y, x)$ and $E(Y|Z = 1, X = x) = E(Y|X = x)$. Following Huber (1981) and Horowitz and Manski (1995), this assumption is referred to as data contamination. Horowitz and Manski (1995) show that under assumption 2 the following bounds on the conditional CDF and the conditional mean apply,

$$F_{y|x}(y, x) \in [0, 1] \cap \left[\frac{F_{o|x}(o, x) - (1 - p(x))}{p(x)}, \frac{F_{o|x}(o, x)}{p(x)} \right] \tag{8}$$

and

$$E(Y|X = x) \in [E(O|O \leq \tau_{o|x}(p(x)), X = x), E(O|O > \tau_{o|x}(1 - p(x)), X = x)] \tag{9}$$

These bounds are a weak subsets of the data corruption bounds. Although the independence assumption of the contamination model can have substantial identifying power, it may not be plausible in many applications. For example, misreporting of income, drug usage or health status is often thought to be related to the true value. This motivates the analysis of the identifying power of alternative assumptions.

Stochastic Dominance Monotonicity Constraints

It is widely known that personally and socially sensitive topics (e.g. drug usage) are prone to underreporting for socially undesirably topics and overreporting for socially desirable behavior and attitudes (e.g. charity variables) (Bound et al., 2001). The presumption of under- and overreporting can be incorporated in the mixture model of data errors by imposing the following stochastic dominance monotonicity assumptions:[6]:

Assumption 3. $F_{y|z,x}(y, 0, x) \geq F_{v|z,x}(v, 0, x)$, $\forall x \in \mathbb{R}^K$ *(Overreporting)*

[6]These assumptions can be interpreted as the distributional equivalent of the monotonicity assumption considered by Dominitz and Sherman (2006, assumption 7) who analyze school performance measures using a mixture model.

Assumption 4. $F_{y|z,x}(y,0,x) \leq F_{v|z,x}(v,0,x)$, $\forall x \in \mathbb{R}^K$ *(Underreporting)*

Since the mean respects stochastic dominance, assumption 3 implies $E(Y|Z=0,X=x) \leq E(V|Z=0,X=x)$ while assumption 4 implies $E(Y|Z=0,X=x) \geq E(V|Z=0,X=x)$.

The applicability of assumptions 3 and 4 is not at all limited to the cases just outlined. Assumptions 3 and 4 might be gainfully invoked if validation studies point at persistent under- or overreporting or in cases where missing values are imputed but the imputation procedure is known to be downwards or upwards biased.

Propositions 1 and 2 present the bounds on the conditional distribution function and on the conditional mean under both stochastic dominance assumptions.

Proposition 1. *Under assumptions 1 and 3,* $F_{y|x}(y,x)$ *is bounded by*

$$F_{y|x}(y,x) \in [0,1] \cap [F_{o|x}(o,x), F_{o|x}(o,x) + (1-p(x))] \tag{10}$$

Under assumptions 1 and 4, $F_{y|x}(y,x)$ *is bounded by*

$$F_{y|x}(y,x) \in [0,1] \cap [F_{o|x}(o,x) - (1-p(x)), F_{o|x}(o,x)] \tag{11}$$

These bounds are sharp given the assumptions.

Proposition 2. *Under assumptions 1 and 3,* $E(Y|X=x)$ *is bounded by*

$$E(Y|X=x) \in [p(x)E(O|O \leq \tau_{o|x}(p(x)), X=x) + (1-p(x))K_0, E(O|X=x)] \tag{12}$$

Under assumptions 1 and 4, $E(Y|X=x)$ *is bounded by*

$$E(Y|X=x) \in [E(O|X=x), p(x)E(O|O > \tau_{o|x}(1-p(x)), X=x) + (1-p(x))K_1] \tag{13}$$

These bounds are sharp given the assumptions.

Note that the bounds in Propositions 1 and 2 are weak subsets of the bounds under data corruption. The stochastic dominance assumptions may be combined with the independence assumption of the data contamination model to further narrow the bounds.

Proposition 3. *Under assumptions 1, 2 and 3, $F_{y|x}(y,x)$ is bounded by*

$$F_{y|x}(y,x) \in [0,1] \cap \left[F_{o|x}(o,x), \frac{F_{o|x}(o,x)}{p(x)} \right] \tag{14}$$

Under assumptions 1, 2 and 4, $F_{y|x}(y,x)$ is bounded by

$$F_{y|x}(y,x) \in [0,1] \cap \left[\frac{F_{o|x}(o,x) - (1-p(x))}{p(x)}, F_{o|x}(o,x) \right] \tag{15}$$

These bounds are sharp given the assumptions.

Proposition 4. *Under assumptions 1, 2 and 3, $E(Y|X=x)$ is bounded by*

$$E(Y|X=x) \in [E(O|O \leq \tau_{o|x}(p(x)),x), E(O|X=x)] \tag{16}$$

Under assumptions 1, 2 and 4, $E(Y|X=x)$ is bounded by

$$E(Y|X=x) \in [E(O|X=x), E(O|O > \tau_{o|x}(1-p(x)), X=x)] \tag{17}$$

These bounds are sharp given the assumptions.

Note that that the bounds in Propositions 3 and 4 are weak subsets of the data contamination bounds.

Intuition behind Propositions 1 - 4 comes from the decomposition of $F_{o|x}(o,x)$ and $F_{y|x}(y,x)$ in equations 4 and 5. Notice that the difference between both distributions originates from the difference between $F_{y|x,z}(y,0,x)$ and $F_{v|x,z}(v,0,x)$. Thus, restricting the relationship between $F_{y|z,x}(y,0,x)$ and $F_{v|z,x}(v,0,x)$ using stochastic dominance assumptions directly restricts the object of interest, $F_{y|x}(y,x)$, to be stochastically larger or smaller than the observed distribution.

To illustrate the identifying power of the stochastic dominance assumptions, I reconsider the example in Horowitz and Manski (1995).

Example 1. *Let the (unconditional) observed distribution be a standard normal with CDF $\Phi(y)$ and assume that the probability of a true report is $p = 0.9$. Then, the following bounds on the CDF of $F_y(y)$ apply,*

Assumptions	$\underline{F}_y(y)$	$\overline{F}_y(y)$
1	$\max[0, \Phi(y) - 0.1]$	$\min[1, \Phi(y) + 0.1]$
1 & 2	$\max[0, (\Phi(y) - 0.1)/0.9]$	$\min[1, \Phi(y)/0.9]$
1 & 3	$\Phi(y)$	$\min[1, \Phi(y) + 0.1]$
1 & 4	$\max[0, \Phi(y) - 0.1]$	$\Phi(y)$
1, 2 & 3	$\Phi(y)$	$\min[1, \Phi(y)/0.9]$
1, 2 & 4	$\max[0, (\Phi(y) - 0.1)/0.9]$	$\Phi(y)$

Figures 1 - 3 provide a graphical illustration of Example 1.

— Insert Figures 1 - 3 around here —

Multiplicative Mean Independence

Until now, I have focused on assumptions that allow to derive bounds on the conditional CDF. Consequently, these assumptions are applicable to the analysis of generalized linear predictors with arbitrary $G(\cdot)$. Here, I discuss the multiplicative mean independence assumption (Kreider and Pepper, 2011) that focuses on conditional means only. Therefore, it's applicability is limited to the important special case of linear link functions. Multiplicative mean independence relaxes the (mean) independence assumption of the data contamination model by allowing the conditional means to differ by a factor of proportionality, denoted by $\gamma(x)$. Formally,

Assumption 5. $E(Y|Z = 0, X = x) = \gamma(x)E(Y|Z = 1, X = x)$, $\forall x \in \mathbb{R}^K$

The proportionality factor $\gamma(x)$ is either known or can be bounded. It is instructive to discuss the implications of assumption 5 and the choice of the proportionality factor $\gamma(x)$ by means of the following example (Kreider and Pepper, 2011) : the use of illicit drugs is thought to be as prevalent among inaccurate reporters as among accurate reporters. Hence, $\gamma(x) \geq 1$ might be plausible while the restriction $\gamma(x) = 1$ (contamination model) is untenable.

Kreider and Pepper (2011) show that under assumption 1 and 5 the following bounds

apply:

$$E(Y|X=x) \in [LB_x(p(x)), UB_x(p(x))] \tag{18}$$

where

$$LB_x(p(x)) = \max\left\{\frac{K_0}{\gamma(x)}, E(O|O \leq \tau_{o|x}(p(x)), X=x)(1+(\gamma(x)-1)(1-p(x)))\right\} \tag{19}$$

and

$$LB_x(p(x)) = \min\left\{\frac{K_1}{\gamma(x)}, E(O|O > \tau_{o|x}(1-p(x)), X=x)(1+(\gamma(x)-1)(1-p(x)))\right\} \tag{20}$$

These bounds reduce to the data contamination bounds if $\gamma(x) = 1$.

The multiplicative mean independence assumption provides an interesting alternative to the assumptions discussed so far. On the one hand it relaxes the data contamination assumption and on the other hand it imposes a complementary type of structure on the measurement error process as opposed to the stochastic dominance assumptions.

Assumptions on the Misreporting Probability

So far, the analysis was based on the assumption that $p(x)$ is known. This assumption is useful in cases where missing data is imputed and the proportion of missing values is known (e.g. Horowitz and Manski, 1995) or when $p(x)$ can be estimated from validation studies. When there is no obvious way to determine an appropriate $p(x)$ it might still be plausible to assume a lower bound on the fraction of draws from the distribution of interest (e.g. Huber, 1981; Horowitz and Manski, 1995; Kreider and Pepper, 2011).

Assumption 6. $p(x) \geq \lambda(x) > 0$, $\forall x \in \mathbb{R}^K$

Horowitz and Manski (1995, Proposition 1.D.) show that incorporating assumption 6 amounts to substituting $\lambda(x)$ for $p(x)$ in expressions for the bounds under data contamination and data corruption.

Combining the multiplicative mean independence assumption with assumption 6 yields the following bounds (Kreider and Pepper, 2011):

$$E(Y|X=x) \in [\inf_{p(x)\in\mathcal{P}(x)} LB_x(p(x)), \sup_{p(x)\in\mathcal{P}(x)} UB_x(p(x))] \tag{21}$$

where $\mathcal{P}(x)$ denotes the set of feasible values for $p(x)$ that is restricted by Equation 6 and the following conditions

$$\frac{K_0}{\gamma(x)} \leq E(O|O > \tau_{o|x}(1-p(x)), X = x) \tag{22}$$

and

$$\frac{K_1}{\gamma(x)} \geq E(O|O \leq \tau_{o|x}(p(x)), X = x). \tag{23}$$

If these conditions are satisfied the lower bound simplifies to $\min\{E(O|X = x), LB(\lambda(x))\}$ for $\gamma(x) \leq 1$ and the upper bound simplifies to $\max\{E(O|X = x), UB(\lambda(x))\}$. If these conditions do not hold, then $E(Y|X = x) = E(O|X = x)$.

As for the bounds under data corruption and data contamination, the bounds under the stochastic dominance assumptions 3 and 4 can be modified by substituting $\lambda(x)$ for $p(x)$. Proposition 5 gives the formal result:

Proposition 5. *Under assumptions 3 and 6, the $F_{y|x}(y, x)$ and $E(Y|X = x)$ are bounded by*

$$F_{y|x}(y, x) \in [0, 1] \cap [F_{o|x}(o, x), F_{o|x}(o, x) + (1 - \lambda(x))] \tag{24}$$

and

$$E(Y|X = x) \in [\lambda(x)E(O|O \leq \tau_{o|x}(\lambda(x)), X = x) + (1 - \lambda(x))K_0, E(O|X = x)] \tag{25}$$

Under assumptions 4 and 6, $F_{y|x}(y, x)$ and $E(Y|X = x)$ are bounded by

$$F_{y|x}(y, x) \in [0, 1] \cap [F_{o|x}(o, x) - (1 - \lambda(x)), F_{o|x}(o, x)] \tag{26}$$

and

$$E(Y|X = x) \in [E(O|X = x), \lambda(x)E(O|O > \tau_{o|x}(1-\lambda(x)), X = x) + (1-\lambda(x))K_1] \tag{27}$$

Under assumptions 2, 3 and 6, $F_{y|x}(y,x)$ and $E(Y|X=x)$ are bounded by

$$F_{y|x}(y,x) \in [0,1] \cap \left[F_{o|x}(o,x), \frac{F_{o|x}(o,x)}{\lambda(x)} \right] \tag{28}$$

and

$$E(Y|X=x) \in [E(O|O \leq \tau_{o|x}(\lambda(x)), X=x), E(O|X=x)] \tag{29}$$

Under assumptions 2, 4 and 6 $F_{y|x}(y,x)$ and $E(Y|X=x)$ are bounded by

$$F_{y|x}(y,x) \in [0,1] \cap [F_{o|x}(o,x) - (1-\lambda(x))\lambda(x), F_{o|x}(o,x)] \tag{30}$$

and

$$E(Y|X=x) \in [E(O|X=x), E(O|O > \tau_{o|x}(1-\lambda(x)), X=x)] \tag{31}$$

These bounds are sharp given the assumptions.

Despite the fact that I only focus on contamination, corruption, multiplicative mean independence and stochastic dominance assumptions in this paper, the framework can be extended to encompass many other potential assumptions that put some structure on the measurement error process.

3.2 Step 2: Bounding the Coefficient Vector in Generalized Linear Predictors

Stoye (2007, Proposition 2.1) shows that if the marginal distribution of X, $F_x(x)$, the bounding functions, $\underline{F}_{y|x}(y,x)$ and $\overline{F}_{y|x}(y,x)$ and the bounds on the conditional mean, $\underline{E}(x)$ and $\overline{E}(x)$, are known, then for any pre-assigned $c \in \mathbb{R}^K$, $c \cdot \theta$ is bounded by

$$c \cdot \left(\int x' x dF_x \right)^{-1} \int x' \underline{g}(x) dF_x \leq c \cdot \theta \leq c \cdot \left(\int x' x dF_x \right)^{-1} \int x' \overline{g}(x) dF_x \tag{32}$$

where $\underline{g}(x)$ and $\overline{g}(x)$ are defined by

$$\underline{g}(x) \equiv \begin{cases} \int G^{-1}(y) d\overline{F}_{y|x} & \text{if } c \cdot (\int x' x dF_x)^{-1} x' > 0 \\ \int G^{-1}(y) d\underline{F}_{y|x} & \text{if } c \cdot (\int x' x dF_x)^{-1} x' \leq 0 \end{cases} \tag{33}$$

and

$$\overline{g}(x) \equiv \begin{cases} \int G^{-1}(y) d\underline{F}_{y|x} & \text{if } c \cdot (\int x'x dF_x)^{-1} x' > 0 \\ \int G^{-1}(y) d\overline{F}_{y|x} & \text{if } c \cdot (\int x'x dF_x)^{-1} x' \leq 0 \end{cases} \tag{34}$$

If G is linear $\underline{g}(x)$ and $\overline{g}(x)$ can be identified by

$$\underline{g}(x) \equiv \begin{cases} G^{-1}(\underline{E}(x)) & \text{if } c \cdot (EX'X)^{-1} x' > 0 \\ G^{-1}(\overline{E}(x)) & \text{if } c \cdot (EX'X)^{-1} x' \leq 0 \end{cases} \tag{35}$$

and

$$\overline{g}(x) \equiv \begin{cases} G^{-1}(\overline{E}(x)) & \text{if } c \cdot (EX'X)^{-1} x' > 0 \\ G^{-1}(\underline{E}(x)) & \text{if } c \cdot (EX'X)^{-1} x' \leq 0 \end{cases} \tag{36}$$

These bounds are tight whenever the bounding functions $\underline{F}_{y|x}(y,x)$ and $\overline{F}_{y|x}(y,x)$ exploit all the available information about $F_{y|x}(y,x)$.

4 Illustration: Earnings Equation

To illustrate the approach proposed in this paper, I analyze the identification of the coefficient in a simple earnings equation. I focus on a scenario where self-reported earnings are assumed to be measured with non-classical errors, while the covariates are assumed to be error-free. The framework is in the spirit of Bound et al. (1994) who use earnings equations and labor supply functions to illustrate the impact of measurement error. I use pooled data from the Swiss Household Panel (SHP) 2009-2010 (waves 11 and 12). The SHP is a yearly panel study following a random sample of households in Switzerland over time, interviewing all household members.[7]

I consider a stylized version of the Mincer earnings equation (Mincer, 1958, 1974). The dependent variable Y is annual gross earnings in Swiss Francs and X contains years of schooling[8] experience in yearst[9], experience squared and a constant. I decide to confine the analysis to male full-time workers to avoid sample selection issues associated with female and part-

[7] Complete information including free access to the data can be acquired trough *www.swisspanel.ch*.

[8] This variable is constructed from the categorical information in the Swiss Household Panel using information about the Swiss education system. See the appendix for more details.

[9] The SHP includes information about actual experience, i.e. the number of years spent in paid work. I use this information instead of construction experience from schooling and age.

time workers. Unfortunately, the data set has missing outcomes and missing covariates. To focus on measurement error, I discard all observations with missing data. When interpreting the results it is important to keep in mind the implicit underlying assumption, namely that Y and X are missing at random. Given this selection scheme, the final sample consists of $N = 2667$ observations. Table 1 contains descriptive statistics.

Section 3 has discussed identification. I provide a brief discussion of estimation and inference here. An obvious estimation approach for θ is to replace the population quantities by their sample analogues, i.e. to consider $\theta_n \equiv (E_n X'X)^{-1} E_n X'Y$. Confidence intervals for each component of the estimated θ_n are computed by means of a bootstrap. The intervals are constructed as suggested by Imbens and Manski (2004) and exhibit a nominal coverage probability with respect to the corresponding component of θ. It is worth noting that the coverage probability applies to the population parameter as opposed to the population bounds (see Imbens and Manski, 2004; Stoye, 2007, for a further discussion of this point).

For the present case of a linear link function $G(\cdot)$, Stoye (2007) suggests that any implementation algorithm should operate on X only and "fill in" the values of Y according to the above formulas. This principle is adapted to the case of erroneous data. In particular, I first compute lower and upper bounds, $\underline{E}(x)$ and $\overline{E}(x)$, for every $X = x$ under the different assumptions discussed in section 3.2. Second, $\underline{E}(x)$ and $\overline{E}(x)$ are "filled in" according to the conditions (35) and (36). For tractability, I consider discretized versions of experience and experience squared and I choose a rather coarse grid for years of schooling (see Tables 2 and 3 in the appendix).

Table 4 shows OLS estimates ignoring measurement error. The results suggest that on average earnings are significantly increasing in schooling. The coefficients of *experience* and $experience^2$ are both significant and imply a concave experience log earnings profile.

— Insert Table 4 around here —

The results from estimating the earnings equation under different non-parametric assumptions about the data error process for different values of $p(x)$ are contained in Tables 5 - 8. These tables also contain 95% confidence intervals for each component of θ_n that are computed using a $N = 100$ bootstrap. I choose K_0 and K_1 to be the minimum and the maximum of the marginal empirical distribution of Y. Although the parameters $p(x)$ and $\gamma(x)$ can in principle be arbitrary functions of x, they are assumed to be constant for the ease of exposition.

— Insert Tables 5 - 8 around here —

The data corruption bounds and the associated 95% confidence intervals are generally wide emphasizing the severity of the identification problem in the presence of arbitrary measurement error even for a small fraction of the population. Coupling the data corruption bounds with the independence assumption of the data contamination model narrows the bounds considerably. For example, if $p(x) = 0.8$ it reduces the width of the bounds by around 80%. The mean independence assumption of the data contamination model can be relaxed by allowing the conditional means to differ by a factor of proportionality. For the purpose of illustration, I consider two values for $\gamma(x)$, $\gamma(x) \in \{0.8, 1.2\}$. The results suggest that relaxing the data contamination assumption by imposing the multiplicative mean independence assumption does not substantially change the width of the bounds and 95% confidence intervals.

If the researcher is not willing to maintain (mean) independence assumptions but instead believes that under- or overreporting might be a realistic feature of the self-reported earnings data, she can formalize such presumptions by imposing the assumption 3 or 4. While powerful compared to the data corruption model, imposing the stochastic dominance assumptions has less identifying power the the (mean) independence assumptions. For $p(x) = 0.8$, the data corruption bounds narrow by around 35% (assumption 3) to 60% (assumption 4). Combining the data contamination model with the stochastic dominance assumptions yields by far the most informative bounds. In particular, for $p(x) = 0.8$ the data corruption bounds are reduced by over 90%.

5 Conclusion

Measurement error is a common problem in empirical research. In this paper, I analyze identification of the coefficients in generalized linear predictors where the dependent variable suffers from non-classical measurement error. I propose a two-step approach to construct identified sets for the coefficient vector of interest. In the first step, I derive bounds on the conditional CDF and on the conditional mean of the outcome variable. This is achieved by conceptualizing measurement error in a mixture model. I consider bounds under alternative sets of assumptions including stochastic dominance assumptions that can be motivated by under- and overreporting in surveys. The second step uses the procedure proposed by Stoye (2007) to translate the identified sets derived in the first step into bounds on the coefficient

vector of interest. This two-step procedure features two main advantages: first, the natural separation of the specification of the data error process increases transparency about the underlying data error process. Second, the first step is very flexible in the sense that it allows to incorporate a lot of alternative assumptions about the structure of the measurement error.

The two-step method is illustrated by analyzing a simple earnings equation using Swiss data. The following conclusions can be drawn from the empirical application. First, if additional assumptions such as data contamination, multiplicative mean independence and stochastic dominance are plausible, they can gainfully be invoked to narrow considerably the data corruption bounds. Second, the stochastic dominance assumptions can have substantial identifying power, in particular, when coupled with the data contamination assumption. Third, allowing for arbitrary measurement error only for a small proportion of the sample causes substantial problems for identification and inference.

The analysis in this paper assumes perfect observability of the covariates and that the data are missing at random. Clearly, these assumption may not hold in many applications. It is thus an important task for future research to investigate identification under in more general setups.

References

Bollinger, C. R., 1996. Bounding mean regressions when a binary regressor is mismeasured. Journal of Econometrics 387-399, 387–399.

Bound, J., Brown, C., Duncan, G. J., Rodgers, W. L., 1994. Evidence on the validity of cross-sectional and longitudinal labor market data. Journal of Labor Economics 12(3), 345–368.

Bound, J., Brown, C., Nancy, A., Mathiowetz, N., 2001. Measurement error in survey data. In: Heckman, J. J., Leamer, E. E. (Eds.), Handbook of Econometrics. Vol. 5. Elsevier Science, North Holland, Ch. 59, pp. 3705–3843.

Chen, X., Hong, H., Nekipleov, D., 2011. Nonlinear models of measurement errors. Journal of Economic Literature 49(4), 901–937.

Dominitz, J., Sherman, R. P., 2004. Sharp bounds under contaminated or corrupted sampling with verification, with an application to environmental pollutant data. Journal of Agricultural, Biological, and Environmental Statistics 9(3), 319–338.

Dominitz, J., Sherman, R. P., 2006. Identification and estimation of bounds on school performance measures: a nonparametric analysis of a mixture model with verification. Journal of Applied Econometrics 21(8), 1295–1326.

Gundersen, C., Kreider, B., Pepper, J. V., 2012. The impact of the national school lunch program on child health: A nonparametric bounds analysis. Journal of Econometrics 166, 79–91.

Horowitz, J. L., Manski, C. F., 1995. Identification and robustness with contaminated and corrupted data. Econometrica 62(2), 281–302.

Horowitz, J. L., Manski, C. F., Ponomareva, M., Stoye, J., 2003. Computation of bounds on population parameters when the data are incomplete. Reliable Computing 9(6), 419–440.

Huber, P., 1981. Robust Statistics. Wiley, New York.

Imbens, G. W., Manski, C. F., 2004. Confidence intervals for partially identified parameters. Econometrica 72(6), 1845–1857.

Kreider, B., 2010. Regression coefficient identification decay in the presence of infrequent classification errors. The Review of Economics and Statistics 92(4), 1017–1023.

Kreider, B., Pepper, J. V., 2011. Identification of expected outcomes in a data error mixing model with multiplicative mean independence. Journal of Business & Economic Statistics 29(1), 49–60.

Kreider, B., Pepper, J. V., Gundersen, C., Jolliffe, D., 2012. Identifying the effects of SNAP (food stamps) on child health outcomes when participation is endogenous and misreported. Journal of the American Statistical Association 107(499), 958–975.

Manski, C. F., 2003. Partial Identification of Probability Distributions. Springer.

Mincer, J., 1958. Investment in human capital and personal income distribution. Journal of Political Economy 66(4), 281–302.

Mincer, J., 1974. Schooling, Experience and Earnings. Columbia University Press for National Bureau of Economic Research, New York.

Molinari, F., 2008. Partial identification of probability distributions with misclassified data. Journal of Econometrics 144, 81–117.

Ponomareva, M., Tamer, E., 2011. Misspecification in moment inequality models: back to moment equalities? The Econometrics Journal 14, 186–203.

Stoye, J., 2007. Bounds on generalized linear predictors with incomplete outcome data. Reliable Computing 13, 293–302.

A Proofs

Proof of Proposition 1. The proof borrows from proof of Theorem 3 in Dominitz and Sherman (2006). Because assumption 3 does not restrict $F_{y|z,x}(y,1,x)$, the identified set for $F_{y|z,x}(y,1,x)$ for a given $p(x)$ follows from Horowitz and Manski (1995, Corollary 1.2),

$$\begin{aligned} F_{y|z,x}(y,1,x) &\in [0,1] \cap [(F_{o|x}(o,x) - (1-p(x)))/p(x), F_{o|x}(o,x)/p(x)] && (37) \\ &\equiv \mathcal{F}_{y|z,x}(y,1,x,p(x)) && (38) \end{aligned}$$

The identified set for $F_{y|x}(y,x)$ is given by

$$\begin{aligned} F_{y|x}(y,x) \in \{&p(x)F_{y|z,x}(y,1,x) + (1-p(x)F_{y|z,x}(y,0,x) : \\ &F_{y|z,x}(y,1,x) \in \mathcal{F}_{y|z,x}(y,1,x,p(x)) \\ &\text{and } F_{y|z,x}(y,0,x) \in \mathcal{F}_{y|z,x}(y,0,x,p(x))\} \end{aligned} \tag{39}$$

where $\mathcal{F}_{y|z,x}(y,0,x,p(x))$ denotes the identified set $F_{y|z,x}(y,0,x)$ for a given $p(x)$. Because the sampling process is uninformative on $F_{y|z,x}(y,0,x)$, without further restrictions, we have that

$$\mathcal{F}_{y|z,x}(y,0,x,p(x)) = [0,1] \tag{40}$$

However, assumption 3 together with Equation 4 implies the following restriction on $F_{y|z,x}(y,0,x)$,

$$\begin{aligned} F_{y|z,x}(y,0,x) &\geq F_{e|z,x}(e,0,x) \\ &= \frac{F_{o|x}(o,x) - p(x)F_{y|z,x}(y,1,x)}{1-p(x)} \end{aligned} \tag{41}$$

Thus, under assumption 3, $\mathcal{F}_{y|z,x}(y,0,x,p(x))$ is given by

$$\mathcal{F}_{y|z,x}(y,0,x,p(x)) = [0,1] \cap \left[\frac{F_{o|x}(o,x) - p(x)F_{y|z,x}(y,1,x)}{1-p(x)}, 1\right] \tag{42}$$

and sharp bounds on $F_{y|x}(y,x)$ are given by

$$\begin{aligned} F_{y|x}(y,x) \in \{&p(x)F_{y|z,x}(y,1,x) + (1-p(x)F_{y|z,x}(y,0,x) : \\ &F_{y|z,x}(y,1,x) \in \mathcal{F}_{y|z,x}(y,1,x,p(x)) && (43) \\ &\text{and } F_{y|z,x}(y,0,x) \in \mathcal{F}_{y|z,x}(y,0,x,p(x))\} \\ = \ &[0,1] \cap [F_{o|x}(o,x), F_{o|x}(o,x) + (1-p(x))] && (44) \end{aligned}$$

The proof under assumption 4 is analogous. □

Proof of Proposition 2. Proposition 1 gives bounds on the cumulative density function $F_{y|x}(y,x)$ under assumptions 1, 3 and 4. The sharp bounds in Proposition 2 follow from the fact that the mean respects stochastic dominance. □

Proof of Proposition 3. The proof borrows from proof of Theorem 3 in Dominitz and Sherman (2006). By assumption 2, $F_{y|z,x}(y,1,x) = F_{y|z,x}(y,0,x) = F_{y|x}(y,x)$. As in the proof of Proposition 1, assumption 3 together with Equation 4 imposes the following additional restriction on $F_{y|z,x}(y,0,x)$,

$$F_{y|z,x}(y,0,x) \geq \frac{F_{o|x}(o,x)-p(x)F_{y|z,x}(y,1,x)}{1-p(x)} \tag{45}$$

By assumption 2, it follows that

$$F_{y|z,x}(y,0,x) \geq \frac{F_{o|x}(o,x)-p(x)F_{y|z,x}(y,1,x)}{1-p(x)} \tag{46}$$

$$= \frac{F_{o|x}(o,x)-p(x)F_{y|z,x}(y,0,x)}{1-p(x)} \tag{47}$$

Solving for $F_{y|z,x}(y,0,x)$ yields

$$F_{y|z,x}(y,0,x) \geq F_{o|x}(o,x) \tag{48}$$

Assumption 2 then implies,

$$F_{y|z,x}(y,1,x) \geq F_{o|x}(o,x) \text{ and } F_{y|x}(y,x) \geq F_{o|x}(o,x) \tag{49}$$

Using similar arguments as in the proof of Proposition 1, one can show that sharp bounds on $F_{y|x}(y,x)$ are given by

$$F_{y|x}(y,x) \in [0,1] \cap [F_{o|x}(o,x), F_{o|x}(o,x)/p(x)] \tag{50}$$

The proof under assumptions 2 and 4 is analogous. □

Proof of Proposition 4. Proposition 3 gives bounds on the cumulative density function $F_{y|x}(y,x)$ under assumptions 1, 2, 3 and 4. The sharp bounds in Proposition 4 follow from the fact that the mean respects stochastic dominance. □

Proof of Proposition 5. Propositions 1 - 4 show that under assumptions 1 and 3 respectively assumptions 1, 2 and 3, the upper bound on $F_{y|x}(y,x)$ and the lower bound on $E(Y|X=x)$ coincide with the bounds under data corruption respectively data contamination. Hence, the sharp bounds in Proposition 5 follows directly from Corollary 1.2 and Proposition 4 in

Horowitz and Manski (1995). Furthermore, notice that the sharp lower bound on $F_{y|x}(y,x)$ and the sharp upper bound on $E(Y|X=x)$ do not depend on $p(x)$ and hence remain unchanged. The proof under assumptions 2, 4 and 6 is analogous. □

B Tables and Figures

Table 1: Summary Statistics

	Mean	St.Dev.	Min.	Max.
annual earnings	109328.200	68540.110	300	2000000
years of schooling	14.763	2.470	9	18
experience	25.053	11.540	0	60

Notes: The sample includes male full-time workers for the years 2009 and 2010. Earnings are measured in Swiss Francs, years of schooling and experience are measured in years. *Source:* Swiss Household Panel (SHP).

Table 2: Schooling categories and assigned years of schooling

Category	Years of Schooling
Compulsory Schooling	9
Apprenticeship/Matura	13
Higher Vocational School	16
University	18

Notes: I use the variable *EDCAT* that contains information about the highest education level achieved (11 categories). These categories are summarized in four groups and linked to years of schooling using information about the Swiss education system, see e.g *http://www.edk.ch/dyn/14861.php* (last accessed 2013, January 7).

Table 3: Discretized Variables

Reported exper., E_{rep}	Discretized exper.	Discretized exper. squared
$E_{rep} < 10$	5	34
$10 \leq E_{rep} < 20$	15	219
$20 \leq E_{rep} < 30$	24	591
$30 \leq E_{rep} < 40$	34	1168
$E_{rep} > 40$	45	2077

Notes: Experience is measured in years. To account for unequal dispersion of the data within cells, within-cell means are chosen as grid points for both, *experience* and $experience^2$.

Table 4: Least Squares Estimates

	Coeff.	Std.Err.	[95% Conf.	Interval]
school.	0.098	0.004	0.091	0.106
exper.	0.054	0.003	0.047	0.060
$exper.^2$	-0.001	0.000	-0.001	-0.001
constant	9.285	0.065	9.158	9.413

Notes: Dependent variable: logarithm of annual gross income in Swiss Francs. *Source:* Swiss Household Panel (SHP).

Table 5: Bounds on the coefficient vector, $p(x) = 0.90$

	Data Corruption				Data Contamination			
	95%	L.B.	U.B.	95%	95%	L.B.	U.B.	95%
school.	-0.106	-0.098	0.274	0.282	0.059	0.066	0.129	0.137
exper.	-0.100	-0.095	0.184	0.191	0.019	0.024	0.075	0.082
exper.^{2}	-0.003	-0.003	0.002	0.002	-0.001	-0.001	0.000	0.000
constant	6.428	6.591	12.257	12.439	8.653	8.822	9.874	10.013
	Mult. Mean. Indep., $\gamma(x) = 1.2$				Mult. Mean. Indep., $\gamma(x) = 0.8$			
	95%	L.B.	U.B.	95%	95%	L.B.	U.B.	95%
school.	0.059	0.067	0.132	0.140	0.057	0.065	0.127	0.135
exper.	0.020	0.025	0.076	0.083	0.018	0.024	0.073	0.080
exper.^{2}	-0.001	-0.001	0.000	0.000	-0.001	-0.001	0.000	0.000
constant	8.811	8.999	10.071	10.234	8.467	8.646	9.676	9.836
	Stoch. Dom. (Ass. 3)				Stoch. Dom. (Ass. 4)			
	95%	L.B.	U.B.	95%	95%	L.B.	U.B.	95%
school.	-0.030	-0.022	0.207	0.215	0.014	0.022	0.165	0.173
exper.	-0.038	-0.032	0.135	0.142	-0.014	-0.009	0.102	0.109
exper.^{2}	-0.003	-0.002	0.001	0.001	-0.002	-0.002	0.000	0.000
constant	7.222	7.374	10.770	10.946	8.332	8.503	10.772	10.931
	Stoch. Dom. (Ass. 3) & Cont.				Stoch. Dom. (Ass. 4) & Cont.			
	95%	L.B.	U.B.	95%	95%	L.B.	U.B.	95%
school.	0.075	0.084	0.112	0.120	0.073	0.080	0.116	0.123
exper.	0.037	0.044	0.064	0.072	0.029	0.034	0.064	0.071
exper.^{2}	-0.001	-0.001	-0.001	0.000	-0.001	-0.001	0.000	0.000
constant	8.845	9.027	9.467	9.637	8.921	9.080	9.692	9.831

Notes: Dependent variable: logarithm of annual gross income in Swiss Francs. K_0 and K_1 are equal to the minimum and the maximum of the the marginal empirical distribution of log earnings. $p(x)$ and $\gamma(x)$ are constant across x. Confidence intervals are computed using a $N = 100$ bootstrap. *Source:* Swiss Household Panel (SHP).

Table 6: Bounds on the coefficient vector, $p(x) = 0.80$

	Data Corruption				Data Contamination			
	95%	L.B.	U.B.	95%	95%	L.B.	U.B.	95%
school.	-0.290	-0.279	0.436	0.447	0.038	0.045	0.151	0.159
exper.	-0.230	-0.226	0.306	0.313	0.004	0.009	0.092	0.099
exper.^{2}	-0.006	-0.006	0.005	0.005	-0.002	-0.002	0.000	0.000
constant	3.873	4.068	14.911	15.140	8.275	8.457	10.213	10.348
	Mult. Mean. Indep., $\gamma(x) = 1.2$				Mult. Mean. Indep., $\gamma(x) = 0.8$			
	95%	L.B.	U.B.	95%	95%	L.B.	U.B.	95%
school.	0.038	0.046	0.157	0.166	0.034	0.043	0.145	0.153
exper.	0.005	0.009	0.095	0.103	0.003	0.009	0.088	0.095
exper.^{2}	-0.002	-0.002	0.000	0.000	-0.002	-0.001	0.000	0.000
constant	8.593	8.795	10.622	10.787	7.931	8.119	9.805	9.974
	Stoch. Dom. (Ass. 3)				Stoch. Dom. (Ass. 4)			
	95%	L.B.	U.B.	95%	95%	L.B.	U.B.	95%
school.	-0.147	-0.137	0.313	0.322	-0.053	-0.044	0.221	0.229
exper.	-0.120	-0.113	0.214	0.221	-0.063	-0.059	0.146	0.152
exper.^{2}	-0.004	-0.004	0.002	0.003	-0.003	-0.003	0.001	0.001
constant	5.364	5.519	12.184	12.390	7.664	7.834	12.012	12.172
	Stoch. Dom. (Ass. 3) & Cont.				Stoch. Dom. (Ass. 4) & Cont.			
	95%	L.B.	U.B.	95%	95%	L.B.	U.B.	95%
school.	0.064	0.072	0.124	0.133	0.063	0.070	0.125	0.133
exper.	0.031	0.038	0.075	0.083	0.020	0.025	0.070	0.077
exper.^{2}	-0.001	-0.001	0.000	0.000	-0.001	-0.001	0.000	0.000
constant	8.580	8.776	9.606	9.779	8.807	8.966	9.893	10.033

Notes: Dependent variable: logarithm of annual gross income in Swiss Francs. K_0 and K_1 are equal to the minimum and the maximum of the the marginal empirical distribution of log earnings. $p(x)$ and $\gamma(x)$ are constant across x. Confidence intervals are computed using a $N = 100$ bootstrap. *Source:* Swiss Household Panel (SHP).

Table 7: Bounds on the coefficient vector, $p(x) = 0.70$

	Data Corruption				Data Contamination			
	95%	L.B.	U.B.	95%	95%	L.B.	U.B.	95%
school.	-0.468	-0.454	0.592	0.606	0.018	0.026	0.171	0.179
exper.	-0.356	-0.352	0.425	0.432	-0.009	-0.004	0.107	0.115
exper.^{2}	-0.008	-0.008	0.007	0.007	-0.002	-0.002	0.000	0.000
constant	1.392	1.631	17.458	17.748	7.934	8.111	10.496	10.643
	Mult. Mean. Indep., $\gamma(x) = 1.2$				Mult. Mean. Indep., $\gamma(x) = 0.8$			
	95%	L.B.	U.B.	95%	95%	L.B.	U.B.	95%
school.	0.018	0.028	0.180	0.188	0.016	0.025	0.161	0.169
exper.	-0.010	-0.004	0.113	0.122	-0.009	-0.004	0.101	0.109
exper.^{2}	-0.002	-0.002	0.000	0.000	-0.002	-0.002	0.000	0.000
constant	8.414	8.627	11.126	11.312	7.437	7.624	9.867	10.035
	Stoch. Dom. (Ass. 3)				Stoch. Dom. (Ass. 4)			
	95%	L.B.	U.B.	95%	95%	L.B.	U.B.	95%
school.	-0.260	-0.248	0.415	0.425	-0.117	-0.108	0.275	0.283
exper.	-0.200	-0.193	0.291	0.297	-0.109	-0.106	0.188	0.194
exper.^{2}	-0.006	-0.005	0.004	0.004	-0.004	-0.003	0.002	0.002
constant	3.566	3.734	13.554	13.793	7.011	7.182	13.189	13.357
	Stoch. Dom. (Ass. 3) & Cont.				Stoch. Dom. (Ass. 4) & Cont.			
	95%	L.B.	U.B.	95%	95%	L.B.	U.B.	95%
school.	0.053	0.063	0.134	0.144	0.054	0.062	0.135	0.143
exper.	0.025	0.032	0.085	0.094	0.013	0.018	0.076	0.083
exper.^{2}	-0.002	-0.001	0.000	0.000	-0.001	-0.001	0.000	0.000
constant	8.344	8.552	9.720	9.897	8.683	8.844	10.061	10.208

Notes: Dependent variable: logarithm of annual gross income in Swiss Francs. K_0 and K_1 are equal to the minimum and the maximum of the the marginal empirical distribution of log earnings. $p(x)$ and $\gamma(x)$ are constant across x. Confidence intervals are computed using a $N = 100$ bootstrap. *Source:* Swiss Household Panel (SHP).

Table 8: Bounds on the coefficient vector, $p(x) = 0.60$

	Data Corruption				Data Contamination			
	95%	L.B.	U.B.	95%	95%	L.B.	U.B.	95%
school.	-0.643	-0.626	0.744	0.761	-0.003	0.006	0.190	0.198
exper.	-0.481	-0.477	0.540	0.547	-0.024	-0.018	0.123	0.132
exper.^{2}	-0.011	-0.010	0.009	0.010	-0.002	-0.002	0.001	0.001
constant	-1.034	-0.726	19.966	20.326	7.604	7.783	10.808	10.981
	Mult. Mean. Indep., $\gamma(x) = 1.2$				Mult. Mean. Indep., $\gamma(x) = 0.8$			
	95%	L.B.	U.B.	95%	95%	L.B.	U.B.	95%
school.	-0.004	0.007	0.201	0.209	-0.004	0.006	0.175	0.183
exper.	-0.025	-0.019	0.132	0.141	-0.022	-0.016	0.113	0.122
exper.^{2}	-0.002	-0.002	0.001	0.001	-0.002	-0.002	0.001	0.001
constant	8.263	8.477	11.673	11.873	6.978	7.160	9.944	10.124
	Stoch. Dom. (Ass. 3)				Stoch. Dom. (Ass. 4)			
	95%	L.B.	U.B.	95%	95%	L.B.	U.B.	95%
school.	-0.371	-0.357	0.515	0.526	-0.180	-0.171	0.326	0.335
exper.	-0.278	-0.272	0.366	0.372	-0.155	-0.152	0.228	0.235
exper.^{2}	-0.007	-0.007	0.006	0.006	-0.004	-0.004	0.003	0.003
constant	1.800	1.990	14.899	15.173	6.392	6.569	14.351	14.532
	Stoch. Dom. (Ass. 3) & Cont.				Stoch. Dom. (Ass. 4) & Cont.			
	95%	L.B.	U.B.	95%	95%	L.B.	U.B.	95%
school.	0.044	0.053	0.145	0.154	0.043	0.051	0.143	0.151
exper.	0.018	0.026	0.095	0.104	0.005	0.010	0.082	0.089
exper.^{2}	-0.002	-0.002	0.000	0.000	-0.001	-0.001	0.000	0.000
constant	8.116	8.324	9.833	10.013	8.583	8.745	10.260	10.420

Notes: Dependent variable: logarithm of annual gross income in Swiss Francs. K_0 and K_1 are equal to the minimum and the maximum of the the marginal empirical distribution of log earnings. $p(x)$ and $\gamma(x)$ are constant across x. Confidence intervals are computed using a $N = 100$ bootstrap. *Source:* Swiss Household Panel (SHP).

Figure 1: Bounds under data contamination and data corruption

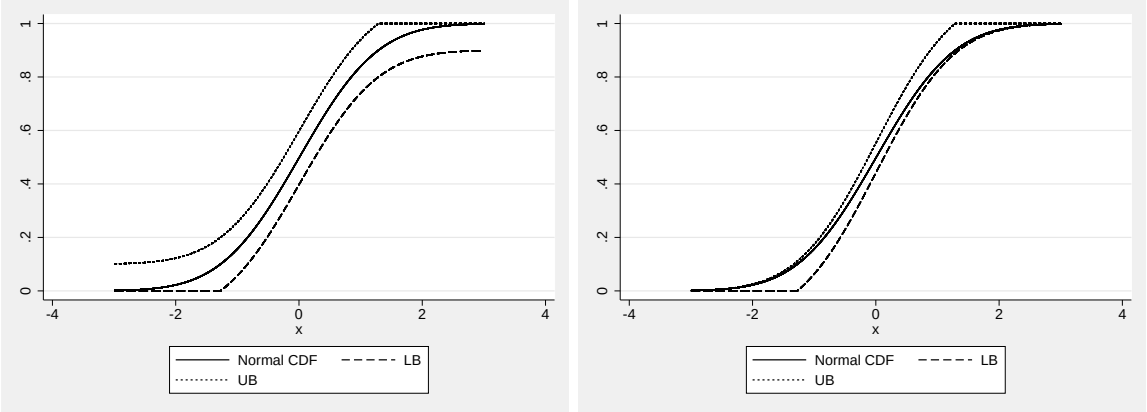

Figure 2: Bounds under stochastic dominance

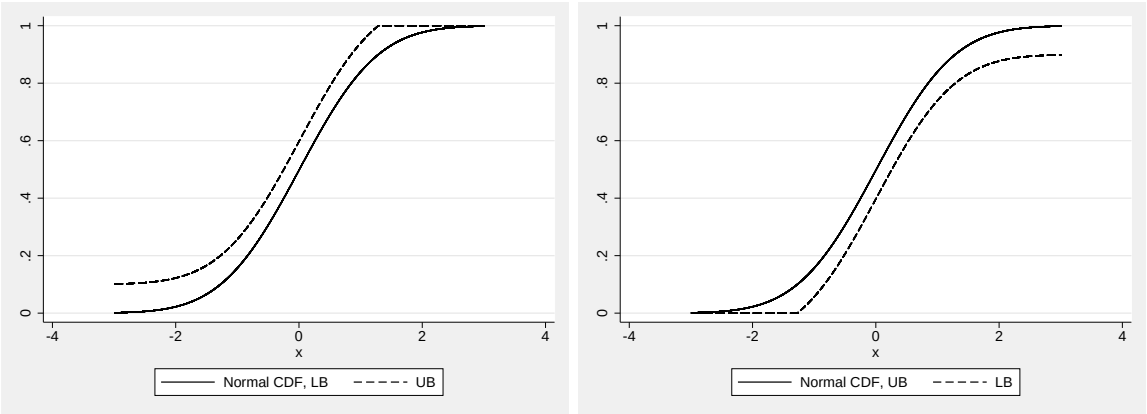

Figure 3: Bounds under stochastic dominance and data contamination