# ERRORS IN THE DEPENDENT VARIABLE OF QUANTILE REGRESSION MODELS

JERRY HAUSMAN, YE LUO, AND CHRISTOPHER PALMER

Abstract. The usual quantile regression estimator of Koenker and Bassett (1978) is biased if there is an additive error term. We analyze this problem as an errors-in-variables problem where the dependent variable suffers from classical measurement error, and we present a sieve maximum-likelihood approach that is robust to left-hand side measurement error. After providing sufficient conditions for identification, we demonstrate that when the number of knots in the quantile grid is chosen to grow at an adequate speed, the sieve maximum-likelihood estimator is consistent and asymptotically normal, permitting inference via bootstrapping. We verify our theoretical results with Monte Carlo simulations and illustrate our estimator with an application to the returns to education highlighting important changes over time in the returns to education that have been obscured in previous work by measurement-error bias.

*Keywords*: Measurement Error, Quantile Regression, Functional Analysis

---

*Date*: August 2016.
Hausman: MIT Department of Economics and NBER; jhausman@mit.edu.
Luo: Univeristy of Florida Department of Economics; yeluo@ufl.edu.
Palmer: Haas School of Business, University of California, Berkeley; cjpalmer@berkeley.edu.
We thank Isaiah Andrews, Colin Cameron, Victor Chernozhukov, Denis Chetverikov, Kirill Evdokimov, Hank Farber, Larry Katz, Brad Larsen, David Levine, Rosa Matzkin, and Shu Shen for helpful discussions, as well as seminar participants at Cornell, Harvard, MIT, Princeton, UC Davis, UCL, and UCLA. Yuqi Song, Jacob Ornelas, and especially Haoyang Liu provided outstanding research assistance.

## 1. Introduction

Economists are aware of problems arising from errors-in-variables in regressors but generally ignore measurement error in the dependent variable. In this paper, we study the consequences of measurement error in the dependent variable of conditional quantile models and propose a maximum likelihood approach to consistently estimate the distributional effects of covariates in such a setting. Quantile regression (Koenker and Bassett, 1978) has become a very popular tool for applied microeconomists to consider the effect of covariates on the distribution of the dependent variable. However, as left-hand side variables in microeconometrics often come from self-reported survey data, the sensitivity of traditional quantile regression to LHS measurement error poses a serious problem to the validity of results from the traditional quantile regression estimator.

The errors-in-variables (EIV) problem has received significant attention in the linear model, including the well-known results that classical measurement error causes attenuation bias if present in the regressors and has no effect on unbiasedness if present in the dependent variable. See Hausman (2001) for an overview. In general, the linear model results do not hold in nonlinear models.[1] In this paper, we focus on the linear quantile regression setting. Hausman (2001) observes that EIV in the dependent variable in quantile regression models generally leads to significant bias, a result very different from the linear model intuition.

In general, EIV in the dependent variable can be viewed as a mixture model.[2] We show that under certain assumptions on the degree of ill-posedness, by choosing the growth speed of the number of knots in the quantile grid, our estimator has fractional polynomial of $n$ convergence speed and asymptotic normality. We suggest using the bootstrap for inference.

[1]Schennach (2008) establishes identification and a consistent nonparametric estimator when EIV exists in an explanatory variable. Carroll and Wei (2009) proposed an iterative estimator for the quantile regression when one of the regressors has EIV. Studies focusing on nonlinear models in which the left-hand side variable is measured with error include Hausman et. al (1998) and Cosslett (2004), who study probit and tobit models, respectively.

[2]A common feature of mixture models under a semiparametric or nonparametric framework is the ill-posed inverse problem, see Fan (1991). We face the ill-posed problem here, and our model specifications are linked to the Fredholm integral equation of the first kind. The inverse of such integral equations is usually ill-posed even if the integral kernel is positive definite. The key symptom of these model specifications is that the high-frequency signal of the objective we are interested in is wiped out, or at least shrunk, by the unknown noise if its distribution is smooth. To uncover these signals is difficult and all feasible estimators have a lower speed of convergence compare to the usual $\sqrt{n}$ case. The convergence speed of our estimator relies on the decay speed of the eigenvalues of the integral operator. We explain this technical problem in more detail in the related section of this paper.

Intuitively, the estimated quantile regression line $x_i\widehat{\beta}(\tau)$ for quantile $\tau$ may be far from the observed $y_i$ because of LHS measurement error or because the unobserved conditional quantile $u_i$ of observation $i$ is far from $\tau$. Our ML framework effectively estimates the likelihood that a given quantile-specific residual ( $\varepsilon_{ij} := y_i - x_i\beta(\tau_j)$) is large because of measurement error rather than observation $i$'s unobserved conditional quantile $u_i$ being far away from $\tau_j$. The estimate of the joint distribution of the conditional quantile and the measurement error allows us to weight the log likelihood contribution of observation $i$ more in the estimation of $\beta(\tau_j)$ where is is more likely that $u_i \approx \tau_j$. We show in simulations that a mixture of normals can accommodate a wide set of EIV distributions.[3] In the case of Gaussian errors in variables, this estimator reduces to weighted least squares, with weights equal to the probability of observing the quantile-specific residual for a given observation as a fraction of the total probability of the same observation's residuals across all quantiles.

An empirical example (extending Angrist et al., 2006) studies heterogeneity in the returns to education across conditional quantiles of the wage distribution. Correcting for likely measurement error in the self-reported wage data, we estimate considerably more heterogeneity across the wage distribution in education-wage gradient. In particular, the returns to education for latently high-wage individuals have been increasing over time and are much higher than previously estimated. By 2000, the return to education for the top of the conditional wage distribution was over three times larger than returns for any other segment of the distribution. We also document that increases in the returns to education between 2000–2010, while still skewed towards top earners, were shared more broadly across the wage distribution.

The rest of the paper proceeds as follows. In Section 2, we introduce model specification and identification conditions. In Section 3, we present our maximum likelihood estimator and analyze its properties. Section 4 operationalizes this estimator through sieve estimation. We present Monte Carlo simulation results in Section 5, and Section 6 contains our empirical application. Section 7 concludes.

Notation: Define $x$ to have dimension $d_x$ and domain $\mathcal{X}$. Define the space of $y$ as $\mathcal{Y}$. Denote $a \wedge b$ as the minimum of $a$ and $b$, and denote $a \vee b$ as the larger of $a$ and $b$. Let $\underset{d}{\to}$ be weak convergence (convergence in distribution), and $\underset{p}{\to}$ stands for convergence in probability. Let $\underset{d}{\to}^*$ be weak convergence in outer probability. Let $f(\varepsilon|\sigma)$

[3]See Burda, Harding, and Hausman (2008, 2011) for other applications demonstrating the flexibility of a finite mixture of normals.

be the p.d.f of the EIV $\varepsilon$ parametrized by $\sigma$ where $\sigma$ has dimension $d_\sigma$ and domain $\Sigma$. Assume the true parameters are $\beta_0(\cdot)$ and $\sigma_0$ for the coefficient of the quantile model and parameter of the density function of the EIV. Define $||(\beta_{0,}\sigma_0)|| := \sqrt{||\beta_0||_2^2 + ||\sigma_0||_2^2}$ as the $L^2$ norm of $(\beta_{0,}\sigma_0)$, where $||\cdot||_2$ is the usual Euclidean norm. For $\beta_k \in \mathbb{R}^k$, define $||(\beta_{k,}\sigma_0)||^2 := \sqrt{||\beta_k||_2^2/k + ||\sigma_0||_2^2}$ and $||(\beta_k, \sigma)||_\infty = ||\beta_k||_\infty + |\sigma_0|_\infty$. Finally, we use the notation $x \precsim y$ for $x = O(y)$ and $x \precsim_p y$ for $x = O_p(y)$.

## 2. Model and Identification

We consider the standard linear conditional quantile model, where the $\tau^{th}$ quantile of the dependent variable $y^*$ is a linear function of $x$

$$Q_{y^*}(\tau|x) = x\beta_0(\tau).$$

However, we are interested in the situation where $y^*$ is not directly observed, and we instead observe $y$ where

$$y = y^* + \varepsilon$$

and $\varepsilon$ is a mean-zero, i.i.d error term independent from $y^*$ and $x$.

Unlike the linear regression case where EIV in the left hand side variable does not matter for consistency and asymptotic normality, EIV in the dependent variable can lead to severe bias in quantile regression. More specifically, with $\rho_\tau(z)$ denoting the check function (plotted in Figure 1)

$$\rho_\tau(z) = z(\tau - 1(z < 0)),$$

the minimization problem in the usual quantile regression

$$\beta(\tau) \in \arg\min_b E[\rho_\tau(y - xb)], \tag{2.1}$$

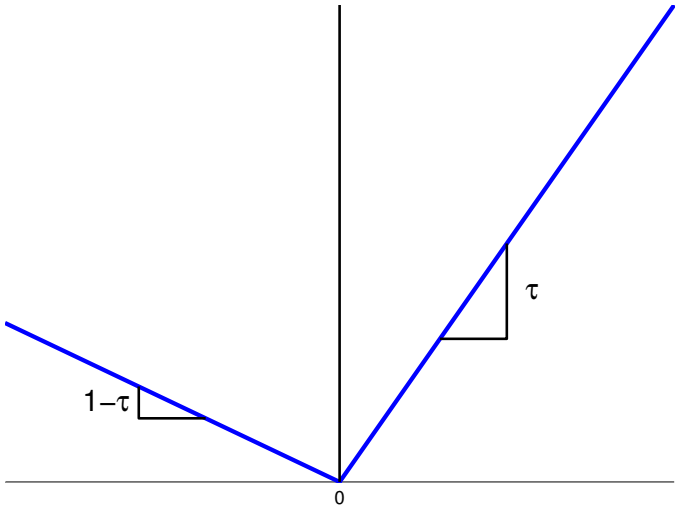
is generally no longer minimized at the true $\beta_0(\tau)$ when EIV exists in the dependent variable. When there exists no EIV in the left-hand side variable, i.e. $y^*$ is observed, the FOC is

$$E[x(\tau - 1(y^* < x\beta(\tau)))] = 0, \tag{2.2}$$

where the true $\beta(\tau)$ is the solution to the above system of first-order conditions as shown by Koenker and Bassett (1978). However, with left-hand side EIV, the first-order condition determining $\widehat{\beta}(\tau)$ becomes

$$E[x(\tau - 1(y^* + \varepsilon < x\beta(\tau)))] = 0. \tag{2.3}$$

FIGURE 1. Check Function $\rho_\tau(z)$




For $\tau \neq 0.5$, the presence of measurement error $\varepsilon$ will result in the FOC being satisfied at a different estimate of $\beta$ than in equation (2.2) even in the case where $\varepsilon$ is symmetrically distributed because of the asymmetry of the check function. In other words, in the minimization problem, observations for which $y^* \geq x\beta(\tau)$ and should therefore get a weight of $\tau$ may end up on the left-hand side of the check function, receiving a weight of $(1-\tau)$ . Thus, equal-sized differences on either side of zero do not cancel each other out.[4]

A straightforward analytical example below demonstrates the intuition behind the problem of left-hand errors in variables for estimators concerned with estimating the distributional parameters. We then provide a simple Monte-Carlo simulation to show the degree of bias in a simple two-factor model with random disturbances on the dependent variable $y$.

**Example 1.** Consider the bivariate data-generating process

$$y_i = \beta_0(u_i) + \beta_1(u_i) \cdot x_i + \varepsilon_i$$

[4]For median regression, $\tau = .5$ and so $\rho_{.5}(\cdot)$ is symmetric around zero. This means that if $\varepsilon$ is symmetrically distributed and $\beta(\tau)$ symmetrically distributed around $\tau = .5$ (as would be the case, for example, if $\beta(\tau)$ were linear in $\tau$), the expectation in equation (2.3) holds for the true $\beta_0(\tau)$. However, for non-symmetric $\varepsilon$, equation (2.3) is not satisfied at the true $\beta_0(\tau)$.

where $x_i \in \{0,1\}$, the measurement error $\varepsilon_i$ is distributed $\mathcal{N}(0,1)$, and the unobserved conditional quantile $u_i$ of observation $i$ follows $u_i \sim U[0,1]$. Let the coefficient function $\beta_0(\tau) = \beta_1(\tau) = \Phi^{-1}(\tau)$, with $\Phi^{-1}(\cdot)$ representing the inverse CDF of the standard normal distribution. Because quantile regression estimates the conditional quantiles of $y$ given $x$, in this simple setting, the estimated slope coefficient function is simply the difference in inverse CDFs for $x = 1$ and $x = 0$. For any quantile $\tau$, $\widehat{\beta}_1(\tau) = \widehat{F}^{-1}_{y|x=1}(\tau) - \widehat{F}^{-1}_{y|x=0}(\tau)$ where $\widehat{F}(\cdot)$ is the estimated CDF of $y$. With no measurement error, the distribution $y|x = 1$ is $\mathcal{N}(0,4)$ and the distribution of $y|x = 0$ is $\mathcal{N}(0,1)$. Then

$$p\lim \widehat{\beta}_1(\tau) = (\sqrt{4} - \sqrt{1})\Phi^{-1}(\tau) = \beta_1(\tau),$$

or the estimated coefficient tends the truth at each $\tau$ by the consistency of quantile regression as an estimator of the conditional distribution of $y$ given $x$. With non-zero measurement error, quantile regression still consistently estimates the conditional distribution of $y$ given $x$ (now $y|x = 1 \sim \mathcal{N}(0,5)$ and $y|x = 0 \sim \mathcal{N}(0,2)$), but the measurement error prevents consistent estimation of the partial effect of $x$ on the conditional quantile of $y$. The probability limit of the estimated coefficient function under measurement error $\tilde{\beta}_1(\cdot)$ is

$$p\lim \tilde{\beta}_1(\tau) = (\sqrt{5} - \sqrt{2})\Phi^{-1}(\tau),$$

which will not equal the truth for any quantile $\tau \neq 0.5$.

This example also illustrates the intuition offered by Hausman (2001) for compression bias for bivariate quantile regression. For the median $\tau = 0.5$, because $\Phi^{-1}(0.5) = 0$, $\beta(\tau) = \widehat{\beta}(\tau) = \tilde{\beta}(\tau) = 0$ such that the median is unbiased. For all other quantiles, however, since $\sqrt{5} - \sqrt{2} < 1$, the coefficient estimated under measurement error will be compressed towards the true coefficient on the median regression $\beta_1(0.5)$.

**Example 2.** We now consider a simulation exercise to illustrate the direction and magnitude of measurement error bias in even simple quantile regression models. The data-generating process for the Monte-Carlo results is

$$y_i = \beta_0(u_i) + x_{1i}\beta_1(u_i) + x_{2i}\beta_2(u_i) + \varepsilon_i$$

with the measurement error $\varepsilon_i$ again distributed as $\mathcal{N}(0,\sigma^2)$ and the unobserved conditional quantile $u_i$ of observation $i$ following $u_i \sim U[0,1]$. The coefficient function $\beta(\tau)$ has components $\beta_0(\tau) = 0$, $\beta_1(\tau) = \exp(\tau)$, and $\beta_2(\tau) = \sqrt{\tau}$. The variables $x_1$ and $x_2$ are

TABLE 1. Monte-Carlo Results: Mean Bias

| Parameter | EIV Distribution | Quantile ( $\tau$) 0.1 | 0.25 | 0.5 | 0.75 | 0.9 |
|---|---|---|---|---|---|---|
| $\beta_1(\tau) = e^\tau$ | $\varepsilon = 0$ | 0.006 | 0.003 | 0.002 | 0.000 | -0.005 |
| | $\varepsilon \sim \mathcal{N}(0, 4)$ | 0.196 | 0.155 | 0.031 | -0.154 | -0.272 |
| | $\varepsilon \sim \mathcal{N}(0, 16)$ | 0.305 | 0.246 | 0.054 | -0.219 | -0.391 |
| | True parameter: | 1.105 | 1.284 | 1.649 | 2.117 | 2.46 |
| $\beta_2(\tau) = \sqrt{\tau}$ | $\varepsilon = 0$ | 0.000 | -0.003 | -0.005 | -0.006 | -0.006 |
| | $\varepsilon \sim \mathcal{N}(0, 4)$ | 0.161 | 0.068 | -0.026 | -0.088 | -0.115 |
| | $\varepsilon \sim \mathcal{N}(0, 16)$ | 0.219 | 0.101 | -0.031 | -0.128 | -0.174 |
| | True parameter: | 0.316 | 0.5 | 0.707 | 0.866 | 0.949 |

Notes: Table reports mean bias (across 500 simulations) of slope coefficients estimated for each quantile $\tau$ from standard quantile regression of $y$ on a constant, $x_1$, and $x_2$ where $y = x_1\beta_1(\tau) + x_2\beta_2(\tau) + \varepsilon$ and $\varepsilon$ is either zero (no measurement error case, i.e. $y^*$ is observed) or $\varepsilon$ is distributed normally with variance 4 or 16. The covariates $x_1$ and $x_2$ are i.i.d. draws from $LN(0, 1)$. $N = 1,000$.

drawn from independent lognormal distributions $LN(0, 1)$. The number of observations is 1,000.

Table 1 presents Monte-Carlo results for three cases: when there is no measurement error and when the variance of $\varepsilon$ equals 4 and 16. The simulation results show that under the presence of measurement error, the quantile regression estimator is severely biased. Furthermore, we find evidence of the attenuation-towards-the-median behavior posited by Hausman (2001), with quantiles above the median biased down and quantiles below the median upwardly biased, understating the distributional heterogeneity in the $\beta(\cdot)$ function. For symmetrically distributed EIV and uniformly distributed $\beta(\tau)$, the median regression results appear unbiased. Comparing the mean bias when the variance of the measurement error increases from 4 to 16 shows that the bias is increasing in the variance of the measurement error. Intuitively, the information of the functional parameter $\beta(\cdot)$ is decaying when the variance of the EIV becomes larger.

2.1. **Identification and Regularity Conditions.** In the linear quantile model, it is assumed that $x\beta(\tau)$ is increasing in $\tau$ for any $x \in \mathcal{X}$, implying that $\beta(\cdot)$ is a component-wise increasing function up to a linear transformation. Therefore, WLOG, we can assume that the coefficients $\beta(\cdot)$ are increasing and refer to the set of functions $\{\beta_k(\cdot)\}_{k=1}^{d_x}$ as co-monotonic functions.

**Condition A1** (Properties of $\beta(\cdot)$)**.** We assume the following properties on the coefficient vectors $\beta(\tau)$:

(1) $\beta(\tau)$ is in the space $M[B_1 \times B_2 \times B_3 ... \times B_{d_x}]$ where the functional space $M$ is defined as the collection of all functions $f = (f_1, ..., f_{d_x}) : [0,1] \to [B_1 \times ... \times B_{d_x}]$ with $B_k \subset \mathbb{R}$ being a closed interval $\forall\, k \in \{1, ..., d_x\}$ such that each entry $f_k : [0,1] \to B_k$ is monotonically increasing in $\tau$.
(2) Let $B_k = [l_k,\, u_k]$ so that $l_k < \beta_{0k}(\tau) < u_k\ \forall\, k \in \{1, ..., d_x\}$ and $\tau \in [0,1]$.
(3) The true parameter $\beta_0$ is a vector of $C^1$ functions with derivative bounded from below by a positive constant, implying that each component of $\beta_0$ is strictly monotonic.
(4) The domain of the parameter $\sigma$ is a compact space $\Sigma$ and the true value $\sigma_0$ is in the interior of $\Sigma$.
(5) Without loss of generality, $\beta_k(0) \geq 0$.

Monotonicity of $\beta_k(\cdot)$ is important for identification because in the log-likelihood function, $f(y|x) = \int_0^1 f(y - x\beta(u)|\sigma)du$ is invariant to a rearrangement of the function $\beta(u)$. The function $\beta(\cdot)$ is therefore unidentified if we do not impose further restrictions. Given the distribution of the random variable $\{\beta(u) \,|\, u \in [0,1]\}$, the vector of functions $\beta : [0,1] \to B_1 \times B_2 \times ... \times B_{d_x}$ is unique under the rearrangement if the functions $\{\beta_k(\cdot)\}_{k=1}^{d_x}$ are co-monotonic, which we assume WLOG as discussed above.

Under assumption A1, it is easy to see that the parameter space $\Theta := M \times \Sigma$ is compact under the $L^\infty$ norm. The following lemma will be instrumental in proving the consistency of our ML estimator.

**Lemma 1.** *The space $M[B_1 \times B_2 \times B_3 ... \times B_k]$ is a compact and complete space under $L^p$, for any $p \geq 1$.*

*Proof.* See Appendix C.1. □

**Condition A2** (Properties of $x$)**.** We assume the following properties of the vectors $x$ that comprise the design matrix $X$:

(1) $E[x'x]$ is non-singular.
(2) The domain of $x$, denoted as $\mathcal{X}$, is bounded continuous on at least one dimension, i.e. there exists $k \in \{1, ..., d_x\}$ such that for every feasible $x_{-k}$, there is a open set $X_k \subset \mathbb{R}$ such that $(X_k, x_{\{-k\}}) \subset \mathcal{X}$.

**Condition A3** (Properties of EIV)**.** We assume the following properties of the measurement error $\varepsilon$. The probability density function of the EIV is denoted $f(\varepsilon|\sigma)$ and the true density is abbreviated $f_0(\varepsilon) := f(\varepsilon|\sigma_0)$.

(1) $f(\varepsilon|\sigma)$ is differentiable in $\varepsilon$ and $\sigma$.
(2) For all $\sigma \in \Sigma$, there exists a uniform constant $C > 0$ such that $\mathbb{E}[|\log f(\varepsilon|\sigma)|] < C$.
(3) $f(\cdot|\sigma)$ is non-zero all over the entire space $\mathbb{R}$ and bounded from above uniformly.
(4) $E[\varepsilon] = \int_{-\infty}^{\infty} \varepsilon f(\varepsilon|\sigma) = 0$.[5]
(5) Define $\phi_\varepsilon(s|\sigma) := \int_{-\infty}^{\infty} \exp(is\varepsilon) f(\varepsilon|\sigma) d\varepsilon$ as the characteristic function of $\varepsilon$ given PDF $f(\varepsilon|\sigma)$. Given assumption A3.4, $\phi'_\epsilon(s|\sigma)|_{s=0} = 0$.
(6) $|\log(f(\varepsilon|\sigma))| \leq C(1 + \varepsilon^\gamma)$ and $E[|\varepsilon^{2\gamma}|f_0(\varepsilon|\sigma)] < \infty$. For some $\gamma > 0$ and for all $\sigma_1, \sigma_2 \in \Sigma$, $|\log(f(\varepsilon|\sigma_1)) - \log(f(\varepsilon|\sigma_2))| \leq C||\sigma_1 - \sigma_2||_2(1 + \varepsilon^\gamma)$
(7) There is a positive constant $C > 0$ such that $|f'(\varepsilon|\sigma)| < C$ and $|\partial_\sigma f(\varepsilon|\sigma)| < C$, for all $\varepsilon$ and $\sigma \in \Sigma$.
(8) For any $\sigma \in \Sigma$, $\int_{-q}^{q} |\phi_\varepsilon(s|\sigma) - \phi_\varepsilon(s|\sigma_0)|^2 ds \geq C_q||\sigma - \sigma_0||_2^2$ for any $q > 0$ and some constant $C_q > 0$.

Note that assumptions A3.6–A3.8 hold for all exponential families.

Lemma 1 and the above conditions on the parameters, covariate matrix, and measurement error allow us to state our main identification result.

**Theorem 1** (Nonparametric Global Identification)**.** *Under Condition A1-A3, for any $\beta(\cdot)$ and $f(\cdot)$ which generate the same density of $y|x$ almost everywhere as the true function $\beta_0(\cdot)$ and $f_0(\cdot)$, it must be that:*

$$\beta(\tau) = \beta_0(\tau) \text{ for all } \tau$$
$$f(\varepsilon) = f_0(\varepsilon) \text{ for all } \varepsilon.$$

*Proof.* See Appendix C.1. □

## 3. Maximum Likelihood Estimator

Denote $\theta := (\beta(\cdot), \sigma) \in \Theta$. For any $\theta$, define the expected log-likelihood function $L(\theta)$ as follows

$$L(\theta) = \mathbb{E}[\log g(y|x, \theta)], \tag{3.1}$$

with the empirical log likelihood being denoted

$$L_n(\theta) = \mathbb{E}_n[\log g(y|x, \theta)]. \tag{3.2}$$

[5]Note that this can always be achieved by a normalization of the constant term $\beta_0(\tau)$.

Using the fact that the unobserved conditional quantile is the CDF of $y|x$ and CDFs are distributed uniformly, the conditional density function $g(y|x,\theta)$ is given by

$$g(y|x,\theta) = \int_0^1 f(y - x\beta(u)|\sigma)du. \tag{3.3}$$

3.1. **Consistency.** The ML estimator is defined as:

$$(\widehat{\beta}(\cdot), \widehat{\sigma}) \in \arg \max_{(\beta(\cdot),\sigma)\in\Theta} E_n[g(y|x,\beta(\cdot),\sigma)]. \tag{3.4}$$

where $g(\cdot|\cdot,\cdot,\cdot)$ is the conditional density of $y$ given $x$ and parameters, as defined in equation (3.3). Defining $E_n$ as the empirical average operator $E_n h(x) := \frac{1}{n}\sum_{i=1}^n h(x_i)$, the following theorem states the consistency property of the ML estimator.

**Lemma 2** (MLE Consistency)**.** *Under conditions A1-A3, the maximum-likelihood estimator*

$$(\widehat{\beta}(\cdot), \widehat{\sigma}) \in \arg \max_{(\beta(\cdot),\sigma)\in\Theta} E_n \left[\log \int_0^1 f(y - x\beta(\tau)|\sigma)d\tau\right]$$

*exists and converges in probability to the true parameter $(\beta_0(\cdot), \sigma_0)$ under the $L^2$ norm in the functional space $M$ and Euclidean norm in $\Sigma$ .*

*Proof.* See Appendix C.2. □

The consistentcy theorem is a special version of a general MLE consistency theorem (Van der Vaart, 2000). Two conditions play critical roles here: the co-monotonicity of the $\beta(\cdot)$ function and the local continuity of at least one right-hand side variable. If we do not restrict the estimator in the family of monotone functions, then we will lose compactness of the parameter space $\Theta$ and the consistency argument will fail.

3.2. **Convergence rate of the parametric component.** As we are trying to estimate $\beta(\cdot)$ and $\varepsilon$ via MLE from the mixture distribution of $y = x\beta(\tau) + \varepsilon$, where $\tau \sim U[0,1]$ and $\varepsilon \sim f(\cdot|\sigma_0)$, the estimation of $\beta(\cdot)$ is ill-posed. However, with the condition that one of the variables in $x$ is continuous, we are able to estimate $\sigma$ at a much faster rate.

**Condition A4** (Variation on Characteristic Function)**.**

(1) $|\phi_{x\beta_0}(s)| \geq C/s$, for some generic constant $C > 0$.
(2) There exists a generic constant $c > 0$ such that for any $(\beta,\sigma) \in \Theta$ and any $s \in \mathbb{R}$,

$$Var_x\left(\frac{\phi_{x\beta}(s)}{\phi_{x\beta_0}(s)}\right) \geq cE_x\left[\left|\frac{\phi_{x\beta}(s) - \phi_{x\beta_0}(s)}{\phi_{x\beta_0}(s)}\right|^2\right].$$

The first condition in the above assumption can be derived from the fact that $x$ and $\beta_0(\tau)$ are bounded and $x\beta_0'(\tau) > c$ for all $\tau \in [0,1]$, where $c > 0$ is a constant. In the Lemma below, we show that $\sigma$ converges to $\sigma_0$ at rate $n^{-\frac{1}{4}}$.

**Lemma 3** (Convergence Rate of $\widehat{\sigma}$)**.** *If conditions A1–A4 hold, the ML estimator $\widehat{\sigma}$ has the following property:*

$$\widehat{\sigma} - \sigma_0 = o_p(n^{-\frac{1}{4}}). \tag{3.5}$$

*Proof.* See Appendix C.2. □

## 4. Sieve Estimation

In the last section we demonstrated that the maximum likelihood estimator restricted to parameter space $\Theta$ converges to the true parameter with probability approaching 1. However, the estimator still lives in a large space with $\beta(\cdot)$ being $d_x$-dimensional co-monotone functions and $\sigma$ being a finite dimensional parameter. Although theoretically such an estimator does exist, in practice it is computationally infeasible to search for the likelihood maximizer within this large space. In this paper, we consider a spline estimator of $\beta(\cdot)$ to mimic the co-monotone functions $\beta(\cdot)$ for their computational advantages in calculating the sieve estimator. The estimator below is easily adapted to the reader's preferred estimator. For simplicity, we use a piecewise constant sieve space, which we define as follows.

**Definition 1** (Sieve Space)**.** Define $\Theta_J = \Omega_J \times \Sigma$, where $\Omega_J$ stands for increasing piecewise constant functions on $[0,1]$ with $J$ knots at $\left\{\frac{j}{J}\right\}$ for $j = 0, 1, ..., J-1$. In other words, for any $\beta(\cdot) \in \Omega_J$, $\beta_k(\cdot)$ is a piecewise constant function on intervals $[\frac{j}{J}, \frac{j+1}{J})$ for $j = 0, \ldots, J-1$ and $k = 1, \ldots, d_x$.

In general, the $L^2$ distance of the space $\Theta_J$ to the true parameter $\theta_0$ satisfies $d_2(\theta_0, \Theta_J) \leq J^{-1}$ for some generic constant $C$, see Chen (2008). It is easy to see that $\Theta_J \subset \Theta$. We propose approximating $\beta(\cdot)$ with a piecewise constant function. While higher-order splines could be used to attain a faster convergence rate under certain assumptions, their computational complexity motivates our use of sieves to estimate the average value of $\beta(\tau)$ over each interval. The sieve estimator is defined as follows:

**Definition 2** (Sieve Estimator)**.**

$$(\beta_J(\cdot), \sigma) = \arg\max_{\theta \in \Theta_{J_n}} \mathbb{E}_n[\log g(y|x, \beta, \sigma)] \tag{4.1}$$

where $J_n \to \infty$ as $n \to \infty$.

For the sieve space $\Theta_J$ and any $(\beta_J, \sigma) \in \Theta_J$, define $\tilde{I}$ as the following matrix:

$$\tilde{I} := E\left[\left(\frac{\int_0^{\frac{1}{J}} f_\tau d\tau, \int_{\frac{1}{J}}^{\frac{2}{J}} f_\tau d\tau, ..., \int_{\frac{J-1}{J}}^{1} f_\tau d\tau}{g}\right)\left(\frac{\int_0^{\frac{1}{J}} f_\tau d\tau, \int_{\frac{1}{J}}^{\frac{2}{J}} f_\tau d\tau, ..., \int_{\frac{J-1}{J}}^{1} f_\tau d\tau}{g}\right)'\right]$$

where $f_\tau := \frac{\partial f(y-x\beta|\sigma)}{\partial \beta}|_{\beta=\beta(\tau)}$. When $J$ goes to infinity, the smallest eigenvalue of $\widetilde{I}$ goes to 0. Therefore, we require the following measure of ill-posedness.

**Condition A5** (Ill-posed Measure)**.** Define $mineigen(I)$ as the minimum eigenvalue for a given matrix $I$. Let one of the following two assumptions on the degree of ill-posedness hold:

(1) Mild ill-posedness: $mineigen(\tilde{I}) \geq C/J^\lambda$ for some $\lambda > 0$ and constant $C > 0$.
(2) Strong ill-posedness: $mineigen(\tilde{I}) \geq C\exp(-\lambda J)$ for some $\lambda > 0$ and constant $C > 0$.

These ill-posed measures are closely related to the smoothness of the PDF of the EIV. A sufficient condition for mild ill-posedness is the following discontinuity assumption on $f$.[6]

**Condition A6** (Discontinuity of $f$)**.** Suppose there exists a positive integer $\lambda$ such that $f \in C^{\lambda-1}(\mathbb{R})$, and the $\lambda^{th}$ order derivative of $f$ equals:

$$f^{(\lambda)}(x) = h(x) + c_\delta \delta(x-a), \tag{4.2}$$

with $h(x)$ being a bounded function and $L^1$ Lipschitz except at $a$, $c_\delta$ being a non-zero constant, and $\delta(x-a)$ is a Dirac $\delta$-function at $a$.

Note that for the symmetric Laplace distribution, $\lambda = 1$. If the PDF of EIV follows a smooth (e.g., Gaussian) distribution then $mineigen(\tilde{I}) \geq C\exp(-\lambda J)$ with $\lambda = 2$.

The following Lemma establishes the consistency of the sieve estimator.

**Lemma 4** (Sieve Estimator Consistency)**.** *If conditions A1-A4 and A5.1 hold and $J_n \to \infty$ slowly enough then the sieve estimator defined in* (4.1) *is consistent.*

*Proof.* See Appendix C.3. □

[6]See Appendix C.3 for a formal statement and proof of this result in Lemma 7, showing that if a function is of the class $C^\lambda$, the minimum eigenvalue of $\tilde{I}$ is of order $O(J^{-\lambda})$ as $J \to \infty$ for $\lambda \in \mathbb{Z}^+$. In general, for smooth functions $f(\cdot)$, the minimum eigenvalue of $\widetilde{I}$ will decay with speed $O(\exp(-J^{-a}))$ for some $a > 0$.

Unlike the usual sieve estimation problem, our problem is ill-posed with decaying eigenvalue with speed $J^\lambda$. However, the curse of dimensionality in $\beta$ is not at play because of the co-monotonicity of $\beta(\cdot)$—each entry of the vector of functions $\beta(\cdot)$ is a function of a single variable $\tau$. It is therefore possible to use sieve estimation to approximate the true functional parameter with the number of intervals in the sieve $J$ growing slower than $\sqrt{n}$.

**Theorem 2.** *Under conditions A1-A4 and A5.1 (the mild ill-posedness case), the following results hold for the sieve-ML estimator:*

(1) Suppose $\gamma \leq \lambda$, where $\gamma$ is defined in Assumption A2. If the number of knots $J_n$ satisfies the growth condition $J_n^{\lambda(1+\gamma)+\gamma}/n^{\frac{1}{4}} \to 0$ as $J_n \to \infty$, then

$$||\beta_{J_n} - \beta_0||_2 = O_p\left(\max\left(\frac{1}{J_n}, \frac{J_n^\lambda}{\sqrt{n}}\right)\right).$$

(2) If $\frac{J_n}{n^{\frac{1}{2\lambda+2}}} \to \infty$, then for every $j = 1, \ldots, J$ there exists a sequence of numbers $\mu_{kjJ_n}$ with $\mu_{kjJ_n} = O\left(\frac{J_n^\lambda}{\sqrt{n}}\right)$, such that

$$\mu_{kjJ_n}(\beta_{k,J}(\tau_j) - \beta_{k,0}(\tau_j)) \underset{d}{\to} \mathcal{N}(0,1).$$

*Proof.* See Appendix C.3. □

By fixing the number of interior points, we can use ML to estimate the sieve estimator. We discuss how to compute the sieve-ML estimator in the next section.

4.1. **Inference via Bootstrap.** In the last section we proved asymptotic normality for the sieve-ML estimator $\theta = (\beta(\tau), \sigma)$. However, computing the convergence speed $\mu_{kjJ}$ for $\beta_{k,J}(\tau_j)$ by explicit formula can be difficult in general. To conduct inference, we recommend using nonparametric bootstrap. Define $(x_i^b, y_i^b)$ as a resampling of data $(x_i, y_i)$ with replacement for bootstrap iteration $b = 1, \ldots B$, and define the estimator

$$\theta^b = \arg\max_{\theta \in \Theta_J} \mathbb{E}_n^b[\log g(y_i^b | x_i^b, \theta)], \tag{4.3}$$

where $\mathbb{E}_n^b$ denotes the operator of empirical average over resampled data for bootstrap iteration $b$. Then our preferred form of the nonparametric bootstrap is to construct the 95% confidence interval pointwise for each covariate $k$ and quantile $\tau$ from the variance $\widehat{Var}(\beta_k(\tau))$ of each vector of bootstrap coefficients $\left\{\beta_k^b(\tau)\right\}_{b=1}^B$ as $\widehat{\beta}_k(\tau) \pm z_{1-\alpha/2} \cdot \sqrt{\widehat{Var}(\beta_k(\tau))}$ where the critical value $z_{1-\alpha/2} \approx 1.96$ for significance level of $\alpha = .05$.

The following lemma establishes the asymptotic normality of the bootstrap estimates and allows us, for example, to use the empirical variance of the bootstrapped parameter estimates to construct bootstrapped confidence intervals.

**Lemma 5** (Validity of the Bootstrap)**.** *Under conditions A1-A5 and choosing the number of knots $J$ according to the condition stated in Theorem 2, the bootstrapped estimates defined in equation* (4.3) *have the following property*

$$\frac{\beta^b_{k,J}(\tau) - \beta_{k,J}(\tau)}{\mu_{kjJ}} \xrightarrow[d]{*} \mathcal{N}(0,1) \tag{4.4}$$

*Proof.* This result follows from Theorem 5.1 of Chen and Pouzo (2013), who establish the validity of the nonparametric bootstrap in semiparametric models for a general functional with finite-dimensional parameterization. □

4.2. **Weighted Least Squares.** Under a normality assumption of the EIV term $\varepsilon$, the maximization of $Q(\cdot|\theta)$ reduces to the minimization of a simple weighted least squares problem. Suppose the disturbance $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Then the maximization problem (4.1) becomes the following, with the parameter vector $\theta = [\beta(\cdot), \sigma]$

$$\max_{\theta'} Q(\theta'|\theta) := \mathbb{E}\left[\log(f(y - x\beta'(\tau))|\theta')\kappa(x, y, \theta)|\theta\right] \tag{4.5}$$

$$= \mathbb{E}\left[\int_\tau^1 \frac{f(y - x\beta(\tau)|\sigma)}{\int_0^1 f(y - x\beta(u)|\sigma)du}\left(-\frac{1}{2}\log(2\pi\sigma'^2) - \frac{(y - x\beta'(\tau))^2}{2\sigma'^2}\right)d\tau\right].$$

It is easy to see from the above equation that the maximization problem of $\beta'(\cdot)|\theta$ is to minimize the sum of weighted least squares. As in standard normal MLE, the FOC for $\beta'(\cdot)$ does not depend on $\sigma'^2$. The $\sigma'^2$ is solved after all the $\beta'(\tau)$ are solved from equation (4.5). Therefore, the estimand can be implemented with an EM algorithm that reduces to iteration on weighted least squares, which is both computationally tractable and easy to implement in practice.

Given an initial estimate of a weighting matrix $W$, the weighted least squares estimates of $\beta$ and $\sigma$ are

$$\begin{aligned} \widehat{\beta}(\tau_j) &= (X'W_jX)^{-1}X'W_jy \\ \widehat{\sigma} &= \sqrt{\frac{1}{NJ}\sum_j\sum_i w_{ij}\widehat{\varepsilon}^2_{ij}} \end{aligned}$$

where $W_j$ is the diagonal matrix formed from the $j^{\text{th}}$ column of $W$, which has elements $w_{ij}$.

Given estimates $\widehat{\varepsilon}_j = y - X\widehat{\beta}(\tau_j)$ and $\widehat{\sigma}$, the weights $w_{ij}$ for observation $i$ in the estimation of $\beta(\tau_j)$ are

$$w_{ij} = \frac{\phi\left(\widehat{\varepsilon}_{ij}/\widehat{\sigma}\right)}{\frac{1}{J}\sum_j \phi\left(\widehat{\varepsilon}_{ij}/\widehat{\sigma}\right)} \tag{4.6}$$

where $\phi(\cdot)$ is the PDF of a standard normal distribution $J$ is the number of $\tau$s in the sieve, e.g. $J = 9$ if the quantile grid is $\{\tau_j\} = \{0.1, 0.2, ..., 0.9\}$.

## 5. Monte-Carlo Simulations

We examine the properties of our estimator empirically in Monte-Carlo simulations. Let the data-generating process be

$$y_i = \beta_0(u_i) + x_{1i}\beta_1(u_i) + x_{2i}\beta_2(u_i) + \varepsilon_i$$

where $n = 100,000$, the conditional quantile $u_i$ of each individual is $u \sim U[0,1]$, and the covariates are distributed as independent lognormal random variables, i.e. $x_{1i},\ x_{2i} \sim LN(0,1)$. The coefficient vector is a function of the conditional quantile $u_i$ of individual $i$

$$\begin{pmatrix} \beta_0(u) \\ \beta_1(u) \\ \beta_2(u) \end{pmatrix} = \begin{pmatrix} 1 + 2u - u^2 \\ \frac{1}{2}\exp(u) \\ u + 1 \end{pmatrix}.$$

In our baseline scenario, we draw mean-zero measurement error $\varepsilon$ from a mixed normal distribution

$$\varepsilon_i \sim \begin{cases} \mathcal{N}(-3,1) & \text{with probability } 0.5 \\ \mathcal{N}(2,1) & \text{with probability } 0.25 \\ \mathcal{N}(4,1) & \text{with probability } 0.25 \end{cases}$$

We also probe the robustness of the mixture specification by simulating measurement error from alternative distributions and testing how well modeling the error distribution as a Gaussian mixture handles alternative scenarios to simulate real-world settings in which the econometrician does not know the true distribution of the residuals.

We use a gradient-based constrained optimizer to find the maximizer of the log-likelihood function defined in Section 3. See Appendix A for a summary of the constraints we impose and analytic characterizations of the log-likelihood gradients for a mixture of three normals. We use quantile regression coefficients for a $\tau$-grid of $J = 9$ knots as start values. For the start values of the distributional parameters, we place

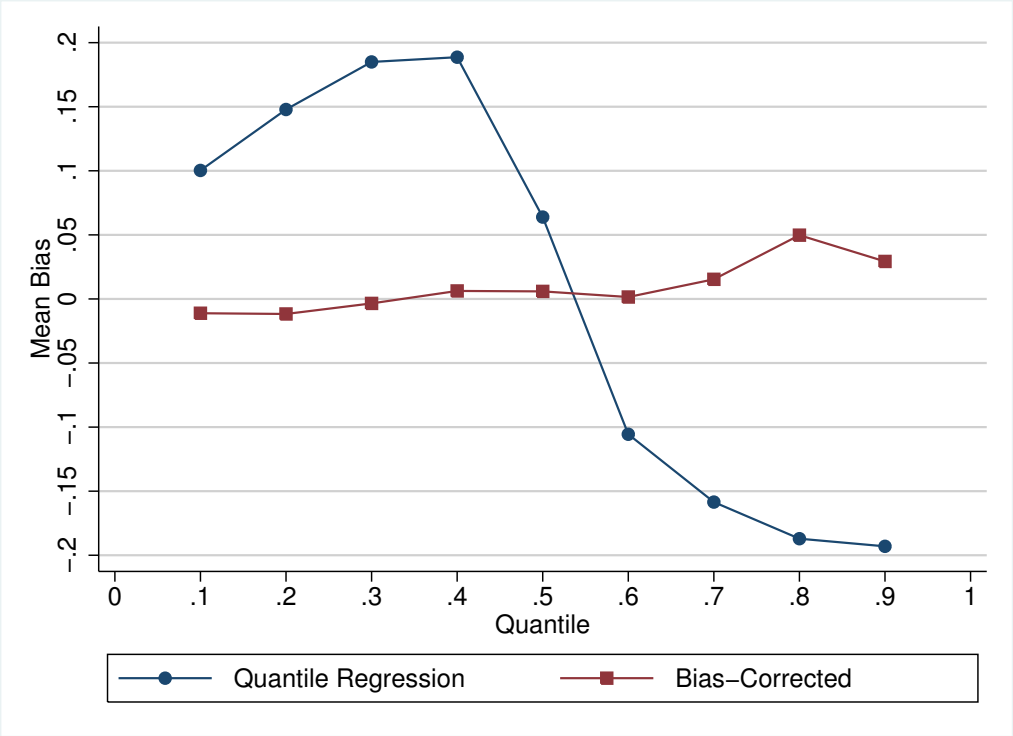
equal 1/3 weights on each mixture component, with unit variance and means -1, 0, and 1.

As discussed in Section 2.1, the likelihood function is invariant to a permutation of the particular quantile labels. For example, the log-likelihood function defined by equations (3.1) and (3.3) would be exactly the same if $\beta(\tau = .2)$ were exchanged with $\beta(\tau = .5)$. Rearrangement helps ensure that the final ordering is consistent with the assumption of $x\beta(\tau)$ being monotonic in $\tau$ and weakly reduces the $L^2$ distance of the estimator $\widehat{\beta}(\cdot)$ with the true parameter functional $\beta(\cdot)$. See Chernozhukov et al. (2009) for further discussion. Accordingly, we sort our estimated coefficient vectors by $\bar{x}\widehat{\beta}(\tau)$ where $\bar{x}$ is the mean of the design matrix across all observations. Given initial estimates $\tilde{\beta}(\cdot)$, we take our final estimates for each simulation to be $\left\{\widehat{\beta}(\tau_j)\right\}$ for $j = 1, ..., J$ where $\widehat{\beta}(\tau_j) = \tilde{\beta}(\tau_r)$ and $r$ is the element of $\tilde{\beta}(\cdot)$ corresponding to the $j^{\text{th}}$ smallest element of the vector $\bar{x}\tilde{\beta}(\cdot)$.

5.1. **Simulation Results.** In Figures 2 and 3, we plot the mean bias (across 500 Monte Carlo simulations) of quantile regression of $y$ (generated with measurement error drawn from a mixture of three normals) on a constant, $x_1$, and $x_2$ and contrast that with the mean bias of our estimator using a sieve for $\beta(\cdot)$ consisting of 9 knots. Quantile regression is badly biased, with lower quantiles biased upwards towards the median-regression coefficients and upper quantiles biased downwards towards the median-regression coefficients. While this pattern of bias towards the median evident in Table 2 still holds, the pattern in Figures 2 and 3 is nonmonotonic for quantiles below the median in the sense that the bias is actually greater for, e.g., $\tau = 0.3$ than for $\tau = 0.1$. Simulations reveal that the monotonic bias towards the median result seems to rely on a symmetric error distribution. Regardless, the bias of the ML estimator is statistically indistinguishable from zero across quantiles of the conditional distribution of $y$ given $x$, with an average mean bias across quantiles of 2% and 1% (for $\beta_1$ and $\beta_2$, respectively) and always less than 5% of the true coefficient magnitude.[7] The mean bias of the quantile regression coefficients, by contrast, is on average over 18% for nonlinear $\beta_1(\cdot)$ and exceeds 27% for some quantiles.

[7]As an example of the validity of the bootstrap (and in particular the asymptotic results in Theorem 2), we varied the sample size in the simulations and calculated how the width of the pointwise 95% confidence intervals changed. Decreasing the sample size from 50,000 to 10,000 observations—an increase in $\sqrt{n}$ by a factor of 2.24—increased the width of the confidence intervals for both $\beta_1$ and $\beta_2$ (averaged across quantiles) by a factor of 2.25.

FIGURE 2. Monte Carlo Simulation Results: Mean Bias of $\widehat{\beta}_1(\tau)$

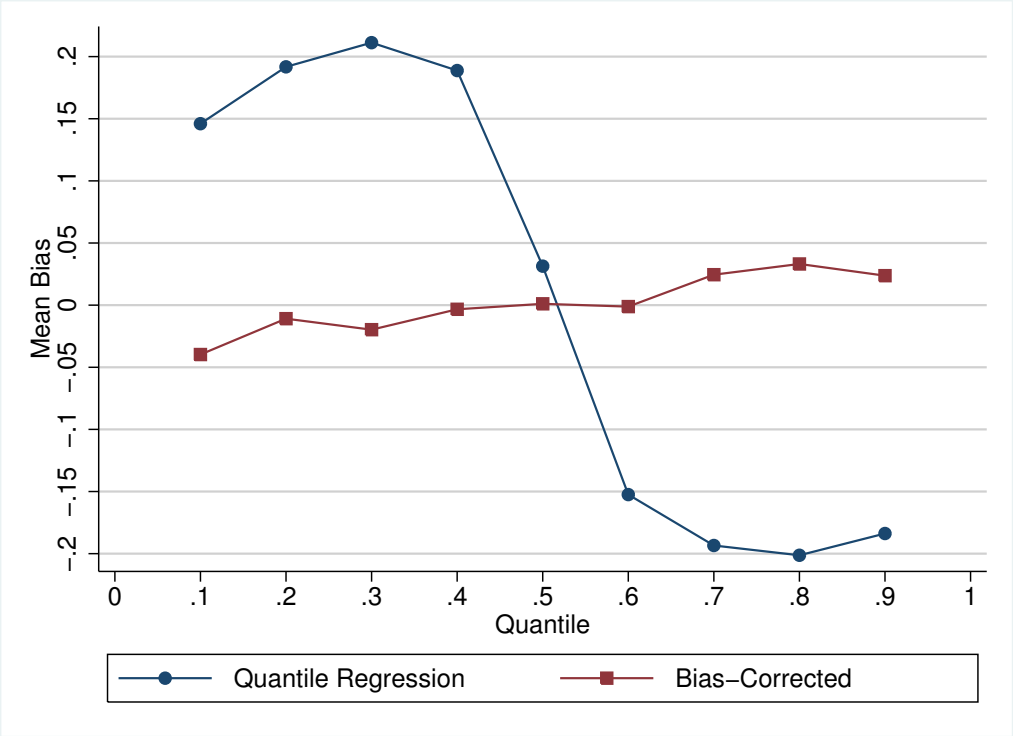



Notes: Figure plots mean bias of estimates of $\beta_1(\tau)$ for classical quantile regression (blue line) and bias-corrected MLE (red line) across 500 MC simulations using the data-generating process described in the text with the measurement error generated as a mixture of three normals.

Figure 4 shows the true mixed-normal distribution of the measurement error $\varepsilon$ as defined above (dashed blue line) plotted with the estimated distribution of the measurement error from the average estimated distributional parameters across all MC simulations (solid red line). The 95% confidence interval of the estimated density (dotted green line) are estimated pointwise as the 2.5th and 97.5th percentiles of EIV densities across all simulations. Despite the bimodal nature of the true measurement error distribution, our algorithm captures the overall features of true distribution very well, with the true density always within the tight confidence interval for the estimated density.

In practice, the econometrician seldom has information on the distribution family to which the measurement error belongs. To probe robustness on this dimension, we demonstrate the flexibility of the Gaussian mixture-of-three specification by showing that it accommodates alternative errors-in-variables data-generating processes well. Table 2 shows that when the errors are distributed with thick tails (as a t-distribution with three degrees of freedom) in panel A or as a mixture of two normals in panel B, the

FIGURE 3. Monte Carlo Simulation Results: Mean Bias of $\widehat{\beta}_2(\tau)$



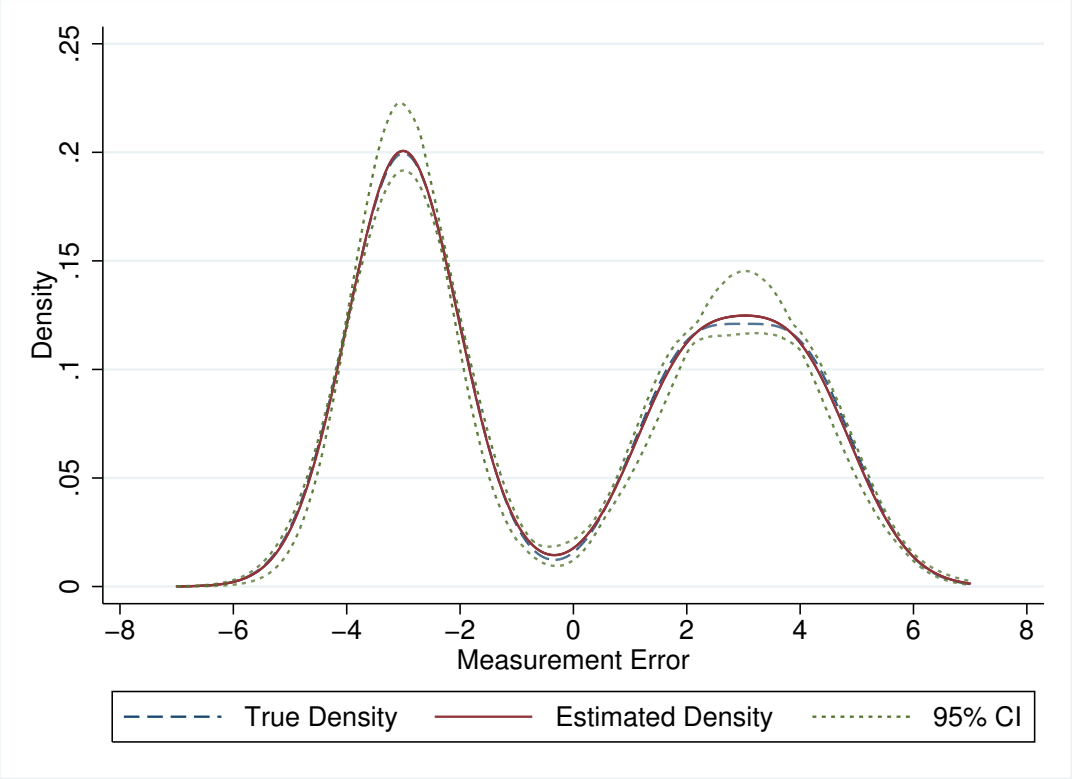

Notes: Figure plots mean bias of estimates of $\beta_2(\tau)$ for classical quantile regression (blue line) and bias-corrected MLE (red line) across 500 MC simulations using the data-generating process described in the text with the measurement error generated as a mixture of three normals.

ML estimates that model the EIV distribution as a mixture of three normals are still unbiased. As expected, quantile regression exhibits typical bias towards the median under both distributions and for both slope coefficients (visible as positive mean bias for quantiles below the median and negative bias for quantiles above the median). By comparison, ML estimates are generally much less biased than quantile regression for both data-generating processes. Our ML framework easily accommodates mixtures of more than three normal components for additional distributional flexibility in a quasi-MLE approach.

## 6. EMPIRICAL APPLICATION

To illustrate the use of our estimator in practice, we examine distributional heterogeneity in the wage returns to education. First, we replicate and extend classical quantile regression results from Angrist et al. (2006) by estimating the quantile-regression analog

FIGURE 4. Monte Carlo Simulation Results: Distribution of Measurement Error




Notes: Figure reports the true measurement error (dashed blue line), a mean-zero mixture of three normals ( $\mathcal{N}(-3,1)$, $\mathcal{N}(2,1)$, and $\mathcal{N}(4,1)$ with weights 0.5, 0.25, and 0.25, respectively) against the average density estimated from the 500 Monte Carlo simulations (solid red line). For each grid point, the dotted green line plots the 2.5th and 97.5th percentile of the EIV density function across all MC simulations.

of a Mincer regression,

$$q_{y|X}(\tau) = \beta_0(\tau) + \beta_1(\tau)education_i + \beta_2(\tau)experience_i + \beta_3(\tau)experience_i^2 + \beta_4(\tau)black_i \tag{6.1}$$

where $q_{y|X}(\tau)$ is the $\tau^{\text{th}}$ quantile of the conditional (on the covariates $X$) log-wage distribution, the *education* and *experience* variables are measured in years, and *black* is an indicator variable. In contrast to the linear Mincer equation, quantile regression assumes that all unobserved heterogeneity enters through the unobserved rank of person $i$ in the conditional wage distribution. The presence of an additive error term, which could include both measurement error and wage factors unobserved by the econometrician, would bias the estimation of the coefficient function $\beta(\cdot)$.

TABLE 2. MC Simulation Results: Robustness to Alternative Data-Generating Processes

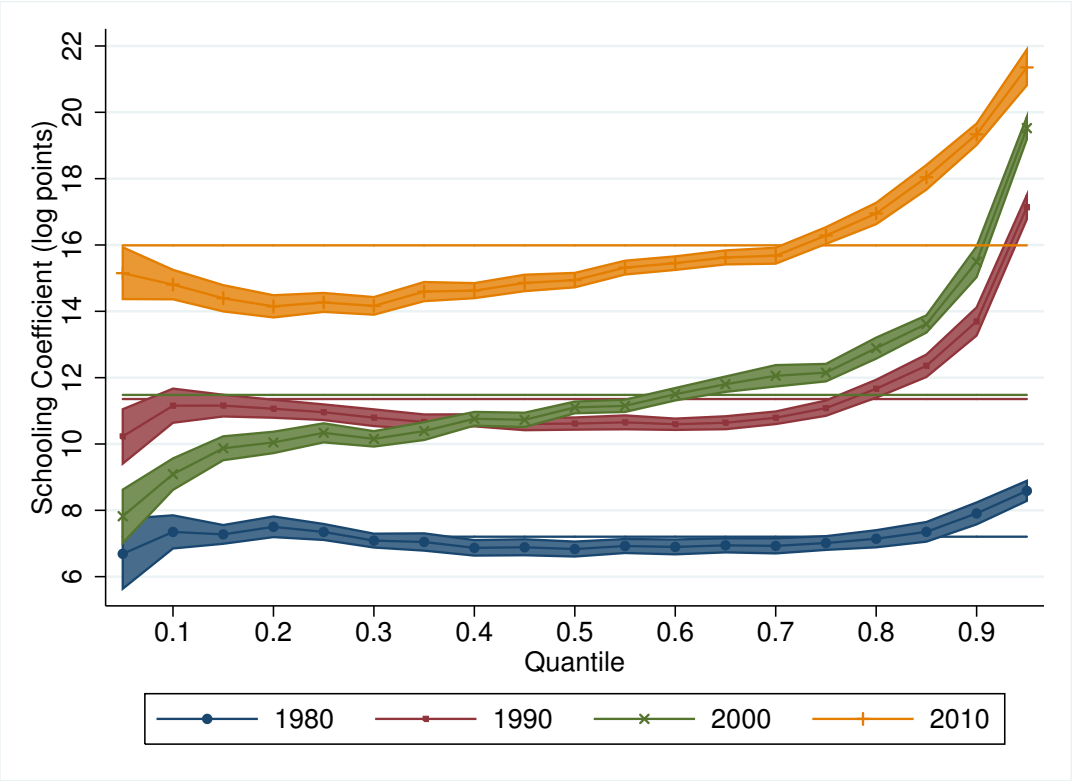
| | A. EIV ∼ T | | | | B. EIV ∼ Mixture of 2 $\mathcal{N}$ | | | |
|---|---|---|---|---|---|---|---|---|
| | $\beta_1$ | | $\beta_2$ | | $\beta_1$ | | $\beta_2$ | |
| Quantile | QReg | MLE | QReg | MLE | QReg | MLE | QReg | MLE |
| 0.1 | 0.07 | 0.01 | 0.09 | -0.03 | 0.14 | 0.04 | 0.18 | 0.03 |
| 0.2 | 0.05 | -0.02 | 0.06 | -0.02 | 0.15 | 0.05 | 0.16 | 0.00 |
| 0.3 | 0.04 | -0.02 | 0.05 | -0.06 | 0.09 | 0.05 | 0.09 | 0.00 |
| 0.4 | 0.03 | 0.00 | 0.03 | -0.02 | 0.03 | 0.05 | -0.01 | 0.01 |
| 0.5 | 0.01 | 0.02 | 0.01 | 0.01 | -0.02 | 0.08 | -0.06 | 0.02 |
| 0.6 | 0.00 | 0.04 | -0.01 | 0.06 | -0.06 | 0.03 | -0.09 | 0.03 |
| 0.7 | -0.03 | 0.05 | -0.03 | 0.05 | -0.09 | 0.05 | -0.11 | 0.00 |
| 0.8 | -0.06 | 0.05 | -0.06 | 0.03 | -0.11 | 0.02 | -0.12 | 0.02 |
| 0.9 | -0.10 | 0.03 | -0.10 | 0.04 | -0.13 | -0.02 | -0.11 | -0.04 |

Note: Table reports mean bias of slope coefficients for estimates from classical quantile regression and bias-corrected MLE across 200 MC simulations of $n = 1,000$ observations each using data simulated from the data-generating process described in the text and the measurement error generated by either a Student's t distribution (left-hand columns) with three degrees of freedom or a mixture of two normals $\mathcal{N}(-2.4, 1)$ and $\mathcal{N}(1.2, 1)$ with weights 1/3 and 2/3, respectively.

Figure 5 plots results of estimating equation (6.1) by quantile regression on census microdata samples from four decennial census years: 1980, 1990, 2000, and 2010, along with simultaneous confidence intervals obtained from 200 bootstrap replications.[8] Horizontal lines in Figure 5 represent OLS estimates of equation 5 for comparison. Consistent with the results in Figure 2 of Angrist et al., we find quantile-regression evidence that heterogeneity in the returns to education across the conditional wage distribution has increased over time. In 1980, an additional year of education was associated with a 7% increase in wages across all quantiles, nearly identical to OLS estimates. In 1990, while the education-wage gradient was mostly constant across the conditional wage distribution, higher conditional (wage) quantiles saw a stronger association between education and wages, especially for top conditional quantiles. By 2000, the education coefficient was roughly seven log points higher for the 95th percentile than for the 5th percentile. Data from 2010 shows a large jump in the returns to education for the entire distribution, with top conditional incomes increasing much less from 2000 to 2010 as bottom

[8]The 1980–2000 data come from Angrist et al.'s IPUMS query, and the 2010 follow their sample selection criteria and again draw from IPUMS (Ruggles et al., 2015). For further details on the data including summary statistics, see Appendix B.

FIGURE 5. Quantile Regression Estimates of the Returns to Education, 1980–2010




Notes: Figure reports quantile regression estimates of log weekly wages (self-reported) on education, a quadratic in experience, and an indicator for blacks for a grid of 19 evenly spaced quantiles from 0.05 to 0.95. Horizontal lines indicate OLS estimates for each year, and bootstrapped 95% simultaneous confidence intervals are plotted for the quantile regression estimates for each year. The data comes from the indicated decennial census year and consist of 40-49 year old white and black men born in America. The number of observations in each sample is 65,023, 86,785, 97,397, and 106,625 in 1980, 1990, 2000, and 2010, respectively.

conditional incomes.[9] Still, the post-1980 convexity of the education-wage gradient is readily visible in the 2010 results, with wages in the top quartile of the conditional distribution being much more sensitive to years of schooling than the rest of the distribution. In 2010, the education coefficient for the 95th percentile percentile was six log points higher than the education coefficient for the 5th percentile. The dependence of the wage-education gradient on the quantile of the wage distribution suggests that average

[9]While some of the increase from 2000–2010 in the returns to education may be driven by selection into employment with the incidence of recession-driven layoffs being more acute on low-wage earners, additional testing shows that results using the 2014 ACS are very similar to the 2010 data. The observation that the OLS estimates also jump from 2000–2010 suggest that this increase is not driven by the difference between the unconditional and conditional distribution of income given education.

or local average treatment effects estimated from linear estimators fail to represent the returns to education for a sizable portion of the population.

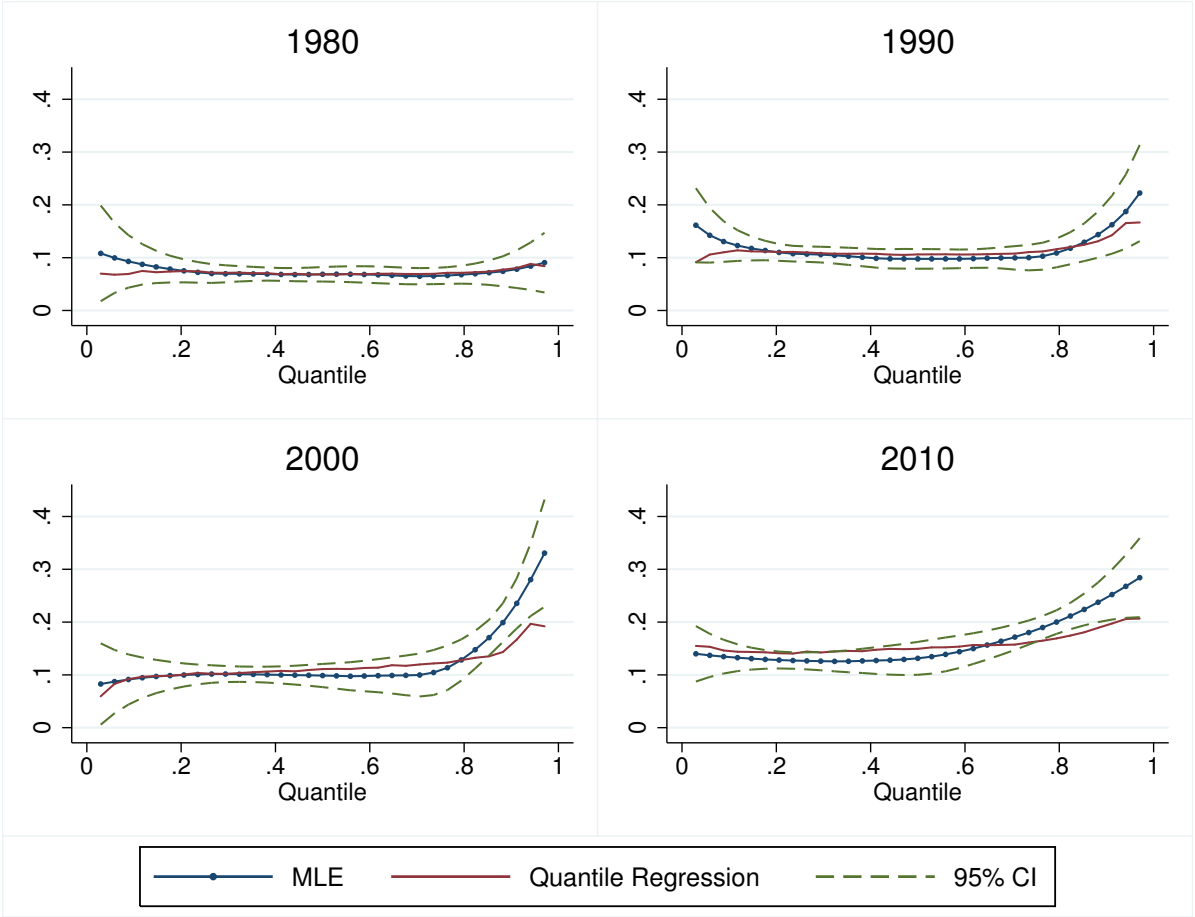
Although quantile regression recovers effects on the conditional distribution of the outcome, it is worth noting two things.[10] First, our exercise here of comparing estimates across years arguably makes the unconditional versus conditional distinction less important. Second, given the substantial variation in wages left unexplained by the Mincer model, the empirical difference between effects on the unconditional and conditional distributions of the dependent variable is likely small. Appendix Figure B1 illustrates this point for the 0.9 quantile estimates, showing that because of the relatively low goodness of fit of equation (6.1) (as is the case in many cross-sectional applied microeconomics settings), over 63% of the observations in the top unconditional decile are also in the top conditional decile.

We observe a different pattern when we correct for measurement-error bias in the self-reported wages used in the census data using ML estimation procedure. We estimate $\beta(\cdot)$ for quantile grid of 33 knots, evenly distributed ( $\tau \in \{j/34\}_{j=1}^{33}$) using our maximum likelihood estimator developed in Section 3 above. As in our simulation results, for the error distribution, we choose a mixture of three normals with the same default distributional start values (equal weights, unit variances, means of -1, 0, 1). For coefficient start values, we run the maximization procedure with start values taken from three alternatives and keep the estimate that results in the higher log-likelihood value: standard quantile regression, the weighted least squares procedure outlined in Section 4.2, and the mean of bootstrapping our ML estimates (using the WLS coefficients as start values for bootstrapping). We again sort our estimates by $\bar{x}\beta(\tau)$ to enforce monotonicity at mean covariate values (see section 5 for details). We smooth our estimates by bootstrapping (following Newton and Raftery, 1994) and then local linear regression of $\widehat{\beta}_1(\tau)$ on $\tau$ to reduce volatility of coefficient estimates across the conditional wage distribution.[11] Finally, we construct 95% bootstrapped confidence intervals pointwise as $\widehat{\beta}_1(\tau) \pm 1.96\widehat{\sigma}_{bs}(\tau)$ for each $\tau$ where $\widehat{\sigma}_{bs}$ is the empirical standard deviation of smoothed bootstrapped estimates of $\widehat{\beta}_1(\tau)$.

[10]See DiNardo et al. (1996) and Powell (2013) for further discussion and methods that recover effects on the unconditional distribution.

[11]Due to ill-posedness, our raw estimates are noisy. Based on the asymptotics of Theorem 1, as long as the bandwidth $h$ of our local-linear estimator is $O(\frac{1}{k})$ where $k$ is the number of knots in $\widehat{\beta}(\cdot)$, our smoothing does not affect asymptotic normality or convergence speed since any additional bias introduced by smoothing is of order $O(h)$ and thus converges to 0 faster than $\widehat{\beta}(\tau) - \beta(\tau)$ for any $\tau$.

FIGURE 6. Returns to Education Correcting for LHS Measurement Error




Notes: Graphs plot education coefficients estimated using quantile regression (red lines) and the ML estimator described in the text (blue line). Green dashed lines plot 95% Confidence Intervals using the bootstrap procedure described in the text. See notes to Figure (5).

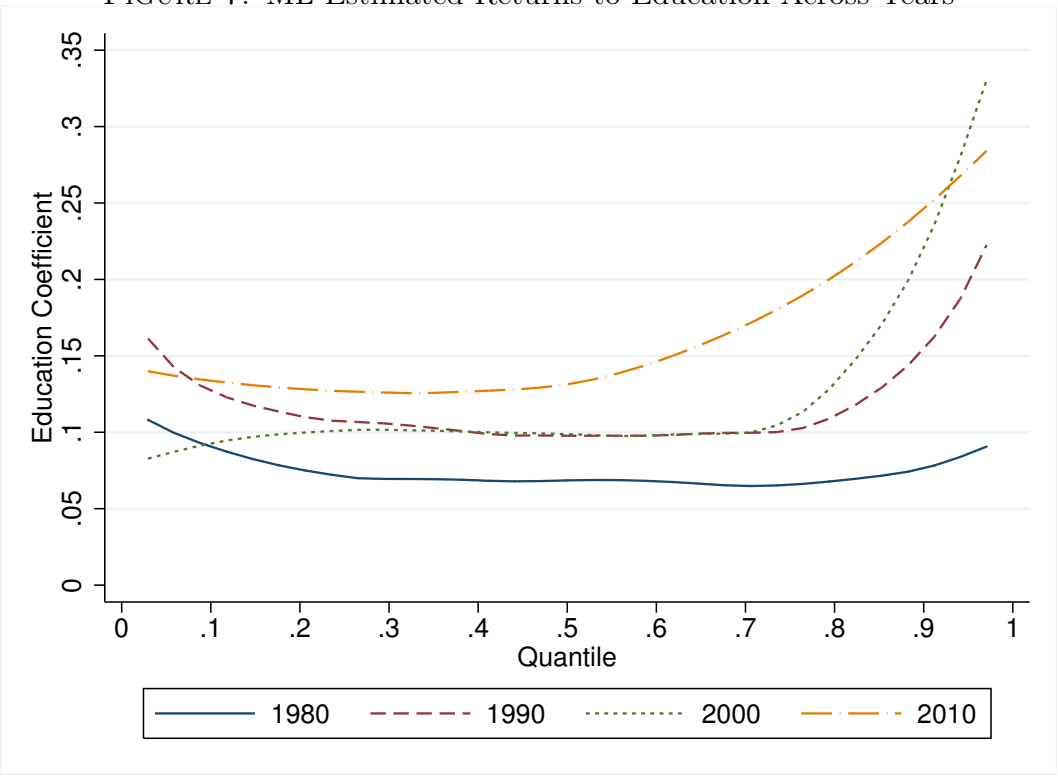
Figure 6 plots the education coefficient $\widehat{\beta}_1(\tau)$ from estimating equation (6.1) by MLE and quantile regression, along with 95% confidence intervals. The results suggest that in 1980, the quantile-regression estimates are relatively unaffected by measurement error in the sense that the classical quantile-regression estimates and bias-corrected ML estimates are nearly indistinguishable. For 1990, the pattern of increasing returns to education for higher quantiles is again visible in the ML estimates with the very highest quantiles seeing an approximately five log point larger increase in the education-wage gradient than suggested by quantile regression, although this difference for top quantiles does not appear statistically significant given typically wide confidence intervals for extremal

quantiles. In the 2000 decennial census, the quantile-regression and ML estimates of the returns to education again diverge for top incomes, with the point estimate suggesting that after correcting for measurement error in self-reported wages, the true returns to an additional year of education for the top of the conditional wage distribution was a statistically significant 13 log points (17 percentage points) higher than estimated by classical quantile regression. This bias correction has a substantial effect on the amount of inequality estimated in the education-wage gradient, with the ML estimates implying that top wage earners gained 23 log points (29 percentage points) more from a year of education than workers in the bottom three quartiles of wage earners. For 2010, both ML and classical quantile-regression estimates agree that the returns to education increased across all quantiles, but again disagree about the marginal returns to schooling for top wage earners. Although the divergence between ML and quantile regression estimates for the top quartile is not as stark as in 2000, the quantile regression estimates at the 95th percentile of the conditional wage distribution are again outside the nonparametric 95% confidence intervals for the ML estimates.

For each year after 1980, the quantile regression lines understate the returns to education in the top tail of the wage distribution. Starting in 1990, correcting for measurement error in self-reported wages significantly increases the estimated returns to education for the top quintile of the conditional wage distribution, a distinction that is missed because of the measurement error in self-reported wage data resulting in compression bias in the quantile regression coefficients. Figure 7 overlays each year's ML estimates to facilitate easier comparisons across years. Over time—especially between 1980 and 1990 and between 2000 and 2010—we see an overall increase in the returns to education, broadly enjoyed across the wage distribution. The increase in the education-wage gradient is relatively constant across the bottom three quartiles and very different for the top quartile.

These two trends—overall moderate increases and acute increases in the schooling coefficient for top earners—are consistent with the observations of Angrist et al. (2006) and other well-known work on inequality that finds significant increases in income inequality post-1980 (e.g. Autor et al., 2008). Nevertheless, the distributional story that emerges from correcting for measurement error suggests that the concentration of education-linked wage gains for top earners is even more substantial than is apparent in previous work. This finding is particularly relevant for recent discussions of top-income inequality

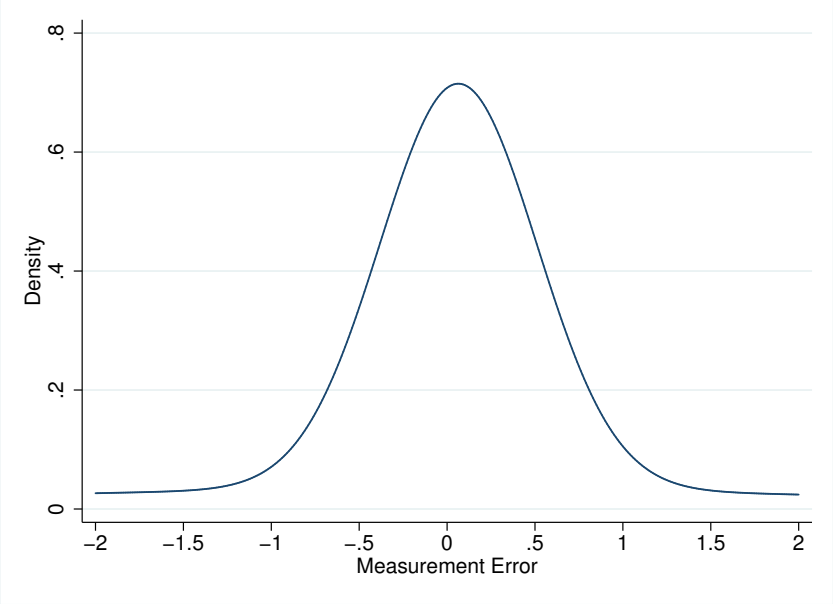
FIGURE 7. ML Estimated Returns to Education Across Years




Notes: Figure overlays ML estimates of the returns to education across the conditional wage distribution from Figure 6. See notes to Figure 6 for details.

(see, for example, Piketty and Saez, 2006) and the increasing returns to cognitive performance (Lin et al., 2016). The time-varying nature of this relationship between the wage distribution and education is suggestive in the role of macroeconomic context in measuring the returns to education. If the wage earnings of highly educated workers at the top of the conditional wage distribution is more volatile, then single-year snapshots of inequality may under or overstate the relationship between wages and education. Judgment based solely on the 2000 pattern of the education gradient would find significantly more inequality in the returns to education than estimates using 2010 data. By 2010, the overall returns to education increased across nearly the entire wage distribution. Whereas the schooling coefficient for the first quartile decreased from 1990-2000 (the only segment of the distribution to do so), by 2010, the first decile was back at 1990-level returns, with the remainder of the distribution below the 90th percentile outpacing the 1990 returns to education. The particularly high education gradient enjoyed by the top quartile in 2000 seems to have been smoothed out and shared by the top half of the

FIGURE 8. Estimated Distribution of Wage Measurement Error



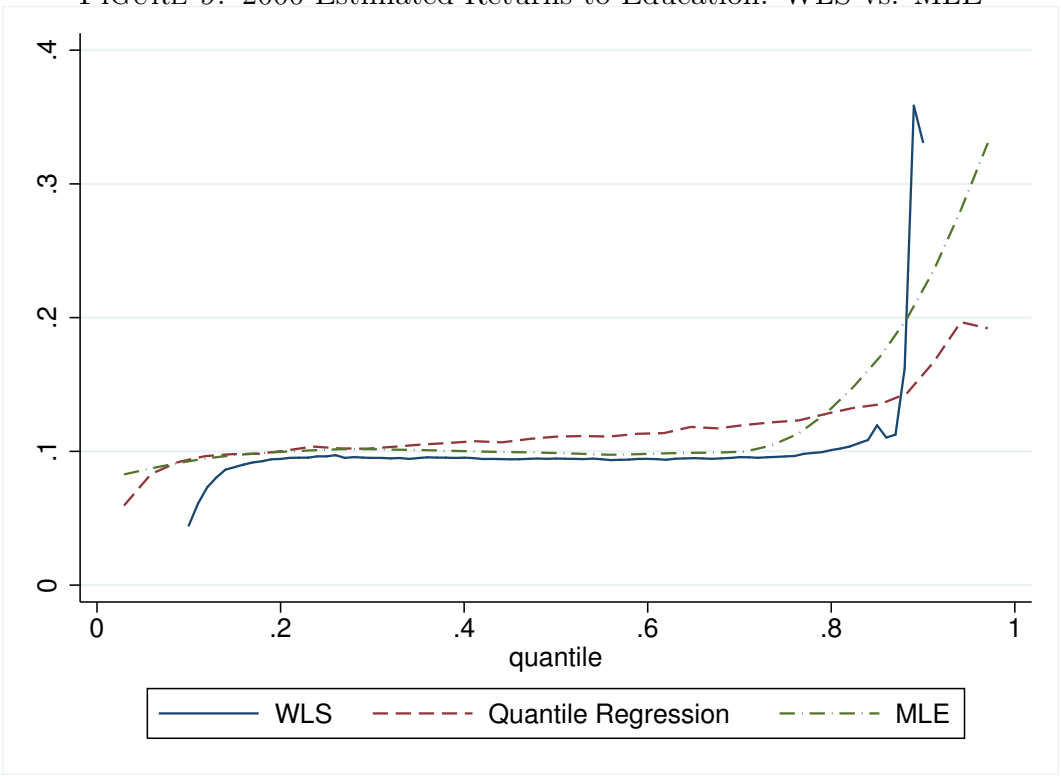

Note: Graph plots the estimated probability density function of the measurement error in 1990 when specified as a mixture of three normal distributions.

wage distribution. Whether the slight decrease in the schooling coefficient for top earners is simply a reflection of their higher exposure to the financial crisis (e.g. hedge-fund managers having larger declines in compensation than average workers) is a question to be asked of future data.

Our methodology also permits a characterization of the distribution of dependent-variable measurement error. Figure 8 plots the estimated distribution of the measurement error in the 1990 data Despite the flexibility afforded by the mixture specification, the estimated density is approximately normal—unimodal and symmetric but with higher kurtosis (fatter tails) than a single normal.

In light of the near-normality of the measurement error distribution estimated in the self-reported wage data, we report results for weighted-least squares estimates of the returns to education (see Section 4.2 for a discussion of the admissibility of the WLS estimator when the EIV distribution is normal). The computational benefits of WLS allow us to estimate the wage gradient over a grid of 99 quantile knots. Figure 9 shows the estimated education-wage gradient across the conditional wage distribution for three estimators—quantile regression, weighted least squares, and MLE. Both the WLS and ML estimates revise the right-tail estimates of the relationship between education and

FIGURE 9. 2000 Estimated Returns to Education: WLS vs. MLE




wages significantly, suggesting that the quantile regression-based estimates for the top quintile of the wage distribution are severely biased from dependent-variable errors in variables. The WLS estimates seem to be particularly affected by the extremal quantile problem (see, e.g. Chernozhukov, 2005), leading us to omit unstable estimates in the top and bottom deciles of the conditional wage distribution. While we prefer our ML estimator, the convenience of the weighted least squares estimator lies in its ability to recover many of the qualitative facts obscured by LHS measurement error bias in quantile regression with a lower computational burden and without the ex-post smoothing (apart from dropping bottom- and top-decile extremal quantile estimates) required to interpret the ML estimates.

## 7. CONCLUSION

In this paper, we develop a methodology for estimating the functional parameter $\beta(\cdot)$ in quantile regression models when there is measurement error in the dependent variable. Assuming that the measurement error follows a distribution that is known up to a finite-dimensional parameter, we establish general convergence speed results for the MLE-based

approach. Under an assumption about the degree of ill-posedness of the problem (A5), we establish the convergence speed of the sieve-ML estimator. When the distribution of the EIV is normal, optimization problem becomes an EM problem that can be computed with iterative weighted least squares. We prove the validity of bootstrapping based on asymptotic normality of our estimator and suggest using a nonparametric bootstrap procedure for inference. Monte Carlo results demonstrate substantial improvements in mean bias of our estimator relative to classical quantile regression when there are modest errors in the dependent variable, highlighted by the ability of our estimator to estimate the simulated underlying measurement error distribution (a bimodal mixture of three normals) with a high-degree of accuracy.

Finally, we revisited the Angrist et al. (2006) question of whether the returns to education across the wage distribution have been changing over time. We find a somewhat different pattern than prior work, highlighting the importance of correcting for errors in the dependent variable of conditional quantile models. When we correct for likely measurement error in the self-reported wage data, we find that top wages have grown much more sensitive to education than wage earners in the bottom three quartiles of the conditional wage distribution, an important source of secular trends in income inequality.

## References


Angrist, J., V. Chernozhukov, and I. Fernández-Val (2006): "Quantile regression under misspecification, with an application to the US wage structure," *Econometrica*, 74(2), 539–563.

Autor, D. H., L. F. Katz, and M. S. Kearney (2008): "Trends in US wage inequality: Revising the revisionists," *The Review of Economics and Statistics*, 90(2), 300–323.

Chen, S. (2002): "Rank estimation of transformation models," *Econometrica*, 70(4), 1683–1697.

Chen, X. (2007): "Large sample sieve estimation of semi-nonparametric models," *Handbook of Econometrics*, 6, 5549–5632.

Chen, X., and D. Pouzo (2013): "Sieve Quasi Likelihood Ratio Inference on Semi/nonparametric Conditional Moment Models," Cowles Foundation Discussion Paper #1897.

Chernozhukov, V. (2005): "Extremal quantile regression," *Annals of Statistics*, pp. 806–839.

Chernozhukov, V., I. Fernandez-Val, and A. Galichon (2009): "Improving point and interval estimators of monotone functions by rearrangement," *Biometrika*, 96(3), 559–575.

Cosslett, S. R. (2004): "Efficient Semiparametric Estimation of Censored and Truncated Regressions via a Smoothed Self-Consistency Equation," *Econometrica*, 72(4), 1277–1293.

——— (2007): "Efficient Estimation of Semiparametric Models by Smoothed Maximum Likelihood," *International Economic Review*, 48(4), 1245–1272.

Dempster, A. P., N. M. Laird, D. B. Rubin, et al. (1977): "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, 39(1), 1–38.

DiNardo, J., N. M. Fortin, and T. Lemieux (1996): "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach," *Econometrica*, 64(5), 1001–1044.

Fan, J. (1991): "On the optimal rates of convergence for nonparametric deconvolution problems," *The Annals of Statistics*, pp. 1257–1272.

Hausman, J. (2001): "Mismeasured variables in econometric analysis: problems from the right and problems from the left," *Journal of Economic Perspectives*, 15(4), 57–68.

Hausman, J. A., J. Abrevaya, and F. M. Scott-Morton (1998): "Misclassification of the dependent variable in a discrete-response setting," *Journal of Econometrics*, 87(2), 239–269.

Hausman, J. A., A. W. Lo, and A. C. MacKinlay (1992): "An ordered probit analysis of transaction stock prices," *Journal of Financial Economics*, 31(3), 319–379.

Horowitz, J. L. (2011): "Applied nonparametric instrumental variables estimation," *Econometrica*, 79(2), 347–394.

Kühn, T. (1987): "Eigenvalues of integral operators with smooth positive definite kernels," *Archiv der Mathematik*, 49(6), 525–534.

Lin, D., R. Lutter, and C. J. Ruhm (2016): "Cognitive Performance and Labor Market Outcomes," NBER Working Paper #22470.

Newey, W. K., and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of Econometrics*, 4, 2111–2245.

Newton, M. A., and A. E. Raftery (1994): "Approximate Bayesian inference with the weighted likelihood bootstrap," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 3–48.

Piketty, T., and E. Saez (2006): "The Evolution of Top Incomes: A Historical and International Perspective," *The American Economic Review*, pp. 200–205.

Powell, D. (2013): "A new framework for estimation of quantile treatment effects: Nonseparable disturbance in the presence of covariates," RAND Working Paper Series WR-824-1.

Ruggles, S., K. Genadek, R. Goeken, J. Grover, and M. Sobek (Minneapolis: University of Minnesota, 2015): "Integrated Public Use Microdata Series Version 6.0 [Machine-readable database]," .

Schennach, S. M. (2008): "Quantile regression with mismeasured covariates," *Econometric Theory*, 24(04), 1010–1043.

Shen, X., et al. (1997): "On methods of sieves and penalization," *The Annals of Statistics*, 25(6), 2555–2591.

Wei, Y., and R. J. Carroll (2009): "Quantile regression with measurement error," *Journal of the American Statistical Association*, 104(487).

## Appendix A. Optimization Details

In this section, for practitioner convenience, we provide additional details on our optimization routine, including analytic characterizations of the gradient of the log-likelihood function. For convenience, we will refer to the log-likelihood $l$ for observation $i$ as

$$l = \log \int_0^1 f_\varepsilon(y - x\beta(\tau)|\sigma)d\tau$$

where $\varepsilon$ is distributed as a mixture of $L$ normal distributions, with probability density function

$$f_\varepsilon(u) = \sum_{\ell=1}^{L} \frac{\pi_\ell}{\sigma_\ell}\phi\left(\frac{u-\mu_\ell}{\sigma_\ell}\right).$$

For the mixture of normals, the probability weights $\pi_\ell$ on each component $\ell$ must sum to unity. Similarly, for the measurement error to be mean zero, $\sum_\ell \mu_\ell \pi_\ell = 0$, where $\mu_\ell$ is the mean of each component. For a three-component mixture, this pins down

$$\mu_3 = -\frac{\mu_1\pi_1 + \mu_2\pi_2}{1-\pi_1-\pi_2}$$

(wherein we already used $\pi_3 = 1 - \pi_1 - \pi_2$). We also need to require that each weight be bounded by $[0, 1]$. To do this, we used a constrained optimizer and require that each of $\pi_1, \pi_2, 1 - \pi_1 - \pi_2 \geq 0.01$. The constraints on the variance of each component are that $\sigma_\ell^2 \geq 0.01$ for each $\ell$.

Using the piecewise constant form of $\beta(\cdot)$, let $\beta(\tau)$ be defined as

$$\beta(\tau) = \begin{cases} \beta_1 & \text{when } \tau_0 \leq \tau < \tau_1 \\ \beta_2 & \text{when } \tau_1 \leq \tau < \tau_2 \\ \dots & \dots \\ \beta_{\mathcal{T}} & \text{when } \tau_{\mathcal{T}-1} \leq \tau < \tau_{\mathcal{T}} \end{cases}$$

where $\tau_0 = 0$ and $\tau_{\mathcal{T}} = 1$. Ignoring the constraints on the weight and mean of the last mixture component for the moment, the first derivatives of $l$ with respect to each coefficient $\beta_j$ and distributional parameter are

$$\frac{\partial l}{\partial \pi_\ell} = \frac{1}{\int_0^1 f_\varepsilon(y - x\beta(\tau))d\tau} \int_0^1 \frac{1}{\sigma_\ell} \phi\left(\frac{y - x\beta(\tau) - \mu_\ell}{\sigma_\ell}\right) d\tau$$

$$\frac{\partial l}{\partial \mu_\ell} = \frac{1}{\int_0^1 f_\varepsilon(y - x\beta(\tau))d\tau} \int_0^1 \frac{\pi_\ell}{\sigma_\ell} \left(\frac{y - x\beta(\tau) - \mu_\ell}{\sigma_\ell^2}\right) \phi\left(\frac{y - x\beta(\tau) - \mu_\ell}{\sigma_\ell}\right) d\tau$$

$$\begin{aligned}\frac{\partial l}{\partial \sigma_\ell} = & \frac{1}{\int_0^1 f_\varepsilon(y - x\beta(\tau))d\tau} \int_0^1 -\frac{\pi_\ell}{\sigma_\ell^2} \phi\left(\frac{y - x\beta(\tau) - \mu_\ell}{\sigma_\ell}\right) d\tau \\ & + \frac{1}{\int_0^1 f_\varepsilon(y - x\beta(\tau))d\tau} \int_0^1 \frac{1}{\sigma_\ell} \left(\frac{(y - x\beta(\tau) - \mu_\ell)^2}{\sigma_\ell^3}\right) \phi\left(\frac{y - x\beta(\tau) - \mu_\ell}{\sigma_\ell}\right) d\tau\end{aligned}$$

$$\begin{aligned}\frac{\partial l}{\partial \beta_j} &= \frac{1}{\int_0^1 f_\varepsilon(y - x\beta(\tau))d\tau} \int_{\tau_{j-1}}^{\tau_j} \frac{\partial f_\varepsilon(y - x\beta_j)}{\partial \beta_j} d\tau \\ &= \frac{1}{\int_0^1 f_\varepsilon(y - x\beta(\tau))d\tau} (\tau_j - \tau_{j-1}) \frac{\partial f_\varepsilon(y - x\beta_j)}{\partial \beta_j} \\ &= \frac{(\tau_j - \tau_{j-1})}{\int_0^1 f_\varepsilon(y - x\beta(\tau))d\tau} f'_\varepsilon(y - x\beta_j)(-x) \\ &= \frac{-(\tau_j - \tau_{j-1})\, x}{\int_0^1 f_\varepsilon(y - x\beta(\tau))d\tau} \sum_{\ell=1}^{3} \frac{\pi_\ell}{\sigma_\ell^2} \phi\left(\frac{y - x\beta_j - \mu_\ell}{\sigma_\ell}\right) \left(-\frac{y - x\beta_j - \mu_\ell}{\sigma_\ell}\right).\end{aligned}$$

Incorporating the constraints on the final $L^{\text{th}}$ mixture weight and mean changes the first-order conditions for the means and weights on the penultimate components. Denoting these constrained parameters $\tilde{\pi}_\ell$ and $\tilde{\mu}_\ell$ for $\ell = 1, \ldots L-1$ strictly less than the number of mixtures, the new first derivatives for the first $L-1$ means and weights are functions of the unconstrained derivatives $\partial l/\partial \pi_\ell$ and $\partial l/\partial \mu_\ell$:

$$\frac{\partial l}{\partial \tilde{\pi}_\ell} = \frac{\partial l}{\partial \pi_\ell} - \frac{\partial l}{\partial \pi_L} - \frac{\mu_\ell(1 - \sum_{\ell=1}^{L-1} \pi_\ell) + \sum_{\ell=1}^{L-1} \pi_\ell \mu_\ell}{(1 - \sum_{\ell=1}^{L-1} \pi_\ell)^2} \frac{\partial l}{\partial \mu_L}$$

$$\frac{\partial l}{\partial \tilde{\mu}_\ell} = \frac{\partial l}{\partial \mu_\ell} - \frac{\pi_\ell}{\sum_{\ell=1}^{L-1} \pi_\ell} \frac{\partial l}{\partial \mu_L}$$

## Appendix B. Data Appendix

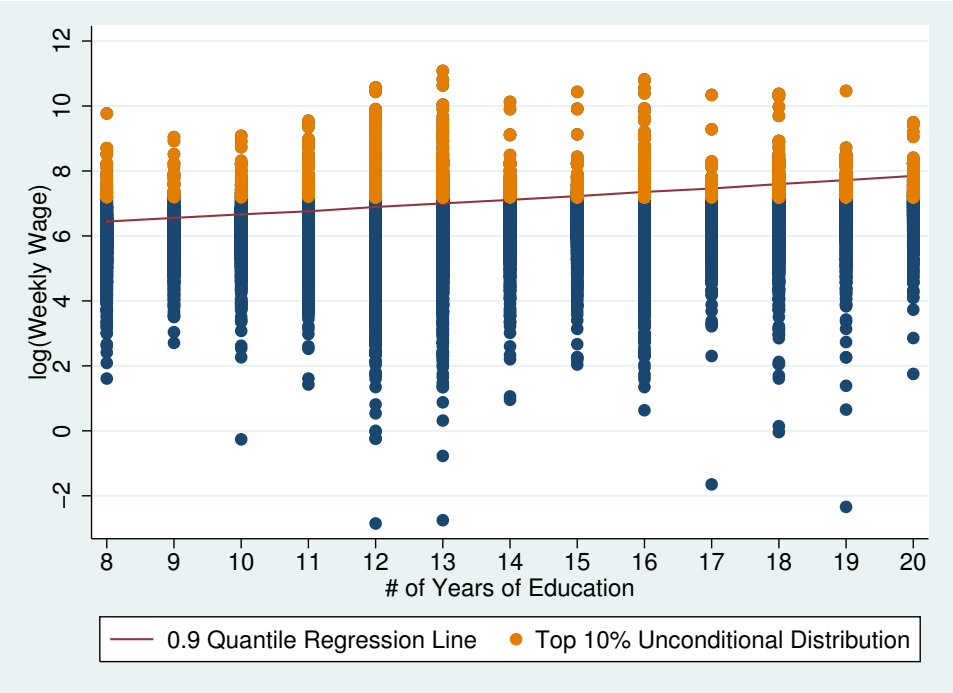
Following the sample selection criteria of Angrist et al. (2006), our data comes from 1% samples of decennial census data available via IPUMS.org (Ruggles et al., 2015) from 1980–2010. From each database, we select annual wage income, education, age, and race data for prime-age (age 40-49) black and white males who have at least five years of education, were born in the United States, had positive earnings and hours worked in the reference year, and whose responses for age, education, and earnings were not imputed (which would have been an additional source of measurement error). Our dependent variable is log weekly wage, obtained as annual wage income divided by weeks worked. For 1980, we take the number of years of education to be the highest grade completed and follow the methodology of Angrist et al. (2006) to convert the categorical education variable in 1990, 2000, and 2010 into a measure of the number of years of schooling. Experience is defined as age minus years of education minus five. For 1980, 1990, and 2000, we use the exact extract of Angrist et al., and draw our own data to extend the data to include the 2010 census. Table B1 reports summary statistics for the variables used in the regressions in the text. Wages for 1980–2000 were expressed in 1989 dollars after deflating using the Personal Consumption Expenditures Index. As slope coefficients in a log-linear quantile regression specification are unaffected by scaling the dependent variable, we do not deflate our 2010 data.

Table B1. Education and Wages Summary Statistics

| Year | 1980 | 1990 | 2000 | 2010 |
|---|---|---|---|---|
| Log weekly wage | 6.40 | 6.46 | 6.47 | 8.34 |
| | (0.67) | (0.69) | (0.75) | (0.78) |
| Education | 12.89 | 13.88 | 13.84 | 14.06 |
| | (3.10) | (2.65) | (2.40) | (2.37) |
| Experience | 25.46 | 24.19 | 24.50 | 24.60 |
| | (4.33) | (4.02) | (3.59) | (3.82) |
| Black | 0.076 | 0.077 | 0.074 | 0.078 |
| | (0.27) | (0.27) | (0.26) | (0.27) |
| Number of Observations | 65,023 | 86,785 | 97,397 | 106,625 |

Notes: Table reports summary statistics for the Census data used in the quantile wage regressions in the text. The 1980, 1990, and 2000 datasets come from Angrist et al. (2006). Following their sample selection, we extended the sample to include 2010 Census microdata from IPUMS.org (Ruggles et al., 2015).

FIGURE B1. Overlap between Unconditional and Conditional Wage Distribution




Notes: Figure plots log weekly wages against years of education from the 1990 decennial Census microdata extract used by Angrist et al. (2006). The regression line plots the average predicted values by year of education from estimating equation (6.1) by classic quantile regression. Lighter colored dots indicate observations in the top 10% of the unconditional wage distribution (individuals with over $1,326 in weekly wages in 1989 dollars).

## Appendix C. Proofs of Lemmas and Theorems

### C.1. **Lemmas and Theorems in Section 2.**

*Proof of Lemma 1.* For bounded monotonic functions, pointwise convergence is equivalent to uniform convergence, making a space of bounded monotonic functions compact under any $L^p$ norm for $p \geq 1$. Hence the product space $B_1 \times B_2 \times ... \times B_k$ is compact. It is complete since the $L^p$ functional space is complete and the limit of monotonic functions is still monotonic. □

*Proof of Theorem 1.* WLOG, under conditions A1-A2, we can assume that the variable $x_1$ is continuous and the conditional distribution of $x_1|x_{-1}$ is continuous. If there exist $\beta(\cdot)$ and $f(\cdot)$ which generate the same density $g(y|x, \beta(\cdot), f(\cdot))$ as the true parameters $\beta_0(\cdot)$ and $f_0(\cdot)$ then by applying a Fourier transformation,

$$\phi(s)\int_0^1 \exp(isx\beta(\tau))d\tau = \phi_0(s)\int_0^1 \exp(isx\beta_0(\tau))d\tau.$$

Denote $m(s) = \frac{\phi(s)}{\phi_0(s)}$. Since $f$ and $f_0$ both satisfy $\int_{-\infty}^{\infty} \epsilon f(\epsilon) = 0$, $m(s) = 1 + O(s^2)$ in an open neighborhood of 0. Therefore,

$$m(s)\int_0^1 \exp(isx_{-1}\beta_{-1}(\tau))\sum_{i=0}^{\infty}\frac{(is)^k x_1^k \beta_1(\tau)^k}{k!}d\tau = \int_0^1 \exp(isx_{-1}\beta_{0,-1}(\tau))\sum_{i=0}^{\infty}\frac{(is)^k x_1^k \beta_{0,1}(\tau)^k}{k!}d\tau.$$

Since $x_1$ is continuous, then it must be that the corresponding polynomials of $x_1$ are the same for both sides. Namely, for any $k \geq 1$,

$$m(s)\frac{(is)^k}{k!}\int_0^1 \exp(isx_{-1}\beta_{-1}(\tau))\beta_1(\tau)^k d\tau = \frac{(is)^k}{k!}\int_0^1 \exp(isx_{-1}\beta_{0.-1}(\tau))\beta_{0,1}(\tau)^k d\tau.$$

Dividing both sides of the above equation by $(is)^k/k!$ and letting $s$ approach 0,

$$\int_0^1 \beta_1(\tau)^k d\tau = \int_0^1 \beta_{0,1}(\tau)^k d\tau$$

for any $k \geq 1$. By Assumption A2, $\beta_1(\cdot)$ and $\beta_{0,1}(\cdot)$ are both strictly monotone, differentiable and greater than or equal to 0. So $\beta_1(\tau) = \beta_{0,1}(\tau)$ for all $\tau \in [0, 1]$.

Now consider the same equation considered above. Dividing both sides by $(is)^k/k!$, we get $m(s)\int_0^1 \exp(isx_{-1}\beta_{-1}(\tau))\beta_{0,1}(\tau)^k d\tau = \int_0^1 \exp(isx_{-1}\beta_{0.-1}(\tau))\beta_{0,1}(\tau)^k d\tau$ for all $k \geq 0$.

As $\beta_{0,1}(\tau)^k, k \geq 0$ is a functional basis of $L^2[0, 1]$, therefore $m(s)\exp(isx_{-1}\beta_{-1}(\tau)) = \exp(isx_{-1}\beta_{0,-1}(\tau))$ for all $s$ in a neighborhood of 0 and all $\tau \in [0, 1]$. If we differentiate both sides with respect to $s$ and evaluate them at 0 (notice that $m'(0) = 0$),

$$x_{-1}\beta_{-1}(\tau) = x_{-1}\beta_{0,-1}(\tau),$$

for all $\tau \in [0, 1]$.

By Assumption A2, $E[x'x]$ is non-singular. Ergo $E[x'_{-1}x_{-1}]$ is also non-singular, and the above equation suggests $\beta_{-1}(\tau) = \beta_{0,-1}(\tau)$ for all $\tau \in [0, 1]$. Therefore, $\beta(\tau) = \beta_0(\tau)$, implying that $\phi(s) = \phi_0(s)$. Consequently, $f(\varepsilon) = f_0(\varepsilon)$. □

C.2. **Lemmas and Theorems in Section 3.** The following additional lemma is used in the proofs of Lemmas 2 and 3.

**Lemma 6** (Donskerness of $\Theta$)**.** *The set of functions* $\mathcal{G} = \{h(y, x, \beta(\cdot), \sigma) := \log(g(y|x, \beta(\cdot), \sigma))|(\beta(\cdot), \sigma) \in \Theta\}$ *is* $\mu$*-Donsker, where* $\mu$ *is the joint PDF of* $(y, x)$*.*

*Proof.* By theorem 2.7.5 of Van der Vaart and Wellner (2000), the space of uniformly bounded monotone functions $\mathcal{F}$ satisfies

$$\log N_{[]}(\varepsilon, \mathcal{F}, L_r(Q)) \leq K\frac{1}{\varepsilon},$$

for every probability measure Q and every $r \geq 1$ and a constant $K$ which depends only on $r$. Consider a collection of functions $\mathcal{F} := q(y, x, \theta)|\theta \in \Theta$ such that

$$|q(y, x, \theta_1) - q(y, x, \theta_2)| \leq ||\theta_1 - \theta_2||_2 w(y, x). \tag{C.1}$$

$$E_Q[|w(y, x)|^2] < \infty, \tag{C.2}$$

where $Q$ is some probability measure on $(y, x)$. Since $\Theta$ is a product space of bounded monotone functions $M$ and a finite-dimensional bounded compact set $\Sigma$, the bracketing number of $\mathcal{F}$ given measure $Q$ is also bounded by

$$\log N_{[]}(\varepsilon, \mathcal{F}, L_2(Q)) \leq K d_x \frac{1}{\varepsilon},$$

where $K$ is a constant only depend on $\Theta$ and $w(y, x)$. Therefore, $\int_0^\delta \sqrt{\log N_{[]}(\varepsilon, \mathcal{F}, Q)} < \infty$, i.e., $\mathcal{F}$ is Donsker.

In particular, let $q = \log g$ and $Q = \mu$, where $\mu$ is the joint pdf of $(x, y)$. By A3.8, equation (3.1) holds with $w(y|x) := \int_0^1 |y - x\beta(\tau)|^\gamma d\tau$. Equation (3.2) is satisfied by Assumption A3.6. Hence, $\mathcal{G}$ is $\mu$-Donsker. □

*Proof of Lemma 2.* To show the consistency of the ML estimator, it is sufficient to prove the satisfaction of the following regularity conditions (Newey and McFadden, 1994).

(1) The parameter space $\Theta = M \times \Sigma$ is compact.
(2) Global identification holds, i.e., there exists no other $\theta' = (\beta', \sigma') \in \Theta$ such that $E[\log \int_0^1 f(y - x\beta'(\tau)|\sigma')d\tau] = E[\log \int_0^1 f(y - x\beta_0(\tau)|\sigma_0)d\tau]$.
(3) The objective function $E[\log \int_0^1 f(y - x\beta'(\tau)|\sigma')d\tau]$ is continuous for all $\theta' = (\beta', \sigma') \in \Theta$.
(4) Stochastic equicontinuity of $E_n[\log \int_0^1 f(y - x\beta(\tau)|\sigma)]$, with $\theta \in \Theta$.

Condition (1) is established by Lemma 1. Condition (2) is provided by Theorem 1. Condition (3) holds under assumptions A2–A3. For the proof of point (4), see Lemma 6 above. Therefore, the ML estimator defined herein is consistent. □

*Proof of Lemma 3.* By Lemma 2, we know that $||\sigma - \sigma_0||_2 \to_p 0$ and $||\beta - \beta_0||_\infty \to_p 0$, i.e., both estimators for $\beta_0$ and $\sigma_0$ are consistent. MLE by definition implies that $E_n[\log(g(y|x, \beta, \sigma))] \leq E_n[\log(g(y|x, \beta_0, \sigma_0))]$.

By the information inequality, $E[\log(g(y|x, \beta_0, \sigma_0))] \leq E[\log(g(y|x, \beta, \sigma))]$.

By Lemma 6, $\mathcal{G} = \{h(y,x,\beta(\cdot),\sigma) := \log(g(y|x,\beta(\cdot),\sigma))|(\beta(\cdot),\sigma) \in \Theta\}$ is Donsker. Thus,

$$E_n[\log(g(y|x,\beta,\sigma))] - E[\log(g(y|x,\beta,\sigma))]$$
$$= E_n[\log(g(y|x,\beta_0,\sigma_0))] - E[\log(g(y|x,\beta_0,\sigma_0))] + o_p(1/\sqrt{n}),$$

implying that

$$E[\log(g(y|x,\beta,\sigma))] - E[\log(g(y|x,\beta_0,\sigma_0))]$$
$$= E_n[\log(g(y|x,\beta,\sigma))] - E_n[\log(g(y|x,\beta_0,\sigma_0))] - o_p(1/\sqrt{n}) \geq -o_p(1/\sqrt{n})$$

and $0 \leq E[\log(g(y|x,\beta_0,\sigma_0))] - E[\log(g(y|x,\beta,\sigma))] \precsim_p 1/\sqrt{n}$.

Let $z(y,x) = g(y|x,\beta,\sigma) - g(y|x,\beta_0,\sigma_0)$ and define $||z(y,x)||_1 := \int_{-\infty}^{\infty} |z(y|x)|dy$. Then by the Scheffe Theorem and Pinsker's Inequality,

$$E_x[||z(y|x)||_1^2] \leq D(g(\cdot|\beta_0)||g(\cdot|\beta)) \tag{C.3}$$
$$\leq 2(E[\log(g(y|x,\beta_0,\sigma_0))] - E[\log(g(y|x,\beta,\sigma))]) \precsim_p \frac{1}{\sqrt{n}},$$

where $D(P|Q)$ is the K-L divergence between two probability distribution $P$ and $Q$. Now consider the characteristic functions of $x\beta(\tau)$ and $x\beta_0(\tau)$ conditional on $x$ and given that $\tau \sim U[0,1]$:

$$\phi_{x\beta}(s) = \frac{\int_{-\infty}^{\infty} g(y|x,\beta,\sigma)e^{isy}dy}{\phi_\varepsilon(s|\sigma)}$$
$$\phi_{x\beta_0}(s) = \frac{\int_{-\infty}^{\infty} g(y|x,\beta_0,\sigma_0)e^{isy}dy}{\phi_\varepsilon(s|\sigma_0)}$$

Then for any $x$ and $s$, $|\phi_{x\beta}(s)\phi_\varepsilon(s|\sigma) - \phi_{x\beta_0}(s)\phi_\varepsilon(s|\sigma_0)| = |\int_{-\infty}^{\infty} z(y|x)e^{isy}dy| \leq ||z(y|x)||_1$. Defining $m(s) := \phi_\varepsilon(s|\sigma_0)/\phi_\varepsilon(s|\sigma)$ and dividing both sides by $\phi_\epsilon(s|\sigma)\phi_{x\beta_0}(s)$,

$$|m(s) - \frac{\phi_{x\beta}(s)}{\phi_{x\beta_0}(s)}| \leq ||z(y|x)||_1/|\phi_{x\beta_0}(s)\phi_\epsilon(s|\sigma)|. \tag{C.4}$$

Plugging in (C.4) back into (C.3) and applying Condition A4.1,

$$E_x[|m(s) - \frac{\phi_{x\beta}(s)}{\phi_{x\beta_0}(s)}|^2] \leq E_x[||z(y|x)||_1^2/|\phi_{x\beta_0}(s)\phi_\epsilon(s|\sigma)|^2] \tag{C.5}$$
$$\leq E_x[||z(y|x)||_1^2]\frac{s^2}{C^2\phi_\epsilon(s|\sigma)} \precsim_p \frac{1}{n^{1/2}}\frac{s^2}{\phi_\epsilon(s|\sigma)}.$$

Using the fact that for any random variable $a$ and any number $b$, $Var(a) \leq E[(a-b)^2]$, we have that $E_x[|m(s) - \frac{\phi_{x\beta}(s)}{\phi_{x\beta_0}(s)}|^2] \geq Var_x(\frac{\phi_{x\beta}(s)}{\phi_{x\beta_0}(s)})$. Inequality (C.5) then implies that

$$Var_x(\frac{\phi_{x\beta}(s)}{\phi_{x\beta_0}(s)}) \precsim_p \frac{1}{n^{1/2}}\frac{s^2}{\phi_\epsilon(s|\sigma)}. \tag{C.6}$$

Applying Condition A4.2, inequality (C.6) implies that

$$E_x\left[\left|\frac{\phi_{x\beta}(s)-\phi_{x\beta_0}(s)}{\phi_{x\beta_0}(s)}\right|^2\right]] \precsim_p \frac{1}{n^{1/2}}\frac{s^2}{\phi_\epsilon(s|\sigma)}. \tag{C.7}$$

We can rewrite $m(s)-\frac{\phi_{x\beta}(s)}{\phi_{x\beta_0}(s)}$ as $(m(s)-1)-\frac{\phi_{x\beta}(s)-\phi_{x\beta_0}(s)}{\phi_{x\beta_0}(s)}$. Using that $\frac{1}{2}a^2-b^2\leq(a-b)^2$ for any $a,b\in\mathbb{R}$, we can bound inequality (3.5) from below and get that

$$\frac{1}{2}E_x\left[|m(s)-1|^2\right]]-E\left[\left|\frac{\phi_{x\beta}(s)-\phi_{x\beta_0}(s)}{\phi_{x\beta_0}(s)}\right|^2\right]]\leq E_x\left[\left|m(s)-\frac{\phi_{x\beta}(s)}{\phi_{x\beta_0}(s)}\right|^2\right]\precsim_p \frac{1}{n^{1/2}}\frac{s^2}{\phi_\epsilon(s|\sigma)}. \tag{C.8}$$

Combining (C.8) with (C.7),

$$E_x\left[|m(s)-1|\right]\precsim_p \frac{1}{n^{1/2}}\frac{s^2}{\phi_\epsilon(s|\sigma)}, \tag{C.9}$$

or, equivalently,

$$|\phi_\epsilon(s|\sigma_0)-\phi_\epsilon(s|\sigma)|^2\precsim_p \frac{s^2}{n^{1/2}}. \tag{C.10}$$

Thus, applying Condition A3.9 along with (C.10), it follows that $||\sigma-\sigma_0||_2\precsim_p n^{-1/4}$. □

C.3. **Lemmas and Theorem in Section 4.** The following lemma establishes mild ill-posedness (Condition A5.1).

**Lemma 7.** *If the function $f$ satisfies Conditions A1–A4 and A6 with degree $\lambda>0$ then*

(1) the minimum eigenvalue of $\tilde{I}$, denoted as $r(\tilde{I})$, satisfies

$$\frac{1}{J^\lambda}\precsim r(\tilde{I}).$$

(2) for any $\theta$, $\frac{1}{J^\lambda}\precsim \sup_{p\in\Theta_J-\theta,p\neq 0}\frac{||p||_d}{||p||}$ where $||p||_d:=\left|p'\widetilde{I}p\right|^{1/2}$.

*Proof.* Suppose $f$ satisfies discontinuity Condition A6 with degree $\lambda>0$, and without loss of generality, we assume that $c_\delta=1$. Denote $l_J=(f_{\tau_1},f_{\tau_2},...,f_{\tau_J},g_\sigma)$. For any $p_J\in\times_J-\theta$, $p_J\tilde{I}p_J'=E[\int_{\mathbb{R}}\frac{(l_Jp_J')^2}{g}dy]\geq CE[(l_Jp_J')^2]$ for some constant $C>0$ since $g$ is bounded from above.

Define $c:=\inf_{x\in\mathcal{X},\tau\in[0,1]}(x\beta_0'(\tau))>0$. Let $S(\lambda):=\sum_{i=0}^{\lambda}\left(\lambda\binom{i}{\lambda}\right)^2$, where $\binom{b}{a}$ stands for the combinatorial number choosing $b$ elements from a set with size $a$. Then

$$S(\lambda)E\left[\int_{\mathbb{R}}(l_Jp_J')^2dy\right]=E\left[\sum_{j=0}^{\lambda}\binom{j}{\lambda}^2\int_{\mathbb{R}}\left(l_J\left(y+\frac{j}{2\lambda uJc}\right)p_J'\left(y+\frac{j}{2\lambda uJc}\right)\right)^2dy\right].$$

where $u>0$ is a constant that will be specified later. By the Cauchy-Schwarz inequality,

$$E\left[\sum_{j=0}^{\lambda}\binom{j}{\lambda}^2\int_{\mathbb{R}}\left(l_J\left(y+\frac{j}{2\lambda uJc}\right)p_J'\left(y+\frac{j}{2\lambda uJc}\right)\right)^2dy\right]$$

$$\geq E\left[\int_{\mathbb{R}}\left(\sum_{j=0}^{\lambda}(-1)^j\binom{j}{\lambda}^2 l_J\left(y+\frac{j}{2\lambda uJc}\right)p_J'\left(y+\frac{j}{2\lambda uJc}\right)\right)^2 dy\right].$$

Defining the interval $Q_J^i := \left[a+x\beta(\frac{i+1/2}{J})-\frac{1}{2\lambda uJ},\ a+x\beta(\frac{i+1/2}{J})+\frac{1}{2\lambda uJ}\right]$,

$$E\left[\sum_{j=0}^{\lambda}\binom{j}{\lambda}^2\int_{\mathbb{R}}\left(l_J\left(y+\frac{j}{2\lambda uJc}\right)p_J'\left(y+\frac{j}{2\lambda uJc}\right)\right)^2 dy\right]$$

$$\geq E\left[\int_{Q_J^i}\left(\sum_{j=0}^{\lambda}(-1)^j\binom{j}{\lambda}^2 l_J\left(y+\frac{j}{2\lambda uJc}\right)p_J'\left(y+\frac{j}{2\lambda uJc}\right)\right)^2 dy\right].$$

By the discontinuity assumption,

$$\sum_{j=0}^{\lambda}(-1)^j\binom{j}{\lambda}^2 l_J\left(y+\frac{j}{2\lambda uJc}\right)p_J'\left(y+\frac{j}{2\lambda uJc}\right)$$

$$=\frac{1}{(uJ)^{\lambda-1}}\left(\frac{c}{(\lambda-1)!}x_i p_i'+\frac{c_i}{uJ}\sum_{j=1,j\neq i}^{J}x_j p_j\right),$$

with $c_j$ uniformly bounded from above since $f^{(\lambda)}$ is $L^1$ Lipschitz except at $a$. Noting that the intervals $\left\{Q_J^i\right\}$ do not intersect each other as $J\to\infty$ and $u\to 0$,

$$S(\lambda)E\left[\int_{\mathbb{R}}(l_J p_J')^2 dy\right]\geq S(\lambda)\sum_{i=1}^{J}E\left[\int_{Q_J^i}(l_J p_J')^2 dy\right]$$

$$\geq E\left[\frac{1}{\lambda uJc}\sum_{i=1}^{J}\left(\frac{1}{(uJ)^{\lambda-1}}\left[\frac{c}{(\lambda-1)!}x_i p_i'+\frac{c_i}{uJ}\sum_{j=1,j\neq i}^{J}x_j p_j\right]\right)^2\right]$$

Finally, when the $u$ is chosen to be large enough (and only depending on $\lambda$, $\sup c_i$ and $c$),

$$E\left[\sum_{i=1}^{J}\left(\frac{1}{(uJ)^{\lambda-1}}\left[\frac{c}{(\lambda-1)!}x_i p_i'+\frac{c_i}{uJ}\sum_{j=1,j\neq i}^{J}x_j p_j\right])\right)^2\right]\geq c(\lambda)\frac{1}{J^{2\lambda-1}}\sum_{j=1}^{J}E[x_j^2 p_j^2]\asymp\frac{1}{J^{2\lambda}}||p||_2^2$$

with the constant $c(\lambda)>0$ only depending on $\lambda$ and $u$. Therefore$p'\widetilde{I}p\gtrsim\frac{1}{J^{2\lambda}}||p||_2^2$. Hence, the smallest eigenvalue of $\widetilde{I}$ is bounded by $\frac{c}{J^\lambda}$ from below with some generic constant $c$ depending on $\lambda$, $\mathcal{X}$, and the $L^1$ Lipschitz coefficient of $f^{(\lambda)}$ at set $\mathbb{R}-[a-\eta,a+\eta]$. □

*Proof of Lemma 4.* The proof of this Lemma is based on Chen (2008) results of sieve extremum estimator consistency. Under the following five conditions (abbreviated CA1–CA5 for "Chen Assumptions 1–5") from in Theorem 3.1 of Chen (2008), the sieve estimator is consistent.

**Condition CA1** (Identification)**.**

(1) $L(\theta_0) < \infty$, and if $L(\theta_0) = -\infty$ then $L(\theta) > -\infty$ for all $\theta \in \Theta_J \setminus \{\theta_0\}$ and for all $J \geq 1$.
(2) There are a nonincreasing positive function $\delta(\cdot)$ and a positive function $m(\cdot)$ such that for all $\varepsilon > 0$ and $J \geq 1$, $L(\theta_0) - \sup_{\theta \in \Theta_J : d(\theta,\theta_0) L(\theta_J) \geq \varepsilon} L(\theta) \geq \delta(J) m(\varepsilon) > 0$.

**Condition CA2** (Sieve Space)**.** $\Theta_J \subset \Theta_{J+1} \subset \Theta$ for all $J \geq 1$ and there exists a sequence $\pi_J \theta_0 \in \Theta_J$ such that $d(\theta_0, \pi_J \theta_0) \to 0$ as $J \to 0$.

**Condition CA3** (Continuity)**.**

(1) $L(\theta)$ is upper semicontinuous on $\Theta_J$ under metric $d(\cdot, \cdot)$
(2) $|L(\theta_0) - L(\pi_J \theta_0)| = o(\delta(J))$.

**Condition CA4** (Compact Sieve Space)**.** $\Theta_J$ is compact under $d(\cdot, \cdot)$.

**Condition CA5** (Uniform Convergence)**.**

(1) For all $J \geq 1$, $\operatorname{plim} \sup_{\theta \in \Theta_J} |L_n(\theta) - L(\theta)| = 0$.
(2) $\widehat{c}(J) = o_p(\delta(J))$ where $\widehat{c}(J) := \sup_{\theta \in \Theta_J} |L_n(\theta) - L(\theta)|$.
(3) $\eta_{J_n} = o(\delta(J_n))$.

We verify each of these assumptions below.

For Identification: Let $d$ be the metric induced by the $L^2$ norm $|| \cdot ||_2$ defined on $\Theta$. By assumption A4 and compactness of $\Theta$, $L(\theta_0) - \sup_{\theta \in \Theta : d(\theta,\theta_0)} L(\theta_k) > 0$ for any $\varepsilon > 0$. So, letting $\delta(k) = 1$, the identification condition holds.

For the Sieve Space: Our Sieve space satisfies $\Theta_J \subset \Theta_{2J} \subset \Theta$ for $J = 1, 2, ....$ In general we can consider the sequence $\Theta_{2^{J(1)}}, \Theta_{2^{J(2)}}, ..., \Theta_{2^{J(n)}} ...$ instead of $\Theta_1, \Theta_2, \Theta_3 ...$ with $J_n$ being an increasing function and $\lim_{n \to \infty} J_n = \infty$.

For the Continuity condition: Since we assume $f$ is continuous and $L^1$ Lipchitz in Assumption A3, (1) is satisfied. (2) is satisfied under the construction of our Sieve space.

The Compact Sieve Space condition is trivial. The Uniform Convergence condition is also easy to verify since the entropy metric of space $\Theta_k$ is finite and uniformly bounded by the entropy metric of $\Theta$. Therefore we have stochastic equicontinuity, i.e., $\sup_{\theta \in \Theta_{J_n}} |L_n(\theta) - L(\theta)| = O_p(\frac{1}{\sqrt{n}})$. In CA5.3, $\eta_J$ is defined as the error in the maximization procedure. Since $\delta$ is constant, as soon as our maximizer $\widehat{\theta}_{J_n}$ satisfies $L_n(\widehat{\theta}_{J_n}) - \sup_{\theta \in \Theta_{J_n}} L_n(\theta_{J_n}) \to 0$, (3) is verified.

Hence, the Sieve ML estimator is consistent. □

*Proof of Theorem 2.* Suppose $\theta = (\beta(\cdot), \sigma) \in \Theta_{J_n}$ is the Sieve estimator. By the consistency of the Sieve estimator established by Lemma 4, $||\theta - \theta_0||_2 \to_p 0$. Denote $\tau_i = \frac{i-1}{J_n}$, $i = 1, 2, ..., J_n$. The first-order conditions give

$$E_n[-x f'(y - x\beta(\tau_i)|\sigma)] = 0 \tag{C.11}$$

$$E_n[g_\sigma(y|x, \beta, \sigma)] = 0. \tag{C.12}$$

It is easy to see that $\Theta_{J_n} \subset \Theta$. By Lemma 3, $\sigma_n$, the estimator of $\sigma_0$, will always converge to $\sigma_0$ at rate of at least $n^{-\frac{1}{4}}$.

By construction of the sieve, there exists a set of parameters $(\beta_n^*, \sigma_0)$ in $\Theta_{J_n}$ such that:

(1) $||\beta_n^* - \beta_0|| = O(\frac{1}{J_n})$,
(2) $\beta_n^* = \arg\min_{(\beta,\sigma_0)\in\Theta_{J_n}} E[\log(g(y|x,\beta,\sigma_0)]$.

Our strategy is to show that our estimator $\beta$ would converge to $\beta_n^*$ with a certain speed. From the first order condition, we know that

$$E_n\left[\left(\frac{f_{\tau_1}(y|x,\theta), f_{\tau_2}(y|x,\theta), ..., f_{\tau_{J_n}}(y|x,\theta)}{g(y|x,\theta)}, \frac{g_\sigma(y|x,\theta)}{g(y|x,\theta)}\right)\right] = 0.$$

Denote $f_{\theta_{J_n}} = (f_{\tau_1}(y|x,\theta), f_{\tau_2}(y|x,\theta), ..., f_{\tau_{J_n}}(y|x,\theta))$. Therefore, $0 = E_n\left[(\frac{f_{\theta_{J_n}}}{g}, \frac{g_\sigma}{g})|_{(\beta_n,\sigma)}\right]$. By definition, $0 = E\left[\frac{f_{\theta_{J_n}}}{g}|_{(\beta_n^*,\sigma_0)}\right]$.

Define $\Sigma := U[0,1]\times\Theta$. Consider a mapping $H : \Sigma \mapsto \mathbb{R}$, $H((u,\theta)) = \sqrt{n}(E_n[\frac{f_u}{g}] - E[\frac{f_u}{g}])$. By Donskerness of $\Theta$ and $U[0,1]$, $\Sigma$ is Donsker. By A1–A5, we have pointwise CLT for $H((u,\theta))$. By Condition A2, the Lipschitz condition guarantees that $E[|\frac{f_u}{g}(\theta) - \frac{f_u}{g}(\theta')|^2] \asymp ||\theta - \theta'||_2^2$. Therefore, stochastic equicontinuity holds: for $\gamma$ small enough,

$$\Pr\left(\sup_{\{|u-u'|\le\gamma, ||\theta-\theta'||_2\le\gamma, \theta,\theta'\in\Theta\}} |H((u,\theta)) - H((u',\theta'))| \ge \gamma\right) \to 0.$$

Similarly, we have stochastic equicontinuity on $\sqrt{n}E_n[\frac{g_\sigma}{g}]$ such that

$$0 = E_n\left[\left(\frac{f_{\theta_{J_n}}}{g}, \frac{g_\sigma}{g}\right)|_{(\beta_n,\sigma)}\right] = E\left[\left(\frac{f_{\theta_{J_n}}}{g}, \frac{g_\sigma}{g}\right)|_{(\beta_n,\sigma)}\right] + n^{-1/2}G_{n,J_n},$$

where $G_{n,J_n} := \sqrt{n}\left(E_n\left[\left(\frac{f_{\theta_{J_n}}}{g}, \frac{g_\sigma}{g}\right)|_{(\beta_n,\sigma)}\right] - E\left[\left(\frac{f_{\theta_{J_n}}}{g}, \frac{g_\sigma}{g}\right)|_{(\beta_n,\sigma)}\right]\right) = O_p(1)$.

Performing a Taylor expansion of $E\left[\frac{f_{\theta_{J_n}}}{g}|_{(\beta_n,\sigma)}\right]$ around $(\beta_n^*, \sigma_0)$, we get

$$\widetilde{I}(\beta_n - \beta_n^*) + O_p(||\beta_n - \beta_n^*||^2) = -\frac{1}{\sqrt{n}}G_{n,J_n}.$$

Notice that by Lemma 7, the minimum eigenvalue of $\widetilde{I}$, denoted as $r(\widetilde{I})$, satisfies $r(\widetilde{I}) \ge \frac{1}{J^\lambda}$. If $J_n^\lambda||\beta_n - \beta_n^*|| \to_p 0$, then we could conclude that $||\beta_n - \beta_n^*||^2 = o_p(\widetilde{I}(\beta_n - \beta_n^*))$, which allows use to obtain the asymptotics of $\beta_n - \beta_n^*$. Such a case is in general difficult to show for ill-posed systems unless certain local quadratic assumptions are made, such as Chen (2008). However, in our specific problem, we use a deconvolution to prove the following lemma to show that $\beta_n$ will converge to $\beta_n^*$ with at least a certain speed.

**Lemma 8.** $J_n^\lambda||\beta_n - \beta_n^*|| \precsim J_n^{\lambda+\gamma/(1+\gamma)}/n^{1/4(1+\gamma)}$.

*Proof.* Our argument follows the proof of Lemma 3. Donskerness of $\Theta$ implies that

$$0 \le \mathbb{E}[\log(g(y|x,\beta_n^*,\sigma_0))] - \mathbb{E}[\log(g(y|x,\beta_n,\sigma))] \precsim_p \frac{1}{\sqrt{n}}. \tag{C.13}$$

By Lemma 3, $|\sigma-\sigma_0| = o_p(n^{-\frac{1}{4}})$. Let $z(y,x) = g(y|x,\beta,\sigma)-g(y|x,\beta_0,\sigma_0)$. Define $||z(y,x)||_1 := \int_{-\infty}^{\infty}|z(y|x)|dy$.

By the Scheffe Theorem and Pinsker's Inequality,

$$E_x[||z(y|x)||_1^2] \leq D(g(\cdot|\beta_0)||g(\cdot|\beta)) \leq 2(E[\log(g(y|x,\beta_0,\sigma_0))] - E[\log(g(y|x,\beta,\sigma))]) \precsim_p \frac{1}{\sqrt{n}}, \tag{C.14}$$

where $D(P|Q)$ is the K-L divergence between two probability distribution $P$ and $Q$. Now consider the characteristic functions of $x\beta_n(\tau)$ and $x\beta_n^*(\tau)$ conditional on $x$, $\tau \sim U[0,1]$

$$\phi_{x\beta_n}(s) = \frac{\int_{-\infty}^{\infty} g(y|x,\beta_n,\sigma)e^{isy}dy}{\phi_\varepsilon(s|\sigma)}$$
$$\phi_{x\beta_n^*}(s) = \frac{\int_{-\infty}^{\infty} g(y|x,\beta_n^*,\sigma_0)e^{isy}dy}{\phi_\varepsilon(s|\sigma_0)}$$

It follows that

$$|\phi_{x\beta_n}(s)\phi_\varepsilon(s|\sigma) - \phi_{x\beta_n^*}(s)\phi_\varepsilon(s|\sigma_0)| = \left|\int_{-\infty}^{\infty} z(y|x)e^{isy}dy\right| \leq ||z(y|x)||_1.$$

Therefore, for some constant $c > 0$,

$$|\phi_{x\beta_n}(s) - \phi_{x\beta_n^*}(s)| \leq_p \frac{||z(y|x)||_1}{\phi_\varepsilon(s|\sigma_0)} + \frac{c}{n^{\frac{1}{4}}}\frac{\partial\phi_\epsilon(s|\sigma_0)}{\partial_\sigma}\frac{\phi_{x\beta_n^*}(s)}{\phi_\epsilon(s|\sigma_0)}.$$

Using the relationship between the CDF and characteristic function of a random variable $x$ ($F_x(w) = \frac{1}{2} - \int_{-\infty}^{\infty}\frac{\exp(iws)\phi_x(s)}{2\pi is}ds$), we have that

$$F_{x\beta_n}(w) - F_{x\beta_n^*}(w) = \lim_{q_n\to\infty}\int_{-q_n}^{q_n}\frac{\exp(iws)}{2\pi is}(\phi_{x\beta_n}(s) - \phi_{x\beta_n^*}(s))ds.$$

Then since in our sieve setting $\beta_n$ and $\beta_n^*$ are step functions with step size $1/J_n$, we know that $\max(\phi_{x\beta_n}(s), \phi_{x\beta_n^*}(s)) \leq J_n\frac{c_\Omega}{s}$ for some constant $c_\Omega > 0$.

Therefore,

$$\begin{aligned} &E_x\left[\left|F_{x\beta_n}(w) - F_{x\beta_n^*}(w)\right|\right] \\ \leq\ & E_x\left[\int_{-q_n}^{q_n}\left|\frac{\exp(iws)}{2\pi is}\right|\frac{||z(y|x)||_1}{|\phi_\varepsilon(s|\sigma_0)|}ds\right] + E_x\left[2\int_{q_n}^{\infty}\frac{1}{2\pi s}\left|\phi_{x\beta_n}(s) - \phi_{x\beta_n^*}(s)\right|ds\right] \\ &+\frac{c}{n^{\frac{1}{4}}}E_x\left[\int_{-q_n}^{q_n}\frac{1}{2\pi s}\left|\frac{\partial\phi_\epsilon(s|\sigma_0)}{\partial_\sigma}\frac{1}{\phi_\epsilon(s|\sigma_0)}\right|\left|\phi_{x\beta_n^*}(s)\right|ds\right]. \end{aligned} \tag{C.15}$$

The first term of the right-hand side of equation (C.15) is weakly bounded from above by

$$\frac{1}{2\pi}\int_{-q_n}^{q_n}\frac{1}{s\phi_\epsilon(s|\sigma_0)}dsE_x[||z(y|x)||_1] \precsim q_n^\lambda/\sqrt{n}.$$

The second term of (C.15) is weakly bounded by $J_n \frac{4c_\Omega}{q_n} \precsim J_n/q_n$. And the third term of(C.15) satisfies

$$\frac{c}{n^{\frac{1}{4}}} E_x \left[ \int_{-q_n}^{q_n} \frac{\left|\phi_{x\beta_n^*}(s)\right|}{2\pi s} \left| \frac{\partial \phi_\epsilon(s|\sigma_0)}{\partial_\sigma} \frac{1}{\phi_\epsilon(s|\sigma_0)} \right| ds \right] \precsim q_n^\gamma / n^{\frac{1}{4}}.$$

Putting these together, since $\gamma \leq \lambda/2$, we know that the right hand side has an upper bound of $O(J_n^{\gamma/(1+\gamma)}/n^{1/4(1+\gamma)})$. Therefore, $J_n^\lambda ||\beta_n - \beta_n^*|| \precsim J_n^\lambda J_n^{\gamma/(1+\gamma)}/n^{1/4(1+\gamma)} \to 0$.

□

With $||\beta_n - \beta_n^*||^2 = o_p\left(\widetilde{I}(\beta_n - \beta_n^*)\right)$, we are able to conclude from equation (C.13) that $\widetilde{I}(\beta_n - \beta_n^*) = -\frac{1}{\sqrt{n}} G_{n,J_n}(1 + o_p(1))$, or, equivalently, $\beta_n - \beta_n^* = -\frac{1}{\sqrt{n}} \widetilde{I}^{-1} G_{n,J_n}(1 + o_p(1))$, which gives pointwise asymptotic normality. Since the maximum eigenvalue of $\widetilde{I} \precsim J_n^\lambda$, we know that

$$||\beta_n - \beta_0|| \leq ||\beta_n - \beta_n^*|| + ||\beta_n^* - \beta_0|| = O_p(\max(J_n^\lambda/n^{1/2}, 1/J_n)).$$

Finally, if $J_n/n^{\frac{1}{2(1+\lambda)}} \to 0$ then the stochastic component $-\frac{1}{\sqrt{n}} \widetilde{I}^{-1} G_{n,J_n}$ dominates the bias $\beta_n^* - \beta_0$ and the second statement in Theorem 2 holds. □