# Bounds on Generalized Linear Predictors With Incomplete Outcome Data[*]

Jörg Stoye
Department of Economics, New York University
269 Mercer Street, New York, NY 10003, USA
j.stoye@nyu.edu

July 6, 2006

**Abstract**

This paper develops easily computed, tight bounds on Generalized Linear Predictors and instrumental variable estimators when outcome data are partially identified. A salient example is given by Best Linear Predictors under square loss, or Ordinary Least Squares regressions, with missing outcome data, in which case the setup specializes the more general but intractable problem examined by Horowitz et al. [9]. The result is illustrated by re-analyzing the data used in that paper.

[*] I am indebted to Chuck Manski for his guidance and feedback and to two anonymous referees for extremely helpful suggestions. Of course, all errors are mine. Financial support through the Robert Eisner Memorial Fellowship, Department of Economics, Northwestern University, as well as a Dissertation Year Fellowship, The Graduate School, Northwestern University, is gratefully acknowledged.

# 1 Introduction

This paper provides an exact characterization of bounds on Generalized Linear Predictors when conditional distributions, and in some cases expectations, of outcome data are interval identified. Thus, it extends research on "partial identification" in econometrics – see the recent survey by Manski [11] and specifically the analysis in Horowitz et al. [9], to which the present contribution is related. The problem solved here can also arise in the context of interval computations (as in Ferson et al. [4]) as well as in Robust Bayesian analysis (as in Wasserman and Kadane [14]). A similar scenario has been considered by Vansteelandt and Gothgebeur [13], but whilst their analysis is in several ways more general, it does not yield the (essentially) closed-form results presented here. Zhilin [15] analyzes interval observations on outcomes, which is a special case of the present setup, but he is rather interested in the relative performance, under various distributional assumptions, of different methods of selecting a predictor from within the bounds.

Consider a scalar outcome variable $Y$ and a row vector of explanatory variables $X \in \mathbb{R}^J$. (I will use capital letters for random variables and minor letters for their realizations.) Let the population distribution of $(Y, X)$ be characterized by the cumulative distribution function $F_{yx}$ with marginal distribution $F_x$ of $X$ and conditional distribution $F_{y|x}$ of $(Y|X)$. Assume a researcher has observed a realization of $X$ and needs to predict the corresponding realization of $Y$. Substantively, the problem could be to predict reactions to medical treatments or – as in this paper's application – job market outcomes from personal characteristics. One approach is to use a Generalized Linear Predictor $\widehat{Y}$ of $Y$ from $X$, which is given by

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

where $G : \mathbb{R} \to \mathbb{R}$ is some pre-assigned, strictly increasing function, the prime symbol is used to denote transposes, and the vector of coefficients $\theta$ is often of independent interest. Important examples include Best Linear Prediction under square loss or Ordinary Least Squares regression, where

$$\begin{aligned} \widehat{Y} &= x\theta \\ \theta &= \left(\mathbf{E}X'X\right)^{-1} \mathbf{E}X'Y, \end{aligned} \tag{1}$$

and Best Logit Prediction under square loss, where $Y$ is binary and

$$\begin{aligned} \widehat{Y} &= \frac{\exp(x\theta)}{1+\exp(x\theta)} \\ \theta &= \left(\mathbf{E}X'X\right)^{-1} \mathbf{E}X' \log\left(\mathbf{P}(Y=1|X)/\mathbf{P}(Y=0|X)\right). \end{aligned}$$

Generalized Linear Predictors are appropriate in the sense that they are the best possible decision rule if the loss function – that is, the penalty incurred for misprediction – is the square of the prediction error, and if one either assumes that the data-generating process is described by $\mathbf{E}(Y|X=x)=G(x\theta)$, or for practical reasons restricts the predictor to be of the generalized linear form; see, e.g., [5].

But $\theta$ and $\widehat{Y}$ can be computed only if $F_{yx}$ is known or can be estimated; in econometric parlance, it has to be identified. This condition often fails in practice, in which case it is of interest to compute bounds on these objects that exploit as much as possible of the existing information. The specific case I will consider is the one in which the conditional distribution $F_{y|x}$ is interval identified, hence one knows or can estimate (in the statistical sense) conditional cumulative density functions $\underline{F}_{y|x}(y,x)$ and $\overline{F}_{y|x}(y,x)$ s.t. $\underline{F}_{y|x}(y,x) \leq F_{y|x}(y,x) \leq \overline{F}_{y|x}(y,x)$ for all $(y,x)$. If the link function $G$ is linear, this condition can be replaced with bounds on conditional expectations, i.e. functions $\underline{\mathbf{E}}(x)$ and $\overline{\mathbf{E}}(x)$ s.t. $\underline{\mathbf{E}}(x) \leq \mathbf{E}(Y|X=x) \leq \overline{\mathbf{E}}(x)$ for any $x$. For these setups, I will provide bounds that can be instantly computed, as well as sufficient conditions for them to be tight.

A salient application is the case of missing observations. Specifically, assume that conditional on $X=x$, a fraction $m(x)$ of realizations of $Y$ are missing. Then

$$(1-m(x))\cdot F^{*}_{y|x}(y,x) \leq F_{y|x}(y,x) \leq (1-m(x))\cdot F^{*}_{y|x}(y,x)+m(x), \tag{2}$$

where $F^{*}_{y|x}(y,x)$ is the conditional distribution of those $Y$ that are observed. The intuition behind this expression is that by simple probability calculus,

$$\begin{aligned}\mathbf{P}(Y &\leq y|x) = \\ \mathbf{P}(Y &\leq y|x, y \text{ is observable})\mathbf{P}(y \text{ is observable}|x) + \mathbf{P}(Y \leq y|x, y \text{ is missing})\mathbf{P}(y \text{ is missing}|x),\end{aligned} \tag{3}$$

and one knows that $\mathbf{P}(Y \leq y|x, y \text{ is observable}) = F^{*}_{y|x}(y,x)$, $\mathbf{P}(y \text{ is missing}|x) = m(x)$, $\mathbf{P}(y \text{ is observable}|x) = 1-m(x)$, and that $\mathbf{P}(Y \leq y|x, y \text{ is missing}) \in [0,1]$; substituting these facts into (3) justifies (2). If one additionally restricts attention to Best Linear Predictors, i.e. $\theta$ as in (1), the setup becomes a simplification of [9], who allow for arbitrary patterns of missingness in the data vector $(Y,X)$. However, their bounds are analytically intractable, and the authors are only able to approximate them by a very expensive genetic algorithm and to provide cheap but, as it turns out, extremely slack outer bounds. Apart from exactly solving an important special case, I also show how to improve analysis of the general problem by reducing its dimensionality; specifically, any optimization algorithm should operate on $X$ only, with missing values of $Y$ being "filled in" according to the results developed below. Furthermore, the present analysis makes explicit the "worst-case scenarios" under which the bounds obtain. Since these scenarios will often be substantively implausible, one can then attempt to find plausible, e.g. nonparametric, assumptions that exclude them and thus lead to sharper inference.[1]

Whilst the empirical part of this paper will consider the application just outlined, it is important to realize that interval identification of distributions or conditional expectations is a much more general problem. The scenario considered here will also occur

- whenever one can directly bound the joint distribution function $F_{yx}$,
- whenever one has interval data on $Y$, e.g. due to bounded measurement error, to interval-valued elicitation of $Y$ (as with income brackets), or any other application cited by [4],
- if one only observes bounds on the probability measure of $(Y|X)$, in which case bounds on $F_{y|x}$ are immediate and bounds on $\mathbf{E}(Y|X)$ follow from results in [2], or
- in Robust Bayesian analysis, where sets of priors may induce bounds on either $F_{y|x}$ or $\mathbf{E}(Y|X)$ [14],

and the above remarks extend to all of these applications.

Finally, the present result applies to regression analysis, where the issue of substantive interest is the causal connection – here captured by $\theta$ – between $X$ and $Y$. If the model specifies $Y = G(X\theta) + \varepsilon$, where the error term $\varepsilon$ obeys $\mathbf{E}(\varepsilon|X) = 0$, then $\theta$ can be estimated consistently by its sample analog, i.e. by replacing population expectations with sample means. This will not work if $\mathbf{E}(\varepsilon|X) \neq 0$. However, assume that one has available a vector of instruments $Z \in \mathbb{R}^K, K \geq J$, s.t. $Z$ is correlated with $X$ but not with $\varepsilon$. (Recall that $J$ is the dimensionality of $X$.) Then $\theta$ can be estimated consistently by the instrumental variables estimator

$$\widehat{\theta}_{IV} = \left(\mathbf{E}_n X'ZWZ'X\right)^{-1} \mathbf{E}_n X'ZWZ'G^{-1}(Y),$$

where $\mathbf{E}_n$ denotes sample means and $W$ is a pre-assigned matrix that weights the instruments; $W$ is relevant only if $K > J$.[2] To keep the presentation simple, I will not work with this more elaborate term, but the below result can be easily adapted to find bounds on $\widehat{\theta}_{IV}$ that arise from incomplete samples.

The remainder of this paper is structured as follows. The identification problem is formally stated and solved in section 2, the empirical example from Horowitz et al. [9] is revisited in section 3, and I conclude in section 4.

[1] A nonparametric assumption is one that avoids imposition of functional forms. In the example, one could impose that $F^*_{y|x}(y,x)$ is ordered relative to $F_{y|x}(y,x)$ in some way, say with respect to first-order stochastic dominance.

[2] This estimator is known as “Generalized Method of Moments” (GMM) estimator; see Hansen [6] for the canonical reference and Hayashi [7] for a textbook treatment. The case of pre-assigned $W$ holds in all one-step GMM methods, including two-step least squares, seemingly unrelated regressions, multivariate regression, fixed effect, and random effect estimators. $W$ is not pre-assigned, and the result does not apply, in efficient (“two-step”) GMM.

## 2 Analysis of the Identification Problem

I will begin with the identification analysis and postpone the consideration of finite sample problems until section 3. Hence, let $F_x$, $\underline{F}_{y|x}$, and $\overline{F}_{y|x}$ be known. Then finding bounds on $\theta$ amounts to solving a constrained optimization problem, namely to find $F_{y|x}$ so as to extremize $\theta$ in a certain direction. This problem is amenable to closed-form analysis since $\theta$ increases in $\mathbf{E}(G^{-1}(y)|X=x)$ for some values of $x$ and decreases in it otherwise. Hence, $\theta$ is maximized by maximizing $\mathbf{E}(G^{-1}(y)|X=x)$ for certain $x$ and minimizing it for others. Such maximization [minimization] is possible since it corresponds to minimization [maximization] of $F_{y|x}$. The remaining question is when to maximize $F_{y|x}$ and when to minimize it.

Proposition 1 formalizes this intuition, and answers the last question, for extremization of any inner product $c\cdot\theta$. Thus, the proposition immediately yields bounds on the actual predictor $\widehat{Y}$ given any $x$. Bounds on the components of $\theta$ can be recovered by identifying $c$ with the corresponding base vectors.

**Proposition 1** *Let $F_x$ and the bounding functions $\left(\underline{F}_{y|x},\overline{F}_{y|x}\right)$ be known. Then for any pre-assigned $c\in\mathbb{R}^J$, $c\cdot\theta$ is bounded by*

$$c\cdot\left(\int x'xdF_x\right)^{-1}\int x'\underline{g}(x)dF_x\leq c\cdot\theta\leq c\cdot\left(\int x'xdF_x\right)^{-1}\int x'\overline{g}(x)dF_x, \tag{4}$$

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

$$\underline{g}(x) \equiv \begin{cases} \int G^{-1}(y)\,d\overline{F}_{y|x}, & c\cdot\left(\int x'xdF_x\right)^{-1}x'>0 \\ \int G^{-1}(y)\,d\underline{F}_{y|x}, & c\cdot\left(\int x'xdF_x\right)^{-1}x'\leq 0 \end{cases} \tag{5}$$

$$\overline{g}(x) \equiv \begin{cases} \int G^{-1}(y)\,d\underline{F}_{y|x}, & c\cdot\left(\int x'xdF_x\right)^{-1}x'>0 \\ \int G^{-1}(y)\,d\overline{F}_{y|x}, & c\cdot\left(\int x'xdF_x\right)^{-1}x'\leq 0 \end{cases}. \tag{6}$$

*These bounds are tight if for any $d\in\mathbb{R}^J$, it is conceivable that $d\cdot x\geq 0\Rightarrow F_{y|x}=\underline{F}_{y|x}$ and $d\cdot x<0\Rightarrow F_{y|x}=\overline{F}_{y|x}$. If $G$ is linear, they can be computed from $\left(\underline{\mathbf{E}}(x),\overline{\mathbf{E}}(x)\right)$ by identifying $\underline{g}(x)$ and $\overline{g}(x)$ with*

$$\underline{g}(x) \equiv \begin{cases} G^{-1}\left(\underline{\mathbf{E}}(x)\right), & c\cdot(\mathbf{E}X'X)^{-1}x'>0 \\ G^{-1}\left(\overline{\mathbf{E}}(x)\right), & c\cdot(\mathbf{E}X'X)^{-1}x'\leq 0 \end{cases} \tag{7}$$

$$\overline{g}(x) \equiv \begin{cases} G^{-1}\left(\overline{\mathbf{E}}(x)\right), & c\cdot(\mathbf{E}X'X)^{-1}x'>0 \\ G^{-1}\left(\underline{\mathbf{E}}(x)\right), & c\cdot(\mathbf{E}X'X)^{-1}x'\leq 0 \end{cases}. \tag{8}$$

**Proof.** Define

$$\Pi\equiv\left\{\widetilde{F}_{yx}:\widetilde{F}_x=F_x,\underline{F}_{y|x}\leq\widetilde{F}_{y|x}\leq\overline{F}_{y|x}\right\},$$

the set of distributions of $(Y,X)$ that are consistent with $F_x$ as well as $\left(\underline{F}_{y|x},\overline{F}_{y|x}\right)$. Then

$$c\cdot\theta\leq\sup_{\widetilde{F}_{yx}\in\Pi}\left\{c\cdot\left(\int x'xdF_x\right)^{-1}\int x'G^{-1}(y)d\widetilde{F}_{yx}\right\}. \tag{9}$$

Define $\alpha \equiv c \cdot \left(\int x'xdF_x\right)^{-1}$. Knowledge of $F_x$ implies knowledge of $\int x'xdF_x$ and hence $\alpha$, so $(c \cdot \theta)$ can be bounded from above as follows:

$$\begin{aligned}
c \cdot \theta &\leq \sup_{\widetilde{F}_{yx} \in \Pi} \left\{ \alpha \int x' G^{-1}(y) d\widetilde{F}_{yx} \right\} && (10)\\
&= \sup_{\widetilde{F}_{yx} \in \Pi} \left\{ \alpha \int \int x' G^{-1}(y) d\widetilde{F}_{y|x} d\widetilde{F}_x \right\} && (11)\\
&= \sup_{\widetilde{F}_{yx} \in \Pi} \left\{ \alpha \int x' \int G^{-1}(y) d\widetilde{F}_{y|x} dF_x \right\} && (12)\\
&\leq \int \sup_{\widetilde{F}_{yx} \in \Pi} \left\{ \alpha x' \int G^{-1}(y) d\widetilde{F}_{y|x} \right\} dF_x && (13)\\
&= \int \alpha x' \overline{g}(x) dF_x, && (14)
\end{aligned}$$

where (11) holds because of the Laws of Total Probability and Iterated Expectations, (12) uses the facts that $\widetilde{F}_x = F_x$ and that the integral operator is linear, and (13) holds due to the objective function's separability. To see the last step, consider the problem

$$\max_{\underline{F}_{y|x} \leq F_{y|x} \leq \overline{F}_{y|x}} \left\{ \alpha x' \int G^{-1}(y) dF_{y|x} \right\}. \tag{15}$$

For any pre-assigned $x$, the objective is linear in the expectation $\int G^{-1}(y)dF_{y|x}$ and the problem is therefore solved by maximizing [minimizing] this term if $\alpha x'$ is positive [negative]. Since $G^{-1}$ is strictly increasing and expectations of strictly increasing functions of random variables increase with first-order stochastic dominance, this is achieved by pointwise minimization [maximization] of $F_{y|x}$. Hence, $\overline{g}(x)$ as in (6) solves (15) for every $x$, and (14) bounds (13) from above.

This establishes the upper bound's validity. Notice that $\Pi$ contains the distribution characterized by

$$F_{y|x}(y,x) = \begin{cases} \underline{F}_{y|x}, & c \cdot \left(\int x'xdF_x\right)^{-1} x' > 0 \\ \overline{F}_{y|x}, & c \cdot \left(\int x'xdF_x\right)^{-1} x' \leq 0 \end{cases}.$$

This distribution solves (15) for every $x$, hence (14) is an equality. Equality of (13), however, obtains only if this distribution is also consistent with any knowledge about $F_{yx}$ that is not reflected in the definition of $\Pi$. The additional condition is sufficient for this. (A stronger sufficient condition is, of course, that $\Pi$ exhausts the available information about $F_{yx}$.) The arguments extend to lower bounds by replacing $c$ with $(-c)$.

Finally, let $G$ and hence $G^{-1}$ be linear. Then (15) is solved by any conditional distribution $F_{y|x}$ that maximizes [minimizes] $\mathbf{E}(Y|X=x)$ if $\alpha x'$ is positive [negative], thus the bounds can be computed from $\overline{g}(x)$ as in (8). ■

It turns out that identification of bounds on $c \cdot \theta$ with incomplete outcome data is a computationally trivial exercise. It is also clear why more general patterns of partial identification exponentiate the problem's complexity: The simplicity of the result crucially depends on identification of the "denominator" $\left(\int x'xdF_x\right)^{-1}$, which renders the problem linear.

The proof not only establishes the bounds' validity, but also identifies distributions of $(Y|X)$ that achieve them. Firstly, this implies that if the bounding functions $\left(\underline{F}_{y|x}, \overline{F}_{y|x}\right)$ exploit the available information about $F_{y|x}$, then the bounds are tight. Secondly, with the worst-case configuration of $F_{y|x}$ being known, one can attempt to find plausible assumptions that exclude it and hence tighten the bounds. Refinements of bounds in this spirit are discussed extensively in [11].

# 3 An Empirical Illustration

To illustrate the result, I re-analyze the dataset from [12] that was also used in [9]. The data concern worker expectations of job loss and were collected between 1994 and 1998 in the Survey of Economic Expectations.[3] Survey respondents answered the following question:

*I would like you to think about your employment prospects over the next 12 months. What do you think is the percent chance that you will lose your job during the next 12 months?*

These answers are taken as outcome of interest $Y$. Responses could be any number in $[0, 100]$; with extremely few exceptions near the extremal values, integers were chosen. The survey also elicited covariates, of which age, race, and income (the first two coded as multivalued indicator variables) will be considered here. A question of obvious interest is the relation between these covariates and expectations about job loss. If the latter varied systematically by race, say, this fact would be intrinsically interesting as well as help to predict expectations.

To analyze this question, Horowitz et al. [9] used Best Linear Prediction or OLS regression, i.e. this paper's framework with $\theta = (\mathbf{E}X'X)^{-1}\mathbf{E}X'Y$ as in (1). A component of $\theta$ can be interpreted as the statistical effect of the respective covariate – membership in a given age group, say – on expected job loss within a linear probability model. Following [9], I will separately consider the regression on all three covariates and on age and race only.

Unfortunately, the data set has missing data on $Y$ as well as on all covariates, so that $\theta$ is only partially identified. Horowitz et al. [9] investigate bounds that take into account the full missingness pattern. In contrast, I will ignore missing components of $X$ by discarding observations with incomplete $X$. The resulting data set lends itself to an application of proposition 1. This data reduction is to some degree illustrative and meant to generate results that compare to [9]. But one might also substantively entertain the underlying assumption, namely that observations of $X$ are missing at random, in which case it suffices to focus on selective missingness in $Y$.

[3] The data are publicly available at http://www.faculty.econ.northwestern.edu/faculty/manski/. The MATLAB code used for the present application is available from the author.

Tables 1 and 2 display bounds on the components of $\theta$, i.e. the different covariates' coefficients. The tables are comparable to tables 2 and 3 in [9], but computation of the bounds is so much faster – a few seconds as opposed to several hours – that two important changes were possible. Firstly, whilst [9] discretize outcomes as well as income to keep computations tractable, this is not needed here, and the original income variable was used. More importantly, [9] were only able to bound the sample analog of $\theta$, i.e. the object $\theta_n \equiv \left(\mathbf{E}_n X'X\right)^{-1} \mathbf{E}_n X'Y$. Whilst $\theta_n$ is the usual estimator for $\theta$, no clear indication of the impact of sampling variation, and thus the likely divergence of $\theta_n$ from $\theta$, was given. In contrast, tables 1 and 2 here present bounds on components of $\theta_n$ (in the columns labelled "L.B." for lower bound respectively "U.B." for upper bound) but also confidence intervals that were computed by means of an $N = 10000$ bootstrap.[4] The intervals spanned by the lower and upper 95% confidence points exhibit a nominal 95% coverage probability with respect to the corresponding component of $\theta$.[5] As in [9], the bounds are computed under two different conditions, once by assuming that $Y = 50\%$ – a very frequent answer – is valid, and once by treating it as missing. The reason for this is that as explained in [9], it has been argued that a response of $Y = 50\%$ expresses of complete ignorance and should not be taken at face value.

| | **Y=50% treated as non-missing** | | | | **Y=50% treated as missing** | | | |
|---|---|---|---|---|---|---|---|---|
| | **95%** | **L.B.** | **U.B.** | **95%** | **95%** | **L.B.** | **U.B.** | **95%** |
| **constant** | 7.71 | 11.22 | 23.16 | 28.43 | 2.74 | 5.76 | 28.62 | 34.65 |
| **age 18-24** | -14.30 | -8.66 | 9.24 | 13.53 | -24.51 | -18.18 | 18.76 | 22.83 |
| **age 25-49** | -14.70 | -9.33 | 5.98 | 9.51 | -24.00 | -17.99 | 14.64 | 17.78 |
| **age 50-64** | -17.54 | -12.08 | 4.96 | 8.96 | -25.91 | -19.77 | 12.65 | 16.30 |
| **black** | 1.98 | 4.87 | 15.14 | 18.52 | -8.79 | -6.02 | 26.03 | 30.38 |
| **other nonwhite** | -4.96 | -2.37 | 7.84 | 11.19 | -11.72 | -9.21 | 14.68 | 18.58 |

Table 1: Bounds on components of $\theta$ for the regression without income.

[4] The bootstrap method is to estimate a statistic's sampling distribution by a Monte Carlo experiment in which the statistic is computed from many (here, 10000) artifical samples which have been generated by drawing with replacement from the original sample. Thus, the sample distribution of the data is used as an estimator of their population distribution. See Efron [3] for the pioneering contribution and Horowitz [8] for a recent overview.

[5] The intervals have been defined as suggested by [10], thus their coverage probability applies to the population parameter as opposed to the "true" (population) bounds on it. Confidence regions for the population bounds would be larger.

Horowitz et al. [9] point out that in general, the bootstrap can be inconsistent for the present type of problem because parameters are estimated subject to inequality constraints. But given proposition 1, the present estimation problem can be rewritten in terms of equality constraints, and the problem is therefore avoided. See also Andrews [1].

|  | **Y=50 treated as non-missing** |  |  |  | **Y=50 treated as missing** |  |  |  |
|---|---|---|---|---|---|---|---|---|
|  | **95%** | **L.B.** | **U.B.** | **95%** | **95%** | **L.B.** | **U.B.** | **95%** |
| **constant** | 8.92 | 13.02 | 22.84 | 30.11 | 2.82 | 6.59 | 29.26 | 38.21 |
| **income** | -0.1151 | -0.0171 | -0.0087 | -0.0048 | -0.1654 | -0.0236 | -0.0022 | 0.0128 |
| **age 18-24** | -14.43 | -7.49 | 7.30 | 12.17 | -26.01 | -17.71 | 17.52 | 22.27 |
| **age 25-49** | -14.40 | -8.07 | 4.55 | 9.04 | -24.80 | -17.32 | 13.80 | 17.77 |
| **age 50-64** | -16.85 | -10.46 | 3.45 | 8.75 | -26.38 | -18.60 | 11.59 | 16.81 |
| **black** | 2.15 | 5.35 | 13.98 | 17.47 | -9.16 | -6.04 | 25.37 | 29.94 |
| **other nonwhite** | -5.02 | -2.13 | 7.69 | 11.46 | -11.61 | -8.78 | 14.34 | 18.62 |

Table 2: Bounds on components of $\theta$ for the regression with income.

As with previous analyses of this dataset, worst-case bounds on most coefficients are quite large and include zero, so that the coefficients' signs are not identified. The only unambiguous effects are that African-Americans have a higher perceived job insecurity, and higher income is predictive of a low perceived insecurity. Even these conclusions only obtain when $Y = 50\%$ is considered a valid response, and a caveat to the first one will be stated below. The "estimation penalty" – that is, the width of confidence intervals above and beyond the sample bounds – is surprisingly large for such a big data set. This fact presumably attests to the objective function's nonlinearity and local volatility, the same features that make the unrestricted identification problem so hard. All in all, the substantive result is negative: With few exceptions, the bounds do not suffice to establish that covariates explain or predict job loss expectations.

Comparison of the bounds to tables 2 and 3 in [9] allows some interesting observations. One might expect some agreement between table 1 here and table 2 in [9] because only 43 (of 3860) data points were deleted, so that these tables are generated from very similar data. The alignment, whilst not perfect, is indeed quite good. This suggests that the discretization of outcomes in [9] as well as the inevitable optimization error encountered there had a limited effect on their results. Regarding table 2 here and table 3 in [9], larger discrepancies are both expected and discovered.

The results also suggest that one observation in [9] must be qualified when sampling uncertainty is taken into account. Their table 3 shows a lower bound on the coefficient for *black* of 0.58, thus this coefficient's sample analog is known to be positive. But the estimation penalty on *black* in table 2 equals $5.35 - 2.15 = 3.20$, which much exceeds 0.58. Although this number cannot be applied directly to table 3 in [9], it leaves no reasonable doubt that the 95% confidence region for their coefficient on *black* would include zero. (Conversely, the coefficient's statistically significant positive sign in table 2 above must be an artifact of dropping the observations with missing covariates.)

# 4 Conclusion

This paper added to the growing literature on partial identification by discussing worst-case bounds on Generalized Linear Predictors or instrumental variable estimators, e.g. OLS regressions, when the outcome variable $Y$ is only partially identified. The intuition behind the bounds is to specify the distribution of $Y$ as high as possible for certain realizations of the regressor and as low as possible otherwise.

I illustrated the bounds with a real-world dataset. Their computation is so fast that some limitations of previous analyses could be overcome; most importantly, I was able to investigate the effect of sampling uncertainty on the bounds. Since regression with observable (or randomly missing) covariates but selectively missing outcomes is a reasonably generic setup, I believe that the results can be useful in a large number of applied settings. They also significantly improve solution algorithms for more general formulations of the problem because these algorithms can use the present result to optimize over $F_{y|x}$. Finally, the analysis opens avenues for further research by identifying the worst-case scenarios that generate the bounds, allowing one to assess their plausibility and perhaps formulate credible assumptions that lead to tighter identification.

# References

[1] Andrews, D.W.K.: Inconsistency of the Bootstrap When a Parameter is on the Boundary of the Parameter Space, *Econometrica* **68** (2) (2000), pp. 399-405.

[2] DeRobertis, L. and J.A. Hartigan: Bayesian Inference Using Intervals of Measures, *Annals of Statistics* **9** (2) (1981), pp. 235-244.

[3] Efron, B.: Bootstrap Methods: Another Look at the Jackknife, *Annals of Statistics* **7** (1) (1979), pp. 1-26.

[4] Ferson, S., L. Ginzburg, V. Kreinovich, L. Longpré, and M. Aviles: Exact Bounds on Finite Populations of Interval Data, *Reliable Computing* **11** (3) (2005), pp. 207-233.

[5] Goldberger, A.S.: *A Course in Econometrics*, Harvard University Press, Cambridge, 1991.

[6] Hansen, L.P.: Large Sample Properties of Generalized Method of Moments Estimators, *Econometrica* **50** (4) (1982), pp. 1029-1054

[7] Hayashi, F.: *Econometrics.* Princeton University Press, Princeton, 2000.

[8] Horowitz, J.L.: The Bootstrap, in: Heckman, J.J. and E. Leamer (ed.), *Handbook of Econometrics (Volume 5)*, Elsevier, Amsterdam, 2001, pp. 3159-3228.

[9] Horowitz, J.L., C.F. Manski, M. Ponomareva, and J. Stoye: Computation of Bounds on Population Parameters When the Data are Incomplete, *Reliable Computing* **9** (6) (2003), pp. 419-440.

[10] Imbens, G. and C.F. Manski: Confidence Intervals for Partially Identified Parameters, *Econometrica* **72** (6) (2004), pp. 1845-1857.

[11] Manski, C.F.: *Partial Identification of Probability Distributions*, Springer Verlag, New York, 2003.

[12] Manski, C.F. and J. Straub: Worker Perceptions of Job Insecurity in the Mid-1990s: Evidence from the Survey of Economic Expectations, *Journal of Human Resources* **35** (3) (2000), pp. 447-479.

[13] Vansteelandt, S. and E. Goethgebeur: Analyzing the Sensitivity of Generalized Linear Models to Incomplete Outcomes via the IDE Algorithm, *Journal of Computational and Graphical Statistics* **10** (4) (2001), pp. 656-672.

[14] Wasserman, L. and J.B. Kadane: Computing Bounds on Expectations, *Journal of the American Statistical Association* **87** (418) (1992), pp. 516-522.

[15] Zhilin, S.I.: On Fitting Empirical Data under Interval Error, *Reliable Computing* **11** (5) (2005), pp. 433-442.