Sensitivity to Missing Data Assumptions: Theory and An Evaluation of the U.S. Wage Structure

Patrick Kline
Department of Economics
UC Berkeley / NBER
pkline@econ.berkeley.edu

Andres Santos*
Department of Economics
UC San Diego
a2santos@ucsd.edu

April, 2012

Abstract

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 non-ignorable selection governing the missingness process by the maximal Kolmogorov-Smirnov (KS) 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.

*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 et al. (2006). A previous version of this paper circulated under the title "Interval Estimation of Potentially Misspecified Quantile Models in the Presence of Missing Data."

1 Introduction

Despite major advances in the design and collection of survey and administrative data, missingness remains a pervasive feature of virtually every modern economic dataset. Hirsch and Schumacher (2004), for instance, find that nearly 30% of the earnings observations in the Outgoing Rotation Groups of the Current Population Survey 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 non-response is approximately ignorable, the extent of missingness in modern economic data suggests that economists ought to assess the sensitivity of their conclusions to small deviations from this assumption.

Previous work on non-ignorable missing data processes has either relied upon parametric models of missingness in conjunction with exclusion restrictions to obtain point identification (Greenlees et al. (1982) and Lillard et al. (1986)) or considered the "worst case" bounds on population moments that result when all assumptions regarding the missingness process are abandoned (Manski (1994, 2003)). Neither approach has garnered much popularity.[1] It is typically quite difficult to find variables which shift the probability of missingness 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 processes that the majority of researchers would consider to be implausible in modern datasets.

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 (MAR) 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 upon 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 missingness process, we investigate the level of non-ignorable 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

[1]See DiNardo et al. (2006) for an applied example comparing these two approaches.

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 yields a natural parameterization of deviations from ignorability, with a distance of zero corresponding to MAR and a distance of one encompassing the totally unrestricted missingness processes considered in Manski (1994). Between these extremes lie a continuum of selection mechanisms which may be studied to determine a critical level of selection above which conclusions obtained under an analysis predicated upon MAR may be overturned. By reporting the minimal level of selection necessary to undermine a hypothesis, we allow the reader to decide for themselves which inferences to draw based upon his or her 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. In order to facilitate the analysis of datasets with many covariates, results are also developed summarizing 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 missingness, 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 governing 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 governing 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 et al. (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

[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 the reader to strike this balance on their own.

the MAR assumption. We begin by revisiting the results of Angrist et al. (2006) regarding changes across Decennial Censuses 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 extensive missingness 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 non-response to the CPS within demographic covariate bins, with very high and very low earning individuals most likely to have invalid CPS earnings records. 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 U.S. 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 briefly conclude in Section 6.

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 one if Y is observable and zero 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 upon alternative assumptions regarding the process generating D. In particular, we will concern ourselves with the sensitivity of conclusions regarding $q(\tau|X)$, the conditional τ-quantile of Y given X, which is often of more direct interest than the distribution function itself. Towards 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 τ-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 et al. (2005)) has relied upon 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 studying 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:

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

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 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

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

A provides a numerical example mapping the parameters of a bivariate normal selection model into values of $\mathcal{S}(F)$ and plots of the corresponding observed and missing data 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 examining 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 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. Towards 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 non-ignorable 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:

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

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 zero and k, as in (4). Formally, the conclusions we can draw from the formulations in (4) and (5) are equivalent – the result of $\mathcal{S}(F) \leq k$ accommodating (5) for any $\tilde{F}_{y|x}$, and that for any $v \in [0, k)$ there is a $\tilde{F}_{y|x}$ such that $\mathcal{S}(F) = v$ when (5) holds.

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

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:

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

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:

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

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 λ 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 λ in the corrupted sampling model. ■

2.3 Conditional Quantiles

For $q(\tau|X)$ the conditional τ-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 $\mathcal{S}(F) \leq k$ being part of a hypothetical exercise only.

The following Lemma provides a sharp characterization of the nominal identified set:

Lemma 2.1. *Suppose Assumptions 2.1(ii)-(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:*

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

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

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

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

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

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

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):

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

Toward this end, first note that in order 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:

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

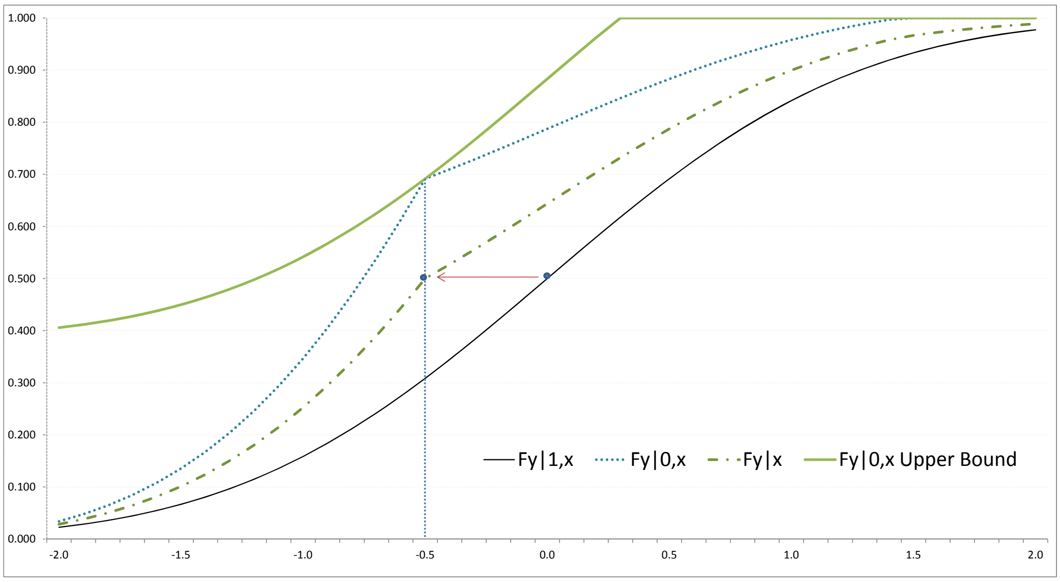
then $\tilde{F}_{y|0,x}$ will satisfy: (I) provided $\Psi_x(c)$ is continuous, increasing and satisfies $\lim_{c \to -\infty} \Psi_x(c) = 0$, (II) provided $0 \leq \Psi_x(c) \leq \kappa(x) \leq k$ for all c, and (III) if $\Psi_x(\theta(x)) = \kappa(x) - F_{y|1,x}(\theta(x))$. For $(a)_+ \equiv \max\{a, 0\}$, 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 zero, and $p(x) = 1/2$. These parameters yield an

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



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 $F_{y|0,x}$ which 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 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 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:

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

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) may be used to evaluate the sensitivity of conclusions obtained under MAR. For simplicity, we let X be binary so that the conditional τ-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 τ_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:

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

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 τ 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 τ. As in Example 2.1, we can define the quantile specific critical level of selection:

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

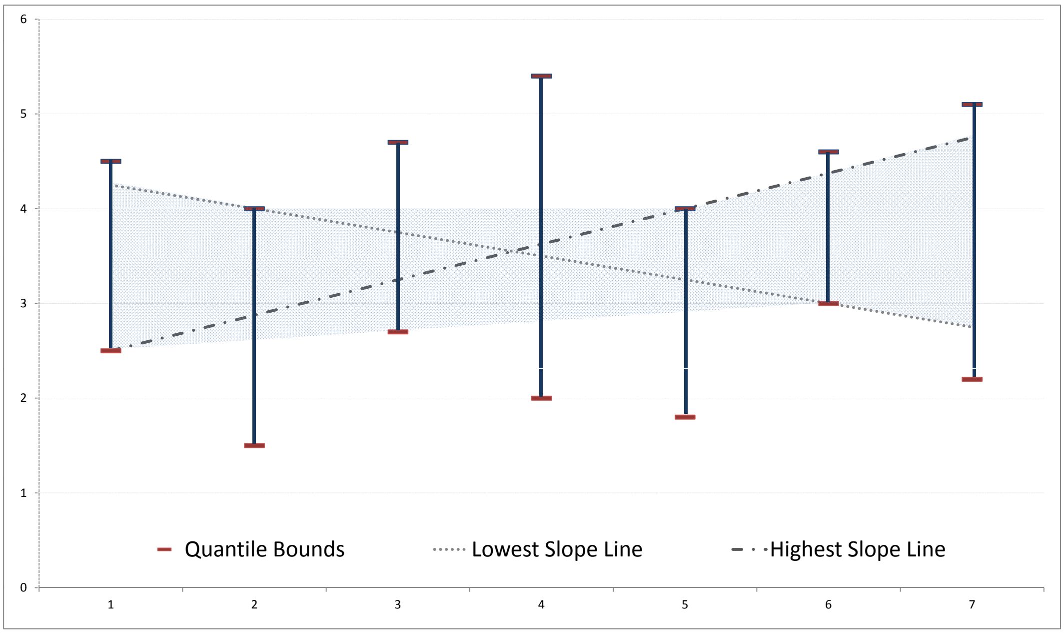
By considering $\kappa_0(\tau)$ at different values of τ, we implicitly define a "breakdown" function $\kappa_0(\cdot)$ that reveals the differential sensitivity of the initial conjecture at each quantile $\tau \in (0,1)$. ■

3 Parametric Modeling

Analysis of the conditional τ-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 obeying the bounds provide a poor description

Figure 2: Linear Conditional Quantile Functions (Shaded Region) as a Subset of the Identified Set



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 et al. (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.

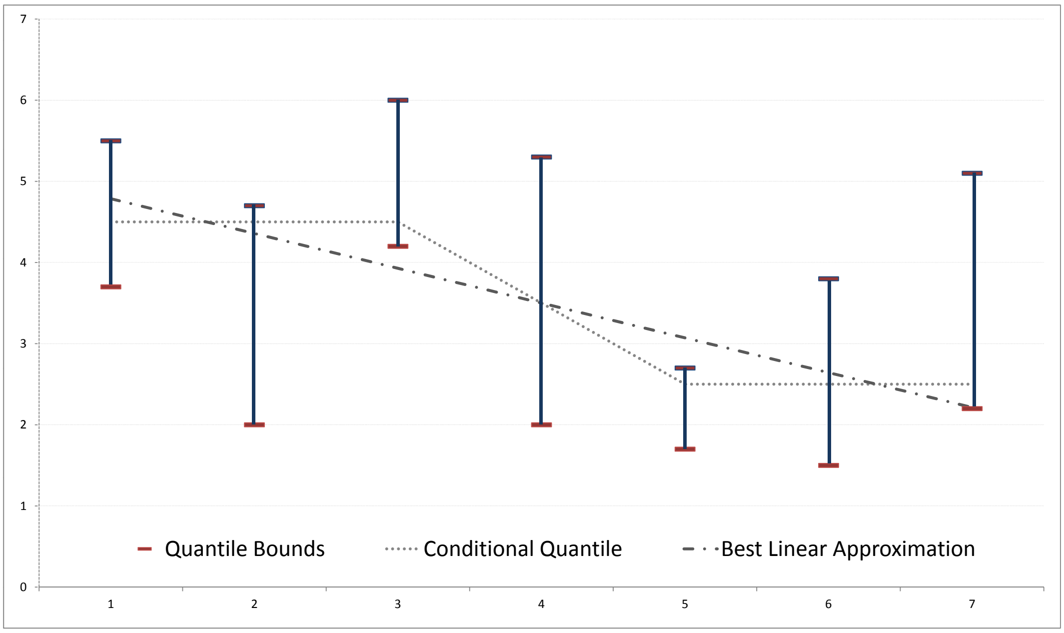
We focus on linear models and approximations that minimize a known quadratic loss function. For S a known measure on $\mathcal{X}$ and $E_S[g(X)]$ denoting the expectation of $g(X)$ when X is distributed according to S, we define the parameters governing a best linear approximation (BLA) as:[5]

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

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

[5]The measure S weights the squared deviations in each covariate bin. Its specification is an inherently context-specific task depending entirely upon 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



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:

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

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.

Lemma 3.1. *If Assumptions 2.1(ii)-(iii), $\mathcal{S}(F) \leq k$ and $E_S[XX']$ is invertible, then it follows that:*

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

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]

[6]Since X has discrete support, we can characterize the function θ 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 θ is a finite dimensional vector over which the criterion is optimized.

Corollary 3.1. *Suppose Assumptions 2.1(ii)-(iii), $\mathcal{S}(F) \leq k$, $E_S[XX']$ is invertible and define:*

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

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

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 τ under the restriction that $\mathcal{S}(F) \leq k$. 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 following set:

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

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 $\mathcal{S}(F) \leq k$. Fortunately, $\pi_L(\cdot, k)$ and $\pi_U(\cdot, k)$ *are* in the identified set over the range of (τ, 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 τ, (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 in order 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 (cont.) Suppose that an analysis under MAR reveals $\beta(\tau_0) \neq 0$ at a specific quantile τ_0. We may then define the critical level of k_0 necessary to cast doubt on this conclusion as:

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

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

Example 2.2 (cont.) In a parametric analogue of first order stochastic dominance of $F_{y|x}$ over

[7] For example, under our assumptions $\lambda'\beta(\cdot)$ is a continuous function of τ. 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 $\mathcal{S}(F) \leq k$.

$F_{y|x'}$ for $x < x'$, a researcher examines whether $\beta(\tau) \leq 0$ for all $\tau \in (0,1)$. Suppose that a MAR analysis reveals that $\beta(\tau) > 0$ for multiple values of τ. The functions (π_L, π_U) enable her to assess what degree of selection is necessary to undermine her conclusions by considering:

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

Note that finding $\pi_L(\tau, k_0) \leq 0$ for all $\tau \in (0,1)$ does in fact cast doubt on the conclusion that $\beta(\tau) > 0$ for some τ 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 (cont.) 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 τ. In this manner, we obtain a quantile specific critical level of selection:

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

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 which 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 (τ, b, x)) given by:

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

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 (τ, 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 (τ, k) pairs such that $\mathcal{S}(F) \leq k$ proves more informative than the restriction that $F_{y|0,x}$ lie between zero and one. Succinctly, for an arbitrary fixed $\zeta > 0$, these conditions are satisfied by values of (τ, k) in the set:

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

Heuristically, by restricting attention to $(\tau, k) \in \mathcal{B}_\zeta$, we are imposing that large or small values of τ be accompanied by small values of k. This simply reflects that the fruitful study of quantiles close to one or zero 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:

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

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:

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

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

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

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:

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

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

We introduce the following additional assumption in order 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 one 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. Assumptions 4.1(ii)-(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 towards constructing confidence intervals for $\pi_L(\tau,k)$ and $\pi_U(\tau,k)$ that are uniform in (τ,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, 4.1 hold and $\{Y_iD_i, X_i, D_i\}_{i=1}^n$ is an 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:

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

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 π_L and π_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 exploit that $\pi_L(\tau,k)$ and $\pi_U(\tau,k)$ are smooth functionals to obtain asymptotically normal estimators *pointwise* in (τ,k) without a discrete support requirement on X – Newey (1994) and Chen et al. (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 – e.g. a finite set of (τ,k). ■

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

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

[8] $F_{y|1,x}$ being 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.

[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 $(\pi_L,\pi_U)=K(q_L,q_U)$.

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 ζ defining $\mathcal{B}_\zeta$ when it is allowed to be arbitrarily close to zero. Intuitively, small values of ζ imply $q_L(\tau, k|x)$ 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 τ_n quantile when $n\tau_n \to c > 0$ is nonnormal, implying 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 Section 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.

Example 2.1 (cont.) 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 (cont.) 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 τ. First, let us define:

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

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

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

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 ω_L allows the researcher to account for the fact that the variance of $G^{(1)}(\tau, k)$ may depend heavily on (τ, 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 (cont.) Employing Theorem 4.1 it is possible to construct a two sided confidence interval for the function $\kappa_0(\cdot)$. Towards this end, we exploit uniformity in τ and k by defining:

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

where as in Example 2.2, ω_L and ω_U are positive weight functions. In addition, we also let:

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

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

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 τ 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 2.1-2.2) 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 requires 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 ω_L and ω_U (as in (42), (45)). We will allow the weight functions ω_L and ω_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 (ω_L, ω_U), let G_ω 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_ω. 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 4.3. *(i) W is independent of (Y, X, D), with $W > 0$ a.s., $E[W] = 1$, $Var(W) = 1$ 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 algorithm:

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

$$\tilde{Q}_{x,n}(c|\tau, b) \equiv \Big(\frac{1}{n}\sum_{i=1}^n W_i\{1\{Y_i \leq c, D_i = 1, X_i = x\} + b1\{D_i = 0, X_i = x\} - \tau 1\{X_i = x\}\}\Big)^2 . \quad (51)$$

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

$$\tilde{q}_L(\tau, k|x) \in \arg\min_{c\in\mathbf{R}} \tilde{Q}_{x,n}(c|\tau, \tau + k\tilde{p}(x)) \qquad \tilde{q}_U(\tau, k|x) \in \arg\min_{c\in\mathbf{R}} \tilde{Q}_{x,n}(c|\tau, \tau - k\tilde{p}(x)) \quad (52)$$

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 missingness 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:

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

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

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

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

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 $\{W_i\}_{i=1}^n$):

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

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 $\tilde{r}_{1-\alpha}$ 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, 4.3 hold, the CDF of $L(G_\omega)$ is strictly increasing and continuous at $r_{1-\alpha}$ and $\{Y_iD_i, X_i, D_i, W_i\}_{i=1}^n$ 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 et al. (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 et al. (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 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 et al. (2006) yields similar results.

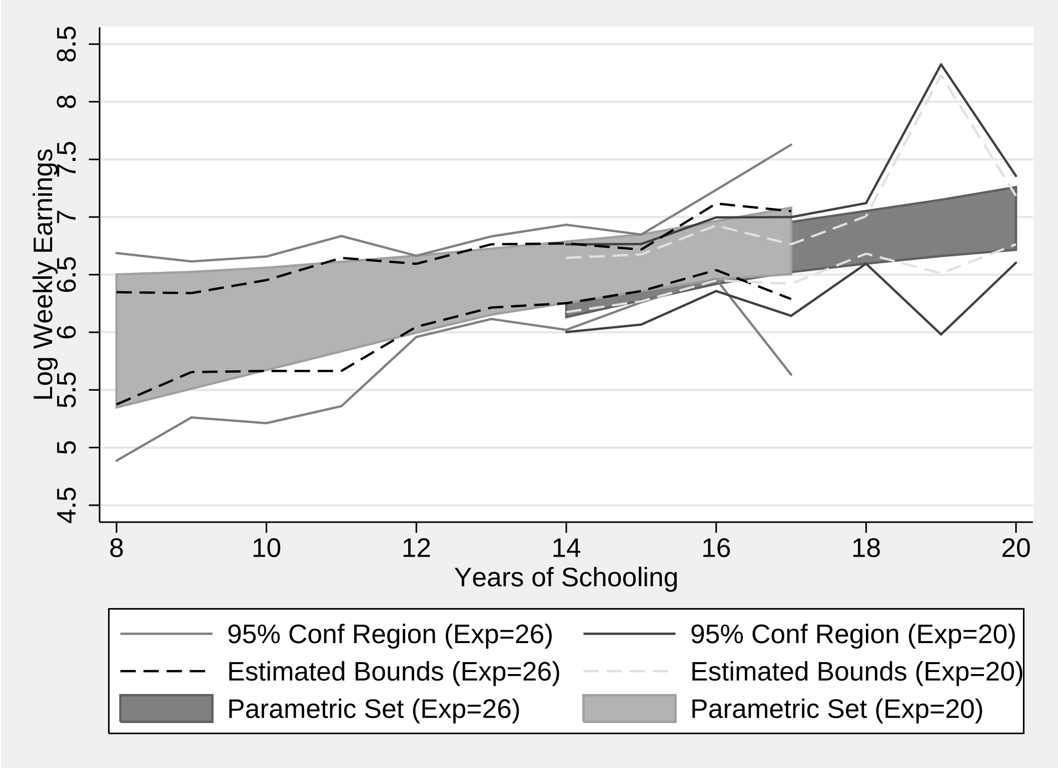
Table 1: Frequency of Missing Weekly Earnings in Census Estimation Sample by Year and Cause

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%

The sample consists of native born black and white men ages 40-49 with six or more years of schooling who worked at least one week in the past year. Details are provided in the Data Appendix. Like Angrist et al. (2006), we use average 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 falling into demographic cells with less than 20 observations are dropped. 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.[10]

[10]It is interesting to note that only 7% of the men in our sample report working no weeks in the past year. Hence,

Figure 4: Worst Case Nonparametric Bounds on 1990 Medians and Linear Model Fits for Two Experience Groups of White Men.



Like Angrist et al. (2006), we estimate linear conditional quantile models for log earnings per week of the form:

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

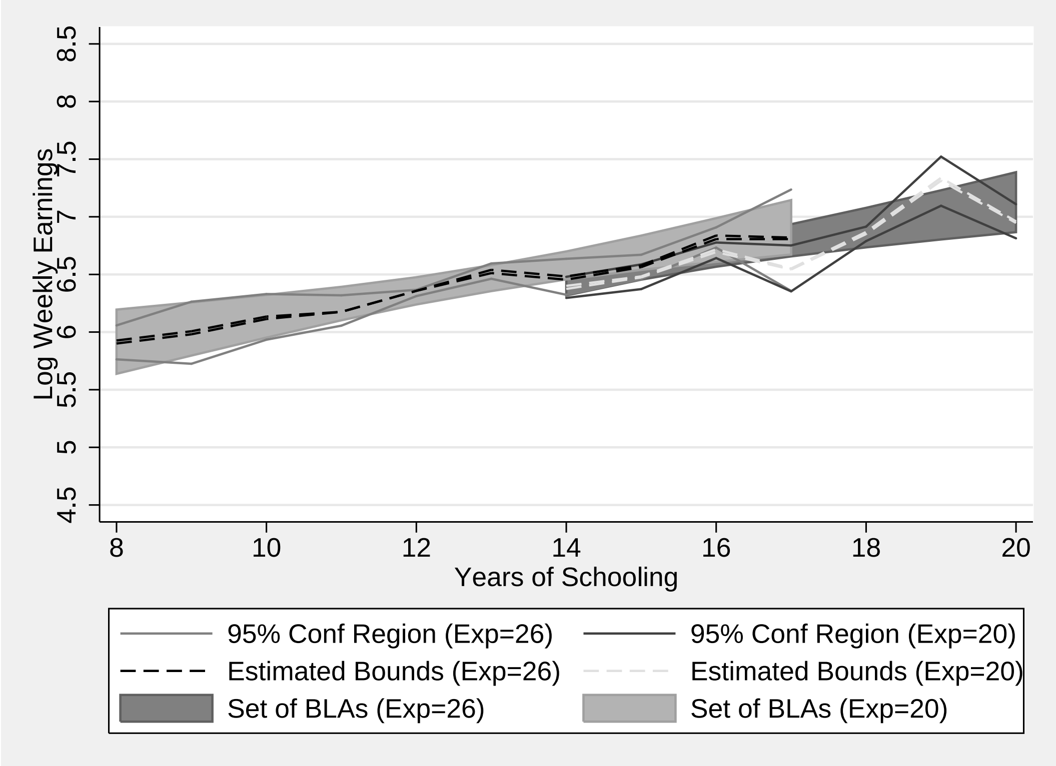
where X consists of an intercept, a black dummy, and a quadratic in potential experience, and E represents years of schooling. Our analysis focuses on the quantile specific "returns" to a year of schooling $\beta(\tau)$ though we note that, particularly in the context of quantile regressions, these Mincerian earnings coefficients need not map into any proper economic concept of individual returns (Heckman et al. (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 et al. (2006), we illustrate the methods developed so far by analyzing the median wages of the 227 demographic groups in our 1990 sample. We begin by considering the worst case nonparametric bounds on these medians. Because the covariates are of dimension three, 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

at least for this population of men, assumptions regarding the determinants of non-response appear to be more important for drawing conclusions regarding the wage structure than assumptions regarding non-participation in the labor force.

Figure 5: Nonparametric Bounds on 1990 Medians and Best Linear Approximations for Two Experience Groups of White Men Under $\mathcal{S}(F) \leq 0.05$.



bounds were obtained using $\hat{q}_L(0.5, 1|x)$ and $\hat{q}_U(0.5, 1|x)$, which are the sample analogues 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%.[11] 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 non-random 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

[11] 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 bins. We exploit independence across x to find a critical value delivering coverage of the conditional median function with asymptotic probability of 0.95. See the implementation appendix for details.

obeying the bounds for the conditional median implies 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. 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 et al. (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 $\mathcal{S}(F) \leq 0.05$.[12] 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 $[.058, .163]$. 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 $[0.102, 0.118]$.[13] 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 et al. (2006) concerning changes across Censuses in the structure of wages 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 the Implementation Appendix. To ensure comparability with Angrist et al. (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).[14] Notably, with the MAR restriction, our estimation procedure is equivalent to the classical minimum distance estimator studied by Chamberlain (1994).

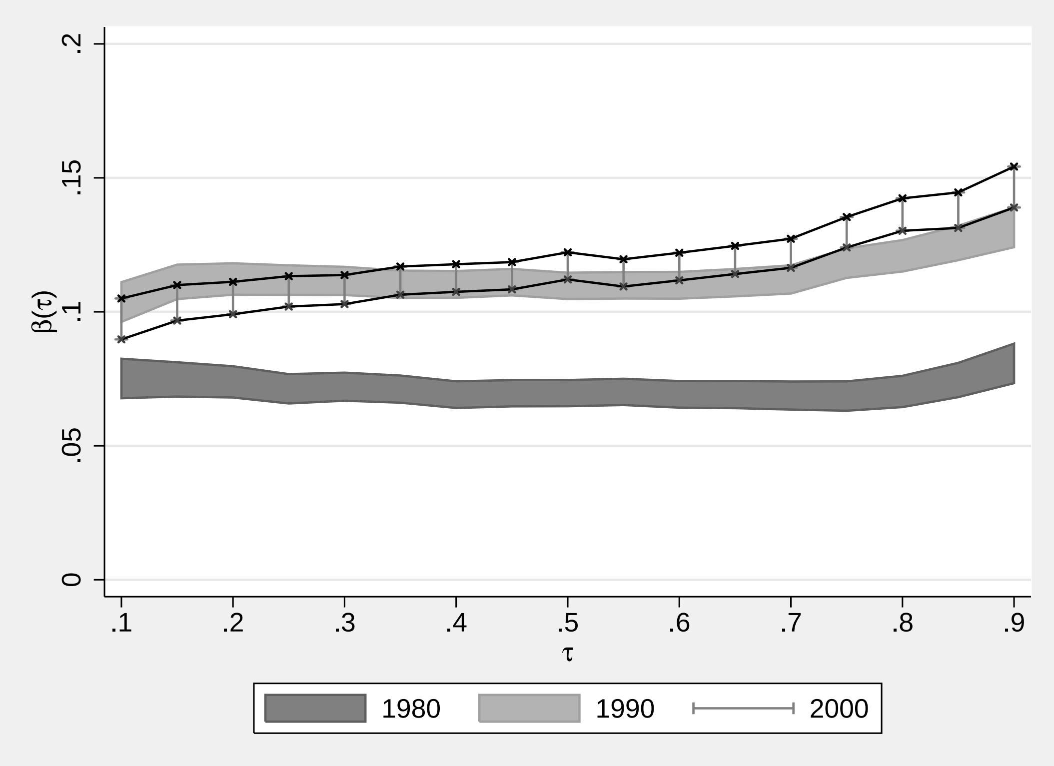
Figure 6 plots estimates of the approximate returns functions $\beta(\cdot)$ in 1980, 1990, and 2000 along with uniform confidence intervals. Our MAR results are similar to those found in Figure 2A of Angrist et al. (2006). They suggest that the returns function increased uniformly across quantiles between 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

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

[13] 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 $[\hat{\pi}_L(0.5, .05) - \hat{c}_L/\sqrt{n}, \hat{\pi}_U(.5, .05) + \hat{c}_U/\sqrt{n}]$.

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

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 true conditional quantile function.

conjectured by Juhn et al. (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 multi-factor 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 limitations 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 missingness 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 capturing 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 heterogeneity 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:

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

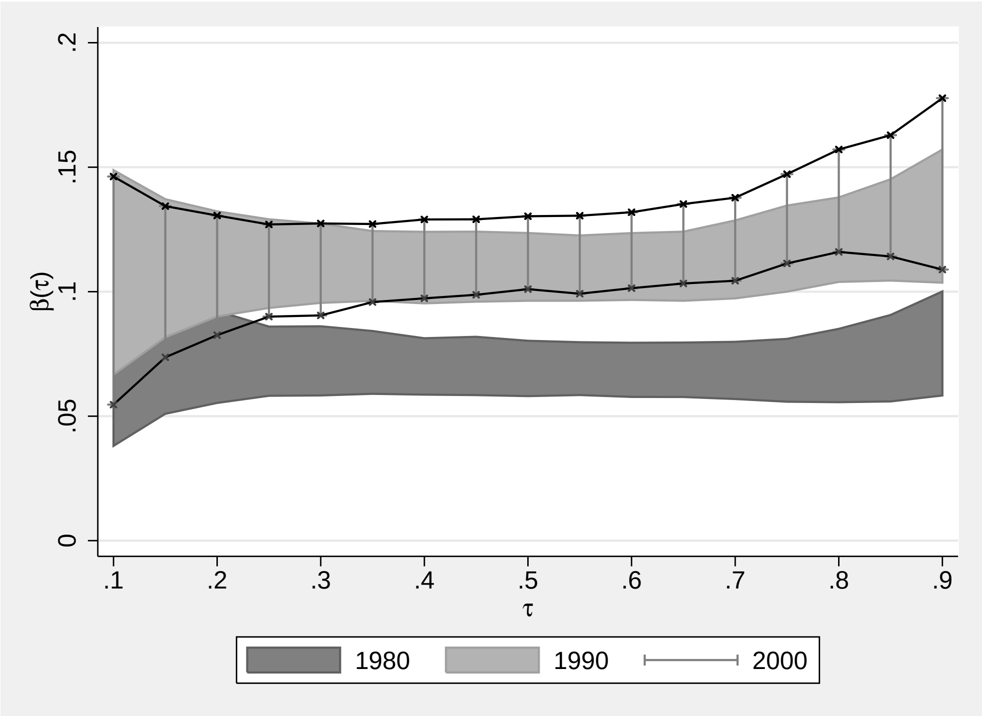
As in Example 2.2, we are particularly interested in k_0, the smallest value of k such that (58) holds, as it will hold trivially for all $k \geq k_0$. A search for the smallest value of k such that the 95% uniform confidence intervals for these two decades overlap at all quantiles between 0.2 and 0.8 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 the evidence is quite strong that the schooling coefficient process changed between 1980 and 1990. Figure 8 plots the uniform confidence regions corresponding 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. Towards this end, we generalize Example 2.3 and define $\kappa_0(\tau)$ to be the smallest level of k such that:

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

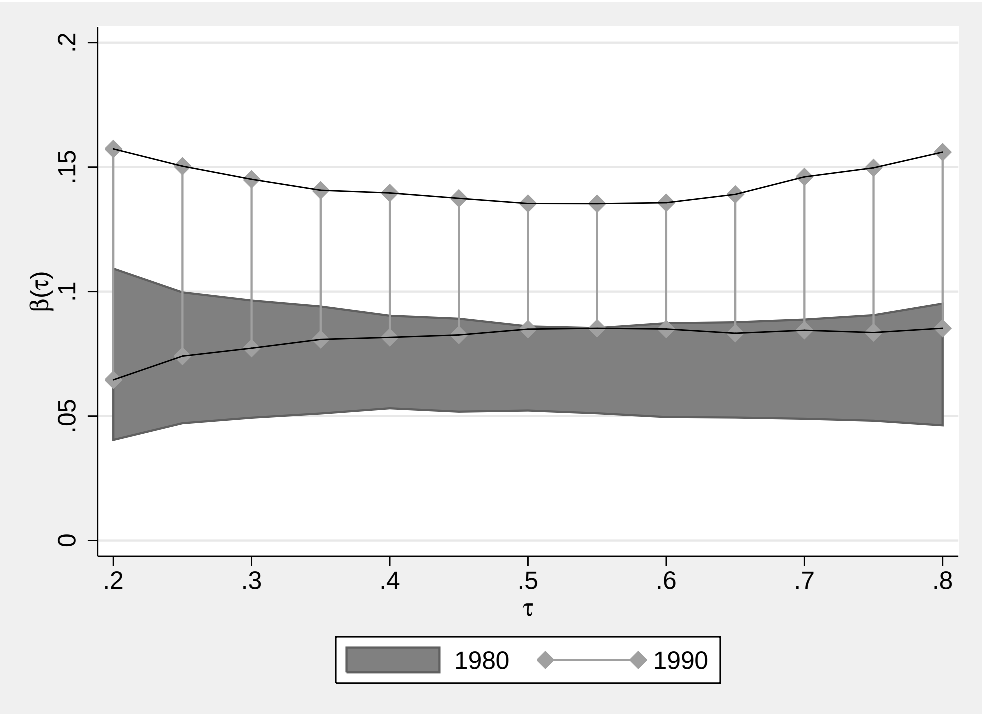
The function $\kappa_0(\cdot)$ summarizes the level of robustness of each quantile-specific conclusion. In this manner, the "breakdown" function $\kappa_0(\cdot)$ reveals the differential sensitivity of the entire conditional

Figure 7: Uniform Confidence Regions for Schooling Coefficients by Quantile and Year Under $\mathcal{S}(F) \leq 0.05$.



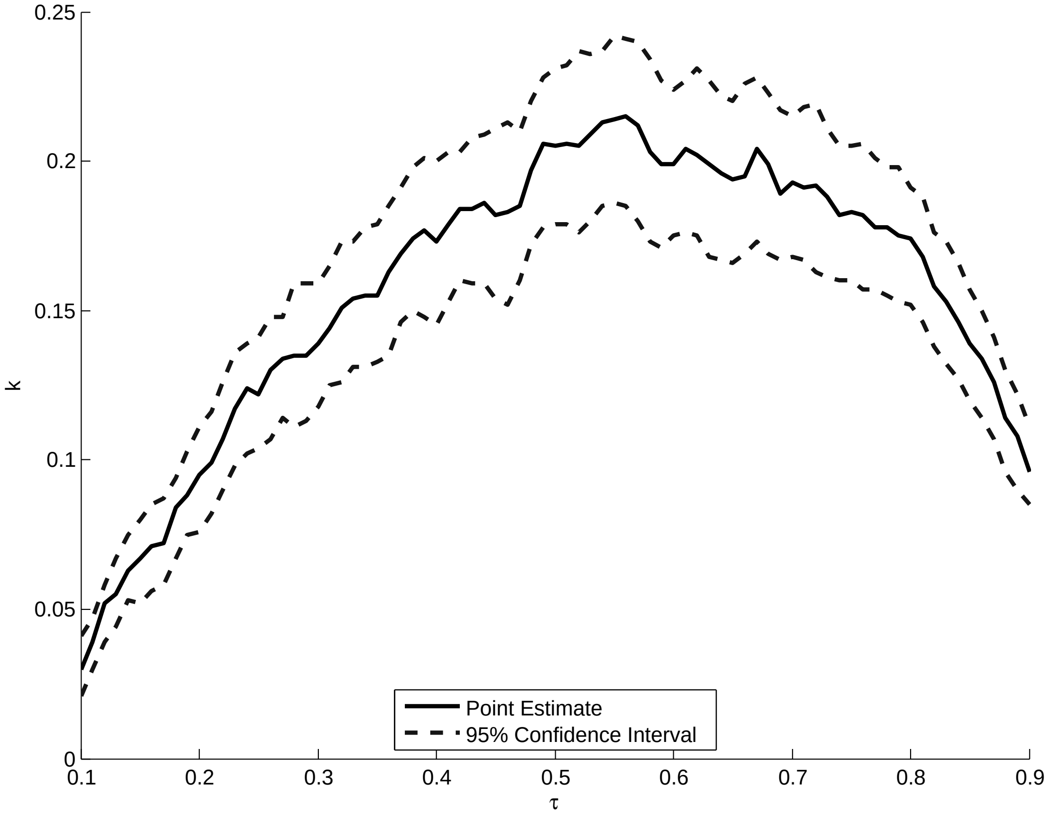
Note: Model coefficients provide minimum mean square approximation to true conditional quantile function.

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 minimum mean square approximation to true conditional quantile function as in Chamberlain (1994).

Figure 9: Breakdown Curve (1980 vs 1990).



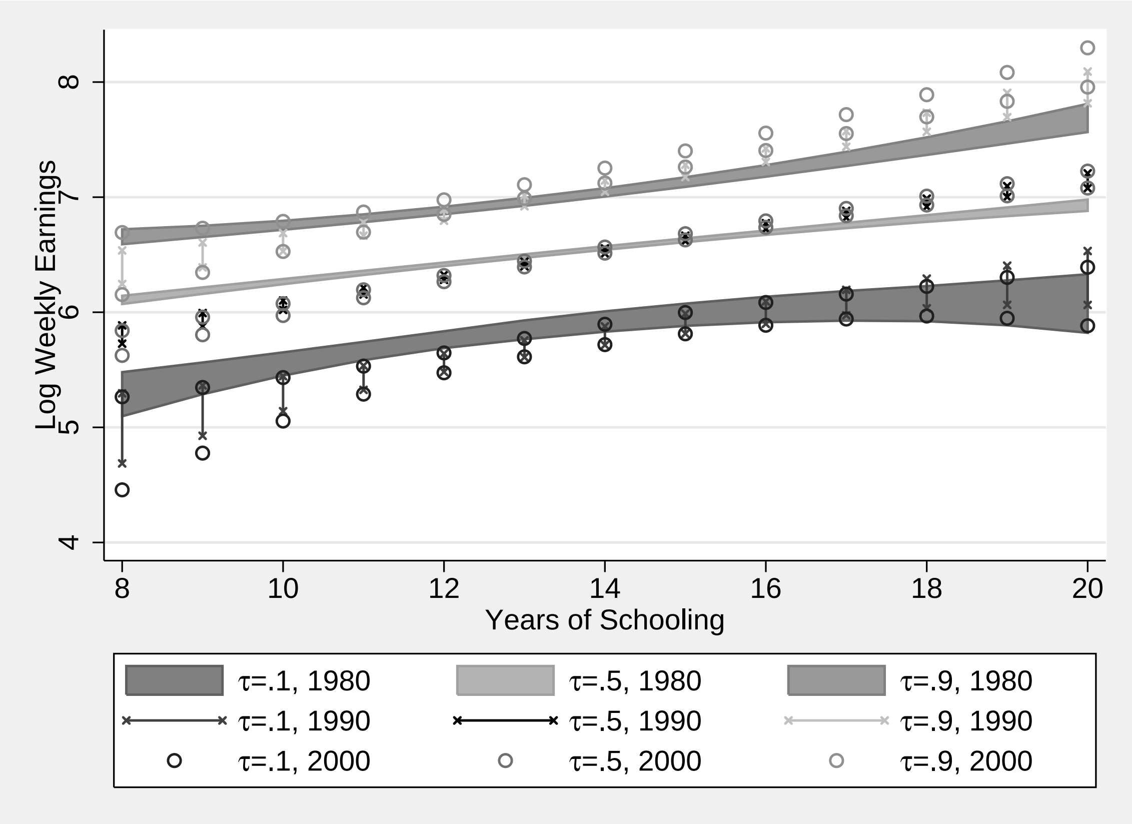
Note: Each point on this curve indicates the minimal level of $\mathcal{S}(F)$ necessary to undermine the conclusion that the schooling coefficient increased between 1980 and 1990 at the quantile of interest.

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 τ we first construct 95% uniform two sided confidence intervals in τ 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, changes in reporting behavior between 1980 and 1990 pose the greatest threat to hypotheses regarding changes at the bottom quantiles of the earnings distribution.

Figure 10: Confidence Intervals for Fitted Values Under $\mathcal{S}(F) \leq 0.05$.



Note: Earnings quantiles modeled using quadratic specification in Lemieux (2006). Model coefficients provide minimum mean square approximation to true conditional quantile function. Covariates other than education set to sample mean.

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.[15] 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 appear 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.

[15] 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.

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 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.

In order 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. Towards this end, let us define the following conditional probabilities:

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

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

By applying Bayes' Rule, it is then possible to express the distance between the distribution of missing and non-missing observations at a given quantile as a function of the selection probabilities:[16]

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

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 τ and x) of $\mathcal{S}(F)$ as well.[17] For this reason, our efforts focus on estimating this function in a dataset with information on the earnings of survey non-respondents.

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. In order 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 five 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 topcode 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 identifying themselves as self-employed to the CPS. Further details are provided in the Data Appendix.

As in our study of the Decennial Census, we take the relevant covariates to be age, years of

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

[17] Note that $P(D=0, F_{y|x}(Y) \leq \tau | X=x) = \int_0^\tau P(D=0|F_{y|x}(Y)=u, X=x)du$ 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 $p(x) = \int_0^1 P(D=0|F_{y|x}(Y)=u, X=x)du$.

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.[18] 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.[19] As a result, we are only able to examine biases generated by earnings non-response.

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 missingness rate of 7.2%. We approximate the probability of non-response with the following sequence of increasingly flexible logistic models:

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

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

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

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 (γ_1, γ_2) but restricts these interactions to be linear in the cell effects δ_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.[20] A comparison of the model log likelihoods reveals that the introduction of the interaction terms (γ_1, γ_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.

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, missingness follows a U-shaped response pattern with very low and very high wage men least likely to provide

[18] We use five-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.

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

[20] 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 1973 CPS-IRS Sample

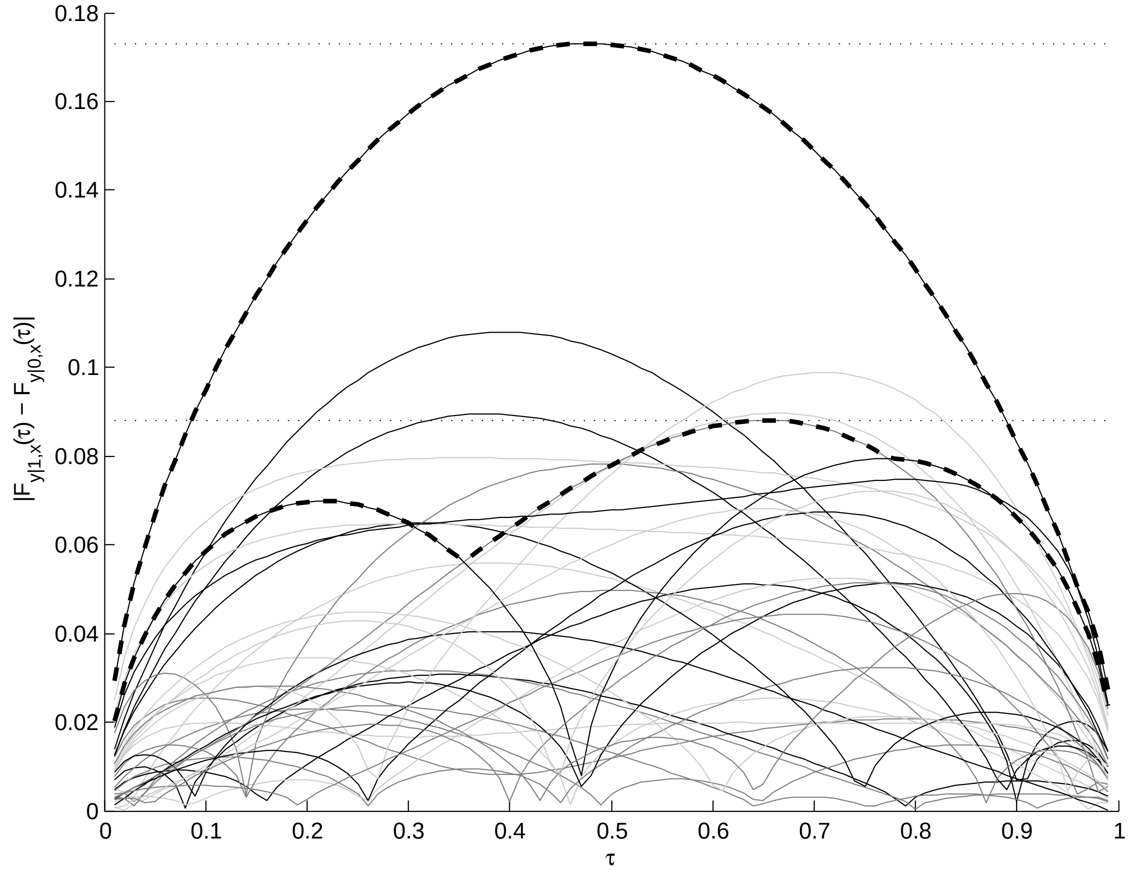
	Model 1	Model 2	Model 3
b_1	-1.06	0.05	
	(0.43)	(5.44)	
b_2	1.09	3.75	
	(0.41)	(4.08)	
γ_1		0.45	
		(2.30)	
γ_2		1.15	
		(1.73)	
Log-Likelihood	-3,802.91	-3798.48	-3759.97
Parameters	37	39	105
Number of observations	15,027	15,027	15,027
Demographic Cells	35	35	35
$E[\frac{\partial P(D=0\|X=x,F_{y\|x}(Y)=\tau)}{\partial \tau}\|_{\tau=0.2}]$	-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 in parentheses.

valid earnings information – a pattern conjectured (but not directly verified) by Lillard et al. (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 which in some cases yields response patterns that are monotonic rather than U-shaped.

It is straightforward to estimate the distance between missing and nonmissing earnings distributions in each demographic bin 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 bins 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,

Figure 11: Logit Based Estimates of Distance Between Missing and Non-Missing CDFs by Quantile of IRS Earnings and Demographic Cell.



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 bins 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 missingness 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 exhibiting first order stochastic dominance. It is interesting to note that the demographic cell obtaining 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 which 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 non-response. 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 datasets.

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.

Appendix A. The bivariate normal selection model and KS.

To develop intuition for our metric $\mathcal{S}(F)$ of deviations from missing at random we provide here a mapping between the parameters of a standard bivariate selection model, the resulting CDFs of observed and missing outcomes, and the implied values of $\mathcal{S}(F)$. Using the notation of Section 2, our DGP of interest is:

$$(Y_i, v_i) \sim N(0, \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)) \qquad D_i = 1\{\mu + v_i > 0\} \ . \tag{63}$$

In this model, the parameter ρ indexes the degree of non-ignorable selection in the outcome variable Y_i. We choose $\mu = .6745$ to ensure a missing fraction of 25% which is approximately the degree of missingness found in our analysis of earnings data in the US Census. We computed the distributions of missing and observed outcomes for various values of ρ by simulation, some of which are plotted in Figures A.1 and A.2. The resulting values of $\mathcal{S}(F)$, which correspond to the maximum vertical distance between the observed and missing CDFs across all points of evaluation, are given in the table below:

Table A.1: $\mathcal{S}(F)$ as a function of ρ

ρ	$\mathcal{S}(F)$	ρ	$\mathcal{S}(F)$	ρ	$\mathcal{S}(F)$
0.05	0.0337	0.35	0.2433	0.65	0.4757
0.10	0.0672	0.40	0.2778	0.70	0.5165
0.15	0.1017	0.45	0.3138	0.75	0.5641
0.20	0.1355	0.50	0.3520	0.80	0.6158
0.25	0.1721	0.55	0.3892	0.85	0.6717
0.30	0.2069	0.60	0.4304	0.90	0.7377

Figure A.1: Missing and Observed Outcome CDFs

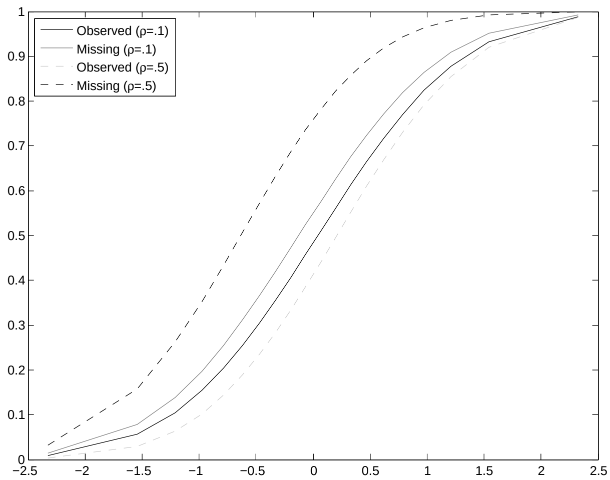
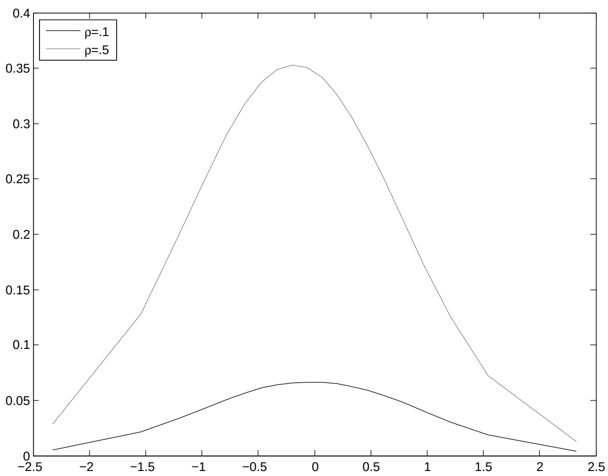


Figure A.2: Vertical Distance Between CDFs



Appendix B. Data Details

Census Data

Our analysis of Decennial Census data uses 1% unweighted extracts obtained from the Minnesota IPUMS website http://usa.ipums.org/ on August 28th, 2009. Our extract contained all native born black and white men ages 40-49. We drop all men with less than five years of schooling and recode the schooling variable according to the scheme described by Angrist et al. (2006) in their online appendix available at: http://econ-www.mit.edu/files/385. Our IPUMS extract, along with Stata code used to impose our sample restrictions is available online at http://www.econ.berkeley.edu/~pkline/papers/web_supplement_KS_missing.zip.

Like Angrist et al. (2006), we drop from the sample all individuals with allocated age or education information and all individuals known not to have worked or generated earnings. We also drop observations in small demographic cells. The following Table lists the effects of these drops on the sample sizes by Census year:

Table A.2: Census Sample Sizes by Year After Imposing Restrictions

	1980	1990	2000
Native Born Black and White Men Ages 40-49			
w/ >= 5 Years of Schooling	97,900	131,667	168,909
Drop Observations w/Imputed Age	96,403	130,806	165,505
Drop Observations w/Imputed Education	90,064	124,161	155,158
Drop Observations w/Unallocated (earnings or weeks worked)=0	80,800	111,366	131,711
Drop Observations in Cells w/ < 20 observations	80,128	111,070	131,265

Following Angrist et al. (2006) we choose as our wage concept log average weekly earnings – the log of total annual wage and salary income divided by annual weeks worked. Earnings in all years are converted to 1989 dollars using the Personal Consumption Expenditure (PCE) price index. We recode to missing all weekly earnings observations with allocated earnings or weeks worked.

CPS Data

For the analysis in Section 5.2 we used ICPSR archive 7616 – "Current Population Survey, 1973, and Social Security Records: Exact Match Data." We extract from this file a sample of white and black men ages 25-54 with six or more years of schooling who reported working at least one week in the last year. We then drop from the sample individuals who are self-employed and those working in industries or occupations identified by Bound and Krueger (1991) as likely to receive tips underreported to the IRS.

Annual IRS wage and salary earnings are topcoded at $\$50,000$ dollars. There are also a small fraction of observations with very low IRS earnings below $\$1,000$. We drop observations falling into either group. We also drop observations with allocated weeks worked. Finally we drop observations falling into demographic cells with less than 50 observations. An itemization of the effect of these decisions on sample size is provided by the following Table:

Table A.3: CPS Sample Sizes After Imposing Restrictions

	1980
Black and White Men Ages 25-54 w/ >= 5 Years of Schooling	
and One or More Weeks Worked	19,693
Drop Self-Employed	17,665
Drop Bound-Krueger Industries/Occupations	17,138
Drop Topcoded IRS Earnings and Outliers	15,632
Drop Observations w/Allocated Weeks Worked	15,355
Drop Cells w/ < 50 observations	15,027

Weeks worked are reported categorically in the 1973 CPS. We code unallocated weeks responses to the midpoint of their interval. Our average weekly earnings measure is constructed by dividing IRS wage and salary earnings by the recoded weeks worked variable.

Appendix C. Implementation Details

We outline here the implementation of our estimation and inference procedure in the Decennial Census data. The MATLAB code employed is available online at `http://www.econ.berkeley.edu/~pkline/papers/web_supplement_KS_missing.zip`.

Our estimation and inference routine consists of three distinct procedures: (i) Obtaining point estimates, (ii) Obtaining bootstrap estimates, and (iii) Constructing confidence regions from such estimates. Below we outline in detail the algorithms employed in each procedure.

(i) Point Estimates: We examine a grid of quantiles, denoted $\mathcal{T}$, with lower and upper limits of $\underline{\tau}$ and $\bar{\tau}$ – for example $\mathcal{T} = \{0.1, 0.15 \ldots 0.85, 0.9\}$. For each τ in this grid, we let

$$K_u(\tau) \equiv \min\left\{\frac{\min\{\tau,(1-\tau)\}}{\max_{x\in\mathcal{X}}(1-\hat{p}(x))}, \frac{\min\{\tau,(1-\tau)\}}{\max_{x\in\mathcal{X}}\hat{p}(x)}, 0.3\right\} - 0.001 \ , \tag{64}$$

and examine for each $\tau \in \mathcal{T}$ a grid of restrictions k, denoted $\mathcal{K}(\tau)$, with a lower bound of 0 and upper bound $K_u(\tau)$ – for example, $\mathcal{K}(\tau) = \{0, 0.01, \ldots, \lfloor \frac{K_u(\tau)}{0.01} \rfloor \times 0.01\}$. These grids approximate:

$$\hat{\mathcal{B}} \equiv \{(\tau, k) \in [0,1]^2 : \underline{\tau} \leq \tau \leq \bar{\tau} \text{ and } 0 \leq k \leq K_u(\tau)\} \ , \tag{65}$$

which is with probability tending to one a subset of $\mathcal{B}_\zeta$ for some ζ such that $\mathcal{B}_\zeta \neq \emptyset$. For each pair (τ, k) in our constructed grid, we then perform the following operations:

STEP 1: For each $x \in \mathcal{X}$, we find $\hat{q}_L(\tau, k|x)$ and $\hat{q}_U(\tau, k|x)$, which respectively are just the $\tau - k(1 - \hat{p}(x))$ and $\tau + k(1 + \hat{p}(x))$ quantiles of observed Y for the demographic group with $X = x$. ■

STEP 2: We obtain $\hat{\pi}_L(\tau, k)$ and $\hat{\pi}_U(\tau, k)$ by solving the linear programming problems in (36) and (37) employing the `linprog` routine in MATLAB. ■

(ii) Bootstrap Estimates: We generate an i.i.d. sample $\{W_i\}_{i=1}^n$ with W exponentially distributed and $E[W] = \text{Var}(W) = 1$. For each (τ, k) in the grid employed to find point estimates, we then perform the following operations:

STEP 1: For each $x \in \mathcal{X}$ we find $\tilde{q}_L(\tau, k|x)$ and $\tilde{q}_U(\tau, k|x)$. Computationally, they respectively equal the $\tau - k(1 - \tilde{p}(x))$ and $\tau + k(1 - \tilde{p}(x))$ weighted quantiles of observed Y for the group $X = x$, where each observation receives weight $W_i/(\sum W_i 1\{D_i = 1, X_i = x\})$ rather than $1/n$. ■

STEP 2: We obtain $\tilde{\pi}_L(\tau, k)$ and $\tilde{\pi}_U(\tau, k)$ by solving the linear programming problems in (53) and (54) employing the `linprog` routine in MATLAB. ■

(iii) Confidence Regions: Throughout we set $\omega_L(\tau, k) = \omega_U(\tau, k) = \omega(\tau)$ where $\omega(\tau) \equiv \phi(\Phi^{-1}(\tau))^{-\frac{1}{2}}$, with $\phi(\cdot)$ and $\Phi(\cdot)$ equal to the standard normal density and CDF. In computing confidence regions, we employ the point estimates $(\hat{\pi}_L^t, \hat{\pi}_U^t)$ for $t \in \{80, 90, 00\}$ from (i), and 1000 bootstrap estimates $\{(\tilde{\pi}_{b,L}^t, \tilde{\pi}_{b,U}^t)\}_{b=1}^{1000}$ computed according to (ii) based on 1000 independent i.i.d. samples $\{W_i\}_{i=1}^n$.

The specifics underlying the computation of each Figure's confidence regions are as follows:

FIGURE 4: For this figure, we compute $(\hat{q}_L(\tau, k|x), \hat{q}_U(\tau, k|x))$ evaluated at $(\tau, k) = (0.5, 1)$ for all $x \in \mathcal{X}$. Employing the bootstrap analogues $(\tilde{q}_L(0, 5, 1|x), \tilde{q}_U(0.5, 1|x))$, as in (52), we obtain estimates $\hat{\sigma}_L$ and $\hat{\sigma}_U$ of the asymptotic variances of $\hat{q}_L(0.5, 1|x)$ and $\hat{q}_U(0.5, 1|x)$ and construct:

$$[\hat{q}_L(0.5, 1|x) - \frac{z_{1-\alpha}\hat{\sigma}_L}{\sqrt{n}}, \hat{q}_U(0.5, 1|x) + \frac{z_{1-\alpha}\hat{\sigma}_U}{\sqrt{n}}] \tag{66}$$

for all $x \in \mathcal{X}$, where $z_{1-\alpha}$ is the $1 - \alpha$ quantile of a standard normal. By independence, the product of the intervals (66) evaluated at $\alpha = 0.95^{\frac{1}{227}}$ is a nonparametric confidence region with asymptotic probability of covering $q(\tau|x)$ of at least 0.95 – Imbens and Manski (2004). Employing `linprog` in MATLAB, we obtain the bounds for the "parametric set" by maximizing/minimizing the coefficient on schooling subject to the constraint that the Mincer specification lie within (66) for all $x \in \mathcal{X}$. ■

FIGURE 5: The nonparametric confidence region was obtained as in Figure 4, but employing $k = .05$ instead. The bounds on the set of BLPs were obtained by solving a linear programming problem as in (24) and (25) but employing the endpoints of (66) in place of $q_L(0.5, 1|x)$ and $q_U(0.5, 1|x)$ – here λ equals one for the coordinate corresponding to the coefficient on education, and zero elsewhere. ■

FIGURE 6: We employ a grid for τ equal to $\mathcal{T} = \{0.1, 0.15, \ldots, 0.85, 0.9\}$. For each $t \in \{80, 90, 00\}$, we compute the $1 - \alpha$ quantile across bootstrap samples of:

$$\max_{\tau \in \{0.1, 0.15, \ldots, 0.85, 0.9\}} \max \left\{ \frac{\tilde{\pi}_L^t(\tau, 0) - \hat{\pi}_L^t(\tau, 0)}{\omega(\tau)}, \frac{\hat{\pi}_U^t(\tau, 0) - \tilde{\pi}_U^t(\tau, 0)}{\omega(\tau)} \right\}, \tag{67}$$

which we denote by $\tilde{r}_{1-\alpha}^t(0)$. For each $t \in \{80, 90, 00\}$ the two sided uniform confidence region is then given by $[\hat{\pi}^t_L(\tau,0) - \tilde{r}^t_{1-\alpha}(0)\omega(\tau), \hat{\pi}^t_U(\tau,0) + \tilde{r}^t_{1-\alpha}(0)\omega(\tau)]$, where for τ outside our grid $\{0.1, 0.15, \ldots, 0.85, 0.9\}$ we obtain a number by linear interpolation. ■

FIGURE 7: The procedure is identical to Figure 6, except k is set at 0.05 instead of at 0. ■

FIGURE 8: For $k_s = 0.05$ we compute the $1-\alpha$ quantile across bootstrap samples of:

$$\max_{\tau \in \{0.2, 0.25, \ldots, 0.75, 0.8\}} \frac{\tilde{\pi}^{80}_U(\tau, k_s) - \hat{\pi}^{80}_U(\tau, k_s)}{\omega(\tau)} , \tag{68}$$

which we denote by $\tilde{r}^{80}_{1-\alpha}(k_s)$. Similarly, we find the $1-\alpha$ quantile across bootstrap samples of:

$$\max_{\tau \in \{0.2, 0.25, \ldots, 0.75, 0.8\}} \frac{\tilde{\pi}^{90}_L(\tau, k_s) - \hat{\pi}^{90}_L(\tau, k_s)}{\omega(\tau)} , \tag{69}$$

which we denote by $\tilde{r}^{90}_{1-\alpha}(k_s)$. Unlike in Figures 6 and 7, we employ a shorter grid for τ as the bounds corresponding to extreme quantiles become unbounded for large k. Next, we examine if:

$$\min_{\tau \in \{0.2, 0.25, \ldots, 0.75, 0.8\}} \{(\hat{\pi}^{80}_U(\tau, k_s) + \tilde{r}^{80}_{1-\alpha}(k_s)\omega(\tau)) - (\hat{\pi}^{90}_L(\tau, k_s) - \tilde{r}^{90}_{1-\alpha}(k_s)\omega(\tau))\} \geq 0 . \tag{70}$$

If (70) holds, we set $k^*_0 = k_s$, otherwise repeat (68)-(70) with $k_s + 0.005$. Hence, $k^*_0 = 0.175$ was the smallest k (under steps of size 0.005) for which the upper confidence interval for $\pi^{80}_U(\tau, k^*_0)$ lied above the lower confidence interval for $\pi^{90}_L(\tau, k^*_0)$ for all $\tau \in \{0.2, 0.25, \ldots, 0.75, 0.8\}$. ■

FIGURE 9: For this figure, we employ a grid of size 0.01 for τ – e.g. $\mathcal{T} = \{0.1, 0.11, \ldots, 0.89, 0.9\}$, and compute $K_u(\tau)$ (as in (64)) using the 1990 Decennial Census. In turn, for each τ we employ a k-grid $\mathcal{K}(\tau) = \{0, 0.001, \ldots, \lfloor \frac{K_u(\tau)}{0.001} \rfloor \times 0.001\}$ and obtain $\hat{\pi}^{80}_U(\tau, k)$ and $\hat{\pi}^{80}_L(\tau, k)$ at each (τ, k) pair. Finally for each τ in our grid, we let $k_I(\tau)$ denote the smallest value of k in its grid such that:

$$\hat{\pi}^{80}_U(\tau, k) \geq \hat{\pi}^{90}_L(\tau, k) , \tag{71}$$

and define $\hat{\kappa}(\tau) \equiv (\hat{\pi}^{80}_U(\tau, k_I(\tau)) + \hat{\pi}^{90}_L(\tau, k_I(\tau)))/2$, which constitutes our estimate of the "breakdown curve". To obtain a confidence region we compute the $1-\alpha$ quantile across bootstrap samples of:

$$\max_{\tau \in \{0.1, 0,15, \ldots, 0.85, 0.9\}} \max_{k \in \{0, 0.01, \ldots, \lfloor \frac{K_u(\tau)}{0.01} \rfloor \times 0.01\}} \frac{|\hat{\pi}^{80}_U(\tau, k) - \tilde{\pi}^{80}_U(\tau, k)|}{\omega(\tau)} , \tag{72}$$

which we denote by $\tilde{r}^{80}_{1-\alpha}$. Similarly, we compute the $1-\alpha$ quantile across bootstrap samples of:

$$\max_{\tau \in \{0.1, 0,15, \ldots, 0.85, 0.9\}} \max_{k \in \{0, 0.01, \ldots, \lfloor \frac{K_u(\tau)}{0.01} \rfloor \times 0.01\}} \frac{|\hat{\pi}^{90}_L(\tau, k) - \tilde{\pi}^{90}_L(\tau, k)|}{\omega(\tau)} , \tag{73}$$

which we denote by $\tilde{r}^{90}_{1-\alpha}$. We then let $k_{I,L}(\tau)$ be the smallest value of k in the grid, such that:

$$\hat{\pi}^{80}_U(\tau, k) + \tilde{r}^{80}_{1-\alpha}\omega(\tau) \geq \hat{\pi}^{90}_L(\tau, k) - \tilde{r}^{90}_{1-\alpha}\omega(\tau) . \tag{74}$$

The lower confidence band is then $\hat{\kappa}_L(\tau) \equiv (\hat{\pi}^{80}_U(\tau, k_{I,L}(\tau)) + \tilde{r}^{80}_{1-\alpha}\omega(\tau) + \hat{\pi}^{90}_L(\tau, k_{I,L}(\tau)) - \tilde{r}^{90}_{1-\alpha}\omega(\tau))/2$. Analogously, we let $k_{I,U}(\tau)$ be the smallest value of $k \in \{0, 0.001, \ldots, \lfloor \frac{K_u(\tau)}{0.001} \rfloor \times 0.001\}$, such that:

$$\hat{\pi}^{80}_U(\tau, k) - \tilde{r}^{80}_{1-\alpha}\omega(\tau) \geq \hat{\pi}^{90}_L(\tau, k) + \tilde{r}^{90}_{1-\alpha}\omega(\tau) , \tag{75}$$

and get the upper confidence band $\hat{\kappa}_U(\tau) \equiv (\hat{\pi}_U^{80}(\tau, k_{I,U}(\tau)) - \tilde{r}_{1-\alpha}^{80}\omega(\tau) + \hat{\pi}_L^{90}(\tau, k_{I,U}(\tau)) + \tilde{r}_{1-\alpha}^{90}\omega(\tau))/2$. Figure 9 is a graph of $\hat{\kappa}_L(\tau)$, $\hat{\kappa}_U(\tau)$ and $\hat{\kappa}(\tau)$. Our bootstraps were conducted over a coarser grid than the one used to obtain point estimates in order to save on computational cost. ■

FIGURE 10: Here λ is set to different levels of education and all other coordinates are set equal to the sample mean of $\{X_i\}_{i=1}^n$. The procedure is otherwise identical to the one employed in the construction of Figure 7, with the exception that a quantile specific critical value is employed. ■

Appendix D. Derivations of Section 5.2.

The following Appendix provides a justification for the derivations in Section 5.2, in particular of the representation derived in equation (62). Towards this end, observe first that by Bayes' rule:

$$\begin{aligned} F_{y|1,x}(c) &= \frac{P(D=1|X=x, Y \leq c) \times F_{y|x}(c)}{p(x)} \\ &= \frac{P(D=1|X=x, F_{y|x}(Y) \leq F_{y|x}(c)) \times F_{y|x}(c)}{p(x)} \ , \end{aligned} \tag{76}$$

where the second equality follows from $F_{y|x}$ being strictly increasing. Evaluating (76) at $c = q(\tau|x)$, employing the definition of $p_L(x, \tau)$ in (60), and noting that $F_{y|x}(q(\tau|x)) = \tau$ yields:

$$F_{y|1,x}(q(\tau|x)) = \frac{p_L(\tau, x) \times \tau}{p(x)} \ . \tag{77}$$

Moreover, by identical arguments, but working instead with the definition of $p_U(\tau, x)$, we derive:

$$1 - F_{y|1,x}(q(\tau|x)) = \frac{P(D=1|Y > q(\tau|x), X=x) \times (1 - F_{y|1,x}(q(\tau|x)))}{p(x)} = \frac{p_U(\tau, x) \times (1-\tau)}{p(x)} \tag{78}$$

Finally, we note that the same manipulations applied to $F_{y|0,x}$ instead of $F_{y|1,x}$ enable us to obtain:

$$F_{y|0,x}(q(\tau|x)) = \frac{(1 - p_L(\tau, x)) \times \tau}{1 - p(x)} \qquad 1 - F_{y|0,x}(q(\tau|x)) = \frac{(1 - p_U(\tau, x)) \times (1-\tau)}{1 - p(x)} \ . \tag{79}$$

Hence, we can obtain by direct algebra from the results (76) and (79) that we must have:

$$|F_{y|1,x}(q(\tau|x)) - F_{y|0,x}(q(\tau|x))| = \frac{|p(x) - p_L(x, \tau)| \times \tau}{p(x)(1 - p(x))} \ . \tag{80}$$

Analogously, exploiting (76) and (79) once again, we can also obtain:

$$\begin{aligned} |F_{y|1,x}(q(\tau|x)) - F_{y|0,x}(q(\tau|x))| &= |(1 - F_{y|1,x}(q(\tau|x))) - (1 - F_{y|0,x}(q(\tau|x)))| \\ &= \frac{|p(x) - p_U(x, \tau)| \times (1-\tau)}{p(x)(1 - p(x))} \ . \end{aligned} \tag{81}$$

Result (62) then follows from taking the square root of the product of (80) and (81).

Appendix E. Proof of Results.

Proof of Lemma 3.1: For any $\theta \in \mathcal{C}(\tau, k)$, the first order condition of the norm minimization problem yields $\beta(\tau) = (E_S[X_i X_i'])^{-1} E_S[X_i \theta(X_i)]$. The Lemma then follows from Corollary 2.1. ■

Proof of Corollary 3.1: Since $\mathcal{P}(\tau, k)$ is convex by Lemma 3.1, it follows that the identified set for $\lambda'\beta(\tau)$ is a convex set in $\mathbf{R}$ and hence an interval. The fact that $\pi_L(\tau, k)$ and $\pi_U(\tau, k)$ are the endpoints of such interval follows directly from Lemma 3.1. ■

Lemma .1. *Let Assumption 2.1 hold, W_i be independent of (Y_i, X_i, D_i) with $E[W_i] = 1$ and $E[W_i^2] < \infty$ and positive almost surely. If $\{Y_i D_i, X_i, D_i, W_i\}$ is an i.i.d. sample, then the following class is Donsker:*

$$\mathcal{M} \equiv \{m_c : m_c(y, x, d, w) \equiv w1\{y \le c, d = 1, x = x_0\} - P(Y_i \le c,\ D_i = 1,\ X_i = x_0),\ c \in \mathbf{R}\} .$$

Proof: For any $1 > \epsilon > 0$, there is an increasing sequence $-\infty = y_0 \le \ldots \le y_{\lceil \frac{8}{\epsilon} \rceil} = +\infty$ such that for any $1 \le j \le \lceil \frac{8}{\epsilon} \rceil$ we have $F_{y|1,x}(y_j) - F_{y|1,x}(y_{j-1}) < \epsilon/4$. Next, define the functions:

$$l_j(y, x, d, w) \equiv w1\{y \le y_{j-1},\ d = 1,\ x = x_0\} - P(Y_i \le y_j,\ D_i = 1,\ X_i = x_0) \tag{82}$$

$$u_j(y, x, d, w) \equiv w1\{y \le y_j,\ d = 1,\ x = x_0\} - P(Y_i \le y_{j-1},\ D_i = 1,\ X_i = x_0) \tag{83}$$

and notice the brackets $\{[l_j, u_j]\}_{j=1}^{\lceil \frac{8}{\epsilon} \rceil}$ form a partition of the class $\mathcal{M}_c$ (since $w \in \mathbf{R}_+$). In addition, note:

$$\begin{aligned} & E[(u_j(Y_i, X_i, D_i, W_i) - l_j(Y_i, X_i, D_i, W_i))^2] \\ & \quad \le 2E[W_i^2 1\{y_{j-1} \le Y_i \le y_j,\ D_i = 1,\ X_i = x_0\}] + 2P^2(y_{j-1} \le Y_i \le y_j,\ D_i = 1,\ X_i = x_0) \\ & \quad \le 4E[W_i^2] \times (F_{y|1,x}(y_j) - F_{y|1,x}(y_{j-1})) , \end{aligned} \tag{84}$$

and hence each bracket has norm bounded by $\sqrt{E[W_i^2]\epsilon}$. Therefore, $N_{[\,]}(\epsilon, \mathcal{M}, \|\cdot\|_{L^2}) \le 16E[W_i^2]/\epsilon^2$, and the Lemma follows by Theorem 2.5.6 in van der Vaart and Wellner (1996). ■

Lemma .2. *Let Assumption 2.1 hold, W_i be independent of (Y_i, X_i, D_i) with $E[W_i] = 1$, $E[W_i^2] < \infty$ and positive almost surely. Also let $\mathcal{S}_\epsilon \equiv \{(\tau, b) \in [0, 1]^2 : b\{1 - p(x)\} + \epsilon \le \tau \le p(x) + b\{1 - p(x)\} - \epsilon\ \forall x \in \mathcal{X}\}$ for some ϵ satisfying $0 < 2\epsilon < \inf_{x \in \mathcal{X}} p(x)$ and denote the minimizers:*

$$s_0(\tau, b, x) = \arg\min_{c \in \mathbf{R}} Q_x(c|\tau, b) \qquad \hat{s}_0(\tau, b, x) \in \arg\min_{c \in \mathbf{R}} \tilde{Q}_{x,n}(c|\tau, b) . \tag{85}$$

Then $s_0(\tau, b, x)$ is bounded in $(\tau, b, x) \in \mathcal{S}_\epsilon \times \mathcal{X}$ and if $\{Y_i D_i, X_i, D_i, W_i\}$ is i.i.d. then for some $M > 0$:

$$P\Big(\sup_{x \in \mathcal{X}} \sup_{(\tau, b) \in \mathcal{S}_\epsilon} |\hat{s}_0(\tau, b, x)| > M\Big) = o(1) .$$

Proof: First note that Assumption 2.1(ii) implies $s_0(\tau, b, x)$ is uniquely defined, while $\hat{s}_0(\tau, b, x)$ may be one of multiple minimizers. By Assumption 2.1(ii) and the definition of $\mathcal{S}_\epsilon$, it follows that the equality:

$$P(Y_i \le s_0(\tau, b, x), D_i = 1 | X_i = x) = \tau - bP(D_i = 0 | X_i = x) \tag{86}$$

implicitly defines $s_0(\tau, b, x)$. Let $\bar{s}(x)$ and $\underline{s}(x)$ be the unique numbers satisfying $F_{y|1,x}(\bar{s}(x)) \times p(x) = p(x) - \epsilon$

and $F_{y|1,x}(\underline{s}(x)) \times p(x) = \epsilon$. By result (86) and the definition of $\mathcal{S}_\epsilon$ we then obtain that for all $x \in \mathcal{X}$:

$$-\infty < \underline{s}(x) \leq \inf_{(\tau,b)\in\mathcal{S}_\epsilon} s_0(\tau,b,x) \leq \sup_{(\tau,b)\in\mathcal{S}_\epsilon} s_0(\tau,b,x) \leq \bar{s}(x) < +\infty \; . \tag{87}$$

Hence, we conclude that $\sup_{(\tau,b)\in\mathcal{S}_\epsilon} |s_0(\tau,b,x)| = O(1)$ and the first claim follows by $\mathcal{X}$ being finite.

In order to establish the second claim of the Lemma, we define the functions:

$$R_x(\tau,b) \equiv bP(D_i = 0, X_i = x) - \tau P(X_i = x) \tag{88}$$

$$R_{x,n}(\tau,b) \equiv \frac{1}{n}\sum_{i=1}^{n} W_i(b1\{D_i = 0, X_i = x\} - \tau 1\{X_i = x\}) \tag{89}$$

as well as the maximizers and minimizers of $R_{x,n}(\tau,b)$ on the set $\mathcal{S}_\epsilon$, which we denote by:

$$(\underline{\tau}_n(x), \underline{b}_n(x)) \in \arg\max_{(\tau,b)\in\mathcal{S}_\epsilon} R_{x,n}(\tau,b) \qquad (\bar{\tau}_n(x), \bar{b}_n(x)) \in \arg\min_{(\tau,b)\in\mathcal{S}_\epsilon} R_{x,n}(\tau,b) \; . \tag{90}$$

Also denote the set of maximizers and minimizers of $\tilde{Q}_{x,n}(c|\tau,b)$ at these particular choices of (τ,b) by:

$$\underline{S}_n(x) \equiv \Big\{\underline{s}_n(x) \in \mathbf{R} : \underline{s}_n(x) \in \arg\min_{c\in\mathbf{R}} \tilde{Q}_{x,n}(c|\underline{\tau}_n(x), \underline{b}_n(x))\Big\} \tag{91}$$

$$\bar{S}_n(x) \equiv \Big\{\bar{s}_n(x) \in \mathbf{R} : \bar{s}_n(x) \in \arg\min_{c\in\mathbf{R}} \tilde{Q}_{x,n}(c|\bar{\tau}_n(x), \bar{b}_n(x))\Big\} \tag{92}$$

From the definition of $\tilde{Q}_{x,n}(c|\tau,b)$, we then obtain from (90), (91) and (92) that for all $x \in \mathcal{X}$:

$$\inf_{\underline{s}_n(x)\in\underline{S}_n(x)} \underline{s}_n(x) \leq \inf_{(\tau,b)\in\mathcal{S}_\epsilon} \hat{s}_0(\tau,b,x) \leq \sup_{(\tau,b)\in\mathcal{S}_\epsilon} \hat{s}_0(\tau,b,x) \leq \sup_{\bar{s}_n(x)\in\bar{S}_n(x)} \bar{s}_n(x) \; . \tag{93}$$

We establish the second claim of the Lemma, by exploiting (93) and showing that for some $0 < M < \infty$:

$$P\Big(\inf_{\underline{s}_n(x)\in\underline{S}_n(x)} \underline{s}_n(x) < -M\Big) = o(1) \qquad P\Big(\sup_{\bar{s}_n(x)\in\bar{S}_n(x)} \bar{s}_n(x) > M\Big) = o(1) \; . \tag{94}$$

To prove that $\inf_{\underline{s}_n(x)\in\underline{S}_n(x)} \underline{s}_n(x)$ is larger than $-M$ with probability tending to one, note that:

$$|R_{x,n}(\underline{\tau}_n(x), \underline{b}_n(x)) + \epsilon P(X_i = x)| = |R_{x,n}(\underline{\tau}_n(x), \underline{b}_n(x)) - \max_{(\tau,b)\in\mathcal{S}_\epsilon} R_x(\tau,b)| = o_p(1) \; , \tag{95}$$

where the second equality follows from the Theorem of the Maximum and the continuous mapping theorem. Therefore, using the equality $a^2 - b^2 = (a-b)(a+b)$, result (95) and Lemma .1, it follows that:

$$\sup_{c\in\mathbf{R}} |\tilde{Q}_{x,n}(c|\underline{\tau}_n(x), \underline{b}_n(x)) - (F_{y|1,x}(c)p(x) - \epsilon)^2 P^2(X_i = x)| = o_p(1) \; . \tag{96}$$

Fix $\delta > 0$ and note that since $F_{y|1,x}(\underline{s}(x))p(x) = \epsilon$ and $\epsilon/p(x) < 1$, Assumption 2.1(ii) implies that:

$$\eta \equiv \inf_{|c-\underline{s}(x)|>\delta} (F_{y|1,x}(c)p(x) - \epsilon)^2 > 0 \; . \tag{97}$$

Therefore, it follows from direct manipulations and the definition of $\underline{S}_n(x)$ in (91) and of $\underline{s}(x)$ that:

$$\begin{aligned} P\Big(|\inf_{\underline{s}_n(x)\in\underline{S}_n(x)} \underline{s}_n(x) - \underline{s}(x)| > \delta\Big) &\leq P\Big(\inf_{|c-\underline{s}(x)|>\delta} \tilde{Q}_{x,n}(c|\underline{\tau}_n(x), \underline{b}_n(x)) \leq \tilde{Q}_{x,n}(\underline{s}(x)|\underline{\tau}_n(x), \underline{b}_n(x))\Big) \\ &\leq P\Big(\eta \leq \sup_{c\in\mathbf{R}} 2|\tilde{Q}_{x,n}(c|\underline{\tau}_n(x), \underline{b}_n(x)) - (F_{y|1,x}(c)p(x) - \epsilon)^2 P^2(X_i = x)|\Big) \; . \end{aligned}$$

We hence conclude from (96) that $\inf_{\underline{s}_n(x)\in\underline{S}_n(x)} \underline{s}_n(x) \xrightarrow{p} \underline{s}(x)$, which together with (87) implies that

$\inf_{\underline{s}_n(x)\in\underline{S}_n(x)} \underline{s}_n(x)$ is larger than $-M$ with probability tending to one for some $M > 0$. By similar arguments it can be shown that $\sup_{\bar{s}_n(x)\in\bar{S}_n(x)} \bar{s}_n(x) \xrightarrow{p} \bar{s}(x)$ which together with (87) establishes (94). The second claim of the Lemma then follows from (93), (94) and $\mathcal{X}$ being finite. ■

Lemma .3. *Let Assumption 2.1 hold, W_i be independent of (Y_i, X_i, D_i) with $E[W_i] = 1$, $E[W_i^2] < \infty$ and positive almost surely. Also let $\mathcal{S}_\epsilon \equiv \{(\tau, b) \in [0,1]^2 : b\{1-p(x)\} + \epsilon \leq \tau \leq p(x) + b\{1-p(x)\} - \epsilon \ \forall x \in \mathcal{X}\}$ for some ϵ satisfying $0 < 2\epsilon < \inf_{x\in\mathcal{X}} p(x)$ and denote the minimizers:*

$$s_0(\tau, b, x) = \arg\min_{c\in\mathbf{R}} Q_x(c|\tau, b) \qquad \hat{s}_0(\tau, b, x) \in \arg\min_{c\in\mathbf{R}} \tilde{Q}_{x,n}(c|\tau, b) \ . \tag{98}$$

If $\{Y_iD_i, X_i, D_i, W_i\}$ is an i.i.d. sample, then $\sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in\mathcal{S}_\epsilon} |\hat{s}_0(\tau, b, x) - s_0(\tau, b, x)| = o_p(1)$.

Proof: First define the criterion functions $M : L^\infty(\mathcal{S}_\epsilon \times \mathcal{X}) \to \mathbf{R}$ and $M_n : L^\infty(\mathcal{S}_\epsilon \times \mathcal{X}) \to \mathbf{R}$ by:

$$M(\theta) \equiv \sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in\mathcal{S}_\epsilon} Q_x(\theta(\tau, b, x)|\tau, b) \qquad M_n(\theta) \equiv \sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in\mathcal{S}_\epsilon} \tilde{Q}_{x,n}(\theta(\tau, b, x)|\tau, b) \ . \tag{99}$$

For notational convenience, let $s_0 \equiv s_0(\cdot,\cdot,\cdot)$ and $\hat{s}_0 \equiv \hat{s}_0(\cdot,\cdot,\cdot)$. By Lemma .2, $s_0 \in L^\infty(\mathcal{S}_\epsilon \times \mathcal{X})$ while with probability tending to one $\hat{s}_0 \in L^\infty(\mathcal{S}_\epsilon \times \mathcal{X})$. Hence, (98) implies that with probability tending to one:

$$\hat{s}_0 \in \arg\min_{\theta\in L^\infty(\mathcal{S}_\epsilon\times\mathcal{X})} M_n(\theta) \qquad s_0 = \arg\min_{\theta\in L^\infty(\mathcal{S}_\epsilon\times\mathcal{X})} M(\theta) \ . \tag{100}$$

By Assumption 2.1(ii) and (86), $Q_x(\cdot|\tau, b)$ is strictly convex in a neighborhood of $s_0(\tau, b, x)$. Furthermore, since by (86) and the implicit function theorem $s_0(\tau, b, x)$ is continuous in $(\tau, b) \in \mathcal{S}_\epsilon$ for every $x \in \mathcal{X}$:

$$\begin{aligned}\inf_{\|\theta-s_0\|_\infty\geq\delta} M(\theta) &\geq \inf_{x\in\mathcal{X}} \inf_{(\tau,b)\in\mathcal{S}_\epsilon} \inf_{|c-s_0(\tau,b,x)|\geq\delta} Q_x(c|\tau, b) \\ &= \inf_{x\in\mathcal{X}} \inf_{(\tau,b)\in\mathcal{S}_\epsilon} \min\{Q_x(s_0(\tau, b, x) - \delta|\tau, b), Q_x(s_0(\tau, b, x) + \delta|\tau, b)\} > 0 \ ,\end{aligned} \tag{101}$$

where the final inequality follows by compactness of $\mathcal{S}_\epsilon$ which together with continuity of $s_0(\tau, b, x)$ implies the inner infimum is attained, while the outer infimum is trivially attained due to $\mathcal{X}$ being finite. Since (101) holds for any $\delta > 0$, s_0 is a well separated minimum of $M(\theta)$ in $L^\infty(\mathcal{S}_\epsilon \times \mathcal{X})$. Next define:

$$G_{x,i}(c) \equiv W_i 1\{Y_i \leq c, D_i = 1, X_i = x\} \tag{102}$$

and observe that compactness of $\mathcal{S}_\epsilon$, a regular law of large numbers, Lemma .1 and finiteness of $\mathcal{X}$ yields:

$$\begin{aligned}&\sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in\mathcal{S}_\epsilon} \sup_{c\in\mathbf{R}} |\frac{1}{n}\sum_{i=1}^n G_{x,i}(c) + R_{x,n}(\tau, b) - E[G_{x,i}(c)] - R_x(\tau, b)| \\ &\leq \sup_{x\in\mathcal{X}} \sup_{c\in\mathbf{R}} |\frac{1}{n}\sum_{i=1}^n G_{x,i}(c) - E[G_{x,i}(c)]| + \sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in\mathcal{S}_\epsilon} |R_{x,n}(\tau, b) - R_x(\tau, b)| = o_p(1) \ ,\end{aligned} \tag{103}$$

where $R_x(\tau, b)$ and $R_{x,n}(\tau, b)$ are as in (88) and (89) respectively. Hence, using (103), the equality $a^2 - b^2 = (a-b)(a+b)$ and $Q_x(c|\tau, b)$ uniformly bounded in $(c, \tau, b) \in \mathbf{R} \times \mathcal{S}_\epsilon$ due to the compactness of $\mathcal{S}_\epsilon$, we obtain:

$$\sup_{\theta\in L^\infty(\mathcal{S}_\epsilon\times\mathcal{X})}|M_n(\theta)-M(\theta)|\leq \sup_{\theta\in L^\infty(\mathcal{S}_\epsilon\times\mathcal{X})}\sup_{x\in\mathcal{X}}\sup_{(\tau,b)\in\mathcal{S}_\epsilon}|\tilde{Q}_{x,n}(\theta(\tau,b,x)|\tau,b)-Q_x(\theta(\tau,b,x)|\tau,b)|$$
$$\leq \sup_{x\in\mathcal{X}}\sup_{(\tau,b)\in\mathcal{S}_\epsilon}\sup_{c\in\mathbf{R}}|\tilde{Q}_{x,n}(c|\tau,b)-Q_x(c|\tau,b)|=o_p(1)\ . \quad (104)$$

The claim of the Lemma then follows from results (100), (101) and (104) together with Corollary 3.2.3 in van der Vaart and Wellner (1996). ■

Lemma .4. *Let Assumptions 2.1, 4.1 hold, W_i be independent of (Y_i, X_i, D_i) with $E[W_i]=1$, $E[W_i^2]<\infty$ and positive a.s. Also let $\mathcal{S}_\epsilon\equiv\{(\tau,b)\in[0,1]^2 : b\{1-p(x)\}+\epsilon\leq\tau\leq p(x)+b\{1-p(x)\}-\epsilon\ \forall x\in\mathcal{X}\}$ for some ϵ satisfying $0<2\epsilon<\inf_{x\in\mathcal{X}}p(x)$ and denote the minimizers:*

$$s_0(\tau,b,x)=\arg\min_{c\in\mathbf{R}}Q_x(c|\tau,b)\qquad\qquad \hat{s}_0(\tau,b,x)\in\arg\min_{c\in\mathbf{R}}\tilde{Q}_{x,n}(c|\tau,b)\ . \quad (105)$$

For $G_{x,i}(c)\equiv W_i1\{Y_i\leq c, D_i=1, X_i=x\}$ and $R_{x,n}(\tau,b)$ as defined in (89), denote the criterion function:

$$\tilde{Q}^s_{x,n}(c|\tau,b)\equiv\Big(\frac{1}{n}\sum_{i=1}^n\{E[G_{x,i}(c)-G_{x,i}(s_0(\tau,b,x))]+G_{x,i}(s_0(\tau,b,x))\}+R_{x,n}(\tau,b)\Big)^2\ . \quad (106)$$

If $\{Y_iD_i, X_i, D_i, W_i\}$ is an i.i.d. sample, it then follows that:

$$\sup_{x\in\mathcal{X}}\sup_{(\tau,b)\in\mathcal{S}_\epsilon}\Big|\frac{d\tilde{Q}^s_{x,n}(\hat{s}_0(\tau,b,x)|\tau,b)}{dc}\Big|=o_p(n^{-\frac{1}{2}})\ . \quad (107)$$

Proof: We first introduce the criterion function $M_n^s : L^\infty(\mathcal{S}_\epsilon\times\mathcal{X})\to\mathbf{R}$ to be given by:

$$M_n^s(\theta)\equiv\sup_{x\in\mathcal{X}}\sup_{(\tau,b)\in\mathcal{S}_\epsilon}\tilde{Q}^s_{x,n}(\theta(\tau,b,x)|\tau,b)\ . \quad (108)$$

We aim to characterize and establish the consistency of an approximate minimizer of $M_n^s(\theta)$ on $L^\infty(\mathcal{S}_\epsilon\times\mathcal{X})$. Observe that by Lemma .1, compactness of $\mathcal{S}_\epsilon$, finiteness of $\mathcal{X}$ and the law of large numbers:

$$\sup_{x\in\mathcal{X}}\sup_{(\tau,b)\in\mathcal{S}_\epsilon}|\frac{1}{n}\sum_{i=1}^n\{G_{x,i}(s_0(\tau,b,x))-E[G_{x,i}(s_0(\tau,b,x))]\}+R_{x,n}(\tau,b)-R_x(\tau,b)|$$
$$\leq\sup_{x\in\mathcal{X}}\sup_{c\in\mathbf{R}}|\frac{1}{n}\sum_{i=1}^n\{G_{x,i}(c)-E[G_{x,i}(c)]\}|+\sup_{x\in\mathcal{X}}\sup_{(\tau,b)\in\mathcal{S}_\epsilon}|R_{x,n}(\tau,b)-R_x(\tau,b)|=o_p(1)\ , \quad (109)$$

where $R_x(\tau,b)$ is as in (88). Hence, by definition of $\mathcal{S}_\epsilon$ and $R_x(\tau,b)$, with probability tending to one:

$$\frac{\epsilon}{2}P(X_i=x)\leq\frac{1}{n}\sum_{i=1}^n\{E[G_{x,i}(s_0(\tau,b,x))]-G_{x,i}(s_0(\tau,b,x))\}-R_{x,n}(\tau,b)$$
$$\leq(p(x)-\frac{\epsilon}{2})P(X_i=x)\qquad\qquad \forall(\tau,b,x)\in\mathcal{S}_\epsilon\times\mathcal{X}\ . \quad (110)$$

By Assumption 2.1(ii), whenever (110) holds, we may implicitly define $\hat{s}^s_0(\tau,b,x)$ by the equality:

$$P(Y_i\leq\hat{s}^s_0(\tau,b,x), D_i=1, X_i=x)=\frac{1}{n}\sum_{i=1}^n\{E[G_{x,i}(s_0(\tau,b,x))]-G_{x,i}(s_0(\tau,b,x))\}-R_{x,n}(\tau,b)\ , \quad (111)$$

for all $(\tau, b, x) \in \mathcal{S}_\epsilon \times \mathcal{X}$ and set $\hat{s}_0^s(\tau, b, x) = 0$ for all $(\tau, b, x) \in \mathcal{S}_\epsilon \times \mathcal{X}$ whenever (110) does not hold. Thus,

$$\sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in\mathcal{S}_\epsilon} |\tilde{Q}^s_{x,n}(\hat{s}_0^s(\tau, b, x)|\tau, b) - \inf_{c\in\mathbf{R}} \tilde{Q}^s_{x,n}(c|\tau, b)| = o_p(n^{-1}) \ . \tag{112}$$

Let $\hat{s}_0^s \equiv \hat{s}_0^s(\cdot,\cdot,\cdot)$ and note that by construction $\hat{s}_0^s \in L^\infty(\mathcal{S}_\epsilon \times \mathcal{X})$. From (112) we then obtain that:

$$M_n^s(\hat{s}_0^s) \leq \sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in\mathcal{S}_\epsilon} \inf_{c\in\mathbf{R}} \tilde{Q}^s_{x,n}(c|\tau, b) + o_p(n^{-1}) \leq \inf_{\theta\in L^\infty(\mathcal{S}_\epsilon\times\mathcal{X})} M_n^s(\theta) + o_p(n^{-1}) \ . \tag{113}$$

In order to establish $\|\hat{s}_0^s - s_0\|_\infty = o_p(1)$, let $M(\theta)$ be as in (99) and notice that arguing as in (104) together with result (109) and Lemma .1 implies that:

$$\sup_{\theta\in L^\infty(\mathcal{S}_\epsilon\times\mathcal{X})} |M_n^s(\theta) - M(\theta)| \leq \sup_{\theta\in L^\infty(\mathcal{S}_\epsilon\times\mathcal{X})} \sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in\mathcal{S}_\epsilon} |\tilde{Q}^s_{x,n}(\theta(\tau, b, x)|\tau, b) - Q_x(\theta(\tau, b, x)|\tau, b)|$$
$$\leq \sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in\mathcal{S}_\epsilon} \sup_{c\in\mathbf{R}} |\tilde{Q}^s_{x,n}(c|\tau, b) - Q_x(c|\tau, b)| = o_p(1) \ . \tag{114}$$

Hence, by (101), (113), (114) and Corollary 3.2.3 in van der Vaart and Wellner (1996) we obtain:

$$\sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in\mathcal{S}_\epsilon} |\hat{s}_0^s(\tau, b, x) - s_0(\tau, b, x)| = o_p(1) \ . \tag{115}$$

Next, define the random mapping $\Delta_n : L^\infty(\mathcal{S}_\epsilon \times \mathcal{X}) \to L^\infty(\mathcal{S}_\epsilon \times \mathcal{X})$ to be given by:

$$\Delta_n(\theta)(\tau, b, x) \equiv \frac{1}{n}\sum_{i=1}^n \{(G_{x,i}(\theta(\tau, b, x)) - E[G_{x,i}(\theta(\tau, b, x))]) - (G_{x,i}(s_0(\tau, b, x)) - E[G_{x,i}(s_0(\tau, b, x))])\} \ , \tag{116}$$

and observe that Lemma .1 and finiteness of $\mathcal{X}$ implies that $\|\Delta_n(\bar{s})\|_\infty = o_p(n^{-\frac{1}{2}})$ for any $\bar{s} \in L^\infty(\mathcal{S}_\epsilon \times \mathcal{X})$ such that $\|\bar{s} - s_0\|_\infty = o_p(1)$. Since $\tilde{Q}_{x,n}(\hat{s}_0(\tau, b, x)|\tau, b) \leq \tilde{Q}_{x,n}(s_0(\tau, b, x)|\tau, b)$ for all $(\tau, b, x) \in \mathcal{S}_\epsilon \times \mathcal{X}$, and by Lemma .1 and finiteness of $\mathcal{X}$, $\sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in\mathcal{S}_\epsilon} \tilde{Q}_{x,n}(s_0(\tau, b, x)|\tau, b) = O_p(n^{-1})$, we conclude that:

$$\sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in\mathcal{S}_\epsilon} \{\tilde{Q}^s_{x,n}(\hat{s}_0(\tau, b, x)|\tau, b) - \tilde{Q}^s_{x,n}(\hat{s}_0^s(\tau, b, x)|\tau, b)\}$$
$$\leq \sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in\mathcal{S}_\epsilon} \{\tilde{Q}_{x,n}(\hat{s}_0(\tau, b, x)|\tau, b) - \tilde{Q}^s_{x,n}(\hat{s}_0^s(\tau, b, x)|\tau, b)\} + \|\Delta_n^2(\hat{s}_0)\|_\infty + 2\|\Delta_n(\hat{s}_0)\|_\infty \times M_n^{\frac{1}{2}}(\hat{s}_0)$$
$$\leq \sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in\mathcal{S}_\epsilon} \{\tilde{Q}_{x,n}(\hat{s}_0^s(\tau, b, x)|\tau, b) - \tilde{Q}^s_{x,n}(\hat{s}_0^s(\tau, b, x)|\tau, b)\} + o_p(n^{-1}) \ , \tag{117}$$

where $M_n(\theta)$ is as in (99). Furthermore, since by (113) we have $M_n^s(\hat{s}_0^s) \leq M_n^s(s_0) + o_p(n^{-1})$ and by Lemma .1 and finiteness of $\mathcal{X}$ we have $M_n^s(s_0) = O_p(n^{-1})$, similar arguments as in (117) imply that:

$$\sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in\mathcal{S}_\epsilon} \{\tilde{Q}_{x,n}(\hat{s}_0^s(\tau, b, x)|\tau, b) - \tilde{Q}^s_{x,n}(\hat{s}_0^s(\tau, b, x)|\tau, b)\}$$
$$\leq \|\Delta_n(\hat{s}_0^s)\|^2_\infty + 2\|\Delta_n(\hat{s}_0^s)\|_\infty \times [M_n^s(\hat{s}_0^s)]^{\frac{1}{2}} = o_p(n^{-1}) \ . \tag{118}$$

Therefore, by combining the results in (112), (117) and (118), we are able to conclude that:

$$\sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in\mathcal{S}_\epsilon} \{\tilde{Q}^s_{x,n}(\hat{s}_0(\tau, b, x)|\tau, b) - \inf_{c\in\mathbf{R}} \tilde{Q}^s_{x,n}(c|\tau, b)\}$$
$$\leq \sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in\mathcal{S}_\epsilon} \{\tilde{Q}^s_{x,n}(\hat{s}_0(\tau, b, x)|\tau, b) - \tilde{Q}^s_{x,n}(\hat{s}_0^s(\tau, b, x)|\tau, b)\} + o_p(n^{-1}) \leq o_p(n^{-1}) \ . \tag{119}$$

Let $\epsilon_n \searrow 0$ be such that $\epsilon_n = o_p(n^{-\frac{1}{2}})$ and in addition satisfies:

$$\sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in\mathcal{S}_\epsilon} |\tilde{Q}^s_{x,n}(\hat{s}_0(\tau,b,x)|\tau,b) - \inf_{c\in\mathbf{R}} \tilde{Q}^s_{x,n}(c|\tau,b)| = o_p(\epsilon_n^2) \ , \tag{120}$$

which is possible by (119). A Taylor expansion at each $(\tau,b,x)\in\mathcal{S}_\epsilon\times\mathcal{X}$ then implies:

$$\begin{aligned} 0 &\leq \sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in\mathcal{S}_\epsilon} \{\tilde{Q}^s_{x,n}(\hat{s}_0(\tau,b,x)+\epsilon_n|\tau,b) - \tilde{Q}^s_{x,n}(\hat{s}_0(\tau,b,x)|\tau,b)\} + o_p(\epsilon_n^2) \\ &= \sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in S} \Big\{\epsilon_n \times \frac{d\tilde{Q}^s_{x,n}(\hat{s}_0(\tau,b,x)|\tau,b)}{dc} + \frac{\epsilon_n^2}{2} \times \frac{d^2\tilde{Q}^s_{x,n}(\bar{s}(\tau,b,x)|\tau,b)}{dc^2}\Big\} + o_p(\epsilon_n^2) \ , \end{aligned} \tag{121}$$

where $\bar{s}(\tau,b,x)$ is a convex combination of $\hat{s}_0(\tau,b,x)$ and $\hat{s}_0(\tau,b,x)+\epsilon_n$. Since Lemma .3 and $\epsilon_n\searrow 0$ imply that $\|\bar{s}-s_0\|_\infty = o_p(1)$, the mean value theorem, $f_{y|1,x}(c)$ being uniformly bounded and (104) yield:

$$\begin{aligned} &\sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in\mathcal{S}_\epsilon} \Big|\frac{1}{n}\sum_{i=1}^n \{E[G_{x,i}(\bar{s}(\tau,b,x)) - G_{x,i}(s_0(\tau,b,x))] + G_{x,i}(s_0(\tau,b,x))\} + R_{x,n}(\tau,b)\Big| \\ &\qquad\qquad \leq \sup_{c\in\mathbf{R}} f_{y|1,x}(c)p(x)P(X_i=x)\times\|\bar{s}-s_0\|_\infty + M_n^{\frac{1}{2}}(s_0) = o_p(1) \ . \end{aligned} \tag{122}$$

Therefore, exploiting (122), $f'_{y|1,x}(c)$ being uniformly bounded and by direct calculation we conclude:

$$\begin{aligned} &\sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in\mathcal{S}_\epsilon} \Big|\frac{d^2\tilde{Q}^s_{x,n}(\bar{s}(\tau,b,x)|\tau,b)}{dc^2} - 2f^2_{y|1,x}(\bar{s}(\tau,b,x))p^2(x)P^2(X_i=x)\Big| \\ &\qquad\qquad \leq \sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in\mathcal{S}_\epsilon} |f'_{y|1,x}(\bar{s}(\tau,b,x))p(x)P(X_i=x)| \times o_p(1) = o_p(1). \end{aligned} \tag{123}$$

Thus, combining results (121) together with (123) and $f_{y|1,x}(c)$ uniformly bounded, we conclude:

$$0 \leq \epsilon_n \times \sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in\mathcal{S}_\epsilon} \frac{d\tilde{Q}^s_{x,n}(\hat{s}_0(\tau,b,x)|\tau,b)}{dc} + O_p(\epsilon_n^2) \ . \tag{124}$$

In a similar fashion, we note that by exploiting (120) and proceeding as in (121)-(124) we obtain:

$$\begin{aligned} 0 &\leq \inf_{x\in\mathcal{X}} \inf_{(\tau,b)\in\mathcal{S}_\epsilon} \{\tilde{Q}^s_{x,n}(\hat{s}_0(\tau,b,x)-\epsilon_n|\tau,b) - \tilde{Q}^s_{x,n}(\hat{s}_0(\tau,b,x)|\tau,b)\} + o_p(\epsilon_n^2) \\ &\leq \inf_{x\in\mathcal{X}} \inf_{(\tau,b)\in\mathcal{S}_\epsilon} \Big\{-\epsilon_n \times \frac{d\tilde{Q}_{x,n}(\hat{s}_0(\tau,b,x)|\tau,b)}{dc} + \frac{\epsilon_n^2}{2} \times \frac{d^2\tilde{Q}^s_{x,n}(\bar{s}(\tau,b,x)|\tau,b)}{dc^2}\Big\} + o_p(\epsilon_n^2) \\ &\leq -\epsilon_n \times \sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in\mathcal{S}_\epsilon} \frac{d\tilde{Q}_{x,n}(\hat{s}_0(\tau,b,x)|\tau,b)}{dc} + O_p(\epsilon_n^2) \ . \end{aligned} \tag{125}$$

Therefore, since $\epsilon_n = o_p(n^{-\frac{1}{2}})$, we conclude from (124) and (125) that we must have:

$$\sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in\mathcal{S}_\epsilon} \frac{d\tilde{Q}^s_{x,n}(\hat{s}_0(\tau,b,x)|\tau,b)}{dc} = O_p(\epsilon_n) = o_p(n^{-\frac{1}{2}}) \ . \tag{126}$$

By similar arguments, but reversing the sign of ϵ_n in (121) and (125) it possible to establish that:

$$\sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in\mathcal{S}_\epsilon} -\frac{d\tilde{Q}^s_{x,n}(\hat{s}_0(\tau,b,x)|\tau,b)}{dc} = o_p(n^{-\frac{1}{2}}) \ . \tag{127}$$

The claim of the Lemma then follows from (126) and (127). ■

Lemma .5. *Let Assumptions 2.1, 4.1 hold, W_i be independent of (Y_i,X_i,D_i) with $E[W_i]=1$, $E[W_i^2]<\infty$*

and positive a.s. Also let $\mathcal{S}_\epsilon \equiv \{(\tau, b) \in [0,1]^2 : b\{1-p(x)\} + \epsilon \leq \tau \leq p(x) + b\{1-p(x)\} - \epsilon \ \forall x \in \mathcal{X}\}$ *for some* ϵ *satisfying* $0 < \epsilon < \inf_{x\in\mathcal{X}} p(x)$ *and denote the minimizers:*

$$s_0(\tau, b, x) = \arg\min_{c\in\mathbf{R}} Q_x(c|\tau, b) \qquad \hat{s}_0(\tau, b, x) \in \arg\min_{c\in\mathbf{R}} \tilde{Q}_{x,n}(c|\tau, b) \ . \tag{128}$$

If $G_{x,i}(c)$ *is as in* (102), $\inf_{x\in\mathcal{X}} \inf_{(\tau,b)\in\mathcal{S}_\epsilon} f_{y|1,x}(s_0(\tau, b, x))p(x) > 0$ *and* $\{Y_iD_i, X_i, D_i, W_i\}$ *is i.i.d., then:*

$$\sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in\mathcal{S}_\epsilon} \Big|(\hat{s}_0(\tau, b, x) - s_0(\tau, b, x)) - \frac{1}{n}\sum_{i=1}^{n} \frac{G_{x,i}(s_0(\tau, b, x)) + W_i(b1\{D_i = 0, X_i = x\} - \tau 1\{X_i = x\})}{P(X_i = x)p(x)f_{y|1,x}(s_0(\tau, b, x))}\Big| = o_p(n^{-\frac{1}{2}}) \ . \tag{129}$$

PROOF: For $\tilde{Q}^s_{x,n}(c|\tau, b)$ as in (106), note that the mean value theorem and Lemma .4 imply:

$$\sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in\mathcal{S}_\epsilon} \Big|(\hat{s}_0(\tau, b, x) - s_0(\tau, b, x)) \times \frac{d^2\tilde{Q}^s_{x,n}(\bar{s}(\tau, b, x)|\tau, b)}{dc^2} + \frac{d\tilde{Q}^s_{x,n}(s_0(\tau, b, x)|\tau, b)}{dc}\Big| = o_p(n^{-\frac{1}{2}}) \tag{130}$$

for $\bar{s}(\tau, b, x)$ a convex combination of $s_0(\tau, b, x)$ and $\hat{s}_0(\tau, b, x)$. Also note that Lemma .1 implies:

$$\sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in\mathcal{S}_\epsilon} \Big|\frac{d\tilde{Q}^s_{x,n}(s_0(\tau, b, x)|\tau, b)}{dc}\Big| = \sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in\mathcal{S}_\epsilon} \Big|2f_{y|1,x}(s_0(\tau, b, x))p(x)P(X_i = x) \times \{\frac{1}{n}\sum_{i=1}^{n} G_{x,i}(s_0(\tau, b, x)) + R_n(\tau, b)\}\Big| = O_p(n^{-\frac{1}{2}}) \ . \tag{131}$$

In addition, by (123), the mean value theorem and $f_{y|1,x}(c)$ being uniformly bounded we conclude that:

$$\sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in\mathcal{S}_\epsilon} \Big|\frac{d^2\tilde{Q}_{x,n}(\bar{s}(\tau, b, x)|\tau, b)}{dc^2} - 2f^2_{y|1,x}(s_0(\tau, b, x))p^2(x)P^2(X_i = 1)\Big| \lesssim \sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in\mathcal{S}_\epsilon} |f^2_{y|1,x}(\bar{s}(\tau, b, x)) - f^2_{y|1,x}(s_0(\tau, b, x))| + o_p(1) \lesssim \sup_{c\in\mathbf{R}} |f'_{y|1,x}(c)| \times \|\bar{s} - s_0\|_\infty + o_p(1) \ . \tag{132}$$

Since by assumption $f_{y|1,x}(s_0(\tau, b, x))p(x)$ is bounded away from zero uniformly in $(\tau, b, x) \in \mathcal{S}_\epsilon \times \mathcal{X}$, it follows from (132) and $\|\bar{s} - s_0\|_\infty = o_p(1)$ by Lemma .3 that for some $\delta > 0$:

$$\inf_{x\in\mathcal{X}} \inf_{(\tau,b)\in\mathcal{S}_\epsilon} \frac{d^2\tilde{Q}_{x,n}(\bar{s}(\tau, b, x)|\tau, b)}{dc^2} > \delta \tag{133}$$

with probability approaching one. Therefore, we conclude from results (130), (131) and (133) that we must have $\|\hat{s}_0 - s_0\|_\infty = O_p(n^{-\frac{1}{2}})$. Hence, by (130) and (132) we conclude that:

$$\sup_{x\in\mathcal{X}} \sup_{(\tau,b)\in\mathcal{S}_\epsilon} \Big|2(\hat{s}_0(\tau, b, x) - s_0(\tau, b, x))f^2_{y|1,x}(s_0(\tau, b, x))p^2(x)P^2(X_i = 1) + \frac{d\tilde{Q}^s_{x,n}(s_0(\tau, b, x)|\tau, b)}{dc}\Big| = o_p(n^{-\frac{1}{2}}) \tag{134}$$

The claim of the Lemma is then established by (131), (133) and (134). ■

Lemma .6. *Let Assumptions 2.1, 4.1(ii)-(iii) hold,* W_i *be independent of* (Y_i, X_i, D_i) *with* $E[W_i] = 1$, $E[W_i^2] < \infty$ *and positive a.s. Let* $\mathcal{S}_\epsilon \equiv \{(\tau, b) \in [0,1]^2 : b\{1-p(x)\} + \epsilon \leq \tau \leq p(x) + b\{1-p(x)\} - \epsilon \ \forall x \in \mathcal{X}\}$ *for some* ϵ *satisfying* $0 < 2\epsilon < \inf_{x\in\mathcal{X}} p(x)$ *and for some* $x_0 \in \mathcal{X}$, *denote the minimizers:*

$$s_0(\tau, b, x_0) = \arg\min_{c\in\mathbf{R}} Q_{x_0}(c|\tau, b) \ .$$

If $\inf_{(\tau,b)\in\mathcal{S}_\epsilon} f_{y|1,x}(s_0(\tau,b,x_0))p(x_0) > 0$ *and* $\{Y_iD_i, X_i, D_i, W_i\}$ *is i.i.d., then the following class is Donsker:*

$$\mathcal{F} \equiv \left\{f_{\tau,b}(y,x,d,w) = \frac{w1\{y \le s_0(\tau,b,x_0), d=1, x=x_0\} + bw1\{d=0, x=x_0\} - \tau w1\{x=x_0\}}{P(X_i=x_0)p(x_0)f_{y|1,x}(s_0(\tau,b,x_0))} : (\tau,b)\in\mathcal{S}_\epsilon\right\}$$

PROOF: For $\delta > 0$, let $\{B_j\}$ be a collection of closed balls in $\mathbf{R}^2$ with diameter δ covering $\mathcal{S}_\epsilon$. Further notice that since $\mathcal{S}_\epsilon \subseteq [0,1]^2$, we may select $\{B_j\}$ so its cardinality is less than $4/\delta^2$. On each B_j define:

$$\begin{aligned}
&\underline{\tau}_j = \min_{(\tau,b)\in\mathcal{S}_\epsilon\cap B_j}\tau && \bar{\tau}_j = \max_{(\tau,b)\in\mathcal{S}_\epsilon\cap B_j}\tau\\
&\underline{b}_j = \min_{(\tau,b)\in\mathcal{S}_\epsilon\cap B_j} b && \bar{b}_j = \max_{(\tau,b)\in\mathcal{S}_\epsilon\cap B_j} b\\
&\underline{s}_j = \min_{(\tau,b)\in\mathcal{S}_\epsilon\cap B_j} s_0(\tau,b,x_0) && \bar{s}_j = \max_{(\tau,b)\in\mathcal{S}_\epsilon\cap B_j} s_0(\tau,b,x_0)\\
&\underline{f}_j = \min_{(\tau,b)\in\mathcal{S}_\epsilon\cap B_j} f_{y|1,x}(s_0(\tau,b,x_0)) && \bar{f}_j = \max_{(\tau,b)\in\mathcal{S}_\epsilon\cap B_j} f_{y|1,x}(s_0(\tau,b,x_0))\ ,
\end{aligned} \tag{135}$$

where we note that all minima and maxima are attained due to compactness of $\mathcal{S}_\epsilon \cap B_j$, continuity of $s_0(\tau,b,x_0)$ by (86) and the implicit function theorem and continuity of $f_{y|1,x}(c)$ by assumption 4.1(iii). Next, for $1 \le j \le \#\{B_j\}$ define the functions:

$$l_j(y,x,d,w) \equiv \frac{w1\{y\le \underline{s}_j,\ d=1,\ x=x_0\} + \underline{b}_j w1\{d=0, x=x_0\}}{P(X_i=x_0)p(x_0)\bar{f}_j} - \frac{\bar{\tau}_j w1\{x=x_0\}}{P(X_i=x_0)p(x_0)\underline{f}_j} \tag{136}$$

$$u_j(y,x,d,w) \equiv \frac{w1\{y\le \bar{s}_j,\ d=1,\ x=x_0\} + \bar{b}_j w1\{d=0, x=x_0\}}{P(X_i=x_0)p(x_0)\underline{f}_j} - \frac{\underline{\tau}_j w1\{x=x_0\}}{P(X_i=x_0)p(x_0)\bar{f}_j} \tag{137}$$

and note that the brackets $[l_j, u_j]$ cover the class $\mathcal{F}$. Since $\bar{f}_j^{-1} \le \underline{f}_j^{-1} \le [\inf_{(\tau,b)\in\mathcal{S}_\epsilon} f_{y|1,x}(s_0(\tau,b,x_0))]^{-1} < \infty$ for all j, there is a finite constant M not depending on j so that $M > 3E[W_i^2]P^{-2}(X_i=x_0)p^{-2}(x_0)\underline{f}_j^{-2}\bar{f}_j^{-2}$ uniformly in j. To bound the norm of the bracket $[l_j,u_j]$ note that for such a constant M it follows that:

$$\begin{aligned}
E[(u_j(Y_i,X_i,D_i,W_i) - l_j(Y_i,X_i,D_i,W_i))^2] &\le M\times(\bar{b}_j\bar{f}_j - \underline{b}_j\underline{f}_j)^2 + M\times(\bar{\tau}_j\bar{f}_j - \underline{\tau}_j\underline{f}_j)^2\\
&+ M\times E[(1\{Y_i\le \underline{s}_j, D_i=1, X_i=x_0\}\underline{f}_j - 1\{Y_i\le\bar{s}_j, D_i=1, X_i=x_0\}\bar{f}_j)^2]
\end{aligned} \tag{138}$$

Next observe that by the implicit function theorem and result (86) we can conclude that for any $(\tau,b)\in\mathcal{S}_\epsilon$:

$$\frac{ds_0(\tau,b,x_0)}{d\tau} = \frac{1}{f_{y|1,x}(s_0(\tau,b,x_0))} \qquad \frac{ds_0(\tau,b,x_0)}{db} = -\frac{1-p(x_0)}{f_{y|1,x}(s_0(\tau,b,x_0))}\ . \tag{139}$$

Since the minima and maxima in (135) are attained, it follows that for some $(\tau_1,b_1),(\tau_2,b_2)\in B_j\cap\mathcal{S}_\epsilon$ we have $s_0(\tau_1,b_1.x_0) = \bar{s}_j$ and $s_0(\tau_2,b_2,x_0) = \underline{s}_j$. Hence, the mean value theorem and (139) imply:

$$|\bar{s}_j - \underline{s}_j| = \Big|\frac{1}{f_{y|1,x}(s_0(\tilde{\tau},\tilde{b},x_0))}(\tau_1-\tau_2) + \frac{1-p(x)}{f_{y|1,x}(s_0(\tilde{\tau},\tilde{b},x_0))}(b_1-b_2)\Big| \le \frac{\sqrt{2}\delta}{\inf_{(\tau,b)\in\mathcal{S}_\epsilon} f_{y|1,x}(s_0(\tau,b,x_0))} \tag{140}$$

where $(\tilde{\tau},\tilde{b})$ is between (τ_1,b_1) and (τ_2,b_2) and the final inequality follows by $(\tilde{\tau},\tilde{b})\in\mathcal{S}_\epsilon$ by convexity of $\mathcal{S}_\epsilon$, $(\tau_1,b_1),(\tau_2,b_2)\in B_j\cap\mathcal{S}_\epsilon$ and B_j having diameter δ. By similar arguments, and (140) we conclude:

$$|\bar{f}_j - \underline{f}_j| \le \sup_{c\in\mathbf{R}}|f'_{y|1,x}(c)|\times|\bar{s}_j - \underline{s}_j| \le \sup_{c\in\mathbf{R}}|f'_{y|1,x}(c)|\times\frac{\sqrt{2}\delta}{\inf_{(\tau,b)\in\mathcal{S}_\epsilon} f_{y|1,x}(s_0(\tau,b,x_0))}\ . \tag{141}$$

Since $\underline{b}_j \le \bar{b}_j \le 1$ due to $\bar{b}_j \in [0,1]$ and $|\bar{b}_j - \underline{b}_j| \le \delta$ by B_j having diameter δ, we further obtain that:

$$(\bar{b}_j \bar{f}_j - \underline{b}_j \underline{f}_j)^2 \le 2\bar{f}_j^2(\bar{b}_j - \underline{b}_j)^2 + 2\underline{b}_j^2(\bar{f}_j - \underline{f}_j)^2 \le 2\sup_{c\in\mathbf{R}} f_{y|1,x}^2(c) \times \delta^2 + \frac{4\delta^2}{\inf_{(\tau,b)\in\mathcal{S}_\epsilon} f_{y|1,x}^2(s_0(\tau,b,x_0))} \ , \quad (142)$$

where in the final inequality we have used result (141). By similar arguments, we also obtain:

$$(\bar{\tau}_j \bar{f}_j - \underline{\tau}_j \underline{f}_j)^2 \le 2\sup_{c\in\mathbf{R}} f_{y|1,x}^2(c) \times \delta^2 + \frac{4\delta^2}{\inf_{(\tau,b)\in\mathcal{S}_\epsilon} f_{y|1,x}^2(s_0(\tau,b,x_0))} \ . \quad (143)$$

Also note that by direct calculation, the mean value theorem and results (140) and (141) it follows that:

$$\begin{aligned} E[(1\{Y_i \le \underline{s}_j, D_i = 1, &X_i = x_0\}\underline{f}_j - 1\{Y_i \le \bar{s}_j, D_i = 1, X_i = x_0\}\bar{f}_j)^2] \\ &\le 2(\bar{f}_j - \underline{f}_j)^2 + \sup_{c\in\mathbf{R}} f_{y|1,x}^2(c) \times P(X_i = x_0)p(x_0)(F_{y|1,x}(\bar{s}_j) - F_{y|1,x}(\underline{s}_j)) \\ &\le \frac{4\delta^2}{\inf_{(\tau,b)\in\mathcal{S}_\epsilon} f_{y|1,x}^2(s_0(\tau,b,x_0))} + \sup_{c\in\mathbf{R}} f_{y|1,x}^3(c) \times \frac{\sqrt{2}\delta}{\inf_{(\tau,b)\in\mathcal{S}_\epsilon} f_{y|1,x}(s_0(\tau,b,x_0))} \ . \quad (144) \end{aligned}$$

Thus, from (138) and (142)-(143), it follows that for $\delta < 1$ and some constant K not depending on j:

$$E[(u_j(Y_i,X_i,D_i,W_i) - l_j(Y_i,X_i,D_i,W_i))^2] \le K\delta \ . \quad (145)$$

Since $\#\{B_j\} \le 4/\delta^2$, we can therefore conclude that $N_{[\,]}(\delta,\mathcal{F},\|\cdot\|_{L^2}) \le 4K^2/\delta^2$ and hence by Theorem 2.5.6 in van der Vaart and Wellner (1996) it follows that the class $\mathcal{F}$ is Donsker. ∎

Lemma .7. *Let Assumptions 2.1, 4.1(ii)-(iii) hold, W_i be independent of (Y_i,X_i,D_i) with $E[W_i]=1$, $E[W_i^2]<\infty$, positive a.s., $\mathcal{S}_\zeta \equiv \{(\tau,b)\in[0,1]^2 : b\{1-p(x)\}+\zeta \le \tau \le p(x)+b\{1-p(x)\}-\zeta \forall x\in\mathcal{X}\}$ and*

$$\tilde{p}(x) \equiv \frac{\sum_{i=1}^n W_i 1\{D_i=1, X_i=x\}}{\sum_{i=1}^n W_i 1\{X_i = x\}} \qquad p(x) \equiv P(D_i = 1|X_i = x) \qquad s_0(\tau,b,x) = \arg\min_{c\in\mathbf{R}} Q_x(c|\tau,b) \ .$$

If $\inf_{(\tau,b,x)\in\mathcal{S}_\zeta\times\mathcal{X}} f_{y|1,x}(s_0(\tau,b,x))p(x) > 0$ and $\{Y_iD_i, X_i, D_i, W_i\}$ is an i.i.d. sample, then for $a\in\{-1,1\}$:

$$\begin{aligned} s_0(\tau,\tau+ak\tilde{p}(x),x) &- s_0(\tau,\tau+akp(x),x) \\ &= -\frac{(1-p(x))ka}{f_{y|1,x}(s_0(\tau,\tau+akp(x),x))P(X=x)} \times \frac{1}{n}\sum_{i=1}^n R(X_i,W_i,x) + o_p(n^{-\frac{1}{2}}) \ , \quad (146) \end{aligned}$$

where $R(W_i,X_i,x) = p(x)\{P(X=x) - W_i 1\{X_i=x\}\} + W_i 1\{D_i=1,X_i=x\} - P(D=1,X=x)$ and (146) holds uniformly in $(\mathcal{B}_\zeta\times\mathcal{X})$. Moreover, the right hand side of (146) is Donsker.

Proof: First observe that $(\tau,k)\in\mathcal{B}_\zeta$ implies $(\tau,\tau+akp(x))\in\mathcal{S}_\zeta$ for all $x\in\mathcal{X}$, and that with probability tending to one $(\tau,\tau+ak\tilde{p}(x))\in\mathcal{S}_\zeta$ for all $(\tau,k)\in\mathcal{B}_\zeta$. In addition, also note that

$$\tilde{p}(x) - p(x) = \frac{1}{nP(X=x)}\sum_{i=1}^n R(X_i,W_i,x) + o_p(n^{-\frac{1}{2}}) \quad (147)$$

by an application of the Delta method and $\inf_{x\in\mathcal{X}} P(X=x) > 0$ due to X having finite support. Moreover, by the mean value theorem and (139) we obtain for some $\bar{b}(\tau, k)$ between $\tau + ak\tilde{p}(x)$ and $\tau + akp(x)$

$$s_0(\tau, \tau + ak\tilde{p}(x), x) - s_0(\tau, \tau + akp(x), x) = -\frac{(1-p(x))ka}{f_{y|1,x}(s_0(\tau, \bar{b}(\tau,k), x))}(\tilde{p}(x) - p(x))$$
$$= -\frac{(1-p(x))ka}{f_{y|1,x}(s_0(\tau, \tau + akp(x), x))}(\tilde{p}(x) - p(x)) + o_p(n^{-\frac{1}{2}}) \quad (148)$$

where the second equality follows from $(\tau, \bar{b}(\tau, k)) \in \mathcal{S}_\zeta$ for all (τ, k) with probability approaching one by convexity of $\mathcal{S}_\zeta$, $\inf_{(\tau,b,x)\in\mathcal{S}_\zeta\times\mathcal{X}} f_{y|1,x}(s_0(\tau, b, x))p(x) > 0$ and $\sup_{(\tau,k)\in\mathcal{B}_\zeta} |ak(\tilde{p}(x) - p(x))| = o_p(1)$ uniformly in $\mathcal{X}$. The first claim of the Lemma then follows by combining (147) and (148).

Finally, observe that the right hand side of (146) is trivially Donsker since $R(X_i, W_i, x)$ does not depend on (k, τ) and the function $(1 - p(x))ka/(f_{y|1,x}(s_0(\tau, \tau + akp(x), x))P(X = x))$ is uniformly continuous on $(\tau, k) \in \mathcal{B}_\zeta$ due to $\inf_{(\tau,b,x)\in\mathcal{S}_\zeta\times\mathcal{X}} f_{y|1,x}(s_0(\tau, b, x))p(x) > 0$. ■

Proof of Theorem 4.1: Throughout the proof we exploit Lemmas .5 and .6 applied with $W_i = 1$ with probability one, so that $\tilde{Q}_{x,n}(c|\tau, b) = Q_{x,n}(c|\tau, b)$ for all (τ, b) in $\mathcal{S}_\zeta$, where

$$\mathcal{S}_\zeta \equiv \{(\tau, b) \in [0, 1]^2 : b\{1 - p(x)\} + \zeta \leq \tau \leq p(x) + b\{1 - p(x)\} - \zeta \ \forall x \in \mathcal{X}\} . \quad (149)$$

Also notice that for every $(\tau, k) \in \mathcal{B}_\zeta$ and all $x \in \mathcal{X}$, the points $(\tau, \tau + kp(x)), (\tau, \tau - kp(x)) \in \mathcal{S}_\zeta$, while with probability approaching one $(\tau, \tau + k\hat{p}(x))$ and $(\tau, \tau - k\hat{p}(x))$ also belong to $\mathcal{S}_\zeta$. Therefore for $s_0(\tau, b, x)$ and $\hat{s}_0(\tau, b, x)$ as defined in (128) we obtain from Lemmas .5 and .6, applied with $W_i = 1$ a.s., that:

$$|(\hat{s}_0(\tau, \tau + akp(x), x) - s_0(\tau, \tau + akp(x), x)) - (\hat{s}_0(\tau, \tau + ak\hat{p}(x), x) - s_0(\tau, \tau + ak\hat{p}(x), x))| = o_p(n^{-\frac{1}{2}}) \quad (150)$$

uniformly in $(\tau, k, x) \in \mathcal{B}_\zeta \times \mathcal{X}$ and $a \in \{-1, 1\}$. Moreover, by Lemma .7 applied with $W_i = 1$ a.s.

$$s_0(\tau, \tau + ak\hat{p}(x), x) - s_0(\tau, \tau + akp(x), x)$$
$$= -\frac{(1-p(x))ka}{f_{y|1,x}(s_0(\tau, \tau + akp(x), x))P(X = x)} \times \frac{1}{n}\sum_{i=1}^{n} R(X_i, x) + o_p(n^{-\frac{1}{2}}) , \quad (151)$$

where $R(X_i, x) = p(x)\{P(X = x) - 1\{X_i = x\}\} + 1\{D_i = 1, X_i = x\} - P(D = 1, X = x)$ again uniformly in $(\tau, k, x) \in \mathcal{B}_\zeta \times \mathcal{X}$. Also observe that since $(\tau, \tau + k\hat{p}(x))$ and $(\tau, \tau - k\hat{p}(x))$ belong to $\mathcal{S}_\zeta$ with probability approaching one, we obtain uniformly in $(\tau, k, x) \in \mathcal{B}_\zeta \times \mathcal{X}$ that:

$$\begin{aligned} q_L(\tau, k|x) &= s_0(\tau, \tau + kp(x), x) & q_U(\tau, k|x) &= s_0(\tau, \tau - kp(x), x) \\ \hat{q}_L(\tau, k|x) &= \hat{s}_0(\tau, \tau + k\hat{p}(x), x) + o_p(n^{-\frac{1}{2}}) & \hat{q}_U(\tau, k|x) &= \hat{s}_0(\tau, \tau - k\hat{p}(x), x) + o_p(n^{-\frac{1}{2}}) . \end{aligned} \quad (152)$$

Therefore, combining results (150)-(152) and exploiting Lemmas .5, .6 and .7 and the sum of Donsker classes being Donsker we conclude that for J a Gaussian process on $L^\infty(\mathcal{B}_\zeta \times \mathcal{X}) \times L^\infty(\mathcal{B}_\zeta \times \mathcal{X})$:

$$\sqrt{n}C_n \xrightarrow{L} J \qquad C_n(\tau, k, x) \equiv \begin{pmatrix} \hat{q}_L(\tau, k|x) - q_L(\tau, k|x) \\ \hat{q}_U(\tau, k|x) - q_U(\tau, k|x) \end{pmatrix} . \quad (153)$$

To establish the second claim of the Theorem, observe that since X has finite support, we may denote

$\mathcal{X} = \{x_1, \ldots, x_{|\mathcal{X}|}\}$ and define the matrix $B = (P(X_i = x_1)x_1, \ldots, P(X_i = x_{|\mathcal{X}|})x_{|\mathcal{X}|})$ as well as the vector:

$$w \equiv \lambda' \; (E_S[X_i X_i'])^{-1} B \; . \tag{154}$$

Since w is also a function on $\mathcal{X}$, we refer to its coordinates by $w(x)$. Solving the linear programming problems in (24) and (25), it is then possible to obtain the closed form solution:

$$\begin{aligned}\pi_L(\tau, k) &= \sum_{x \in \mathcal{X}} \{1\{w(x) \geq 0\} w(x) q_L(\tau, k|x) + 1\{w(x) \leq 0\} w(x) q_U(\tau, k|x)\} \\ \pi_U(\tau, k) &= \sum_{x \in \mathcal{X}} \{1\{w(x) \geq 0\} w(x) q_U(\tau, k|x) + 1\{w(x) \leq 0\} w(x) q_L(\tau, k|x)\}\end{aligned} \tag{155}$$

with a similar representation holding for $(\hat{\pi}_L(\tau, k), \hat{\pi}_U(\tau, k))$ but with $(\hat{q}_L(\tau, k|x), \hat{q}_U(\tau, k|x))$ in place of $(q_L(\tau, k|x), q_U(\tau, k|x))$. We hence define the linear map $K : L^\infty(\mathcal{B}_\zeta \times \mathcal{X}) \times L^\infty(\mathcal{B}_\zeta \times \mathcal{X}) \to L^\infty(\mathcal{B}_\zeta) \times L^\infty(\mathcal{B}_\zeta)$, to be given by:

$$K(\theta)(\tau, k) \equiv \begin{pmatrix} \sum_{x \in \mathcal{X}} \{1\{w(x) \geq 0\} w(x) \theta^{(1)}(\tau, k, x) + 1\{w(x) \leq 0\} w(x) \theta^{(2)}(\tau, k, x)\} \\ \sum_{x \in \mathcal{X}} \{1\{w(x) \geq 0\} w(x) \theta^{(2)}(\tau, k, x) + 1\{w(x) \leq 0\} w(x) \theta^{(1)}(\tau, k, x)\} \end{pmatrix} \tag{156}$$

where for any $\theta \in L^\infty(\mathcal{X} \times \mathcal{B}_\zeta) \times L^\infty(\mathcal{B}_\zeta \times \mathcal{X})$, $\theta^{(i)}(\tau, k, x)$ denotes the i^{th} coordinate of the two dimensional vector $\theta(\tau, k, x)$. It then follows from (153), (155) and (156) that:

$$\sqrt{n} \begin{pmatrix} \hat{\pi}_L - \pi_L \\ \hat{\pi}_U - \pi_U \end{pmatrix} = \sqrt{n} K(C_n) \; . \tag{157}$$

Moreover, employing the norm $\| \cdot \|_\infty + \| \cdot \|_\infty$ on the product spaces $L^\infty(\mathcal{B}_\zeta \times \mathcal{X}) \times L^\infty(\mathcal{B}_\zeta \times \mathcal{X})$ and $L^\infty(\mathcal{B}_\zeta) \times L^\infty(\mathcal{B}_\zeta)$, we can then obtain by direct calculation that for any $\theta \in L^\infty(\mathcal{B}_\zeta \times \mathcal{X}) \times L^\infty(\mathcal{B}_\zeta \times \mathcal{X})$:

$$\|K(\theta)\|_\infty \leq 2 \sum_{x \in \mathcal{X}} |w(x)| \times \sup_{x \in \mathcal{X}} \sup_{(\tau, b) \in \mathcal{S}_\zeta} |\theta(\tau, b, x)| = 2 \sum_{x \in \mathcal{X}} |w(x)| \times \|\theta\|_\infty \; , \tag{158}$$

which implies the linear map K is continuous. Therefore, the theorem is established by (153), (157), the linearity of K and the continuous mapping theorem. ■

Proof of Theorem 4.2: For a metric space $\mathbb{D}$, let $BL_c(\mathbb{D})$ denote the set of real valued bounded Lipschitz functions with supremum norm and Lipschitz constant less than or equal to c. We first aim to show that:

$$\sup_{h \in BL_1(\mathbf{R})} |E[h(L(\tilde{G}_\omega)) | \mathcal{Z}_n] - E[h(L(G_\omega))]| = o_p(1) \; , \tag{159}$$

where $\mathcal{Z}_n = \{Y_i D_i, X_i, D_i\}_{i=1}^n$ and $E[h(\tilde{Z}) | \mathcal{Z}_n]$ denotes outer expectation over $\{W_i\}_{i=1}^n$ with $\mathcal{Z}_n$ fixed. Let

$$\hat{s}_0(\tau, b, x) \in \arg\min_{c \in \mathbf{R}} Q_{x,n}(c|\tau, b) \qquad \tilde{s}_0(\tau, b, x) \in \arg\min_{c \in \mathbf{R}} \tilde{Q}_{x,n}(c|\tau, b) \qquad s_0(\tau, b, x) \in \arg\min_{c \in \mathbf{R}} Q_x(c|\tau, b) \; . \tag{160}$$

Also note that with probability approaching one the points $(\tau, \tau + ak\tilde{p}(x)) \in \mathcal{S}_\zeta$ for all $(\tau, k, x) \in \mathcal{B}_\zeta \times \mathcal{X}$

and $a \in \{-1, 1\}$ for $\mathcal{S}_\zeta$ as in (149). Hence, arguing as in (150) and (151) we obtain:

$$\tilde{q}_L(\tau, k|x) - \hat{q}_L(\tau, k|x) = \tilde{s}_0(\tau, \tau + kp(x), x) - \hat{s}_0(\tau, \tau + kp(x), x) - \frac{(1 - p(x))k}{f_{y|1,x}(s_0(\tau, \tau + kp(x), x))P(X = x)} \times \frac{1}{n}\sum_{i=1}^{n} \Delta R(X_i, W_i, x) + o_p(n^{-\frac{1}{2}}) \quad (161)$$

$$\tilde{q}_U(\tau, k|x) - \hat{q}_U(\tau, k|x) = \tilde{s}_0(\tau, \tau - kp(x), x) - \hat{s}_0(\tau, \tau - kp(x), x) + \frac{(1 - p(x))k}{f_{y|1,x}(s_0(\tau, \tau - kp(x), x))P(X = x)} \times \frac{1}{n}\sum_{i=1}^{n} \Delta R(X_i, W_i, x) + o_p(n^{-\frac{1}{2}}) \quad (162)$$

where $\Delta R(X_i, W_i, x) = (1 - W_i)(1\{X_i = x\}p(x) - 1\{D_i = 1, X_i = x\})$ and both statements hold uniformly in $(\tau, k, x) \in \mathcal{B}_\zeta \times \mathcal{X}$. Also note that for the operator K as defined in (156), we have:

$$\sqrt{n}\begin{pmatrix} \tilde{\pi}_L - \hat{\pi}_L \\ \tilde{\pi}_U - \hat{\pi}_U \end{pmatrix} = \sqrt{n}K(\tilde{C}_n) \qquad \tilde{C}_n(\tau, k, x) \equiv \begin{pmatrix} \tilde{q}_L(\tau, k|x) - \hat{q}_L(\tau, k|x) \\ \tilde{q}_U(\tau, k|x) - \hat{q}_U(\tau, k|x) \end{pmatrix} . \quad (163)$$

By Lemmas .5, .6 and .7, results (161) and (162) and Theorem 2.9.2 in van der Vaart and Wellner (1996), the process $\sqrt{n}\tilde{C}_n$ converges unconditionally to a tight Gaussian process on $L^\infty(\mathcal{B}_\zeta \times \mathcal{X})$. Hence, by the continuous mapping theorem, $\sqrt{n}K(\tilde{C}_n)$ is asymptotically tight. Define,

$$\bar{G}_\omega \equiv \sqrt{n}\begin{pmatrix} (\tilde{\pi}_L - \hat{\pi}_L)/\omega_L \\ (\tilde{\pi}_U - \hat{\pi}_U)/\omega_U \end{pmatrix} , \quad (164)$$

and notice that $\omega_L(\tau, k)$ and $\omega_U(\tau, k)$ being bounded away from zero, $\hat{\omega}_L(\tau, k)$ and $\hat{\omega}_U(\tau, k)$ being uniformly consistent by Assumption 4.2(ii) and $\sqrt{n}K(\tilde{C}_n)$ being asymptotically tight imply that:

$$|L(\tilde{G}_\omega) - L(\bar{G}_\omega)| \leq \sup_{(\tau,k)\in\mathcal{B}_\zeta} M_0 \Big| \frac{\sqrt{n}(\hat{\pi}_L(\tau, k) - \tilde{\pi}_L(\tau, k))}{\hat{\omega}_L(\tau, k)} - \frac{\sqrt{n}(\hat{\pi}_L(\tau, k) - \tilde{\pi}_L(\tau, k))}{\omega_L(\tau, k)} \Big| + \sup_{(\tau,k)\in\mathcal{B}_\zeta} M_0 \Big| \frac{\sqrt{n}(\hat{\pi}_U(\tau, k) - \tilde{\pi}_U(\tau, k))}{\hat{\omega}_U(\tau, k)} - \frac{\sqrt{n}(\hat{\pi}_U(\tau, k) - \tilde{\pi}_U(\tau, k))}{\omega_U(\tau, k)} \Big| = o_p(1) , \quad (165)$$

for some constant M_0 due to L being Lipschitz. By definition of BL_1, all $h \in BL_1$ have Lipschitz constant less than or equal to 1 and are also bounded by 1. Hence, for any $\eta > 0$ Markov's inequality implies:

$$P\Big(\sup_{h\in BL_1(\mathbf{R})} |E[h(L(\tilde{G}_\omega))|\mathcal{Z}_n] - E[h(L(\bar{G}_\omega))|\mathcal{Z}_n]| > \eta\Big) \leq P\Big(2P(|L(\tilde{G}_\omega) - L(\bar{G}_\omega)| > \frac{\eta}{2}|\mathcal{Z}_n) + \frac{\eta}{2}P(|L(\tilde{G}_\omega) - L(\bar{G}_\omega)| \leq \frac{\eta}{2}|\mathcal{Z}_n) > \eta\Big) \leq \frac{4}{\eta}E\Big[E\Big[1\Big\{|L(\tilde{G}_\omega) - L(\bar{G}_\omega)| > \frac{\eta}{2}\Big\}|\mathcal{Z}_n\Big]\Big] . \quad (166)$$

Therefore, by (165), (166) and Lemma 1.2.6 in van der Vaart and Wellner (1996), we obtain:

$$P\Big(\sup_{h\in BL_1(\mathbf{R})} |E[h(L(\tilde{G}_\omega))|\mathcal{Z}_n] - E[h(L(\bar{G}_\omega))|\mathcal{Z}_n]| > \eta\Big) \leq \frac{4}{\eta}P\Big(|L(\tilde{G}_\omega) - L(\bar{G}_\omega)| > \frac{\eta}{2}\Big) = o(1) . \quad (167)$$

Next, let $\stackrel{L}{=}$ stands for "equal in law" and notice that for J the Gaussian process in (153):

$$L(G_\omega) \stackrel{L}{=} T \circ K(J) \qquad L(\bar{G}_\omega) = \sqrt{n}L \circ K(\tilde{C}_n) , \quad (168)$$

due to the continuous mapping theorem and (163). For $w(x)$ as defined in (151) and $C_0 \equiv 2\sum_{x\in\mathcal{X}} |w(x)|$, it follows from linearity of K and (156), that K is Lipschitz with Lipschitz constant C_0. Therefore, for any $h \in BL_1(\mathbf{R})$, result (168) implies that $h \circ L \circ K \in BL_{C_0 M_0}(L^\infty(\mathcal{B}_\zeta \times \mathcal{X}))$ for some $M_0 > 0$ and hence

$$\sup_{h\in BL_1(\mathbf{R})} |E[h(L(\bar{G}_\omega))|\mathcal{Z}_n] - E[h(L(G_\omega))]| \leq \sup_{h\in BL_{C_0 M_0}(L^\infty(\mathcal{B}_\zeta\times\mathcal{X}))} |E[h(\bar{G}_\omega)|\mathcal{Z}_n] - E[h(J)]| = o_p(1) \ , \quad (169)$$

where the final equality follows from (161), (162), (168), arguing as in (166)-(167) and Lemmas .6, .7 and Theorem 2.9.6 in van der Vaart and Wellner (1996). Hence, (167) and (169) establish (159).

Next, we aim to show that for all $t \in \mathbf{R}$ at which the CDF of $L(G_\omega)$ is continuous, and any $\eta > 0$:

$$P(|P(L(\tilde{G}_\omega) \leq t|\mathcal{Z}_n) - P(L(G_\omega) \leq t)| > \eta) = o(1) \ . \quad (170)$$

Towards this end, for every $\lambda > 0$, and t at which the CDF of $L(G_\omega)$ is continuous define the functions:

$$h^U_{\lambda,t}(u) = 1 - 1\{u > t\}\min\{\lambda(u-t), 1\} \qquad h^L_{\lambda,t}(u) = 1\{u < t\}\min\{\lambda(t-u), 1\} \ . \quad (171)$$

Notice that by construction, $h^L_{\lambda,t}(u) \leq 1\{u \leq t\} \leq h^U_{\lambda,t}(u)$ for all $u \in \mathbf{R}$, the functions $h^L_{\lambda,t}$ and $h^U_{\lambda,t}$ are both bounded by one and they are both Lipschitz with Lipschitz constant λ. Also by direct calculation:

$$0 \leq E[h^U_{\lambda,t}(L(G_\omega)) - h^L_{\lambda,t}(L(G_\omega))] \leq P(t - \lambda^{-1} \leq L(G_\omega) \leq t + \lambda^{-1}) \ . \quad (172)$$

Therefore, exploiting that $h^L_{\lambda,t}, h^U_{\lambda,t} \in BL_\lambda(\mathbf{R})$ and that $h \in BL_\lambda(\mathbf{R})$ implies $\lambda^{-1}h \in BL_1(\mathbf{R})$, we obtain:

$$\begin{aligned}
|P(L(\tilde{G}_\omega) \leq & t|\mathcal{Z}_n) - P(L(G_\omega) \leq t)| \\
& \leq |E[h^L_{\lambda,t}(L(\tilde{G}_\omega))|\mathcal{Z}_n] - E[h^U_{\lambda,t}(L(G_\omega))]| + |E[h^U_{\lambda,t}(L(\tilde{G}_\omega))|\mathcal{Z}_n] - E[h^L_{\lambda,t}(L(G_\omega))]| \\
& \leq 2 \sup_{h\in BL_\lambda(\mathbf{R})} |E[h(L(\tilde{G}_\omega))|\mathcal{Z}_n] - E[h(L(G_\omega))]| + 2P(t-\lambda^{-1} \leq L(G_\omega) \leq t + \lambda^{-1}) \\
& = 2\lambda \sup_{h\in BL_1(\mathbf{R})} |E[h(L(\tilde{G}_\omega))|\mathcal{Z}_n] - E[h(L(G_\omega))]| + 2P(t-\lambda^{-1} \leq L(G_\omega) \leq t + \lambda^{-1}) \ . \quad (173)
\end{aligned}$$

for any $\lambda > 0$. Moreover, we may select a λ_η sufficiently large, so that $2P(t-\lambda_\eta^{-1} \leq L(G_\omega) \leq t+\lambda_\eta^{-1}) < \eta/2$ due to t being a continuity point of the CDF of $L(G_\omega)$. Therefore, from (173) we obtain:

$$\begin{aligned}
P(|P(L(\tilde{G}_\omega) \leq t|\mathcal{Z}_n) - P(L(G_\omega) \leq t)| > \eta) \\
\leq P(2\lambda_\eta \sup_{h\in BL_1(\mathbf{R})} |E[h(L(\tilde{G}_\omega))|\mathcal{Z}_n] - E[h(L(G_\omega))]| > \frac{\eta}{2}) = o(1) \ , \quad (174)
\end{aligned}$$

where the final equality follows from (159).

Finally, note that since the CDF of $L(G_\omega)$ is strictly increasing and continuous at $r_{1-\alpha}$, we obtain that:

$$P(L(G_\omega) \leq r_{1-\alpha} - \epsilon) < 1 - \alpha < P(L(G_\omega) \leq r_{1-\alpha} + \epsilon) \quad (175)$$

$\forall \epsilon > 0$. Define the event $A_n \equiv \{P(L(\tilde{G}_\omega) \leq r_{1-\alpha} - \epsilon|\mathcal{Z}_n) < 1 - \alpha < P(L(\tilde{G}_\omega) \leq r_{1-\alpha} + \epsilon|\mathcal{Z}_n)\}$ and notice

$$P(|\tilde{r}_{1-\alpha} - r_{1-\alpha}| \leq \epsilon) \geq P(A_n) \to 1 \ , \quad (176)$$

where the inequality follows by definition of $\tilde{r}_{1-\alpha}$ and the second result is implied by (170) and (175). ■

References

Acemoglu, D. and Autor, D. H. (2011). Skills, tasks, and technology: Implications for employment and earnings. In *Handbook of Labor Economics 4b*, vol. 4. Elsevier, 1043–1171.

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

Altonji, J. G., Elder, T. and Taber, C. (2008). Using selection on observed variables to assess bias from unobservables when evaluating swan-ganz catheterization. *American Economic Review, Paper and Proceedings*, **98** 145–350.

Angrist, J., Chernozhukov, V. and Fernández-Val, I. (2006). Quantile regression under misspecification, with an application to the u.s. wage structure. *Econometrica*, **74** 539–563.

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

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

Chamberlain, G. (1994). Quantile regression, censoring and the structure of wages. In *Advances in Econometrics, Sixth World Congress* (C. Sims, ed.). Elsevier, 171–209.

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

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

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

Galambos, J. (1973). A general poisson limit theorem of probability theory. *Duke Mathematical Journal*, **40** 581–586.

Greenlees, J. S., Reece, W. S. and Zieschang, K. D. (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.

Heckman, J. J., Lochner, L. and Todd, P. (2006). Earnings functions, rates of return and treatment effects: The mincer equation and beyond. In *Handbook of Education Economics* (E. Hanushek and F. Welch, eds.), vol. 1, chap. 7. Elsevier.

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

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

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

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

Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review, Papers and Proceedings*, **93** 126–132.

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

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

Lemieux, T. (2006). Postsecondary education and increasing wage inequality. *The American Economic Review*, **96** 195–199.

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

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

Manski, C. F. (1994). The selection problem. In *Advances in Econometrics, Sixth World Congress* (C. Sims, ed.). Cambridge University Press, Cambridge, UK, 143–170.

Manski, C. F. (2003). *Partial Identification of Probability Distributions.* Springer-Verlag, New York.

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

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

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

Rosenbaum, P. and Rubin, D. (1983). Assessing sensitivity to an unobserved binary covarite in an observational study with binary outcome. *Journal of the Royal Statistical Society, Series B*, **45** 212–218.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* 581–592.

Sawa, T. (1978). Information criteria for discriminating among alternative regression models. *Econometrica*, **46** 1273–1291.

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

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

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: with Applications to Statistics*. Springer, New York.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, **50** 1–25.