Quantitative Economics 4 (2013), 231–267 1759-7331/20130231

# Sensitivity to missing data assumptions: Theory and an evaluation of the U.S. wage structure

Patrick Kline
Department of Economics, University of California, Berkeley and NBER

Andres Santos
Department of Economics, University of California, San Diego

This paper develops methods for assessing the sensitivity of empirical conclusions regarding conditional distributions to departures from the missing at random (MAR) assumption. We index the degree of nonignorable selection governing the missing data process by the maximal Kolmogorov–Smirnov distance between the distributions of missing and observed outcomes across all values of the covariates. Sharp bounds on minimum mean square approximations to conditional quantiles are derived as a function of the nominal level of selection considered in the sensitivity analysis and a weighted bootstrap procedure is developed for conducting inference. Using these techniques, we conduct an empirical assessment of the sensitivity of observed earnings patterns in U.S. Census data to deviations from the MAR assumption. We find that the well documented increase in the returns to schooling between 1980 and 1990 is relatively robust to deviations from the missing at random assumption except at the lowest quantiles of the distribution, but that conclusions regarding heterogeneity in returns and changes in the returns function between 1990 and 2000 are very sensitive to departures from ignorability.

Keywords. Quantile regression, missing data, sensitivity analysis, wage structure.

JEL classification. C01, C80, J31.

## 1. Introduction

Despite major advances in the design and collection of survey and administrative data, missing and incomplete records remain a pervasive feature of virtually every modern

Patrick Kline: pkline@econ.berkeley.edu
Andres Santos: a2santos@ucsd.edu

We thank Elie Tamer and two anonymous referees for comments and suggestions that greatly improved this paper. We are also grateful to David Card, Guido Imbens, Justin McCrary, Azeem Shaikh, Hal White, and seminar participants at UC Berkeley, UC San Diego, the University of Michigan, USC, Stanford, Wisconsin, the 2010 NBER Summer Institute, the 2010 Western Economic Association Summer Meeting, the 2010 Seoul Summer Economics Conference, and the 2010 Econometric Society World Congress for useful comments and corrections. We thank Ivan Fernández-Val for assistance in replicating the results of Angrist, Chernozhukov, and Fernández-Val (2006). A previous version of this paper circulated under the title "Interval estimation of potentially misspecified quantile models in the presence of missing data."


DOI: 10.3982/QE176

economic data set. Hirsch and Schumacher (2004), for instance, find that nearly 30% of the earnings observations in the Outgoing Rotation Groups of the Current Population Survey (CPS) are imputed. Similar allocation rates are present in other major earnings sources such as the March CPS and Decennial Census, with the problem growing worse in more recent years.

The dominant framework for dealing with missing data has been to assume that it is "missing at random" (Rubin (1976)) or "ignorable" conditional on observable demographics; an assumption whose popularity owes more to convenience than plausibility. Even in settings where it is reasonable to believe that nonresponse is approximately ignorable, the prevalence of missing values in modern economic data suggests that economists ought to assess the sensitivity of their conclusions to small deviations from this assumption.

Previous work on nonignorable missing data processes has either relied upon parametric models of nonresponse in conjunction with exclusion restrictions to obtain point identification (Greenlees, Reece, and Zieschang (1982) and Lillard, Smith, and Welch (1986)) or considered the "worst case" bounds on population moments that result when all assumptions regarding the missing data process are abandoned (Manski (1994), Manski (2003)). Neither approach has garnered much popularity.[1] It is typically quite difficult to find variables that shift the probability of nonresponse but are uncorrelated with population outcomes. And for most applied problems, the worst case bounds are overly conservative in the sense that they consider missing data mechanisms that the majority of researchers would consider to be implausible in modern data sets.

Proponents of the bounding approach are well aware of the fact that the worst case bounds may be conservative. As Horowitz and Manski (2006) state, "an especially appealing feature of conservative analysis is that it enables establishment of a domain of consensus among researchers who may hold disparate beliefs about what assumptions are appropriate." However, when this domain of consensus proves uninformative, some researchers may wish to consider stronger assumptions. Thus, a complementary approach is to consider a continuum of assumptions ordered from strongest (ignorability) to weakest (worst case bounds), and to report the conclusions obtained under each one. In this manner, consumers of economic research may draw their own (potentially disparate) inferences, depending on the strength of the assumptions they are willing to entertain.

We propose here a feasible version of such an approach for use in settings where one lacks prior knowledge of the missing data mechanism. Rather than ask what can be learned about the parameters of interest given assumptions on the missing data process, we investigate the level of nonignorable selection necessary to undermine one's conclusions regarding the conditional distribution of the data obtained under a missing at random (MAR) assumption. We do so by making use of a nonparametric measure of selection: the maximal Kolmogorov–Smirnov (KS) distance between the distributions of missing and observed outcomes across all values of the covariates. The KS distance

[1]See DiNardo, McCrary, and Sanbonmatsu (2006) for an applied example comparing these two approaches.

yields a natural parameterization of deviations from ignorability, with a distance of 0 corresponding to MAR and a distance of 1 encompassing the totally unrestricted missing data processes considered in Manski (1994). Between these extremes lies a continuum of selection mechanisms that may be studied to determine a critical level of selection above which conclusions obtained under an analysis predicated on MAR may be overturned. By reporting the minimal level of selection necessary to undermine a hypothesis, we allow readers to decide for themselves which inferences to draw based on their beliefs about the selection process.[2]

To enable such an analysis, we begin by deriving sharp bounds on the conditional quantile function (CQF) under nominal restrictions on the degree of selection present. We focus on the commonly encountered setting where outcome data are missing and covariates are discrete. To facilitate the analysis of data sets with many covariates, results are also developed that summarize the conclusions that can be drawn regarding linear parametric approximations to the underlying nonparametric CQF of the sort considered by Chamberlain (1994). When point identification of the CQF fails due to the presence of missing data, the identified set of corresponding best linear approximations consists of all elements of the parametric family that provide a minimum mean square approximation to some function lying within the CQF bounds.

We obtain sharp bounds on the parameters that govern the linear approximation and propose computationally simple estimators for them. We show that these estimators converge in distribution to a Gaussian process indexed by the quantile of interest and the level of the nominal restriction on selection, and develop a weighted bootstrap procedure for consistently estimating that distribution. This procedure enables inference on the coefficients that govern the approximation when considered as an unknown function of the quantile of interest and the level of the selection bound.

Substantively, these methods allow a determination of the critical level of selection for which hypotheses regarding conditional quantiles, parametric approximations to conditional quantiles, or entire conditional distributions cannot be rejected. For example, we study the "breakdown" function defined implicitly as the level of selection necessary for conclusions to be overturned at each quantile. The uniform confidence region for this function effectively summarizes the differential sensitivity of the entire conditional distribution to violations of MAR. These techniques substantially extend the recent econometrics literature on sensitivity analysis (Altonji, Elder, and Taber (2005, 2008), Imbens (2003), Rosenbaum and Rubin (1983), Rosenbaum (2002)), most of which has focused on the sensitivity of scalar treatment effect estimates to confounding influences, typically by using assumed parametric models of selection.

Having established our inferential procedures, we turn to an empirical assessment of the sensitivity of heavily studied patterns in the conditional distribution of U.S. wages to deviations from the MAR assumption. We begin by revisiting the results of Angrist, Chernozhukov, and Fernández-Val (2006) regarding changes across Decennial Censuses

[2]Our approach has parallels with classical hypothesis testing. It is common practice to report $p$-values rather than the binary results of statistical tests, because readers may differ in the balance they wish to strike between type I and type II errors. By reporting $p$-values, the researcher leaves it to readers to strike this balance on their own.

in the quantile-specific returns to schooling. Weekly earnings information is missing for roughly a quarter of the observations in their study, suggesting the results may be sensitive to small deviations from ignorability. We show that despite the prevalence of missing values in the dependent variable, the well documented increase in the returns to schooling between 1980 and 1990 is relatively robust to deviations from the missing at random assumption except at the lowest quantiles of the conditional distribution. However, deterioration in the quality of Decennial Census data renders conclusions regarding heterogeneity in returns and changes in the returns function between 1990 and 2000 very sensitive to departures from ignorability at all quantiles. We also show, using a more flexible model studied by Lemieux (2006), that the apparent convexification of the earnings–education profile between 1980 and 2000 is robust to modest deviations from MAR, while changes in the wage structure at lower quantiles are more easily obscured by selection.

To gauge the practical relevance of these sensitivity results, we analyze a sample of workers from the 1973 Current Population Survey for whom IRS earnings records are available. This sample allows us to observe the earnings of CPS participants who, for one reason or another, failed to provide valid earnings information to the CPS. We show that IRS earnings predict nonresponse to the CPS within demographic covariate bins, with very high and very low earning individuals most likely to have invalid CPS earnings records. By measuring the degree of selection using our proposed KS metric, we find significant deviations from ignorability with patterns of selection that vary substantially across demographic groups. Given recent trends in survey imputation rates, these findings suggest economists' knowledge of the location and shape of conditional earnings distributions in the United States may be more tentative than previously supposed.

The remainder of the paper is structured as follows: Section 2 describes our index of selection and our general approach to assessing sensitivity. Section 3 develops our approach to assessing the sensitivity of parametric approximations to conditional quantiles. Section 4 obtains the results necessary for estimation and inference on the bounds provided by restrictions on the selection process. In Section 5, we present our empirical study, and we briefly conclude in Section 6. Appendixes are available in a supplementary file on the journal website, http://qeconomics.org/supp/176/supplement.pdf.

## 2. Assessing sensitivity

Consider the random variables $(Y, X, D)$ with joint distribution $F$, where $Y \in \mathbf{R}$, $X \in \mathbf{R}^l$, and $D \in \{0, 1\}$ is a dummy variable that equals 1 if $Y$ is observable and 0 otherwise, that is, only $(YD, X, D)$ is observable. Denote the distribution of $Y$ given $X$ and of $Y$ given $X$ and $D$ by

$$F_{y|x}(c) \equiv P(Y \leq c|X = x), \qquad F_{y|d,x}(c) \equiv P(Y \leq c|D = d, X = x), \tag{1}$$

where $d \in \{0, 1\}$, and further define the probability of $Y$ being observed conditional on $X$ to be

$$p(x) \equiv P(D = 1|X = x). \tag{2}$$

In conducting a sensitivity analysis, the researcher seeks to assess how the identified features of $F_{y|x}$ depend on alternative assumptions regarding the process that generates $D$. In particular, we concern ourselves with the sensitivity of conclusions regarding $q(\tau|X)$, the conditional $\tau$-quantile of $Y$ given $X$, which is often of more direct interest than the distribution function itself. Toward this end, we impose the following assumptions on the data generating process.

ASSUMPTION 2.1. (i) $X \in \mathbf{R}^l$ *has finite support* $\mathcal{X}$; (ii) $F_{y|d,x}(c)$ *is continuous and strictly increasing at all c such that* $0 < F_{y|d,x}(c) < 1$; (iii) *the observable variables are* $(YD, D, X)$.

The discrete support requirement in Assumption 2.1(i) simplifies inference, as it obviates the need to employ nonparametric estimators of conditional quantiles. While this assumption may be restrictive in some environments, it is still widely applicable as illustrated in our study of quantile-specific returns to education in Section 5. It is also important to emphasize that Assumption 2.1(i) is not necessary for our identification results, but only for our discussion of inference. Assumption 2.1(ii) ensures that for any $0 < \tau < 1$, the $\tau$-conditional quantile of $Y$ given $X$ is uniquely defined.

### 2.1 *Index of selection*

Most previous work on sensitivity analysis (e.g., Rosenbaum and Rubin (1983), Altonji, Elder, and Taber (2005)) has relied on parametric models of selection. While potentially appropriate in cases where particular deviations from ignorability are of interest, such approaches risk understating sensitivity by implicitly ruling out a wide class of selection mechanisms. We now develop an alternative approach designed to allow an assessment of sensitivity to arbitrary deviations from ignorability that retains much of the parsimony of parametric methods. Specifically, we propose to study a nonparametric class of selection models indexed by a scalar measure of the deviations from MAR they generate. A sensitivity analysis may then be conducted by considering the conclusions that can be drawn under alternative levels of the selection index, with particular attention devoted to determination of the threshold level of selection necessary to undermine conclusions obtained under an ignorability assumption.

Since ignorability occurs when $F_{y|1,x}$ equals $F_{y|0,x}$, it is natural to measure deviations from MAR in terms of the distance between these two distributions. We propose as an index of selection the maximal Kolmogorov–Smirnov (KS) distance between $F_{y|1,x}$ and $F_{y|0,x}$ across all values of the covariates.[3] Thus, for $\mathcal{X}$, the support of $X$, we define the selection metric

$$\mathcal{S}(F) \equiv \sup_{x\in\mathcal{X}} \sup_{c\in\mathbf{R}} \bigl|F_{y|1,x}(c) - F_{y|0,x}(c)\bigr|. \tag{3}$$

Note that the missing at random assumption may be equivalently stated as $\mathcal{S}(F) = 0$, while $\mathcal{S}(F) = 1$ corresponds to severe forms of selection where the supports of random

[3]The Kolmogorov–Smirnov distance between two distributions $H_1(\cdot)$ and $H_2(\cdot)$ is defined as $\sup_{c\in\mathbf{R}} |H_1(c) - H_2(c)|$.

variables distributed according to $F_{y|1,x}$ and $F_{y|0,x}$ fail to intersect for some $x \in \mathcal{X}$. For illustrative purposes, Appendix A (available in the supplementary file) provides a numerical example that maps the parameters of a bivariate normal selection model into values of $\mathcal{S}(F)$, and maps plots of the corresponding observed and missing data cumulative distribution functions (CDFs).

By indexing a selection mechanism according to the discrepancy $\mathcal{S}(F)$ it generates, we effectively summarize the difficulties it implies for identifying $F_{y|x}$. In what follows, we aim to examine what can be learned about $F_{y|x}$ under a hypothetical bound on the degree of selection present as measured by $\mathcal{S}(F)$. Specifically, we study what conclusions can be obtained under the nominal restriction

$$\mathcal{S}(F) \leq k. \tag{4}$$

We emphasize that knowledge of a true value of $k$ for which (4) holds is not assumed. Rather, we propose to examine the conclusions that can be drawn when we presume the severity of selection, as measured by $\mathcal{S}(F)$, to be no larger than $k$. This hypothetical restriction will be shown to yield sharp tractable bounds on both the conditional distribution ($F_{y|x}$) and the quantile ($q(\cdot|x)$) functions. Such bounds will, in turn, enable us to determine the level of selection $k$ necessary to overturn conclusions drawn under MAR.

## 2.2 *Interpretation of k*

Our motivation for working with $\mathcal{S}(F)$ rather than a parametric selection model is that researchers generally lack prior knowledge of the selection process. It is useful, however, to have in mind a simple class of nonparametric data generating processes that provide an intuitive understanding of what the value $k$ in (4) represents. Toward this end, we borrow from the robust statistics literature (e.g., Tukey (1960), Huber (1964)) in modeling departures from ignorability as a mixture of missing at random and arbitrary nonignorable missing data processes.[4]

Specifically, consider a model where a fraction $k$ of the missing population is distributed according to an arbitrary CDF $\tilde{F}_{y|x}$, while the remaining fraction $1-k$ of that population are missing at random in the sense that they are distributed according to $F_{y|1,x}$. Succinctly, suppose

$$F_{y|0,x}(c) = (1-k)F_{y|1,x}(c) + k\tilde{F}_{y|x}(c), \tag{5}$$

where $\tilde{F}_{y|x}$ is unknown, and (5) holds for all $x \in \mathcal{X}$ and any $c \in \mathbf{R}$. In this setting, we then have

$$\begin{aligned} \mathcal{S}(F) &= \sup_{x\in\mathcal{X}} \sup_{c\in\mathbf{R}} \big|F_{y|1,x}(c) - k\tilde{F}_{y|x}(c) - (1-k)F_{y|1,x}(c)\big| \\ &= k \times \sup_{x\in\mathcal{X}} \sup_{c\in\mathbf{R}} \big|F_{y|1,x}(c) - \tilde{F}_{y|x}(c)\big|. \end{aligned} \tag{6}$$

[4]We thank an anonymous referee for suggesting this interpretation.

Hence, if we consider the mixture model in (5) for unknown $\tilde{F}_{y|x}$, then we must allow for the possibility that $\mathcal{S}(F)$ takes any value between 0 and $k$, as in (4).

Thus, we may interpret the level of $k$ in the restriction $\mathcal{S}(F) \leq k$ as a bound on the fraction of the missing sample that is not well represented by the observed data distribution. This heuristic is particularly helpful in establishing a link to the foundational work on bounds of Manski (1994, 2003). In the presence of missing data, the latter approach exploits that $F_{y|0,x}(c) \leq 1$ to obtain an upper bound for $F_{y|x}(c)$ of the form

$$\begin{aligned} F_{y|x}(c) &= F_{y|1,x}(c) \times p(x) + F_{y|0,x}(c) \times \big(1 - p(x)\big) \\ &\leq F_{y|1,x}(c) \times p(x) + \big(1 - p(x)\big). \end{aligned} \tag{7}$$

Heuristically, the upper bound in (7) follows from a "least favorable" configuration where the entire missing population lies below the point $c$. By contrast, under the mixture specification,

$$\begin{aligned} F_{y|x}(c) &= F_{y|1,x}(c) \times p(x) + \big\{(1-k)F_{y|1,x}(c) + k\tilde{F}_{y|x}(c)\big\} \times \big(1 - p(x)\big) \\ &\leq F_{y|1,x}(c) \times \big(1 - k\big(1 - p(x)\big)\big) + k\big(1 - p(x)\big). \end{aligned} \tag{8}$$

Thus, in this setting we need only worry about a fraction $k$ of the unobserved population being below the point $c$. We can, therefore, interpret $k$ as the proportion of the unobserved population that is allowed to take the least favorable configuration of Manski (1994).

Remark 2.1. The mixture interpretation of (5) also provides an interesting link to the work on "corrupted sampling" of Horowitz and Manski (1995), who derive bounds on the distribution of $Y_1 \in \mathbf{R}$ in a setting where, for $Z \in \{0, 1\}$, $Y_2 \in \mathbf{R}$, and

$$Y = Y_1 Z + Y_2(1 - Z), \tag{9}$$

only $Y$ is observed. The resulting identification region for the distribution of $Y_1$ can be characterized in terms of $\lambda \equiv P(Z = 1)$, and the authors study "robustness" in terms of critical levels of $\lambda$ under which conclusions are as uninformative as when $\lambda = 1$. In our missing data setting, the problematic observations are identified. Hence, it is the unobserved population that is "corrupted," as in equation (5). Our index $k$ then plays a similar role to $\lambda$ in the corrupted sampling model.

### 2.3 *Conditional quantiles*

For $q(\tau|X)$, the conditional $\tau$-quantile of $Y$ given $X$, we now examine what can be learned about the conditional quantile function $q(\tau|\cdot)$ under the nominal restriction $\mathcal{S}(F) \leq k$. In the absence of additional restrictions, the conditional quantile function ceases to be identified under any deviation from ignorability ($k > 0$). Nonetheless, $q(\tau|\cdot)$ may still be shown to lie within a nominal identified set. This set consists of the values of $q(\tau|\cdot)$ that would be compatible with the distribution of observables were the putative restriction $\mathcal{S}(F) \leq k$ known to hold. We qualify such a set as nominal due to the restriction that $\mathcal{S}(F) \leq k$ is part of a hypothetical exercise only.

The following lemma provides a sharp characterization of the nominal identified set.

Lemma 2.1. *Suppose Assumption* 2.1(ii) *and* (iii) *hold,* $\mathcal{S}(F) \leq k$, *and let* $F^{-}_{y|1,x}(c) = F^{-1}_{y|1,x}(c)$ *if* $0 < c < 1$, $F^{-}_{y|1,x}(c) = -\infty$ *if* $c \leq 0$, *and* $F^{-}_{y|1,x}(c) = \infty$ *if* $c \geq 1$. *Defining* $(q_L(\tau, k|x), q_U(\tau, k|x))$ *by*

$$q_L(\tau, k|x) \equiv F^{-}_{y|1,x}\left(\frac{\tau - \min\{\tau + kp(x), 1\}(1 - p(x))}{p(x)}\right), \tag{10}$$

$$q_U(\tau, k|x) \equiv F^{-}_{y|1,x}\left(\frac{\tau - \max\{\tau - kp(x), 0\}(1 - p(x))}{p(x)}\right), \tag{11}$$

*it follows that the identified set for* $q(\tau|\cdot)$ *is* $\mathcal{C}(\tau, k) \equiv \{\theta : \mathcal{X} \to \mathbf{R} : q_L(\tau, k|\cdot) \leq \theta(\cdot) \leq q_U(\tau, k|\cdot)\}$.

Proof. Letting $\mathrm{KS}(F_{y|1,x}, F_{y|0,x}) \equiv \sup_c |F_{y|1,x}(c) - F_{y|0,x}(c)|$, we first observe that

$$\begin{aligned} &\mathrm{KS}(F_{y|1,x}, F_{y|0,x}) \\ &\quad = \frac{1}{p(x)} \times \sup_{c \in \mathbf{R}} \left| F_{y|1,x}(c) \times p(x) + F_{y|0,x}(c) \times \{1 - p(x)\} - F_{y|0,x}(c) \right| \\ &\quad = \frac{1}{p(x)} \times \sup_{c \in \mathbf{R}} \left| F_{y|x}(c) - F_{y|0,x}(c) \right|. \end{aligned} \tag{12}$$

Therefore, it immediately follows from the hypothesis $\mathcal{S}(F) \leq k$ and result (12) that

$$\begin{aligned} \tau &= F_{y|1,x}\big(q(\tau|x)\big) \times p(x) + F_{y|0,x}\big(q(\tau|x)\big) \times \{1 - p(x)\} \\ &\leq F_{y|1,x}\big(q(\tau|x)\big) \times p(x) + \min\{F_{y|x}\big(q(\tau|x)\big) + kp(x), 1\} \times \{1 - p(x)\} \\ &= F_{y|1,x}\big(q(\tau|x)\big) \times p(x) + \min\{\tau + kp(x), 1\} \times \{1 - p(x)\}. \end{aligned} \tag{13}$$

By identical manipulations, $F_{y|1,x}(q(\tau|x)) \times p(x) \leq \tau - \max\{\tau - kp(x), 0\} \times \{1 - p(x)\}$ and, hence, by inverting $F_{y|1,x}$, we conclude that indeed $q(\tau|\cdot) \in \mathcal{C}(\tau, k)$.

To prove the bounds are sharp, we aim to show that for every $\theta \in \mathcal{C}(\tau, k)$ and every $x \in \mathcal{X}$, there is a $\tilde{F}_{y|0,x}$ such that (I) Assumption 2.1(ii) is satisfied, (II) $\sup_c |F_{y|1,x}(c) - \tilde{F}_{y|0,x}(c)| \leq k$, and (III)

$$F_{y|1,x}\big(\theta(x)\big) \times p(x) + \tilde{F}_{y|0,x}\big(\theta(x)\big) \times \big(1 - p(x)\big) = \tau. \tag{14}$$

Toward this end, first note that for (14) to hold, we must set $\tilde{F}_{y|0,x}(\theta(x)) = \kappa(x)$, where

$$\kappa(x) \equiv \frac{\tau - F_{y|1,x}(\theta(x)) \times p(x)}{1 - p(x)}. \tag{15}$$

Moreover, since $\theta \in \mathcal{C}(\tau, k)$, direct calculation reveals that $|\kappa(x) - F_{y|1,x}(\theta(x))| \leq k$. Further assuming $\kappa(x) \geq F_{y|1,x}(\theta(x))$ (the case $\kappa(x) \leq F_{y|1,x}(\theta(x))$ is analogous), we then note that if

$$\tilde{F}_{y|0,x}(c) \equiv \min\{F_{y|1,x}(c) + \Psi_x(c), 1\}, \tag{16}$$

then $\tilde{F}_{y|0,x}$ will satisfy (I) provided $\Psi_x(c)$ is continuous, increasing, and satisfies $\lim_{c\to-\infty} \Psi_x(c) = 0$, (II) provided $0 \leq \Psi_x(c) \leq \kappa(x) \leq k$ for all $c$, and (III) if $\Psi_x(\theta(x)) = \kappa(x) - F_{y|1,x}(\theta(x))$. For $(a)_+ \equiv \max\{a, 0\}$ and $K_0 > \kappa(x)/F_{y|1,x}(\theta(x))$, these conditions are satisfied by the function

$$\Psi_x(c) \equiv \max\{0, \kappa(x) - F_{y|1,x}(\theta(x)) - K_0(F_{y|1,x}(\theta(x)) - F_{y|1,x}(c))_+\}. \tag{17}$$

Therefore, the claim of the lemma follows from (16) and (17). □

Figure 1 provides intuition as to why the bounds in Lemma 2.1 are sharp. In this illustration, the median of the observed distribution $F_{y|1,x}$ is 0 and $p(x) = 1/2$. These parameters yield an upper bound for $F_{y|0,x}(c) \leq \min\{1, F_{y|1,x}(c) + k\}$—the line termed "$F_{y|0,x}$ Upper Bound" in Figure 1. The lower bound $q_L(0.5|x)$ is then given by the point at which the mixture of $F_{y|1,x}$ and the upper bound for $F_{y|0,x}$ crosses 1/2, which in Figure 1 is given by −0.5. Any CDF $F_{y|0,x}$ that equals its upper bound at the point $q_L(0.5|x)$ and whose maximal deviation from $F_{y|1,x}$ occurs at $q_L(0.5|x)$ will then justify $q_L(0.5|x)$ as a possible median. The same logic reveals that a CDF $F_{y|0,x}$ can be constructed that stays below its bound and such that the median of $F_{y|x}$ equals any point in $(q_L(0.5|x), 0]$.

Remark 2.2. A key advantage, for our purposes, of employing Kolmogorov–Smirnov type distances is that they are defined directly in terms of CDFs. Competing metrics such

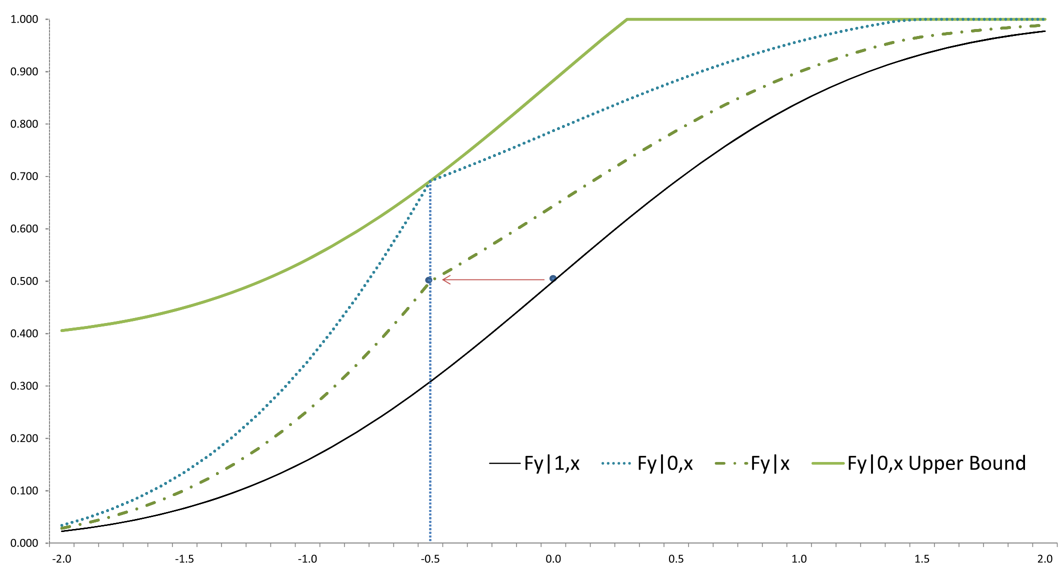


Figure 1. Illustration with $p(x) = 1/2$, $k = 0.38$, and $q_L(0.5|x) = -0.5$.

as Hellinger or Kullback–Leibler are, by contrast, defined on densities. Consequently, the quantile bounds that result from employing these alternative metrics do not take simple analytic forms as in Lemma 2.1.

Remark 2.3. An alternative Kolmogorov–Smirnov type index that delivers tractable bounds is

$$\mathcal{W}(F) \equiv \sup_{x\in\mathcal{X}} w(x) \times \Big\{ \sup_{c\in\mathbf{R}} \big| F_{y|1,x}(c) - F_{y|0,x}(c) \big| \Big\} \tag{18}$$

for weights $w(x) > 0$. The restriction $\mathcal{W}(F) \leq k$ is equivalent to employing the bound $k/w(x)$ on the Kolmogorov–Smirnov distance between $F_{y|1,x}$ and $F_{y|0,x}$ at each $x \in \mathcal{X}$. Thus, the identified set for the conditional quantile $q(\tau|x)$ follows from Lemma 2.1 with $k/w(x)$ in place of $k$ in (10) and (11). This alternative index of selection may prove useful to researchers who suspect particular forms of heterogeneity in the selection mechanism across covariate values.

### 2.4 *Examples*

We conclude this section by illustrating through examples how the bound functions $(q_L, q_U)$ can be used to evaluate the sensitivity of conclusions obtained under MAR. For simplicity, we let $X$ be binary so that the conditional $\tau$-quantile function $q(\tau|\cdot)$ takes only two values.

Example 2.1 (Pointwise Conclusions). Suppose interest centers on whether $q(\tau|X=1)$ equals $q(\tau|X=0)$ for a specific quantile $\tau_0$. A researcher who finds them to differ under a MAR analysis may easily assess the sensitivity of his conclusion to the presence of selection by employing the functions $(q_L(\tau_0|\cdot), q_U(\tau_0|\cdot))$. Concretely, the minimal amount of selection necessary to overturn the conclusion that the conditional quantiles differ is given by

$$\begin{aligned} k_0 \equiv \inf k : q_L(\tau_0, k|X=1) - q_U(\tau_0, k|X=0) \\ \leq 0 \leq q_U(\tau_0, k|X=1) - q_L(\tau_0, k|X=0). \end{aligned} \tag{19}$$

That is, $k_0$ is the minimal level of selection under which the nominal identified sets for $q(\tau_0|X=0)$ and $q(\tau_0|X=1)$ contain a common value.

Example 2.2 (Distributional Conclusions). A researcher is interested in whether the conditional distribution $F_{y|x=0}$ first order stochastically dominates $F_{y|x=1}$ or, equivalently, whether $q(\tau|X=1) \leq q(\tau|X=0)$ for all $\tau \in (0,1)$. She finds under MAR that $q(\tau|X=1) > q(\tau|X=0)$ at multiple values of $\tau$, leading her to conclude that first order stochastic dominance does not hold. She may assess what degree of selection is necessary to cast doubt on this conclusion by examining

$$k_0 \equiv \inf k : q_L(\tau, k|X=1) \leq q_U(\tau, k|X=0) \quad \text{for all } \tau \in (0,1). \tag{20}$$

Here, $k_0$ is the smallest level of selection for which an element of the identified set for $q(\cdot|X=1)$ ($q_L(\cdot, k_0|X=1)$) is everywhere below an element of the identified set for $q(\cdot|X=0)$ ($q_U(\cdot, k_0|X=0)$). Thus, $k_0$ is the threshold level of selection under which $F_{y|x=0}$ may first order stochastically dominate $F_{y|x=1}$.

Example 2.3 (Breakdown Analysis). A more nuanced sensitivity analysis might examine what degree of selection is necessary to undermine the conclusion that $q(\tau|X=1) \neq q(\tau|X=0)$ at each specific quantile $\tau$. As in Example 2.1, we can define the quantile-specific critical level of selection

$$\begin{aligned}\kappa_0(\tau) \equiv \inf k : {} & q_L(\tau, k|X=1) - q_U(\tau, k|X=0) \\ & \leq 0 \leq q_U(\tau, k|X=1) - q_L(\tau, k|X=0).\end{aligned} \tag{21}$$

By considering $\kappa_0(\tau)$ at different values of $\tau$, we implicitly define a "breakdown" function $\kappa_0(\cdot)$ that reveals the differential sensitivity of the initial conjecture at each quantile $\tau \in (0, 1)$.

## 3. Parametric modeling

Analysis of the conditional $\tau$-quantile function $q(\tau|\cdot)$ and its corresponding nominal identified set $\mathcal{C}(\tau, k)$ can be cumbersome when many covariates are present as the resulting bounds will be of high dimension and difficult to visualize. Moreover, it can be arduous even to state the features of a high dimensional CQF one wishes to examine for sensitivity. It is convenient in such cases to be able to summarize $q(\tau|\cdot)$ using a parametric model. Failure to acknowledge, however, that the model is simply an approximation can easily yield misleading conclusions.

Figure 2 illustrates a case where the nominal identified set $\mathcal{C}(\tau, k)$ possesses an erratic (though perhaps not unusual) shape. The set of linear CQFs that obey the bounds provide a poor description of this set, covering only a small fraction of its area. Were the true CQF known to be linear, this reduction in the size of the identified set would be welcome, the benign result of imposing additional identifying information. But in the absence of true prior information, these reductions in the size of the identified set are unwarranted—a phenomenon we term *identification by misspecification*.

The specter of misspecification leaves the applied researcher with a difficult choice. One can either conduct a fully nonparametric analysis of the nominal identified set, which may be difficult to interpret with many covariates, or work with a parametric set likely to overstate what is known about the CQF. Under identification, this tension is typically resolved by estimating parametric models that possess an interpretation as best approximations to the true CQF and adjusting the corresponding inferential methods accordingly as in Chamberlain (1994) and Angrist, Chernozhukov, and Fernández-Val (2006). Following Horowitz and Manski (2006), Stoye (2007), and Ponomareva and Tamer (2009), we extend this approach and develop methods for conducting inference on the best parametric approximation to the true CQF under partial identification.

We focus on linear models and approximations that minimize a known quadratic loss function. For $S$ a known measure on $\mathcal{X}$ and $E_S[g(X)]$ denoting the expectation of

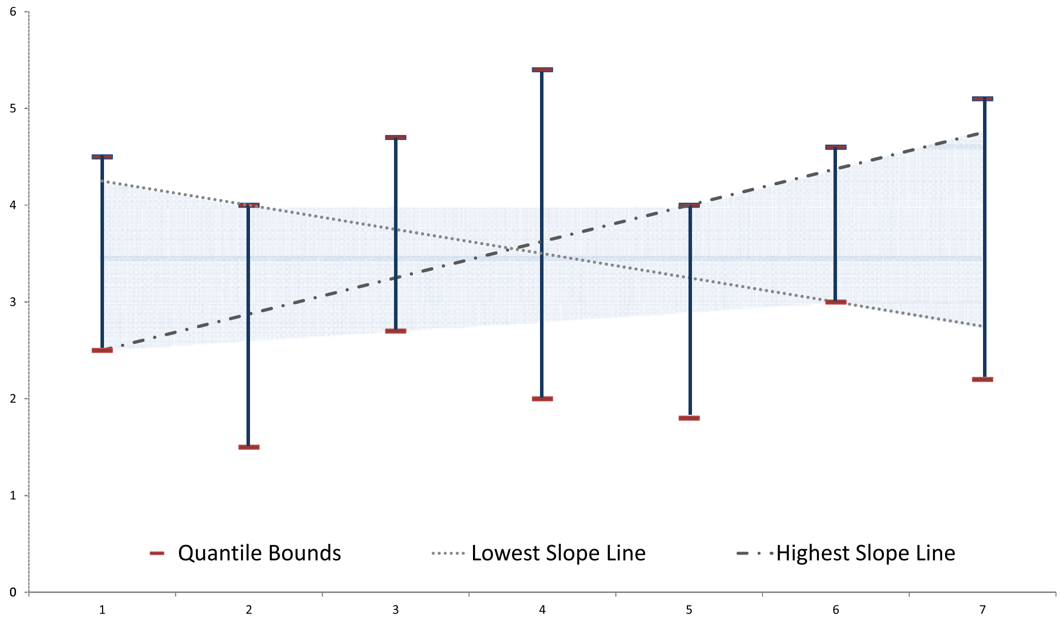


FIGURE 2. Linear conditional quantile functions (shaded region) as a subset of the identified set.

$g(X)$ when $X$ is distributed according to $S$, we define the parameters that govern a best linear approximation (BLA) as[5]

$$\beta(\tau) \equiv \arg\min_{\gamma \in \mathbf{R}^l} E_S\big[\big(q(\tau|X) - X'\gamma\big)^2\big]. \tag{22}$$

In cases where the CQF is actually linear in $X$, it will coincide with its best linear approximation. Otherwise, the BLA will provide a minimum mean square approximation to the CQF. In many settings, such an approximation may serve as a relatively accurate and parsimonious summary of the underlying quantile function.

Lack of identification of the conditional quantile function $q(\tau|\cdot)$ due to missing data implies lack of identification of the parameter $\beta(\tau)$. We therefore consider the set of parameters that correspond to the best linear approximation to *some* CQF in $\mathcal{C}(\tau, k)$. Formally, we define

$$\mathcal{P}(\tau, k) \equiv \Big\{\beta \in \mathbf{R}^l : \beta \in \arg\min_{\gamma \in \mathbf{R}^l} E_S\big[\big(\theta(X) - X'\gamma\big)^2\big] \text{ for some } \theta \in \mathcal{C}(\tau, k)\Big\}. \tag{23}$$

Figure 3 illustrates the approximation generated by an element of $\mathcal{P}(\tau, k)$ graphically. While intuitively appealing, the definition of $\mathcal{P}(\tau, k)$ is not necessarily the most convenient for computational purposes. Fortunately, the choice of quadratic loss and the characterization of $\mathcal{C}(\tau, k)$ in Lemma 2.1 imply a tractable alternative representation for $\mathcal{P}(\tau, k)$, which we obtain in the following lemma.

[5]The measure $S$ weights the squared deviations in each covariate bin. Its specification is an inherently context-specific task that depends entirely on the researcher's objectives. In Section 4, we weight the deviations by sample size. Other schemes (including equal weighting) may also be of interest in some settings.

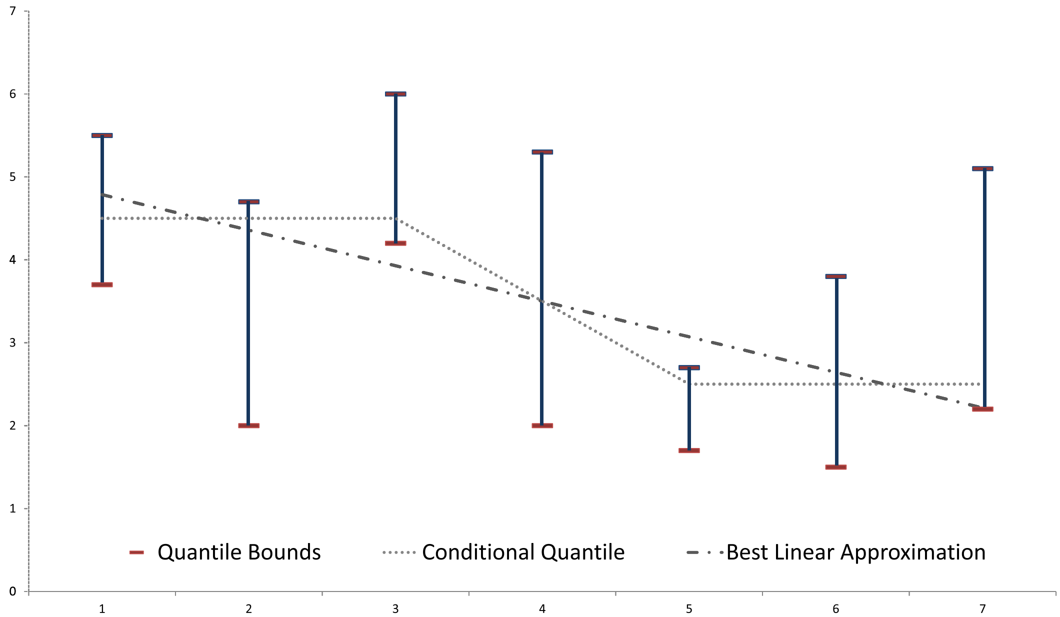


FIGURE 3. Conditional quantile and its best linear approximation.

LEMMA 3.1. *If Assumption* 2.1(ii) *and* (iii) *hold,* $\mathcal{S}(F) \le k$, *and* $E_S[XX']$ *is invertible, then it follows that*

$$\mathcal{P}(\tau, k) = \{\beta \in \mathbf{R}^l : \beta = (E_S[XX'])^{-1} E_S[X\theta(X)] \\ \text{s.t. } q_L(\tau, k|x) \le \theta(x) \le q_U(\tau, k|x) \text{ for all } x \in \mathcal{X}\}.$$

Interest often centers on either a particular coordinate of $\beta(\tau)$ or the value of the approximate CQF at a specified value of the covariates. Both these quantities may be expressed as $\lambda'\beta(\tau)$ for some known vector $\lambda \in \mathbf{R}^l$. Using Lemma 3.1, it is straightforward to show that the nominal identified set for parameters of the form $\lambda'\beta(\tau)$ is an interval with endpoints characterized as the solution to linear programming problems.[6]

COROLLARY 3.1. *Suppose Assumption* 2.1(ii) *and* (iii) *hold,* $\mathcal{S}(F) \le k$, *and* $E_S[XX']$ *is invertible, and define*

$$\pi_L(\tau, k) \equiv \inf_{\beta \in \mathcal{P}(\tau, k)} \lambda'\beta = \inf_{\theta} \lambda'(E_S[XX'])^{-1} E_S[X\theta(X)] \\ \text{s.t. } q_L(\tau, k|x) \le \theta(x) \le q_U(\tau, k|x), \tag{24}$$

[6]Since $X$ has discrete support, we can characterize the function $\theta$ by the finite number of values it may take. Because the weighting scheme $S$ is known, so is $\lambda'(E_S[XX'])^{-1}$, and, hence, the objectives in (24) and (25) are of the form $w'\theta$, where $w$ is a known vector and $\theta$ is a finite dimensional vector over which the criterion is optimized.

$$\pi_U(\tau,k) \equiv \sup_{\beta\in\mathcal{P}(\tau,k)} \lambda'\beta = \sup_{\theta} \lambda'\big(E_S[XX']\big)^{-1}E_S\big[X\theta(X)\big] \quad \text{s.t. } q_L(\tau,k|x) \le \theta(x) \le q_U(\tau,k|x). \tag{25}$$

*The nominal identified set for $\lambda'\beta(\tau)$ is then given by the interval* $[\pi_L(\tau,k), \pi_U(\tau,k)]$.

Corollary 3.1 provides sharp bounds on the quantile process $\lambda'\beta(\cdot)$ at each point of evaluation $\tau$ under the restriction that $\mathcal{S}(F) \le k$. However, sharpness of the bounds at each point of evaluation does not, in this case, translate into sharp bounds on the entire process. To see this, note that Corollary 3.1 implies $\lambda'\beta(\cdot)$ must belong to the set

$$\mathcal{G}(k) \equiv \big\{g:[0,1]\to\mathbf{R}: \pi_L(\tau,k) \le g(\tau) \le \pi_U(\tau,k) \text{ for all } \tau\big\}. \tag{26}$$

While the true $\lambda'\beta(\cdot)$ must belong to $\mathcal{G}(k)$, not all functions in $\mathcal{G}(k)$ can be justified as some distribution's BLA process.[7] Therefore, $\mathcal{G}(k)$ does not constitute the nominal identified set for the process $\lambda'\beta(\cdot)$ under the restriction $\mathcal{S}(F) \le k$. Fortunately, $\pi_L(\cdot,k)$ and $\pi_U(\cdot,k)$ *are* in the identified set over the range of $(\tau,k)$ for which the bounds are finite. Thus, the set $\mathcal{G}(k)$, though not sharp, does retain the favorable properties of (i) sharpness at any point of evaluation $\tau$, (ii) containing the true identified set for the process so that processes not in $\mathcal{G}(k)$ are also known not to be in the identified set, (iii) sharpness of the lower and upper bound functions $\pi_L(\cdot,k)$ and $\pi_U(\cdot,k)$, and (iv) ease of analysis and graphical representation.

### 3.1 *Examples*

We now revisit Examples 2.1–2.3 from Section 2.1 so as to illustrate how to characterize the sensitivity of conclusions drawn under MAR with parametric models. We keep the simplifying assumption that $X$ is scalar, but no longer assume it is binary and instead consider the model

$$q(\tau|X) = \alpha(\tau) + X\beta(\tau). \tag{27}$$

Note that when $X$ is binary, equation (27) provides a nonparametric model of the CQF, in which case our discussion coincides with that of Section 2.1.

Example 2.1 (Continued). Suppose that an analysis under MAR reveals $\beta(\tau_0) \ne 0$ at a specific quantile $\tau_0$. We may then define the critical level of $k_0$ necessary to cast doubt on this conclusion as

$$k_0 \equiv \inf k : \pi_L(\tau_0,k) \le 0 \le \pi_U(\tau_0,k). \tag{28}$$

That is, under any level of selection $k \ge k_0$, it is no longer possible to conclude that $\beta(\tau_0) \ne 0$.

[7] For example, under our assumptions, $\lambda'\beta(\cdot)$ is a continuous function of $\tau$. Hence, any $g\in\mathcal{G}(k)$ that is discontinuous is not in the nominal identified set for $\lambda'\beta(\cdot)$ under the hypothetical that $\mathcal{S}(F)\le k$.

Example 2.2 (Continued). In a parametric analogue of first order stochastic dominance of $F_{y|x}$ over $F_{y|x'}$ for $x < x'$, a researcher examines whether $\beta(\tau) \le 0$ for all $\tau \in (0,1)$. Suppose that a MAR analysis reveals that $\beta(\tau) > 0$ for multiple values of $\tau$. The functions $(\pi_L, \pi_U)$ enable her to assess what degree of selection is necessary to undermine her conclusions by considering

$$k_0 \equiv \inf k : \pi_L(\tau, k) \le 0 \quad \text{for all } \tau \in (0,1). \tag{29}$$

Note that finding $\pi_L(\tau, k_0) \le 0$ for all $\tau \in (0,1)$ does, in fact, cast doubt on the conclusion that $\beta(\tau) > 0$ for some $\tau$ because $\pi_L(\cdot, k_0)$ is itself in the nominal identified set for $\beta(\cdot)$. That is, under a degree of selection $k_0$, the *process* $\beta(\cdot)$ may equal $\pi_L(\cdot, k_0)$.

Example 2.3 (Continued). Generalizing the considerations of Example 2.1, we can examine what degree of selection is necessary to undermine the conclusion that $\beta(\tau) \neq 0$ at each specific $\tau$. In this manner, we obtain a quantile-specific critical level of selection:

$$\kappa_0(\tau) \equiv \inf k : \pi_L(\tau, k) \le 0 \le \pi_U(\tau, k). \tag{30}$$

As in Section 2.1, the resulting breakdown function $\kappa_0(\cdot)$ enables us to characterize the differential sensitivity of the entire conditional distribution to deviations from MAR.

## 4. Estimation and inference

In what follows, we develop methods for conducting sensitivity analysis using sample estimates of $\pi_L(\tau, k)$ and $\pi_U(\tau, k)$. Our strategy for estimating the bounds $\pi_L(\tau, k)$ and $\pi_U(\tau, k)$ consists of first obtaining estimates $\hat{q}_L(\tau, k|x)$ and $\hat{q}_U(\tau, k|x)$ of the conditional quantile bounds, and then employing them in place of $q_L(\tau, k|x)$ and $q_U(\tau, k|x)$ in the linear programming problems given in (24) and (25). Thus, an appealing characteristic of our estimator is the reliability and low computational cost involved in solving a linear programming problem—considerations that become particularly salient when implementing a bootstrap procedure for inference.

Recall that the conditional quantile bounds $q_L(\tau, k|x)$ and $q_U(\tau, k|x)$ may be expressed as quantiles of the observed data (see Lemma 2.1). We estimate these bounds using their sample analogues. For the development of our bootstrap procedure, however, it will be useful to consider a representation of these sample estimates as the solution to a general $M$-estimation problem. Toward this end, we define a family of population criterion functions (as indexed by $(\tau, b, x)$) given by

$$Q_x(c|\tau, b) \equiv \big(P(Y \le c, D = 1, X = x) + bP(D = 0, X = x) - \tau P(X = x)\big)^2. \tag{31}$$

Notice that if $Q_x(\cdot|\tau, b)$ is minimized at some $c^* \in \mathbf{R}$, then $c^*$ must satisfy the first order condition

$$F_{y|1,x}(c^*) = \frac{\tau - b(1 - p(x))}{p(x)}. \tag{32}$$

Therefore, Lemma 2.1 implies that if unique minimizers to $Q_x(\cdot|\tau, b)$ exist for $b = \min\{\tau + kp(x), 1\}$ and $b = \max\{\tau - kp(x), 0\}$, then they must be given by $q_L(\tau, k|x)$ and $q_U(\tau, k|x)$, respectively.

For this approach to prove successful, however, we must focus on values of $(\tau, k)$ such that $Q_x(\cdot|\tau, b)$ has a unique minimizer at the corresponding $b$—which we note are the same values for which the bounds $q_L(\tau, k|x)$ and $q_U(\tau, k|x)$ are finite. Additionally, we focus on $(\tau, k)$ pairs such that $\mathcal{S}(F) \leq k$ proves more informative than the restriction that $F_{y|0,x}$ lies between 0 and 1. Succinctly, for an arbitrary fixed $\zeta > 0$, these conditions are satisfied by values of $(\tau, k)$ in the set

$$\begin{aligned}\mathcal{B}_\zeta \equiv \Big\{&(\tau, k) \in [0,1]^2 : \text{ (i) } kp(x)\big(1 - p(x)\big) + 2\zeta \leq \tau p(x),\\ &\text{(ii) } kp(x)\big(1 - p(x)\big) + 2\zeta \leq (1-\tau)p(x), \text{ (iii) } kp(x) + 2\zeta \leq \tau,\\ &\text{(iv) } kp(x) + 2\zeta \leq 1 - \tau, \forall x \in \mathcal{X}\Big\}.\end{aligned}$$

Heuristically, by restricting attention to $(\tau, k) \in \mathcal{B}_\zeta$, we are imposing that large or small values of $\tau$ must be accompanied by small values of $k$. This simply reflects that the fruitful study of quantiles close to 1 or 0 requires stronger assumptions on the nature of the selection process than the study of, for example, the conditional median.

For any $(\tau, k) \in \mathcal{B}_\zeta$, we then obtain the desired characterization of $q_L(\tau, k|x)$ and $q_U(\tau, k|x)$ as

$$\begin{aligned} q_L(\tau, k|x) &= \arg\min_{c \in \mathbf{R}} Q_x\big(c|\tau, \tau + kp(x)\big),\\ q_U(\tau, k|x) &= \arg\min_{c \in \mathbf{R}} Q_x\big(c|\tau, \tau - kp(x)\big).\end{aligned} \tag{33}$$

These relations suggest estimating the bounds $q_L(\tau, k|x)$ and $q_U(\tau, k|x)$ through the minimizers of an appropriate sample analogue. Toward this end, we define the sample criterion function

$$\begin{aligned} Q_{x,n}(c|\tau, b) \equiv \Bigg(\frac{1}{n}\sum_{i=1}^{n}\big\{&1\{Y_i \leq c, D_i = 1, X_i = x\}\\ &+ b1\{D_i = 0, X_i = x\} - \tau 1\{X_i = x\}\big\}\Bigg)^2 \end{aligned} \tag{34}$$

and, exploiting (31), we consider the extremum estimators for $q_L(\tau, k|x)$ and $q_U(\tau, k|x)$ given by

$$\begin{aligned} \hat{q}_L(\tau, k|x) &\in \arg\min_{c \in \mathbf{R}} Q_{x,n}\big(c|\tau, \tau + k\hat{p}(x)\big),\\ \hat{q}_U(\tau, k|x) &\in \arg\min_{c \in \mathbf{R}} Q_{x,n}\big(c|\tau, \tau - k\hat{p}(x)\big),\end{aligned} \tag{35}$$

where $\hat{p}(x) \equiv (\sum_i 1\{D_i = 1, X_i = x\})/(\sum_i 1\{X_i = x\})$. Finally, employing these estimators, we may solve the sample analogues to the linear programming problems in (24) and (25) to obtain

$$\begin{aligned} \hat{\pi}_L(\tau,k) \equiv \inf_\theta \lambda'\big(E_S[XX']\big)^{-1} E_S\big[X\theta(X)\big] \\ \text{s.t. } \hat{q}_L(\tau,k|x) \leq \theta(x) \leq \hat{q}_U(\tau,k|x), \end{aligned} \tag{36}$$

$$\begin{aligned} \hat{\pi}_U(\tau,k) \equiv \sup_\theta \lambda'\big(E_S[XX']\big)^{-1} E_S\big[X\theta(X)\big] \\ \text{s.t. } \hat{q}_L(\tau,k|x) \leq \theta(x) \leq \hat{q}_U(\tau,k|x). \end{aligned} \tag{37}$$

We introduce the following additional assumption so as to develop our asymptotic theory.

Assumption 4.1. (i) $\mathcal{B}_\zeta \neq \emptyset$; (ii) $F_{y|1,x}(c)$ *has a continuous bounded derivative* $f_{y|1,x}(c)$; (iii) $f_{y|1,x}(c)$ *has a continuous bounded derivative* $f'_{y|1,x}(c)$; (iv) $E_S[XX']$ *is invertible*; (v) $f_{y|1,x}(c)$ *is bounded away from zero uniformly on all* $c$ *satisfying* $\zeta \leq F_{y|1,x}(c)p(x) \leq p(x) - \zeta\ \forall x \in \mathcal{X}$.

Provided that the conditional probability of missing is bounded away from 1 and $\zeta > 0$ is sufficiently small, Assumption 4.1(i) will be satisfied since $\mathcal{B}_\zeta$ contains the MAR analysis as a special case. Assumption 4.1(ii) and (iii) demands that $F_{y|1,x}$ be twice continuously differentiable, while Assumption 4.1(iv) ensures $\pi_L(\tau,k)$ and $\pi_U(\tau,k)$ are well defined; see Corollary 3.1. Assumption 4.1(v) demands that the density $f_{y|1,x}$ be positive at all the quantiles that are estimated—a common requirement in the asymptotic study of sample quantiles. We note that Assumption 4.1(v) is a strengthening of Assumption 2.1(ii), which already imposes strict monotonicity of $F_{y|1,x}$.[8]

As a preliminary result, we derive the asymptotic distribution of the nonparametric bound estimators $\hat{q}_L(\tau,k|x)$ and $\hat{q}_U(\tau,k|x)$ uniformly in $(\tau,k,x) \in \mathcal{B}_\zeta \times \mathcal{X}$. Though potentially of independent interest, this result also enables us to derive the asymptotic distribution of the functions $\hat{\pi}_L$ and $\hat{\pi}_U$, pointwise defined by (36) and (37), as elements of $L^\infty(\mathcal{B}_\zeta)$ (the space of bounded functions on $\mathcal{B}_\zeta$). Such a derivation is a key step toward constructing confidence intervals for $\pi_L(\tau,k)$ and $\pi_U(\tau,k)$ that are uniform in $(\tau,k)$. As we illustrate in Section 4.2, these uniformity results are particularly useful for conducting the sensitivity analyses illustrated in Examples 2.1–2.3.

Theorem 4.1. *If Assumptions* 2.1 *and* 4.1 *hold, and* $\{Y_iD_i, X_i, D_i\}_{i=1}^n$ *is an independent and identically distributed* (*i.i.d.*) *sample, then*

$$\sqrt{n}\begin{pmatrix} \hat{q}_L - q_L \\ \hat{q}_U - q_U \end{pmatrix} \xrightarrow{\mathcal{L}} J, \tag{38}$$

*where* $J$ *is a Gaussian process on* $L^\infty(\mathcal{B}_\zeta \times \mathcal{X}) \times L^\infty(\mathcal{B}_\zeta \times \mathcal{X})$. *Moreover, under the same assumptions,*

$$\sqrt{n}\begin{pmatrix} \hat{\pi}_L - \pi_L \\ \hat{\pi}_U - \pi_U \end{pmatrix} \xrightarrow{\mathcal{L}} G, \tag{39}$$

[8] That $F_{y|1,x}$ is strictly increasing on $C \equiv \{c : 0 < F_{y|1,x}(c) < 1\}$ implies $f_{y|1,x}(c) > 0$ on a dense subset of $C$.

*where G is a Gaussian process on the space* $L^{\infty}(\mathcal{B}_{\zeta}) \times L^{\infty}(\mathcal{B}_{\zeta})$.

We note that since $J$ and $G$ are Gaussian processes, their marginals $J(\tau, k, x)$ and $G(\tau, k)$ are simply bivariate normal random variables. For notational convenience, we let $J^{(i)}(\tau, k, x)$ and $G^{(i)}(\tau, k)$ denote the $i$th component of the vector $J(\tau, k, x)$ and $G(\tau, k)$, respectively. Thus, for instance, $G^{(1)}(\tau, k)$ is the limiting distribution corresponding to the lower bound estimate $\hat{\pi}_L(\tau, k)$, while $G^{(2)}(\tau, k)$ is the limiting distribution of the upper bound estimate $\hat{\pi}_U(\tau, k)$.

Remark 4.1. Our derivations show that $\pi_L$ and $\pi_U$ are linear transformations of the nonparametric bounds $q_L$ and $q_U$ to establish (39).[9] If $X$ does not have discrete support, then result (38) fails to hold and our arguments do not deliver (39). While it would constitute a significant extension to Theorem 4.1, it is, in principle, possible to employ nonparametric estimators for $q_L(\tau, k|\cdot)$ and $q_U(\tau, k|\cdot)$, and to exploit that $\pi_L(\tau, k)$ and $\pi_U(\tau, k)$ are smooth functionals to obtain asymptotically normal estimators *pointwise* in $(\tau, k)$ without a discrete support requirement on $X$ (Newey (1994) and Chen, Linton, and Keilegom (2003)). However, obtaining an asymptotic distribution jointly in *all* $(\tau, k) \in \mathcal{B}_{\zeta}$, as in (39), would present a substantial complication, as standard results in semiparametric estimation concern finite dimensional parameters, for example, a finite set of $(\tau, k)$.

Remark 4.2. Letting $P$ denote the joint distribution of $(YD, D, X)$ and letting $\mathbf{P}$ denote a set of distributions, we note that Theorem 4.1 is not uniform over classes $\mathbf{P}$ such that

$$\inf_{P \in \mathbf{P}} \inf_{x \in \mathcal{X}} P(X = x) = 0. \tag{40}$$

Heuristically, the asymptotically normal approximation for $\hat{q}_L(\tau, k|x)$ and $\hat{q}_U(\tau, k|x)$ in (38) will prove unreliable at $(x, P)$ pairs for which $P(X = x)$ is small relative to $n$. As argued in Remark 4.1, however, a failure of (38) does not immediately translate into a failure of (39).

Remark 4.3. For fixed distribution $P$, Theorem 4.1 is additionally not uniform in the parameter $\zeta$ defining $\mathcal{B}_{\zeta}$ when it is allowed to be arbitrarily close to zero. Intuitively, small values of $\zeta$ imply that $q_L(\tau, k|x)$ and $q_U(\tau, k|x)$ can correspond to extreme quantiles of $F_{y|1,x}$ for certain $(\tau, k) \in \mathcal{B}_{\zeta}$. However, the limiting distribution of a sample $\tau_n$ quantile when $n\tau_n \to c > 0$ is nonnormal, implying that Theorem 4.1 cannot hold under sequences $\zeta_n \downarrow 0$ with $n\zeta_n \to c' > 0$ (e.g., Galambos (1973)).

### 4.1 *Examples*

We now return to the examples of Sections 2.1 and 3.1, and discuss how to conduct inference on the various sensitivity measures introduced there. For simplicity, we assume the relevant critical values are known. In Section 4.2, we develop a bootstrap procedure for their estimation.

[9] Formally, there exists a continuous linear transformation $K : L^{\infty}(\mathcal{B}_{\zeta} \times \mathcal{X}) \times L^{\infty}(\mathcal{B}_{\zeta} \times \mathcal{X}) \to L^{\infty}(\mathcal{B}_{\zeta}) \times L^{\infty}(\mathcal{B}_{\zeta})$ such that $(\pi_L, \pi_U) = K(q_L, q_U)$.

EXAMPLE 2.1 (Continued). Since under any level of selection $k$ larger than $k_0$, it is also not possible to conclude $\beta(\tau_0) \neq 0$, it is natural to construct a one-sided (rather than two-sided) confidence interval for $k_0$. Toward this end, let $r^{(i)}_{1-\alpha}(k)$ be the $1-\alpha$ quantile of $G^{(i)}(\tau_0, k)$ and define

$$k_0^* \equiv \inf k : \hat{\pi}_L(\tau_0, k) - \frac{r^{(1)}_{1-\alpha}(k)}{\sqrt{n}} \leq 0 \leq \hat{\pi}_U(\tau_0, k) + \frac{r^{(2)}_{1-\alpha}(k)}{\sqrt{n}}. \tag{41}$$

The confidence interval $[k_0^*, 1]$ then covers $k_0$ with asymptotic probability at least $1-\alpha$.

EXAMPLE 2.2 (Continued). Construction of a one-sided confidence interval for $k_0$ in this setting is more challenging, as it requires us to employ the uniformity of our estimator in $\tau$. First, let us define

$$r_{1-\alpha}(k) = \inf r : P\left(\sup_{\tau \in \mathcal{B}_\zeta(k)} \frac{G^{(1)}(\tau, k)}{\omega_L(\tau, k)} \leq r\right) \geq 1-\alpha, \tag{42}$$

where $\mathcal{B}_\zeta(k) = \{\tau : (\tau, k) \in \mathcal{B}_\zeta\}$ and $\omega_L$ is a positive weight function chosen by the researcher. For every fixed $k$, we may then construct the function of $\tau$,

$$\hat{\pi}_L(\cdot, k) - \frac{r_{1-\alpha}(k)}{\sqrt{n}} \omega_L(\cdot, k), \tag{43}$$

which lies below $\pi_L(\cdot, k)$ on $\mathcal{B}_\zeta(k)$ with asymptotic probability $1-\alpha$. Hence, (43) provides a one- sided confidence interval for the *process* $\pi_L(\cdot, k)$. The weight function $\omega_L$ allows the researcher to account for the fact that the variance of $G^{(1)}(\tau, k)$ may depend heavily on $(\tau, k)$. Defining

$$k_0^* \equiv \inf k : \sup_{\tau \in \mathcal{B}_\zeta(k)} \hat{\pi}_L(\tau, k) - \frac{r_{1-\alpha}(k)}{\sqrt{n}} \omega_L(\tau, k) \leq 0, \tag{44}$$

it can then be shown that $[k_0^*, 1]$ covers $k_0$ with asymptotic probability at least $1-\alpha$.

EXAMPLE 2.3 (Continued). Employing Theorem 4.1, it is possible to construct a two-sided confidence interval for the function $\kappa_0(\cdot)$. Toward this end, we exploit uniformity in $\tau$ and $k$ by defining

$$r_{1-\alpha} \equiv \inf r : P\left(\sup_{(\tau,k) \in \mathcal{B}_\zeta} \max\left\{\frac{|G^{(1)}(\tau, k)|}{\omega_L(\tau, k)}, \frac{|G^{(2)}(\tau, k)|}{\omega_L(\tau, k)}\right\} \leq r\right) \geq 1-\alpha, \tag{45}$$

where, as in Example 2.2, $\omega_L$ and $\omega_U$ are positive weight functions. In addition, we also let

$$\begin{aligned} \kappa_L^*(\tau) \equiv \inf k : \hat{\pi}_L(\tau, k) - \frac{r_{1-\alpha}}{\sqrt{n}} \omega_L(\tau, k) \leq 0, \quad \text{and} \\ 0 \leq \hat{\pi}_U(\tau, k) + \frac{r_{1-\alpha}}{\sqrt{n}} \omega_U(\tau, k), \end{aligned} \tag{46}$$

$$\kappa_U^*(\tau) \equiv \sup k : \hat{\pi}_L(\tau, k) + \frac{r_{1-\alpha}}{\sqrt{n}} \omega_L(\tau, k) \geq 0, \quad \text{or} \\ 0 \geq \hat{\pi}_U(\tau, k) - \frac{r_{1-\alpha}}{\sqrt{n}} \omega_U(\tau, k). \tag{47}$$

It can then be shown that the functions $(\kappa_L^*(\cdot), \kappa_U^*(\cdot))$ provide a functional confidence interval for $\kappa_0(\cdot)$. That is, $\kappa_L^*(\tau) \leq \kappa_0(\tau) \leq \kappa_U^*(\tau)$ for all $\tau$ with asymptotic probability at least $1 - \alpha$.

Remark 4.4. One could also conduct inference in these examples by employing the sample analogues of $k_0$ (Examples 2.1 and 2.2) or $\kappa_0(\cdot)$ (Example 2.3). While the consistency of such estimators follows directly from Theorem 4.1, their asymptotic distribution and bootstrap consistency require a specialized analysis of the particular definition of "critical $k$" that corresponds to the conjecture under consideration. For this reason, we instead study $\hat{\pi}_L$ and $\hat{\pi}_U$, which, as illustrated by Examples 2.1–2.3, enables us to conduct inference in a wide array of settings.

### 4.2 *Bootstrap critical values*

As illustrated in Examples 2.1–2.3, conducting inference requires use of critical values that depend on the unknown distribution of $G$, the limiting Gaussian process in Theorem 4.1, and possibly on weight functions $\omega_L$ and $\omega_U$ (as in (42) and (45)). We will allow the weight functions $\omega_L$ and $\omega_U$ to be unknown, but require the existence of consistent estimators of them.

Assumption 4.2. (i) *$\omega_L(\tau, k) \geq 0$ and $\omega_U(\tau, k) \geq 0$ are continuous and bounded away from zero on $\mathcal{B}_\zeta$*; (ii) *there exist estimators $\hat{\omega}_L(\tau, k)$ and $\hat{\omega}_U(\tau, k)$ that are uniformly consistent on $\mathcal{B}_\zeta$.*

Given $(\omega_L, \omega_U)$, let $G_\omega$ be the Gaussian process on $L^\infty(\mathcal{B}_\zeta) \times L^\infty(\mathcal{B}_\zeta)$ pointwise defined by

$$G_\omega(\tau, k) \equiv \begin{pmatrix} G^{(1)}(\tau, k)/\omega_L(\tau, k) \\ G^{(2)}(\tau, k)/\omega_U(\tau, k) \end{pmatrix}. \tag{48}$$

The critical values employed in Examples 2.1–2.3 can be expressed in terms of quantiles of some Lipschitz transformation $L : L^\infty(\mathcal{B}_\zeta) \times L^\infty(\mathcal{B}_\zeta) \to \mathbf{R}$ of the random variable $G_\omega$. For instance, in Example 2.2, the relevant critical value, defined in (42), is the $1 - \alpha$ quantile of the random variable

$$L(G_\omega) = \sup_{\tau \in \mathcal{B}_\zeta(k)} G_\omega^{(1)}(\tau, k). \tag{49}$$

Similarly, in Example 2.3 the appropriate critical value defined in (45) is the $1 - \alpha$ quantile of

$$L(G_\omega) = \sup_{(\tau,k) \in \mathcal{B}_\zeta} \max\{G_\omega^{(1)}(\tau, k), G_\omega^{(2)}(\tau, k)\}. \tag{50}$$

We therefore conclude by establishing the validity of a weighted bootstrap procedure for consistently estimating the quantiles of random variables of the form $L(G_\omega)$. The bootstrap procedure is similar to the traditional nonparametric bootstrap with the important difference that the random weights on different observations are independent from each other—a property that simplifies the asymptotic analysis as noted in Ma and Kosorok (2005) and Chen and Pouzo (2009). Specifically, letting $\{W_i\}_{i=1}^n$ be an i.i.d. sample from a random variable $W$, we impose the following assumption.

Assumption 4.3. (i) *$W$ is independent of $(Y, X, D)$, with $W > 0$ almost surely (a.s.), $E[W] = 1$, $\mathrm{Var}(W) = 1$, and $E[|W|^{2+\delta}] < \infty$ for some $\delta > 0$*; (ii) *$L : L^\infty(\mathcal{B}_\zeta) \times L^\infty(\mathcal{B}_\zeta) \to \mathbf{R}$ is Lipschitz continuous.*

A consistent estimator for quantiles of $L(G_\omega)$ may then be obtained through the following algorithm.

Step 1. Generate a sample of i.i.d. weights $\{W_i\}_{i=1}^n$ that satisfy Assumption 4.3(i) and define

$$\begin{aligned}\tilde{Q}_{x,n}(c|\tau, b) \equiv \Bigg(\frac{1}{n}\sum_{i=1}^{n} W_i\big\{1\{Y_i \le c, D_i = 1, X_i = x\} \\ + b1\{D_i = 0, X_i = x\} - \tau 1\{X_i = x\}\big\}\Bigg)^2.\end{aligned} \tag{51}$$

Employing $\tilde{Q}_{x,n}(c|\tau, b)$, obtain the bootstrap estimators for $q_L(\tau, k|x)$ and $q_U(\tau, k|x)$,

$$\begin{aligned}\tilde{q}_L(\tau, k|x) &\in \arg\min_{c\in\mathbf{R}} \tilde{Q}_{x,n}\big(c|\tau, \tau + k\tilde{p}(x)\big), \\ \tilde{q}_U(\tau, k|x) &\in \arg\min_{c\in\mathbf{R}} \tilde{Q}_{x,n}\big(c|\tau, \tau - k\tilde{p}(x)\big),\end{aligned} \tag{52}$$

where $\tilde{p}(x) \equiv (\sum_i W_i 1\{D_i = 1, X_i = x\})/(\sum_i W_i 1\{X_i = x\})$. Note that $\tilde{q}_L(\tau, k|x)$ and $\tilde{q}_U(\tau, k|x)$ are simply the weighted empirical quantiles of the observed data evaluated at a point that depends on the reweighted missing data probability. Note also that if we had used the conventional bootstrap, we would run the risk of drawing a sample for which a covariate bin is empty. This is not a concern with the weighted bootstrap as the weights are required to be strictly positive.

Step 2. Using the bootstrap bounds $\tilde{q}_L(\tau, k|x)$ and $\tilde{q}_U(\tau, k|x)$ from Step 1, obtain the estimators

$$\begin{aligned}\tilde{\pi}_L(\tau, k) \equiv \inf_\theta \lambda'\big(E_S[XX']\big)^{-1} E_S\big[X\theta(X)\big] \\ \text{s.t. } \tilde{q}_L(\tau, k|x) \le \theta(x) \le \tilde{q}_U(\tau, k|x),\end{aligned} \tag{53}$$

$$\begin{aligned}\tilde{\pi}_U(\tau, k) \equiv \sup_\theta \lambda'\big(E_S[XX']\big)^{-1} E_S\big[X\theta(X)\big] \\ \text{s.t. } \tilde{q}_L(\tau, k|x) \le \theta(x) \le \tilde{q}_U(\tau, k|x).\end{aligned} \tag{54}$$

Algorithms for quickly solving linear programming problems of this sort are available in most modern computational packages. The weighted bootstrap process for $G_\omega$ is then defined pointwise by

$$\tilde{G}_\omega(\tau, k) \equiv \sqrt{n}\begin{pmatrix} (\tilde{\pi}_L(\tau, k) - \hat{\pi}_L(\tau, k))/\hat{\omega}_L(\tau, k) \\ (\tilde{\pi}_U(\tau, k) - \hat{\pi}_U(\tau, k))/\hat{\omega}_U(\tau, k) \end{pmatrix}. \tag{55}$$

Step 3. Our estimator for $r_{1-\alpha}$, the $1-\alpha$ quantile of $L(G_\omega)$, is then given by the $1-\alpha$ quantile of $L(\tilde{G}_\omega)$ conditional on the sample $\{Y_i D_i, X_i, D_i\}_{i=1}^n$ (but not $\{W_i\}_{i=1}^n$):

$$\tilde{r}_{1-\alpha} \equiv \inf\{r : P(L(\tilde{G}_\omega) \geq r | \{Y_i D_i, X_i, D_i\}_{i=1}^n) \geq 1-\alpha\}. \tag{56}$$

In applications, $\tilde{r}_{1-\alpha}$ will generally need to be computed through simulation. This can be accomplished by repeating Steps 1 and 2 until the number of bootstrap simulations of $L(\tilde{G}_\omega)$ is large. The estimator $\tilde{r}_{1-\alpha}$ is then well approximated by the empirical $1-\alpha$ quantile of the bootstrap statistic $L(\tilde{G}_\omega)$ across the computed simulations.

We conclude our discussion of inference by establishing that $\tilde{r}_{1-\alpha}$ is indeed consistent for $r_{1-\alpha}$.

Theorem 4.2. *Let $r_{1-\alpha}$ be the $1-\alpha$ quantile of $L(G_\omega)$. If Assumptions 2.1, 4.1, 4.2, and 4.3 hold, the CDF of $L(G_\omega)$ is strictly increasing and continuous at $r_{1-\alpha}$, and $\{Y_i D_i, X_i, D_i, W_i\}_{i=1}^n$ is i.i.d., then*

$$\tilde{r}_{1-\alpha} \xrightarrow{p} r_{1-\alpha}.$$

## 5. Evaluating the U.S. wage structure

We turn now to an assessment of the sensitivity of observed patterns in the U.S. wage structure to deviations from the MAR assumption. A large literature reviewed by (among others) Autor and Katz (1999), Heckman, Lochner, and Todd (2006), and Acemoglu and Autor (2011) documents important changes over time in the conditional distribution of earnings with respect to schooling levels.

In this section, we investigate the sensitivity of these findings to alternative missing data assumptions by revisiting the results of Angrist, Chernozhukov, and Fernández-Val (2006) regarding changes across Decennial Censuses in the quantile-specific returns to schooling. We analyze the 1980, 1990, and 2000 Census samples considered in their study, but to simplify our estimation routine and to correct small mistakes found in the IPUMS (Integrated Public Use Microdata Series) files since the time their extract was created, we use new extracts of the 1% unweighted IPUMS files for each decade rather than their original mix of weighted and unweighted samples. Use of the original extracts analyzed in Angrist, Chernozhukov, and Fernández-Val (2006) yields similar results.

The sample consists of native born black and white men ages 40–49 with 6 or more years of schooling who worked at least one week in the past year. Details are provided in Appendix B. Like Angrist, Chernozhukov, and Fernández-Val (2006), we use average

Table 1. Frequency of missing weekly earnings in census estimation sample by year and cause.

| Census Year | Total Number of Observations | Allocated Earnings | Allocated Weeks Worked | Fraction of Total Missing |
|---|---|---|---|---|
| 1980 | 80,128 | 12,839 | 5,278 | 19.49% |
| 1990 | 111,070 | 17,370 | 11,807 | 23.09% |
| 2000 | 131,265 | 26,540 | 17,455 | 27.70% |
| Total | 322,463 | 56,749 | 34,540 | 23.66% |

weekly earnings as our wage concept, which we measure as the ratio of annual earnings to annual weeks worked. We code weekly earnings as missing for observations with allocated earnings or weeks worked. Observations that fall into demographic cells with less than 20 observations are dropped.[10] The resulting sample sizes and imputation rates for the weekly earnings variable are given in Table 1. As the table makes clear, allocation rates have been increasing across Censuses, with roughly a quarter of the weekly earnings observations missing by 2000. Roughly a third of these allocations result from missing weeks worked information.[11]

Like Angrist, Chernozhukov, and Fernández-Val (2006), we estimate linear conditional quantile models for log earnings per week of the form

$$q(\tau|X, E) = X'\gamma(\tau) + E\beta(\tau), \tag{57}$$

where $X$ consists of an intercept, a black dummy, and a quadratic in potential experience, and $E$ represents years of schooling. Our analysis focuses on the quantile-specific "returns" to a year of schooling $\beta(\tau)$ though we note that, particularly in the context of quantile regressions, these Mincerian earnings coefficients need not map into any proper economic concept of individual returns (Heckman, Lochner, and Todd (2006)). Rather, these coefficients merely provide a parsimonious summary of the within and between group inequality in wages that has been a focus of this literature.

### 5.1 *Analyzing the median*

Before revisiting the main results of Angrist, Chernozhukov, and Fernández-Val (2006), we illustrate the methods developed so far by analyzing the median wages of the 227 demographic cells in our 1990 sample. We begin by considering the worst case nonparametric bounds on these medians. Because the covariates are of dimension 3, the identified set is difficult to visualize directly. Figure 4 reports the upper and lower bounds for two experience groups of white men as a function of their years of schooling. The bounds were obtained using $\hat{q}_L(0.5, 1|x)$ and $\hat{q}_U(0.5, 1|x)$, which are the sample ana-

[10] Demographic cells are defined by the intersection of single digit age, race (black vs. white), and years of schooling.

[11] It is interesting to note that only 7% of the men in our sample report working no weeks in the past year. Hence, at least for this population of men, assumptions regarding the determinants of nonresponse appear to be more important for drawing conclusions regarding the wage structure than assumptions regarding nonparticipation in the labor force.

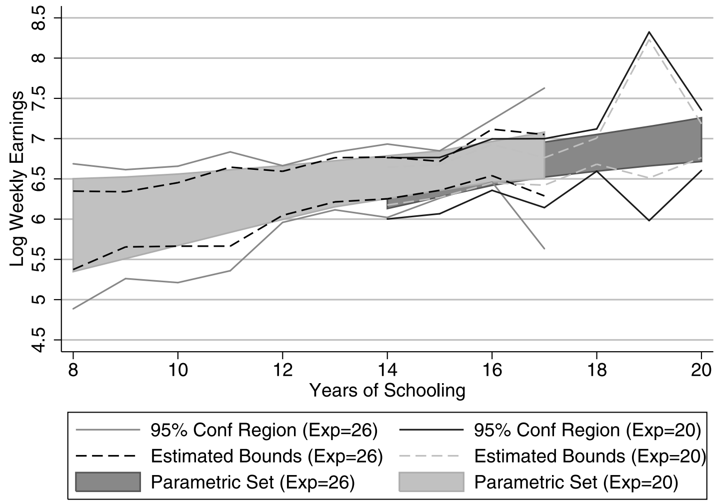


FIGURE 4. Worst case nonparametric bounds on 1990 medians and linear model fits for two experience groups of white men.

logues to the quantiles in Lemma 2.1 for the case when $k = 1$. We also report confidence regions containing the conditional median function with asymptotic probability of 95%.[12] Finally, we show the envelope of parametric Mincer fits that lie within the estimated confidence region.

The estimated worst case bounds on the conditional median are quite wide, with a range of roughly 100 log points for high school dropouts. Accounting for sampling uncertainty widens these bounds substantially despite our use of large Census samples. Unsurprisingly, a wide range of Mincer models fit within the confidence region, with the associated parametric returns to schooling spanning the interval $[1.5\%, 16.3\%]$. Moreover, the set of parametric models in the confidence region clearly overstates our knowledge of the true conditional median function relative to the nonparametric confidence region.

Figure 5 reports the nonparametric bounds and their associated 95% confidence region when allowing for a small amount of nonrandom selection via the nominal restriction that $\mathcal{S}(F) \leq 0.05$. As discussed in Section 2.2, this restriction would be satisfied if 95% of the missing data were missing at random. Sampling uncertainty is relatively more important here than before, as the sample bounds now imply a very narrow identified set. Even after accounting for uncertainty, however, the irregular shape of the bounds prohibits use of a linear model. Formally, our inability to find a linear model that obeys the bounds for the conditional median implies that the Mincer specification may be rejected at the 5% level despite the model being partially identified. Nevertheless, the conditional median function still appears to be approximately linear in schooling.

[12]These regions were obtained by bootstrapping the covariance matrix of upper and lower bounds for each $x \in \mathcal{X}$, where $\mathcal{X}$ is the set of all 227 demographic cells. We exploit independence across $x$ to find a critical value that delivers coverage of the conditional median function with asymptotic probability of 0.95. See Appendix C for details.

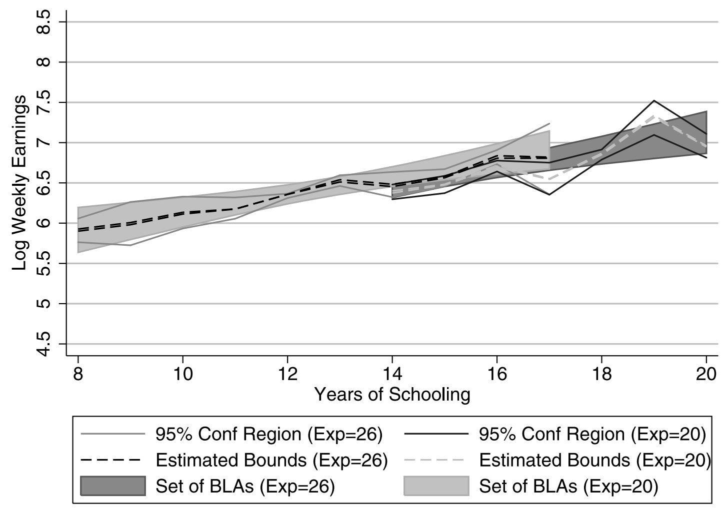


Figure 5. Nonparametric bounds on 1990 medians and best linear approximations for two experience groups of white men under $\mathcal{S}(F) \leq 0.05$.

Were the data known to be missing at random, so that the median was point identified, we would summarize the relationship between schooling and earnings using an approximate parametric model as in Chamberlain (1994) or Angrist, Chernozhukov, and Fernández-Val (2006). As we saw in Section 3, lack of identification presents no essential obstacle to such an exercise.

The shaded regions of Figure 5 report the set of best linear approximations to the set of conditional medians lying within the confidence region obtained under $\mathcal{S}(F) \leq 0.05$.[13] Note that this set provides a reasonably accurate summary of the nonparametric confidence region. The approximate returns to schooling coefficients associated with this set lie in the interval [0.058, 0.163]. Much of this rather wide range results from sampling uncertainty. Using the methods of Section 4.2, we can reduce this uncertainty by constructing a confidence interval for the schooling coefficient $\beta(0.5)$ directly rather than inferring one from the confidence region for the entire nonparametric identified set. Doing so yields a relatively narrow interval for the approximate returns to schooling of [0.102, 0.118].[14] Thus, in our setting, switching to an explicit approximating model not only avoids an inappropriate narrowing of the bounds due to misspecification, but allows for substantial improvements in precision.

### 5.2 *A replication*

We turn now to a replication of the main results in Angrist, Chernozhukov, and Fernández-Val (2006) concerning changes across Censuses in the structure of wages

[13]As in the next section, we weight the squared prediction errors in each demographic bin by sample size when defining the best linear predictor.

[14]We employed the bootstrap procedure of Section 4.2 to obtain estimators of the asymptotic 95% quantiles of $\sqrt{n}(\hat{\pi}_U(0.5, 0.5) - \pi_U(0.5, 0.5))$ and $\sqrt{n}(\pi_L(0.5, 0.5) - \hat{\pi}_L(0.5, 0.5))$, which we denote by $\hat{c}_U$ and $\hat{c}_L$, respectively. The confidence interval reported is then $[\hat{\pi}_L(0.5, 0.05) - \hat{c}_L/\sqrt{n}, \hat{\pi}_U(0.5, 0.05) + \hat{c}_U/\sqrt{n}]$.

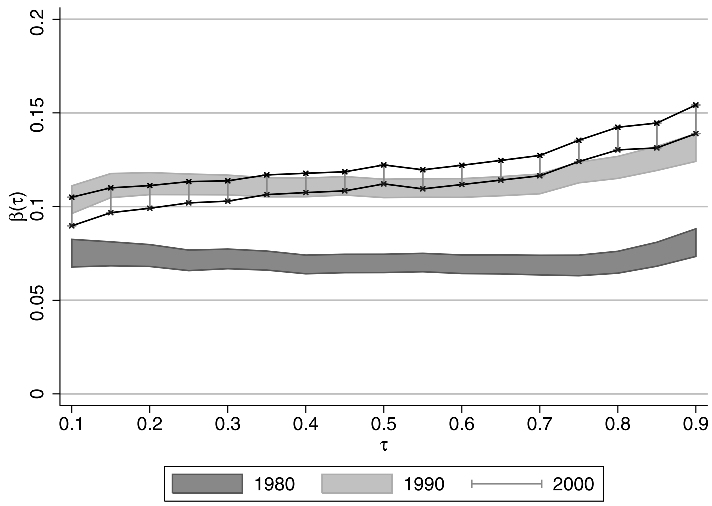


FIGURE 6. Uniform confidence regions for schooling coefficients by quantile and year under missing at random assumption ($\mathcal{S}(F) = 0$). *Note*: Model coefficients provide minimum mean square approximation to the true conditional quantile function.

under the assumption that the data are missing at random. This is accomplished by applying the methods of Section 4 subject to the restriction that $\mathcal{S}(F) = 0$. Details of our algorithm are described in Appendix C. To ensure comparability with Angrist, Chernozhukov, and Fernández-Val (2006), we define our approximation metric as weighting the errors in each demographic bin by sample size (i.e., we choose $S$ equal to empirical measure; see Section 3).[15] Notably, with the MAR restriction, our estimation procedure is equivalent to the classical minimum distance estimator studied by Chamberlain (1994).

Figure 6 plots estimates of the approximate returns functions $\beta(\cdot)$ in 1980, 1990, and 2000 along with uniform confidence intervals. Our MAR results are similar to those found in Figure 2A of Angrist, Chernozhukov, and Fernández-Val (2006). They suggest that the returns function increased uniformly across quantiles between 1980 and 1990, but exhibited a change in slope in 2000. The change between 1980 and 1990 is consistent with a general economy-wide increase in the return to human capital accumulation as conjectured by Juhn, Murphy, and Pierce (1993). However, the finding of a shape change in the quantile process between 1990 and 2000 indicates that skilled workers experienced increases in inequality relative to their less skilled counterparts, a pattern that appears not to have been present in previous decades. This pattern of heteroscedasticity is consistent with recently proposed multifactor models of technical change reviewed in Acemoglu and Autor (2011).

### 5.3 *Sensitivity analysis*

A natural concern is the extent to which some or all of the conclusions regarding the wage structure drawn under a missing at random assumption are compromised by lim-

[15]We also performed the exercises in this section by weighting the set of demographic groups present in all three decades equally and found similar results.

itations in the quality of Census earnings data. As Table 1 shows, the prevalence of earnings imputations increases steadily across Censuses, with roughly a quarter of the observations allocated by 2000. With these levels of missingness, quantiles below the 25th percentile and above the 75th become unbounded in the absence of restrictions on the missing data process.

We now examine the bounds on the schooling coefficients governing our approximating model that result in each year when we allow for families of deviations from MAR indexed by different values of $\mathcal{S}(F)$. These upper and lower bounds may then be compared across years to assess the sensitivity of conclusions regarding changes in the wage structure to violations of MAR. Of course, it is possible for substantial deviations from MAR to be present in each year but for the nature of those deviations to be stable across time. Likewise, in a single cross section, each schooling group may violate ignorability, but those violations may be similar across adjacent groups. If such prior information is available, the bounds on changes in the quantile-specific returns to schooling and their level may be narrowed. While it is, in principle, possible to add a second dimension of sensitivity to capture changes in the selection mechanism across time or demographic groups, we leave such extensions for future work, as they would complicate the analysis considerably. We simply note that if conclusions regarding changes across Censuses are found to be robust to large unrestricted deviations from MAR, adding additional restrictions will not change this assessment.

Figure 7 provides 95% uniform confidence regions for the set $\mathcal{G}(k)$ of coefficients governing the BLA, as defined in (26), that result when we allow for a small amount of selection by setting $\mathcal{S}(F) \leq 0.05$. Though it remains clear that the schooling coefficients increased between 1980 and 1990, we cannot reject the null hypothesis that the quantile process was unchanged from 1990 to 2000. Moreover, there is little evidence of hetero-

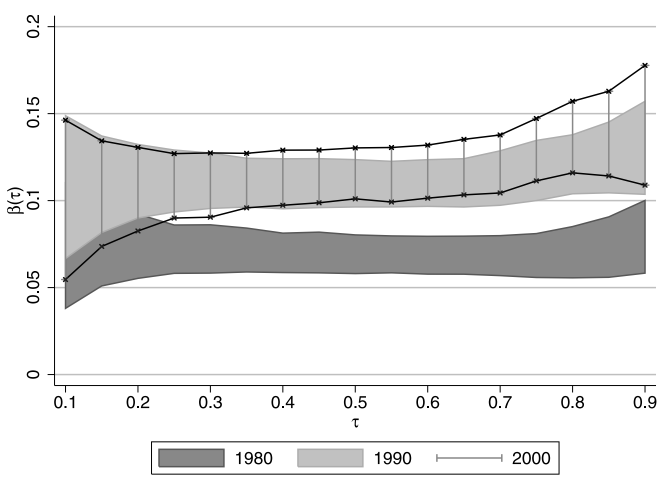


FIGURE 7. Uniform confidence regions for schooling coefficients by quantile and year under $\mathcal{S}(F) \leq 0.05$. *Note*: Model coefficients provide a minimum mean square approximation to the true conditional quantile function.

geneity across quantiles in any of the three Census samples—a straight line can be fit through each sample's confidence region.

To further assess the robustness of our conclusions regarding changes between 1980 and 1990, it is informative to find the level of $k$ necessary to fail to reject the hypothesis that no change in fact occurred between these years under the restriction that $\mathcal{S}(F) \leq k$. Specifically, for $\pi_L^t(\tau, k)$ and $\pi_U^t(\tau, k)$, the lower and upper bounds on the schooling coefficients in year $t$, we aim to obtain a confidence interval for the values of selection $k$ under which

$$\pi_U^{80}(\tau, k) \geq \pi_L^{90}(\tau, k) \quad \text{for all } \tau \in [0.2, 0.8]. \tag{58}$$

As in Example 2.2, we are particularly interested in $k_0$, the smallest value of $k$ such that (58) holds, as it will hold trivially for all $k \geq k_0$. A search for the smallest value of $k$ such that the 95% uniform confidence intervals for these two decades overlap at all quantiles between 0.2 and 0.8 found this "critical $k$" to be $k_0^* = 0.175$. Due to the independence of the samples between 1980 and 1990, the one-sided interval $[k_0^*, 1]$ provides an asymptotic coverage probability for $k_0$ of at least 90%. The lower end of this confidence interval constitutes a large deviation from MAR, indicating that the evidence is quite strong that the schooling coefficient process changed between 1980 and 1990. Figure 8 plots the uniform confidence regions that correspond to the hypothetical $\mathcal{S}(F) \leq k_0^*$.

Though severe selection would be necessary for all of the changes between 1980 and 1990 to be spurious, it is clear that changes at some quantiles may be more robust than others. It is interesting then to conduct a more detailed analysis by evaluating the critical level of selection necessary to undermine the conclusion that the schooling coefficient increased at each quantile. Toward this end, we generalize Example 2.3 and define $\kappa_0(\tau)$ to be the smallest level of $k$ such that

$$\pi_U^{80}(\tau, k) \geq \pi_L^{90}(\tau, k). \tag{59}$$

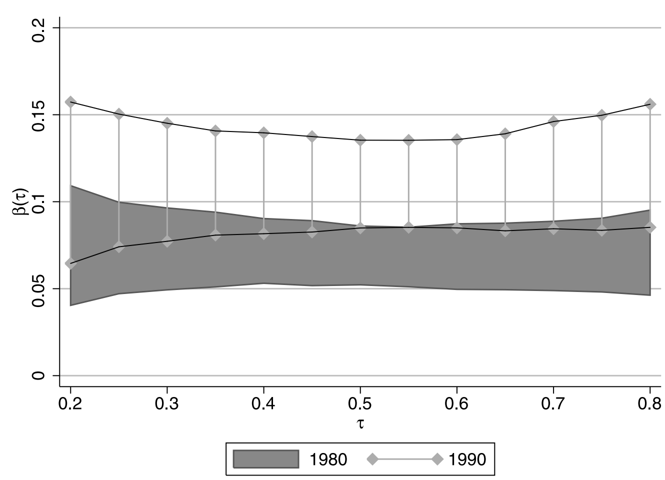


FIGURE 8. Uniform confidence regions for schooling coefficients by quantile and year under $\mathcal{S}(F) \leq 0.175$ (1980 vs. 1990). *Note*: Model coefficients provide a minimum mean square approximation to the true conditional quantile function as in Chamberlain (1994).

The function $\kappa_0(\cdot)$ summarizes the level of robustness of each quantile-specific conclusion. In this manner, the breakdown function $\kappa_0(\cdot)$ reveals the differential sensitivity of the entire conditional distribution to violations of the missing at random assumption.

The point estimate for $\kappa_0(\tau)$ is given by the value of $k$ where $\hat{\pi}_U^{80}(\tau, k)$ intersects with $\hat{\pi}_L^{90}(\tau, k)$. To obtain a confidence interval for $\kappa_0(\tau)$ that is uniform in $\tau$, we first construct 95% uniform two-sided confidence intervals in $\tau$ and $k$ for the 1980 upper bound $\pi_U^{80}(\tau, k)$ and the 1990 lower bound $\pi_L^{90}(\tau, k)$. Given the independence of the 1980 and 1990 samples, the intersection of the true bounds $\pi_U^{80}(\tau, k)$ and $\pi_L^{90}(\tau, k)$ must lie between the intersection of their corresponding confidence regions with asymptotic probability of at least 90%. Since $\kappa_0(\tau)$ is given by the intersection of $\pi_U^{80}(\tau, k)$ with $\pi_L^{90}(\tau, k)$, a valid lower bound for the confidence region of the function $\kappa_0(\cdot)$ is given by the intersection of the upper envelope for $\pi_U^{80}(\tau, k)$ with the lower envelope for $\pi_L^{90}(\tau, k)$ and a valid upper bound is given by the converse intersection.

Figure 9 illustrates the resulting estimates of the breakdown function $\kappa_0(\cdot)$ and its corresponding confidence region. Unsurprisingly, the most robust results are those for quantiles near the center of the distribution for which very large levels of selection would be necessary to overturn the hypothesis that the schooling coefficient increased. However, the curve is fairly asymmetric, with the conclusions at low quantiles being much more sensitive to deviations from ignorability than those at the upper quantiles. Hence,

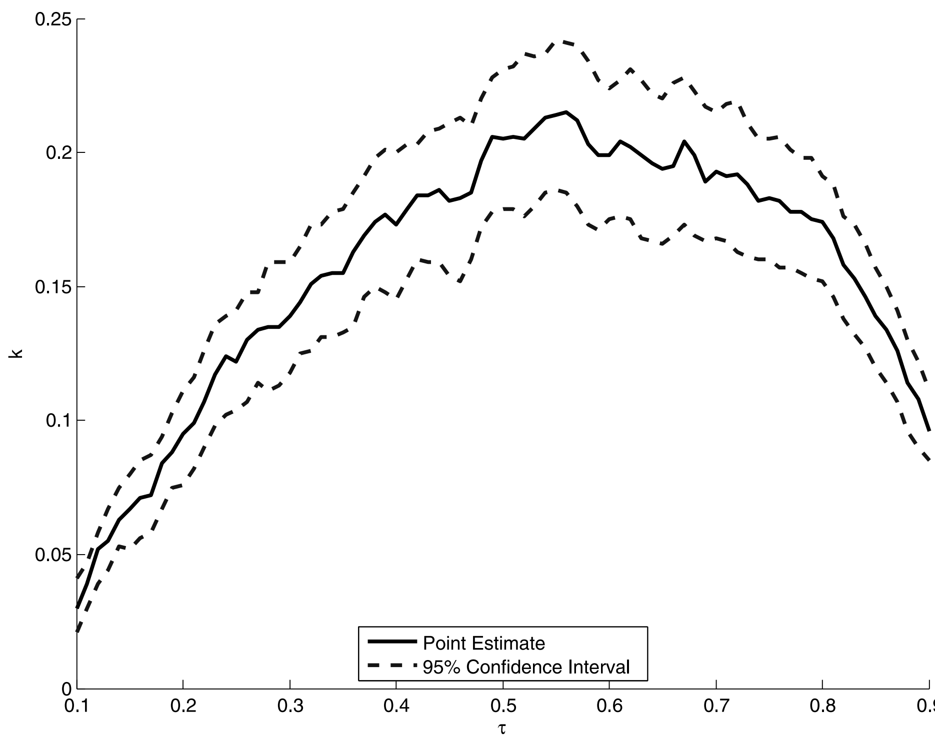


FIGURE 9. Breakdown curve (1980 vs. 1990). *Note*: Each point on this curve indicates the minimal level of $\mathcal{S}(F)$ necessary to undermine the conclusion that the schooling coefficient increased between 1980 and 1990 at the quantile of interest.

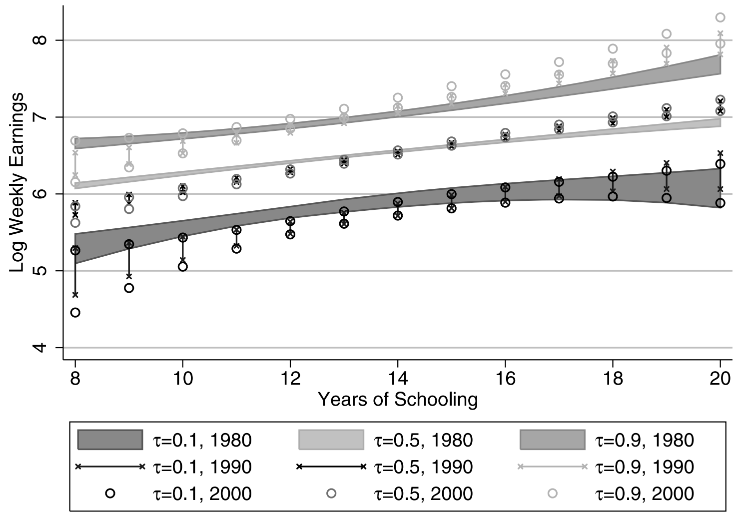


FIGURE 10. Confidence intervals for fitted values under $\mathcal{S}(F) \leq 0.05$. *Note*: Earnings quantiles are modeled using the quadratic specification in Lemieux (2006). Model coefficients provide a minimum mean square approximation to the true conditional quantile function. Covariates other than education are set to the sample mean.

changes in reporting behavior between 1980 and 1990 pose the greatest threat to hypotheses regarding changes at the bottom quantiles of the earnings distribution.

To conclude our sensitivity analysis we also consider the fitted values that result from the more flexible earnings model of Lemieux (2006), which allows for quadratic effects of education on earnings quantiles.[16] Figure 10 provides bounds on the 10th, 50th, and 90th conditional quantiles of weekly earnings by schooling level in 1980, 1990, and 2000 using our baseline hypothetical restriction $\mathcal{S}(F) \leq 0.05$. Little evidence exists of a change across Censuses in the real earnings of workers at the 10th conditional quantile. At the conditional median, however, the slope of the relationship with schooling (which appears roughly linear) increased substantially, leading to an increase in inequality across schooling categories. Uneducated workers witnessed wage losses while skilled workers experienced wage gains, though in both cases these changes seem to have occurred entirely during the 1980s. Finally, we also note that, as observed by Lemieux (2006), the schooling locus appears to have gradually convexified at the upper tail of the weekly earnings distribution, with very well educated workers experiencing substantial gains relative to the less educated.

### 5.4 *Estimates of the degree of selection in earnings data*

Our analysis of Census data revealed that the finding of a change in the quantile-specific schooling coefficients between 1990 and 2000 is easily undermined by small amounts of

[16]The model also includes a quartic in potential experience. Our results differ substantively from those of Lemieux, both because of differences in sample selection and our focus on weekly (rather than hourly) earnings.

selection, while changes between 1980 and 1990 (at least above the lower quantiles of the distribution) appear to be relatively robust. Employing a sample where validation data are present, we now turn to an investigation of what levels of selection, as indexed by $\mathcal{S}(F)$, are plausible in U.S. survey data.

To estimate $\mathcal{S}(F)$, we first derive an alternative representation of the distance between $F_{y|0,x}$ and $F_{y|1,x}$ that illustrates its dependence on the conditional probability of the outcome being missing. Toward this end, let us define the conditional probabilities

$$p_L(x,\tau) \equiv P\big(D=1|X=x, F_{y|x}(Y) \le \tau\big), \tag{60}$$

$$p_U(x,\tau) \equiv P\big(D=1|X=x, F_{y|x}(Y) > \tau\big). \tag{61}$$

By applying Bayes' rule, it is then possible to express the distance between the distribution of missing and nonmissing observations at a given quantile as a function of the selection probabilities[17]

$$\begin{aligned} &\big|F_{y|1,x}\big(q(\tau|x)\big) - F_{y|0,x}\big(q(\tau|x)\big)\big| \\ &\quad = \frac{\sqrt{(p_L(x,\tau)-p(x))(p_U(x,\tau)-p(x))\tau(1-\tau)}}{p(x)(1-p(x))}. \end{aligned} \tag{62}$$

Notice that knowledge of the missing probability $P(D=0|X=x, F_{y|x}(Y)=\tau)$ is sufficient to compute by integration all of the quantities in (62) and (by taking the supremum over $\tau$ and $x$) of $\mathcal{S}(F)$ as well.[18] For this reason, our efforts focus on estimating this function in a data set with information on the earnings of survey nonrespondents.

We work with an extract from the 1973 March Current Population Survey (CPS) for which merged Internal Revenue Service (IRS) earnings data are available. Because we only have access to a single cross section of validation data, our analysis will of necessity be confined to determination of plausible levels of $\mathcal{S}(F)$ in a given year rather than changes in the nature of selection across years. Moreover, the CPS data contain far fewer observations than our earlier Census extracts. To ensure reasonably precise estimates, we broaden our sample selection criteria to include additional age groups. Specifically, our sample consists of black and white men between the ages of 25 and 55 with 5 or more years of schooling who reported working at least one week in the past year and had valid IRS earnings. We drop observations with annual IRS earnings less than \$1,000 or equal to the IRS top code of \$50,000. Following Bound and Krueger (1991), we also drop men employed in agriculture, forestry, and fishing or in occupations likely to receive tips. Finally, because self-employment income may be underreported to the IRS, we drop individuals who identify themselves as self-employed to the CPS. Further details are provided in Appendix B.

[17] See Appendix B for a detailed derivation of (62).

[18] Note that $P(D=0, F_{y|x}(Y) \le \tau | X=x) = \int_0^\tau P(D=0|F_{y|x}(Y)=u, X=x)\,du$ because $F_{y|x}(Y)$ is uniformly distributed on $[0,1]$ conditional on $X=x$. Thus $p_L(x,\tau) = \int_0^\tau P(D=0|F_{y|x}(Y)=u, X=x)\,du/\tau$. Likewise $p_U(x,\tau) = \int_\tau^1 P(D=0|F_{y|x}(Y)=u, X=x)\,du/(1-\tau)$ and $p(x) = \int_0^1 P(D=0|F_{y|x}(Y)=u, X=x)\,du$.

As in our study of the Decennial Census, we take the relevant covariates to be age, years of schooling, and race. However, because our CPS sample is much smaller than our Census sample, we coarsen our covariate categories and drop demographic cells with fewer than 50 observations.[19] This yields an estimation sample of 15,027 observations distributed across 35 demographic cells.

For comparability with our analysis of Census data, we again take average weekly earnings as our wage concept. Because we lack an administrative measure of weeks worked, we construct our wage metric by dividing the IRS based measure of annual wage and salary earnings by the CPS based measure of weeks worked. Observations with allocated weeks information are dropped.[20] As a result, we are only able to examine biases generated by earnings nonresponse.

We take the log of annual IRS earnings divided by weeks worked as our measure of $Y$ and use response to the March CPS annual civilian earnings question as our measure of $D$. This yields a missing data rate of 7.2%. We approximate the probability of nonresponse with the sequence of increasingly flexible logistic models

$$P\big(D=0|X=x, F_{y|x}(Y)=\tau\big)=\Lambda\big(b_1\tau+b_2\tau^2+\delta_x\big), \tag{M1}$$

$$P\big(D=0|X=x, F_{y|x}(Y)=\tau\big)=\Lambda\big(b_1\tau+b_2\tau^2+\gamma_1\delta_x\tau+\gamma_2\delta_x\tau^2+\delta_x\big), \tag{M2}$$

$$P\big(D=0|X=x, F_{y|x}(Y)=\tau\big)=\Lambda\big(b_{1,x}\tau+b_{2,x}\tau^2+\delta_x\big), \tag{M3}$$

where $\Lambda(\cdot)=\exp(\cdot)/(1+\exp(\cdot))$ is the logistic CDF. These models differ primarily in the degree of demographic bin heterogeneity allowed for in the relationship between earnings and the probability of responding to the CPS. Model (M1) relies entirely on the nonlinearities in the index function $\Lambda(\cdot)$ to capture heterogeneity across cells in the response profiles. The model (M2) allows for additional heterogeneity through the interaction coefficients $(\gamma_1, \gamma_2)$ but restricts these interactions to be linear in the cell effects $\delta_x$. Finally, (M3), which is equivalent to a cell specific version of (M1), places no restrictions across demographic groups on the shape of the response profile.

Maximum likelihood estimates from the three models are presented in Table 2.[21] A comparison of the model log likelihoods reveals that the introduction of the interaction terms $(\gamma_1, \gamma_2)$ in Model 2 yields a substantial improvement in fit over the basic separable logit of Model 1 despite the insignificance of the resulting parameter estimates. However, the restrictions of the linearly interacted Model 2 cannot, at conventional significance levels, be rejected relative to its fully interacted generalization in Model 3, which appears to be somewhat overfit.

[19] We use 5-year age categories instead of single digit ages and collapse years of schooling into four categories: <12 years of schooling, 12 years of schooling, 13–15 years of schooling, and 16+ years of schooling. Our more stringent requirement that cells have 50 observations is motivated by our desire to accurately estimate $\mathcal{S}(F)$ while allowing for rich forms of heterogeneity across demographic groups.

[20] Weeks allocations are less common in the 1973 CPS than the Census, comprising roughly 20% of all allocations.

[21] We use the respondent's sample quantile in his demographic cell's distribution of $Y$ as an estimate of $F_{y|x}(Y)$. It can be shown that sampling errors in the estimated quantiles have asymptotically negligible effects on the limiting distribution of the parameter estimates.

TABLE 2. Logit estimates of $P(D=0|X=x, F_{y|x}(Y)=\tau)$ in the 1973 CPS–IRS sample.

| | Model 1 | Model 2 | Model 3 |
|---|---|---|---|
| $b_1$ | −1.06 | 0.05 | |
| | (0.43) | (5.44) | |
| $b_2$ | 1.09 | 3.75 | |
| | (0.41) | (4.08) | |
| $\gamma_1$ | | 0.45 | |
| | | (2.30) | |
| $\gamma_2$ | | 1.15 | |
| | | (1.73) | |
| Log likelihood | −3,802.91 | −3,798.48 | −3,759.97 |
| Parameters | 37 | 39 | 105 |
| Number of observations | 15,027 | 15,027 | 15,027 |
| Demographic cells | 35 | 35 | 35 |
| $E[\frac{\partial P(D=0\|X=x,F_{y\|x}(Y)=\tau)}{\partial \tau}\|_{\tau=0.2}]$ | −0.04 | −0.04 | −0.03 |
| $E[\frac{\partial P(D=0\|X=x,F_{y\|x}(Y)=\tau)}{\partial \tau}\|_{\tau=0.5}]$ | 0.00 | 0.00 | −0.01 |
| $E[\frac{\partial P(D=0\|X=x,F_{y\|x}(Y)=\tau)}{\partial \tau}\|_{\tau=0.8}]$ | 0.05 | 0.04 | 0.02 |
| Min KS distance | 0.02 | 0.02 | 0.01 |
| Median KS distance | 0.02 | 0.05 | 0.12 |
| Max KS distance ($\mathcal{S}(F)$) | 0.02 | 0.17 | 0.67 |
| *Ages 40–49* | | | |
| Min KS distance | 0.02 | 0.02 | 0.01 |
| Median KS distance | 0.02 | 0.05 | 0.08 |
| Max KS distance ($\mathcal{S}(F)$) | 0.02 | 0.09 | 0.39 |

*Note*: Asymptotic standard errors are given in parentheses.

A Wald test of joint significance of the earned income terms $(b_1, b_2)$ in the first model rejects the null hypothesis that the data are missing at random with a $p$-value of 0.03. Evidently, missing data probabilities follow a U-shaped response pattern, with very low and very high wage men least likely to provide valid earnings information—a pattern conjectured (but not directly verified) by Lillard, Smith, and Welch (1986). This pattern is also found in the two more flexible logit models as illustrated in the third panel of the table, which provides the average marginal effects of earnings evaluated at three quantiles of the distribution. These average effects are consistently negative at $\tau = 0.2$ and positive at $\tau = 0.8$. It is important to note, however, that Models 2 and 3 allow for substantial heterogeneity across covariate bins in these marginal effects that in some cases yields response patterns that are monotonic rather than U-shaped.

It is straightforward to estimate the distance between the missing and nonmissing earnings distributions in each demographic cell by integrating our estimates of $P(D=0|X=x, F_{y|x}(Y)=\tau)$ across the relevant quantiles of interest. We implement this integration numerically via one dimensional Simpson quadrature. The third panel of Table 2 shows quantiles of the distribution of resulting cell-specific KS distance estimates. Model 1 is nearly devoid of heterogeneity in KS distances across demographic

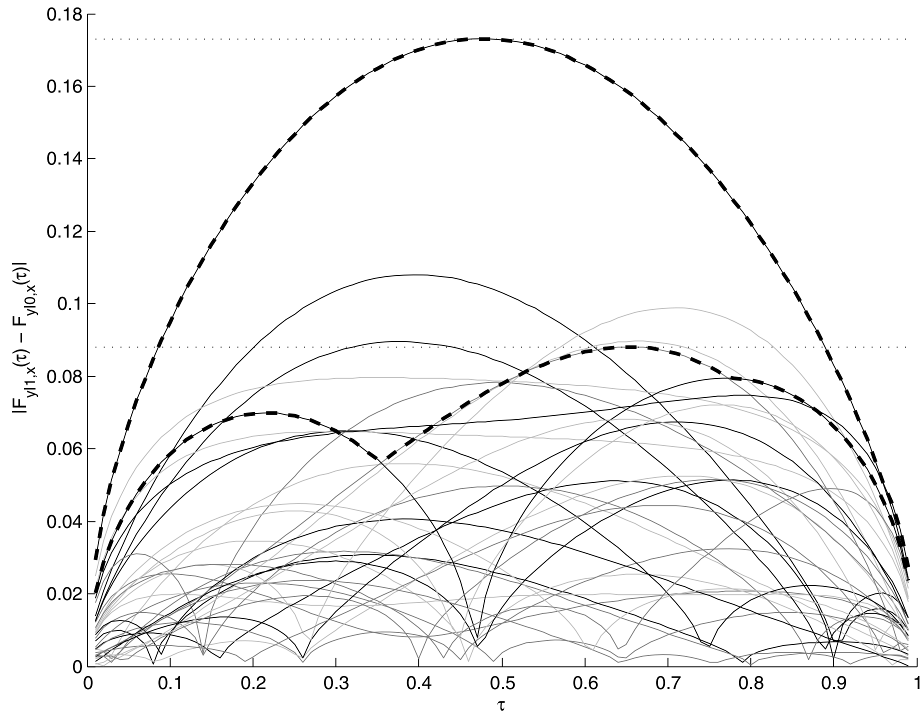


FIGURE 11. Logit based estimates of distance between missing and nonmissing CDFs by quantile of IRS earnings and demographic cell.

cells because of the additive separability implicit in the model. Model 2 yields substantially more heterogeneity, with a minimum KS distance of 0.02 and a maximum distance $\mathcal{S}(F)$ of 0.17. Finally, Model 3, which we suspect has been overfit, yields a median KS distance of 0.12 and an enormous maximum KS distance of 0.67. For comparability with our earlier Census analysis, the bottom panel of Table 2 provides equivalent figures among men ages 40–49. These age groups exhibit somewhat smaller estimates of $\mathcal{S}(F)$, with maximum KS distances of 0.09 and 0.39 in Models 2 and 3, respectively.

Figure 11 provides a visual representation of our estimates from Model 2 of the underlying distance functions $|F_{y|1,x}(q(\tau|x)) - F_{y|0,x}(q(\tau|x))|$ in each of the 35 demographic cells in our sample. The upper envelope of these functions corresponds to the quantile-specific level of selection considered in the breakdown analysis of Figure 9, while the maximum point on the envelope corresponds to $\mathcal{S}(F)$. Note that while some of the distance functions exhibit an unbroken inverted U-shaped pattern, others exhibit double or even triple arches. The pattern of multiple arches occurs when the CDFs are estimated to have crossed at some quantile, which yields a distance of zero at that point. A quadratic relationship between missing data probabilities and earnings can easily yield such patterns. Because of the interactions in Model 2, some cells exhibit effects that are not quadratic and tend to generate CDFs that exhibit first order stochastic dominance. It is interesting to note that the demographic cell that obtains the maximum KS distance of 0.17 corresponds to young (age 25–30), black, high school dropouts for whom more IRS earnings are estimated to monotonically increase the probability of responding to the CPS earnings question. This leads to a distribution of observed earnings that stochastically dominates that of the corresponding unobserved earnings.

The upper envelope of distance functions among men ages 40–49 is also illustrated in Figure 11 and spans three demographic cells. The maximum KS distance in this group of 0.09 is obtained by 45–49-year-old white men with a college degree. These estimates, when compared to the breakdown function of Figure 9, reinforce our earlier conclusion that most of the apparent changes in wage structure between 1980 and 1990 are robust to plausible violations of MAR, but that conclusions regarding lower quantiles could potentially be overturned by selective nonresponse. Likewise, the apparent emergence of heterogeneity in the returns function in 2000 may easily be justified by selection of the magnitude found in our CPS sample. Though our estimates of selection are fairly sensitive to the manner in which cell-specific heterogeneity is modeled, we take the patterns in Table 2 and Figure 11 as suggestive evidence that small, but by no means negligible, deviations from missing at random are likely present in modern earnings data. These deviations may yield complicated discrepancies between observed and missing CDFs about which it is hard to develop strong priors. We leave it to future research to examine these issues more carefully with additional validation data sets.

## 6. Conclusion

We have proposed assessing the sensitivity of estimates of conditional quantile functions with missing outcome data to violations of the MAR assumption by considering the minimum level of selection, as indexed by the maximal KS distance between the distribution of missing and nonmissing outcomes across all covariate values, necessary to overturn conclusions of interest. Inferential methods were developed that account for uncertainty in estimation of the nominal identified set and that acknowledge the potential for model misspecification. We found in an analysis of U.S. Census data that the well documented increase in the returns to schooling between 1980 and 1990 is relatively robust to alternative assumptions on the missing process, but that conclusions regarding heterogeneity in returns and changes in the returns function between 1990 and 2000 are very sensitive to departures from ignorability.

## References

Acemoglu, D. and D. H. Autor (2011), "Skills, tasks, and technology: Implications for employment and earnings." In *Handbook of Labor Economics*, Vol. 4B, 1043–1171, Elsevier, Amsterdam. [252, 256]

Altonji, J. G., T. Elder, and C. Taber (2005), "Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools." *Journal of Political Economy*, 113, 151–184. [233, 235]

Altonji, J. G., T. Elder, and C. Taber (2008), "Using selection on observed variables to assess bias from unobservables when evaluating Swan–Ganz catheterization." *American Economic Review*, 98, 345–350. [233]

Angrist, J., V. Chernozhukov, and I. Fernández-Val (2006), "Quantile regression under misspecification, with an application to the U.S. wage structure." *Econometrica*, 74, 539–563. [231, 233, 241, 252, 253, 255, 256]

Autor, D. H. and L. F. Katz (1999), "Changes in the wage structure and earning inequality." In *Handbook of Labor Economics*, Vol. 3A, 1463–1555, North-Holland, Amsterdam. [252]

Bound, J. and A. Krueger (1991), "The extent of measurement error in longitudinal earnings data: Do two wrongs make a right?" *Journal of Labor Economics*, 9, 1–24. [261]

Chamberlain, G. (1994), "Quantile regression, censoring and the structure of wages." In *Advances in Econometrics, Sixth World Congress* (C. Sims, ed.), 171–209, Cambridge University Press, Cambridge. [233, 241, 255, 256, 258]

Chen, X., O. Linton, and I. V. Keilegom (2003), "Estimation of semiparametric models when the criterion function is not smooth." *Econometrica*, 71, 1591–1608. [248]

Chen, X. and D. Pouzo (2009), "Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals." *Journal of Econometrics*, 152, 46–60. [251]

DiNardo, J., J. McCrary, and L. Sanbonmatsu (2006), "Constructive proposals for dealing with attrition: An empirical example." Working paper, University of Michigan. [232]

Galambos, J. (1973), "A general Poisson limit theorem of probability theory." *Duke Mathematical Journal*, 40, 581–586. [248]

Greenlees, J. S., W. S. Reece, and K. D. Zieschang (1982), "Imputations of missing values when the probability of response depends on the variable being imputed." *Journal of the American Statistical Association*, 77, 251–261. [232]

Heckman, J. J., L. Lochner, and P. Todd (2006), "Earnings functions, rates of return and treatment effects: The Mincer equation and beyond." In *Handbook of Education Economics*, Vol. 1 (E. Hanushek and F. Welch, eds.), Chap. 7, Elsevier, Amsterdam. [252, 253]

Hirsch, B. T. and E. J. Schumacher (2004), "Match bias in wage gap estimates due to earnings imputation." *Journal of Labor Economics*, 22, 689–722. [232]

Horowitz, J. L. and C. F. Manski (1995), "Identification and robustness with contaminated and corrupted data." *Econometrica*, 63, 281–302. [237]

Horowitz, J. L. and C. F. Manski (2006), "Identification and estimation of statistical functionals using incomplete data." *Journal of Econometrics*, 132, 445–459. [232, 241]

Huber, P. J. (1964), "Robust estimation of a location parameter." *Annals of Mathematical Statistics*, 35, 73–101. [236]

Imbens, G. W. (2003), "Sensitivity to exogeneity assumptions in program evaluation." *American Economic Review*, 93, 126–132. [233]

Juhn, C., K. M. Murphy, and B. Pierce (1993), "Wage inequality and the rise in the return to skill." *Journal of Political Economy*, 101, 410–442. [256]

Lemieux, T. (2006), "Postsecondary education and increasing wage inequality." *American Economic Review*, 96, 195–199. [234, 260]

Lillard, L., J. P. Smith, and F. Welch (1986), "What do we really know about wages? The importance of nonreporting and census imputation." *Journal of Political Economy*, 94, 489–506. [232, 263]

Ma, S. and M. R. Kosorok (2005), "Robust semiparametric M-estimation and the weighted bootstrap." *Journal of Multivariate Analysis*, 96, 190–217. [251]

Manski, C. F. (1994), "The selection problem." In *Advances in Econometrics, Sixth World Congress* (C. Sims, ed.), 143–170, Cambridge University Press, Cambridge. [232, 233, 237]

Manski, C. F. (2003), *Partial Identification of Probability Distributions*. Springer, New York. [232, 237]

Newey, W. K. (1994), "The asymptotic variance of semiparametric estimators." *Econometrica*, 62, 1349–1382. [248]

Ponomareva, M. and E. Tamer (2009), "Misspecification in moment inequality models: Back to moment equalities?" Working paper, Northwestern University. [241]

Rosenbaum, P. (2002), *Observational Studies*. Springer, New York. [233]

Rosenbaum, P. and D. Rubin (1983), "Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome." *Journal of the Royal Statistical Society, Ser. B*, 45, 212–218. [233, 235]

Rubin, D. B. (1976), "Inference and missing data." *Biometrika*, 63, 581–592. [232]

Stoye, J. (2007), "Bounds on generalized linear predictors with incomplete outcome data." *Reliable Computing*, 13, 293–302. [241]

Tukey, J. W. (1960), "A survey of sampling from contaminated distributions." In *Contributions in Probability and Statistics: Essays in Honor of Harold Hotelling*, 448–485, Stanford University Press, Stanford. [236]

Submitted July, 2011. Final version accepted April, 2012.