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 in the dependent variable. We analyze this problem as an errors-in-variables problem where the dependent variable suffers from classical measurement error and develop a sieve maximum-likelihood approach that is robust to left-hand side measurement error. After describing sufficient conditions for identification, we show that when the number of knots in the quantile grid is chosen to grow at an adequate speed, the sieve maximum-likelihood estimator is asymptotically normal. 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: April 2016.
Hausman: MIT Department of Economics; 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 Victor Chernozhukov, Denis Chetverikov, Kirill Evdokimov, Brad Larsen, and Rosa Matzkin for helpful discussions, as well as seminar participants at Cornell, Harvard, MIT, UCL, and UCLA. Haoyang Liu, Yuqi Song, and Jacob Ornelas 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] We are particularly interested in the linear quantile regression setting.[2] 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.[3] We show that under certain discontinuity assumptions, 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.

Intuitively, the estimated quantile regression line $x_i\widehat{\beta}(\tau)$ for quantile τ 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 τ. Our ML framework effectively estimates the likelihood that a given quantile-specific residual ($\varepsilon_{ij} \equiv y_i - x_i\beta(\tau_j)$) is large because of measurement error rather than observation i's unobserved conditional quantile u_i being far away from τ_j. The estimate of

[1]Schennach (2008) establishes identification and a consistent nonparametric estimator when EIV exists in an explanatory variable. 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]Carroll and Wei (2009) proposed an iterative estimator for the quantile regression when one of the regressors has EIV.
[3]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.

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 they are likely to have $u_i \approx \tau_j$. 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 the heterogeneity of returns to education across conditional quantiles of the wage distribution. We find that when we correct for likely measurement error in the self-reported wage data, we estimate considerably more heterogeneity across the wage distribution in the returns to education. In particular, the education coefficient for the bottom of the wage distribution is lower than previously estimated, and the returns to education for latently high-wage individuals has been increasing over time and is much higher than previously estimated. By 2000, the returns to education for the top of the conditional wage distribution are over three times larger than returns for any other segment of the distribution.

The rest of the paper proceeds as follows. In Section 2, we introduce model specification and identification conditions. In Section 3, we consider the MLE estimation method and analyzes its properties. In Section 4, we discuss sieve estimation. We present Monte Carlo simulation results in Section 5, and Section 6 contains our empirical application. Section 7 concludes. The Appendix contains an extension using the deconvolution method and additional proofs.

Notation: Define the domain of x as $\mathcal{X}$. 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 $f(\varepsilon|\sigma)$ be the p.d.f of the EIV ε parametrized by σ. Assume the true parameters are $\beta_0(\cdot)$ and σ_0 for the coefficient of the quantile model and parameter of the density function of the EIV. Let d_x be the dimension of x. Let Σ be the domain of σ. Let d_σ be the dimension of parameter σ. Define $||(\beta_0,\sigma_0)|| := \sqrt{||\beta_0||_2^2 + ||\sigma_0||_2^2}$ as the L^2 norm of (β_0,σ_0), where $||\cdot||_2$ is the usual Euclidean norm. For $\beta_k \in \mathbb{R}^k$, define $||(\beta_k,\sigma_0)||^2 := \sqrt{||\beta_k||_2^2/k + ||\sigma_0||_2^2}$.

2. Model and Identification

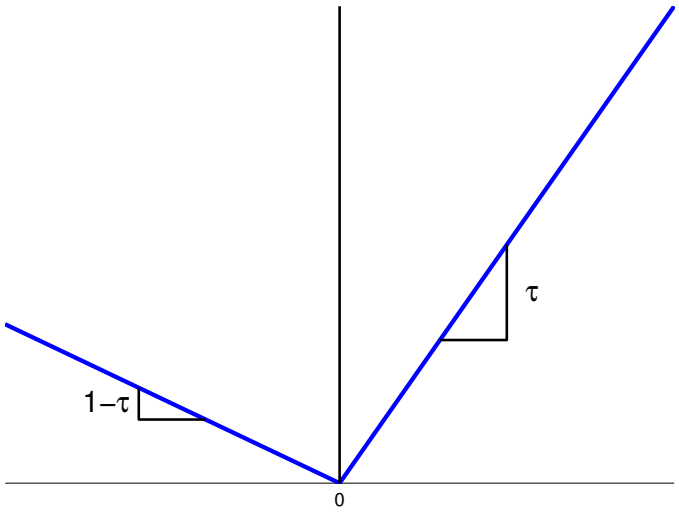
We consider the standard linear conditional quantile model, where the τ^{th} quantile of the dependent variable y^* is a linear function of x

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

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

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

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



and ε is a mean-zero, i.i.d error term independent from y^*, x and τ.

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 FOC conditions as shown by Koenker and Bassett (1978). However, with left-hand side EIV, the FOC condition determining $\widehat{\beta}(\tau)$ becomes

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

For $\tau \neq 0.5$, the presence of measurement error ε will result in the FOC being satisfied at a different estimate of β than in equation (2.2) even in the case where ε 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 τ 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

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

where $x_i \in \{0, 1\}$, the measurement error ε_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 τ, $\widehat{\beta}_1(\tau) = F^{-1}_{y|x=1}(\tau) - F^{-1}_{y|x=0}(\tau)$ where $F(\cdot)$ is the 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 $\mathcal{N}(0, 1)$. In this case,

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

or the estimated coefficient equals the truth at each τ. However, with non-zero measurement error, $y|x = 1 \sim \mathcal{N}(0, 5)$ and $y|x = 0 \sim \mathcal{N}(0, 2)$. The estimated coefficient function under measurement error $\tilde{\beta}_1(\cdot)$ is

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

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

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

TABLE 1. Monte-Carlo Results: Mean Bias

Parameter	EIV Distribution	Quantile (τ) 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 τ 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 ε is either zero (no measurement error case, i.e. y^* is observed) or ε is distributed normally with variance 4 or 16. The covariates x_1 and x_2 are i.i.d. draws from $LN(0,1)$. $N = 1,000$.

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 ε_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 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 ε 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 for any $x \in \mathcal{X}$, $x\beta(\tau)$ is increasing in τ. Suppose $x_1, ..., x_{d_x}$ are d_x-dimensional linearly independent vectors in $\text{int}(\mathcal{X})$. So $Q_{y^*}(\tau|x_i) = x_i\beta(\tau)$ must be strictly increasing in τ. Consider the linear transformation of the model with matrix $A = [x_1, ..., x_{d_x}]'$:

$$Q^*_y(\tau|x) = x\beta(\tau) = (xA^{-1})(A\beta(\tau)). \tag{2.4}$$

Let $\widetilde{x} = xA^{-1}$ and $\widetilde{\beta}(\tau) = A\beta(\tau)$. The transformed model becomes

$$Q_y^*(\tau|\widetilde{x}) = \widetilde{x}\widetilde{\beta}(\tau), \tag{2.5}$$

with every coefficient $\widetilde{\beta}_k(\tau)$ being weakly increasing in τ for $k \in \{1, ..., d_x\}$. 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. We therefore proceed assuming that $\beta_k(\cdot)$ is an increasing function which has the properties of $\widetilde{\beta}(\tau)$. All of the convergence and asymptotic results for $\widetilde{\beta}_k(\tau)$ hold for the parameter β_k after the inverse transform A^{-1}.

*Condition C*1 (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 τ.
(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) β_0 is a vector of C^1 functions with derivative bounded from below by a positive constant.
(4) The domain of the parameter σ is a compact space Σ and the true value σ_0 is in the interior of Σ.

Under assumption C1 it is easy to see that the parameter space $\Theta := M \times \Sigma$ is compact.

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 D.1. □

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 when the distribution of random variable $\beta(u)$ is invariant. 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.

*Condition C*2 (Properties of x). We assume the following properties of the design matrix x:

(1) $E[x'x]$ is non-singular.
(2) The domain of x, denoted as $\mathcal{X}$, is 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 $(X_k, x_{\{-k\}}) \subset \mathcal{X}$.
(3) Without loss of generality, $\beta_k(0) \geq 0$.

*Condition C*3 (Properties of EIV). We assume the following properties of the measurement error ε:

(1) The probability function $f(\varepsilon|\sigma)$ is differentiable in σ.
(2) For all $\sigma \in \Sigma$, there exists a uniform constant $C > 0$ such that $\mathbb{E}[|\log f(\varepsilon|\sigma)|] < C$.
(3) $f(\cdot)$ is non-zero all over the space $\mathbb{R}$, and bounded from above.
(4) $E[\varepsilon] = 0$.
(5) Denote $\phi(s|\sigma) := \int_{-\infty}^{\infty} \exp(is\varepsilon) f(\varepsilon|\sigma) d\varepsilon$ as the characteristic function of ε.
(6) Assume for any σ_1 and σ_2 in the domain of σ, denoted as Σ, there exists a neighborhood of 0, such that $\frac{\phi(s|\sigma_1)}{\phi(s|\sigma_2)}$ can be expanded as $1 + \sum_{k=2}^{\infty} a_k (is)^k$.

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

$$L(\theta) = \mathbb{E}[\log g(y|x, \theta)], \tag{2.6}$$

where the conditional density function $g(y|x, \theta)$ is defined as

$$g(y|x, \theta) = \int_0^1 f(y - x\beta(u)|\sigma) du. \tag{2.7}$$

Define the empirical likelihood as

$$L_n(\theta) = \mathbb{E}_n[\log g(y|x, \theta)], \tag{2.8}$$

The main identification results rely on the monotonicity of $x\beta(\tau)$. The global identification condition is the following:

*Condition C*4 (Identification). There does not exist $(\beta_1, \sigma_1) \neq (\beta_0, \sigma_0)$ in parameter space Θ such that $g(y|x, \beta_1, \sigma_1) = g(y|x, \beta_0, \sigma_0)$ for all (x, y) with positive continuous density or positive mass.

Theorem 1 (Nonparametric Global Identification)**.** *Under condition C1-C3, for any $\beta(\cdot)$ and $f(\cdot)$ which generates the same density of $y|x$ almost everywhere as the true function $\beta_0(\cdot)$ and $f_0(\cdot)$, it must be that:*

$$\beta(\tau) = \beta_0(\tau)$$
$$f(\varepsilon) = f_0(\varepsilon).$$

Proof. See Appendix D.1. □

We also summarize the local identification condition as follows:

Lemma 2 (Local Identification)**.** *Define $p(\cdot)$ and $\delta(\cdot)$ as functions that measure the deviation of a given β or σ from the truth: $p(\tau) = \beta(\tau) - \beta_0(\tau)$ and $\delta(\sigma) = \sigma - \sigma_0$. Then there does not exist a function $p(\tau) \in L^2[0, 1]$ and $\delta \in \mathbb{R}^{d_\sigma}$ such that for almost all (x, y) with positive*

continuous density or positive mass

$$\int_0^1 f(y - x\beta(\tau)|\sigma)xp(\tau)d\tau = \int_0^1 \delta' f_\sigma(y - x\beta(\tau))d\tau,$$

except that $p(\tau) = 0$ *and* $\delta = 0$.

Proof. See Appendix D.1. □

*Condition C*5 (Stronger local identification for σ). For any $\delta \in S^{d_\sigma - 1}$,

$$\inf_{p(\tau)\in L^2[0,1]} \left(\int_0^1 f_y(y - x\beta(\tau)|\sigma)xp(\tau)d\tau - \int_0^1 \delta' f_\sigma(y - x\beta(\tau))d\tau)^2 \right) > 0. \tag{2.9}$$

3. Maximum Likelihood Estimator

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

$$(\widehat{\beta}(\cdot), \widehat{\sigma}) \in \arg \max_{(\beta(\cdot),\sigma)\in\Theta} E_n[g(y|x, \beta(\cdot), \sigma)]. \tag{3.1}$$

where $g(\cdot|\cdot, \cdot, \cdot)$ is the conditional density of y given x and parameters, as defined in equation (2.7)

The following theorem states the consistency property of the ML estimator.

Lemma 3 (MLE Consistency)**.** *Under conditions C1-C3, the random coefficients* $\beta(\cdot)$ *and the parameter* σ *that determines the distribution of* ε *are identified in the parameter space* Θ*. 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 to the true parameter $(\beta_0(\cdot), \sigma_0)$ *under the* L^∞ *norm in the functional space* M *and Euclidean norm in* Σ *with probability approaching 1.*

Proof. See Appendix D.2. □

The identification 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 Θ and the consistency argument will fail.

3.2. **Ill-posed Fredholm Integration of the First Kind.** In the usual MLE setting with the parameter being finite dimensional, the Fisher information matrix I is defined as:

$$I := E \left[\frac{\partial f}{\partial \theta} \frac{\partial f'}{\partial \theta} \right].$$

In our case, since the parameter is continuous, the informational kernel $I(u,v)$ is defined as[5]

$$I(u,v)[p(v),\delta] = E\left[\left(\frac{f_{\beta(u)}}{g},\frac{g_\sigma}{g}\right)'\left(\int_0^1 \frac{f_v}{g}p(v)dv + \frac{g_\sigma}{g}'\delta\right)\right]. \tag{3.2}$$

If we assume that we know the true σ, i.e., $\delta = 0$, then the informational kernel becomes

$$I_0(u,v) := E[\frac{\partial f_u}{g}\frac{\partial f_u}{g}'].$$

By condition C5, the eigenvalue of I and I_0 are both non-zero.

Since $E[X'X] < \infty$, $I_0(u,v)$ is a compact (Hilbert-Schmidt) operator. From the Riesz Representation Theorem, $L(u,v)$ has countable eigenvalues $\lambda_1 \geq \lambda_2 \geq ... > 0$ with 0 being the only limit point of this sequence. It can be written as the following form:

$$I_0(\cdot) = \sum_{i=1}^{\infty} \lambda_i \langle \psi_i, \cdot \rangle \psi_i$$

where ψ_i is the system of orthogonal functional basis, $i = 1, 2, \ldots$. For any function $s(\cdot)$, $\int_0^1 I_0(u,v)h(v)dv = s(u)$ is called the Fredholm integral equation of the first kind. Thus, the first-order condition of the ML estimator is ill-posed. In Section (3.3) below, we establish convergence rate results for the ML estimator using a deconvolution method.[6]

Although the estimation problem is ill-posed for the function $\beta(\cdot)$, the estimation problem is not ill-posed for the finite dimensional parameter σ given that $E[g_\sigma g_\sigma']$ is non-singular. In the Lemma below, we show that σ converges to σ_0 at rate $1/\sqrt{n}$.

Lemma 4 (Estimation of σ)**.** *If $E[g_\sigma g_\sigma']$ is positive definite and conditions C1-C5 hold, the ML estimator $\widehat{\sigma}$ has the following property:*

$$\widehat{\sigma} - \sigma_0 \to O_p(n^{-\frac{1}{4}}). \tag{3.3}$$

Proof. See Appendix D.2. □

3.3. **Bounds for the Maximum Likelihood Estimator.** In this subsection, we use the deconvolution method to establish bounds for the maximum likelihood estimator. Recall that the maximum likelihood estimator is the solution

$$(\beta,\sigma) = \text{argmax}_{(\beta,\sigma)\in\Theta}\mathbb{E}_n[\log(g(y|x,\beta,\sigma))]$$

[5]For notational convenience, we abbreviate $\frac{\partial f(y-x\beta_0(\tau)|\sigma_0)}{\partial\beta(\tau)}$ as $f_{\beta(\tau)}$, $g(y|x,\beta_0(\tau),\sigma_0)$ as g, and $\frac{\partial g}{\partial\sigma}$ as g_σ.

[6]Kuhn (1990) shows that if the integral kernel is positive definite with smooth degree r, then the eigenvalue λ_i is decaying with speed $O(n^{-r-1})$. However, it is generally more difficult to obtain a lower bound for the eigenvalues. The decay speed of the eigenvalues of the information kernel is essential in obtaining convergence rate. From the Kuhn (1990) results, we see that the decay speed is linked with the degree of smoothness of the function f. The less smooth the function f is, the slower the decaying speed is. We show below that by assuming some discontinuity conditions, we can obtain a polynomial rate of convergence.

Define $\chi(s_0) = \sup_{|s|\leq s_0} |\frac{1}{\phi_\varepsilon(s|\sigma_0)}|$. We use the following smoothness assumptions on the distribution of ε (see Evodokimov, 2010).

*Condition C*6 (Ordinary Smoothness (OS)). $\chi(s_0) \leq C(1 + |s_0|^\lambda)$.

*Condition C*7 (Super Smoothness (SS)). $\chi(s_0) \leq C_1(1 + |s_0|^{C_2}) \exp(|s_0|^\lambda / C_3)$.

The Laplace distribution $L(0, b)$ satisfies the Ordinary Smoothness (OS) condition with $\lambda = 2$. The Chi-2 Distribution χ^2_ν satisfies the OS condition with $\lambda = \frac{\nu}{2}$. The Gamma Distribution $\Gamma(\nu, \theta)$ satisfies the OS condition with $\lambda = \nu$. The Exponential Distribution satisfies condition OS with $\lambda = 1$. The Cauchy Distribution satisfies the Super Smoothness (SS) condition with $\lambda = 1$. The Normal Distribution satisfies the SS condition with $\lambda = 2$.

Lemma 5 (MLE Convergence Speed)**.** *Under assumptions C1-C5 and the results in Lemma 4,*

(1) if Ordinary Smoothness holds (C6), then the ML estimator of $\widehat{\beta}$ satisfies for all τ

$$|\widehat{\beta}(\tau) - \beta_0(\tau)| \precsim n^{-\frac{1}{2(1+\lambda)}}$$

(2) if Super Smoothness holds (C7), then the ML estimator of $\widehat{\beta}$ satisfies for all τ

$$|\widehat{\beta}(\tau) - \beta_0(\tau)| \precsim \log(n)^{-\frac{1}{\lambda}}.$$

Proof. See Appendix D.2. □

4. Sieve Estimation

In the last section we demonstrated that the maximum likelihood estimator restricted to parameter space Θ 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 σ 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 Ω_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 $k = 1, \ldots, d_x$.

We know that the L^2 distance of the space Θ_J to the true parameter θ_0 satisfies $d_2(\theta_0, \Theta_J) \leq C\frac{1}{J}$ for some generic constant C.

The sieve estimator is defined as follows:

Definition 2 (Sieve Estimator)**.**

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

Let $d(\theta_1, \theta_2)^2 := E[\int_{\mathbb{R}} \frac{(g(y|x,\theta_1) - g(y|x,\theta_2))^2}{g(y|x,\theta_0)} dy]$ be a pseudo metric on the parameter space Θ. We know that $d(\theta, \theta_0) \leq L(\theta_0|\theta_0) - L(\theta|\theta_0)$.

Let $||\cdot||_d$ be a norm such that $||\theta_0 - \theta||_d \leq Cd(\theta_0, \theta)$. Our metric $||\cdot||_d$ here is chosen to be:

$$||\theta||_d^2 := \langle \theta, I(u, v)[\theta] \rangle . \tag{4.2}$$

By Condition C4, $||\cdot||_d$ is indeed a metric.

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

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

Again, by Condition C4, this matrix $\tilde{I}$ is non-singular. Furthermore, if a certain discontinuity condition is assumed, the smallest eigenvalue of $\tilde{I}$ can be proved to be bounded away from some polynomial of J.

*Condition C*8 (Discontinuity of f). Suppose there exists a positive integer λ such that $f \in C^{\lambda-1}(\mathbb{R})$, and the λ^{th} order derivative of f equals:

$$f^{(\lambda)}(x) = h(x) + \delta(x - a), \tag{4.3}$$

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

Remark. The Laplace distribution satisfies the above assumption with $\lambda = 1$. There is an intrinsic link between the above discontinuity condition and the tail property of the characteristic function stated as the (OS) condition. Because $\phi_{f'}(s) = is\phi_f(s)$, we know that $\phi_f(s) = \frac{1}{(is)^\lambda}\phi_{f^{(\lambda)}}(s)$, while $\phi_{f^{(\lambda)}}(s) = O(1)$ under the above assumption. Therefore, assumption C9 indicates that $\chi_f(s) := \sup|\frac{1}{\phi_f(s)}| \leq C(1 + s^\lambda)$. In general, these results will hold as long as the number of Dirac functions in $f^{(\lambda)}$ are finite in (4.3).

The following Lemma establishes the decay speed of the minimum eigenvalue of $\tilde{I}$.

Lemma 6. *If the function f satisfies condition C8 with degree $\lambda > 0$,*

(1) the minimum eigenvalue of $\tilde{I}$, denoted as $r(\tilde{I})$, has the following property:

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

(2) $\frac{1}{J^\lambda} \precsim \sup_{\theta \in \mathcal{P}_k} \frac{||\theta||_d}{||\theta||}$

Proof. See Appendix D.3. □

The following Lemma establishes the consistency of the sieve estimator.

Lemma 7 (Sieve Estimator Consistency)**.** *If conditions C1-C6 and C9 hold, The sieve estimator defined in* (4.1) *is consistent.*

Given the identification assumptions in the last section, if the problem is identified, then there exists a neighborhood $\cdot$ of θ_0, for any $\theta \in \Delta$, $\theta \neq \theta_0$, we have:

$$L(\theta_0|\theta_0) - L(\theta|\theta_0) \geq E_x \left[\int_{\mathcal{Y}} \frac{(g(y|x,\theta_0) - g(y|x,\theta))^2}{g(y|x,\theta_0)} \right] > 0. \tag{4.4}$$

Proof. See Appendix D.3. □

Unlike the usual sieve estimation problem, our problem is ill-posed with decaying eigenvalue with speed J^{λ}. However, the curse of dimensionality is not at play because of the co-monotonicity: all entries of vector of functions $\beta(\cdot)$ are function of a single variable τ. 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 $\sqrt{n}$.

We summarize the tail property of f in the following condition:

*Condition C*9 (Tail Property of f). Assume there exists a generic constant C such that $Var\left(\frac{f_{\beta_J}}{g_0}, \frac{g_\sigma}{g_0}\right) < C$ for any $\beta_J(\cdot)$ and σ in a fixed neighborhood of $(\beta_0(\cdot), \sigma_0)$.

Remark. The above condition is true for the Normal, Laplace, and Beta Distributions, among others.

*Condition C*10 (Eigenvector of $\tilde{I}$). Suppose the smallest eigenvalue of $\tilde{I}$, $r(\tilde{I}) = \frac{c_J}{J^{\lambda}}$. Suppose v is a normalized eigenvector of $r(\tilde{I})$, and that $J^{-\alpha} \precsim \min |v_i|$ for some fixed $\alpha \geq \frac{1}{2}$.

Theorem 2. *Under conditions C1-C5 and C8-C10, the following results hold for the sieve-ML estimator:*

(1) If the number of knots J satisfies the following growth condition:

(a) $\frac{J^{2\lambda+1}}{\sqrt{n}} \to 0$,

(b) $\frac{J^{\lambda+r}}{\sqrt{n}} \to \infty$,

then $||\theta - \theta_0|| = O_p(\frac{J^{\lambda}}{\sqrt{n}})$.

(2) If condition 11 holds and the following growth conditions hold

(a) $\frac{J^{2\lambda+1}}{\sqrt{n}} \to 0$

(b) $\frac{J^{\lambda+r-\alpha}}{\sqrt{n}} \to \infty$,

then (a) for every $j = 1, \ldots, J$, there exists a number μ_{kjJ}, such that $\frac{\mu_{kjJ}}{J^r} \to 0$, $\frac{J^{\lambda-\alpha}}{\mu_{kjJ}} = O(1)$, and

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

and (b) for the parameter σ, there exists a positive definite matrix V of dimension $d_\sigma \times d_\sigma$ such that the sieve estimator satisfies:

$$\sqrt{n}(\sigma_J - \sigma_0) \to \mathcal{N}(0, V).$$

Proof. See Appendix D.3. □

Once we fix 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 μ_{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 $b = 1, \dots B$, and define the estimator

$$\theta^b = \arg\max_{\theta \in \Theta_J} \mathbb{E}_n^b[\log g(y_i^b | x_i^b, \theta)], \tag{4.5}$$

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 τ 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 $\alpha = .05$.

Chen and Pouzo (2013) establishes results in validity of the nonparametric bootstrap in semiparametric models for a general functional of a parameter θ. The Lemma below is an implementation of theorem (5.1) of Chen and Pouzo (2013), and establishes the asymptotic normality of the bootstrap estimates that allows us, for example, to use their empirical variance to construct bootstrapped confidence intervals.

Lemma 8 (Validity of the Bootstrap)**.** *Under condition C1-C6 and C9, choosing the number of knots J according to the condition stated in Theorem 1, the bootstrap defined in equation* (4.5) *has the following property:*

$$\frac{\beta_{k,J}^b(\tau) - \beta_{k,J}(\tau)}{\mu_{kjJ}} \xrightarrow[d]{*} \mathcal{N}(0,1) \tag{4.6}$$

Proof. See Appendix D.3. □

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

following, with the parameter vector $\theta = [\beta(\cdot), \sigma]$

$$\max_{\theta'} Q(\theta'|\theta) := \mathbb{E}\left[\log(f(y - x\beta'(\tau))|\theta')\kappa(x, y, \theta)|\theta\right] \tag{4.7}$$

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

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 σ'^2. The σ'^2 is solved after all the $\beta'(\tau)$ are solved from equation (4.7). 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 β and σ are

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

where W_j is the diagonal matrix formed from the j^{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

$$w_{ij} = \frac{\phi\left(\widehat{\varepsilon}_{ij}/\widehat{\sigma}\right)}{\frac{1}{J}\sum_j \phi\left(\widehat{\varepsilon}_{ij}/\widehat{\sigma}\right)} \tag{4.8}$$

where $\phi(\cdot)$ is the pdf of a standard normal distribution J is the number of τs 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 ε from a mixed normal distribution

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

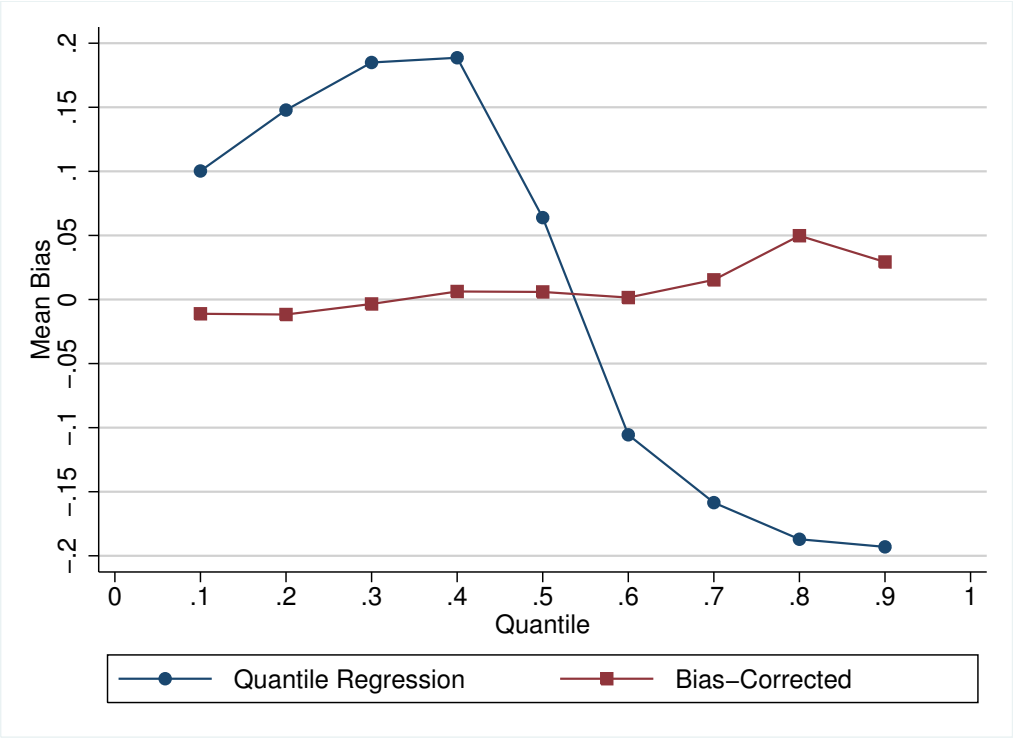
We also probe the robustness of the mixture specification by simulating measurement error from alternative distributions and testing how well modeling the error distribution as a Gaussian mixture handles alternative scenarios to simulate real-world settings in which the econometrician does not know the 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 τ-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 (2.6) and (2.7) 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 τ 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^{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 β_1 and β_2, respectively) and always less

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



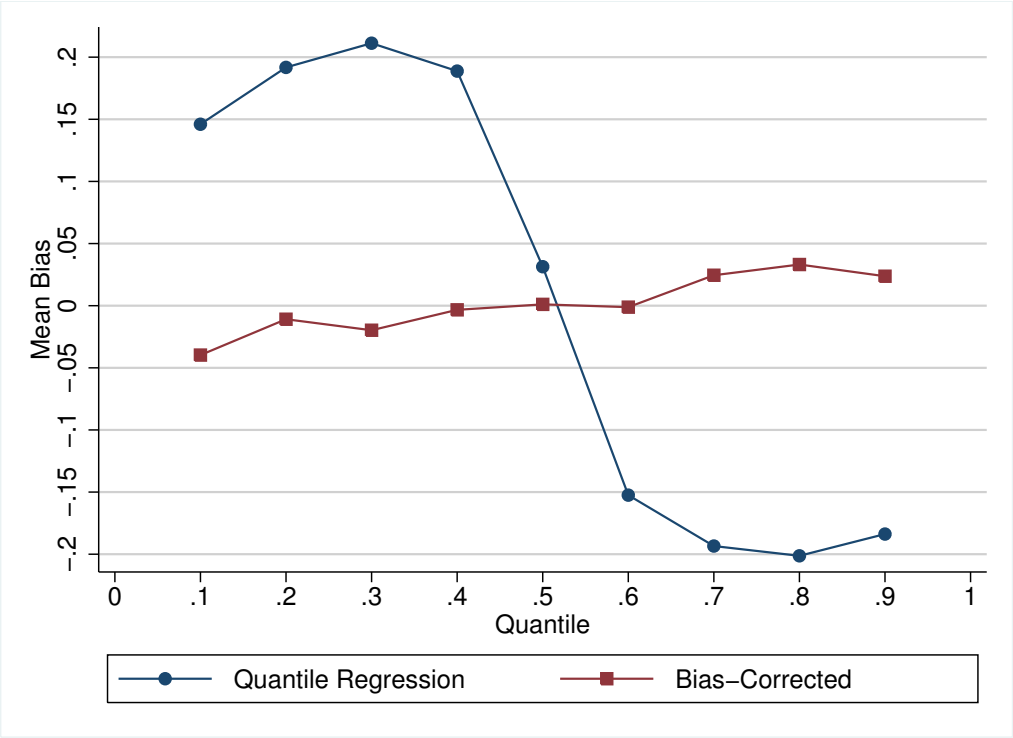
Notes: Figure plots mean bias of estimates of $\beta_1(\tau)$ for classical quantile regression (blue line) and bias-corrected MLE (red line) across 500 MC simulations using the data-generating process described in the text with the measurement error generated as a mixture of three normals.

than 5% of the true coefficient magnitude. 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.

Figure 4 shows the true mixed-normal distribution of the measurement error ε 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 5th and 95th percentile 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

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.

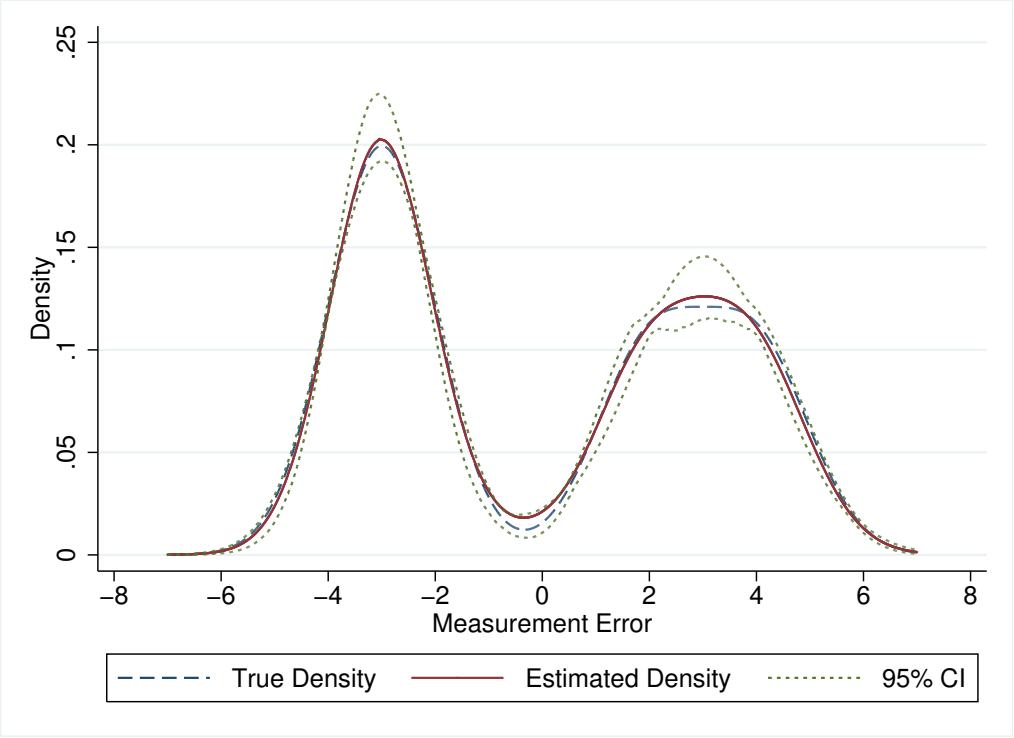
as a mixture of two normals in panel B, the ML estimates that model the EIV distribution as a mixture of three normals are still unbiased. As expected, quantile regression exhibits typical bias towards the median under both distributions and for both slope coefficients (visible as positive mean bias 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 of a Mincer regression,

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

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 5th and 95th percentile of the EIV density function across all MC simulations.

where $q_{y|X}(\tau)$ is the τ^{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.[7] 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

[7]Here we emphasize that, 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 $\beta(\cdot)$.

[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.

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

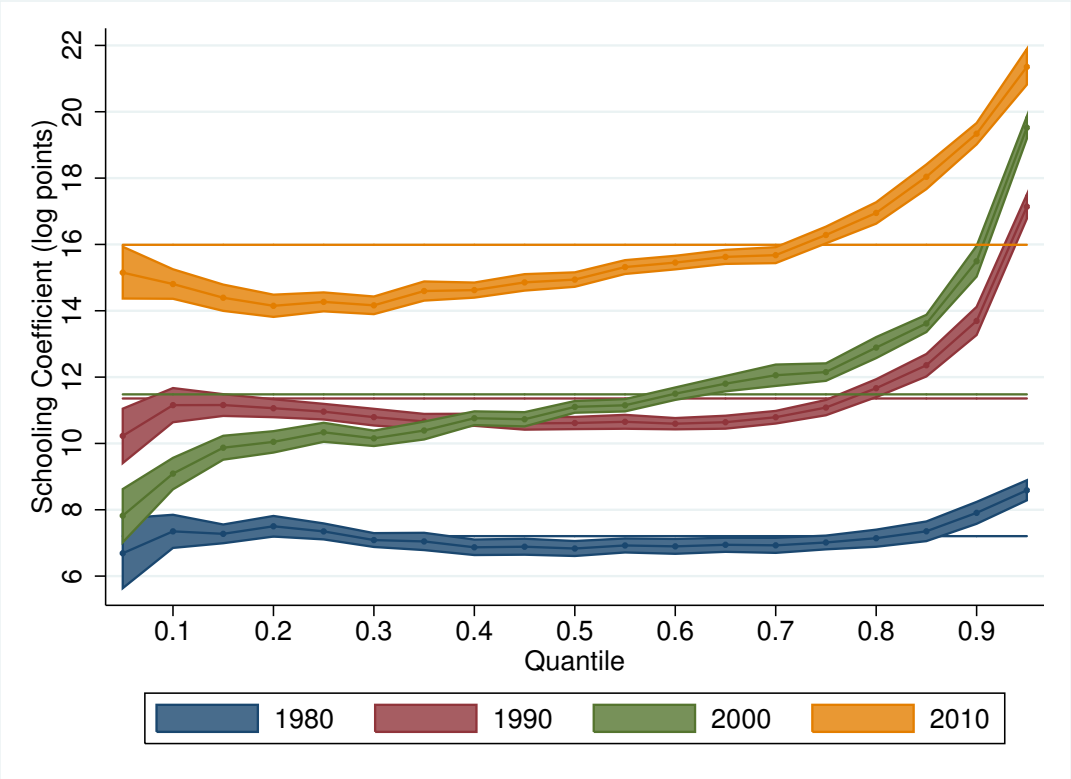
	A. EIV ∼ T				B. EIV ∼ Mixture of 2 $\mathcal{N}$			
	β_1		β_2		β_1		β_2	
Quantile	QReg	MLE	QReg	MLE	QReg	MLE	QReg	MLE
0.1	0.07	0.01	0.09	-0.03	0.14	0.04	0.18	0.03
0.2	0.05	-0.02	0.06	-0.02	0.15	0.05	0.16	0.00
0.3	0.04	-0.02	0.05	-0.06	0.09	0.05	0.09	0.00
0.4	0.03	0.00	0.03	-0.02	0.03	0.05	-0.01	0.01
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.

identical to OLS estimates. In 1990, most of the conditional wage distribution still had a similar education-wage gradient, although higher conditional (wage) quantiles saw a slightly stronger association between education and wages. 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 conditional incomes. 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 or local average treatment effects estimated from linear estimators fail to represent the returns to education for a sizable portion of the population.

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

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

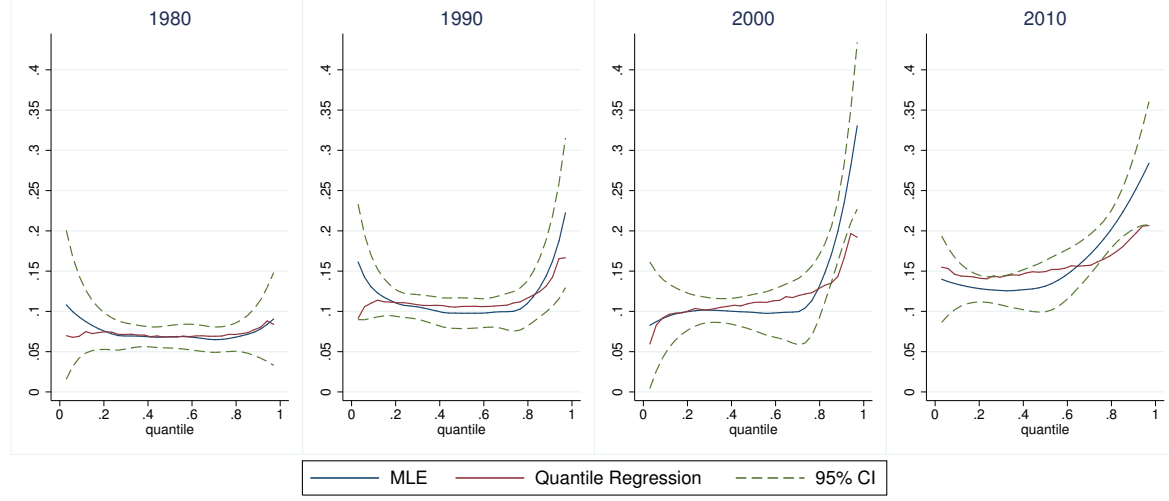


Notes: Figure reports quantile regression estimates of log weekly wages (self-reported) on education, a quadratic in experience, and an indicator for blacks for a grid of 19 evenly spaced quantiles from 0.05 to 0.95. Horizontal lines indicate OLS estimates for each year, and bootstrapped 95% simultaneous 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.

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 τ to reduce volatility of coefficient estimates across the conditional wage distribution. Finally, we construct nonparametric 95% pointwise confidence intervals by bootstrapping and taking the 5th and 95th percentiles of the smoothed education coefficients for each quantile.

Figure 6 plots the education coefficient $\widehat{\beta}_1(\tau)$ from estimating equation (6.1) by MLE and quantile regression, along with nonparametric simultaneous 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

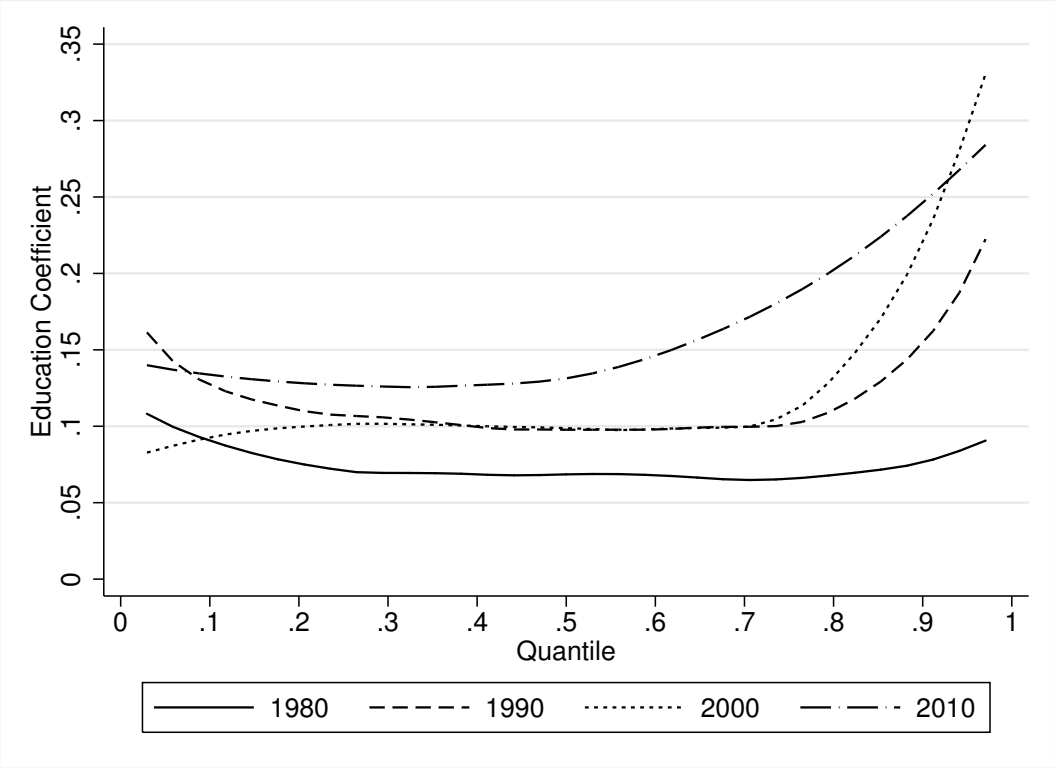
FIGURE 6. Returns to Education Correcting for LHS Measurement Error



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).

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.

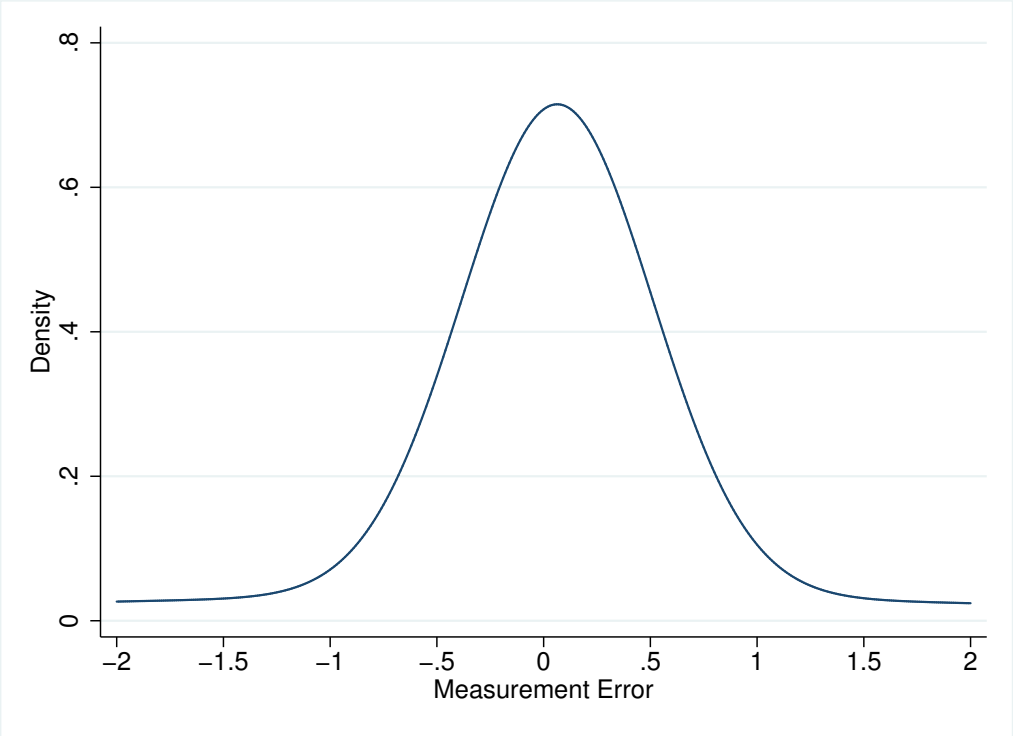
FIGURE 7. ML Estimated Returns to Education Across Years



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.

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 (see, for example, Piketty

FIGURE 8. Estimated Distribution of Wage Measurement Error

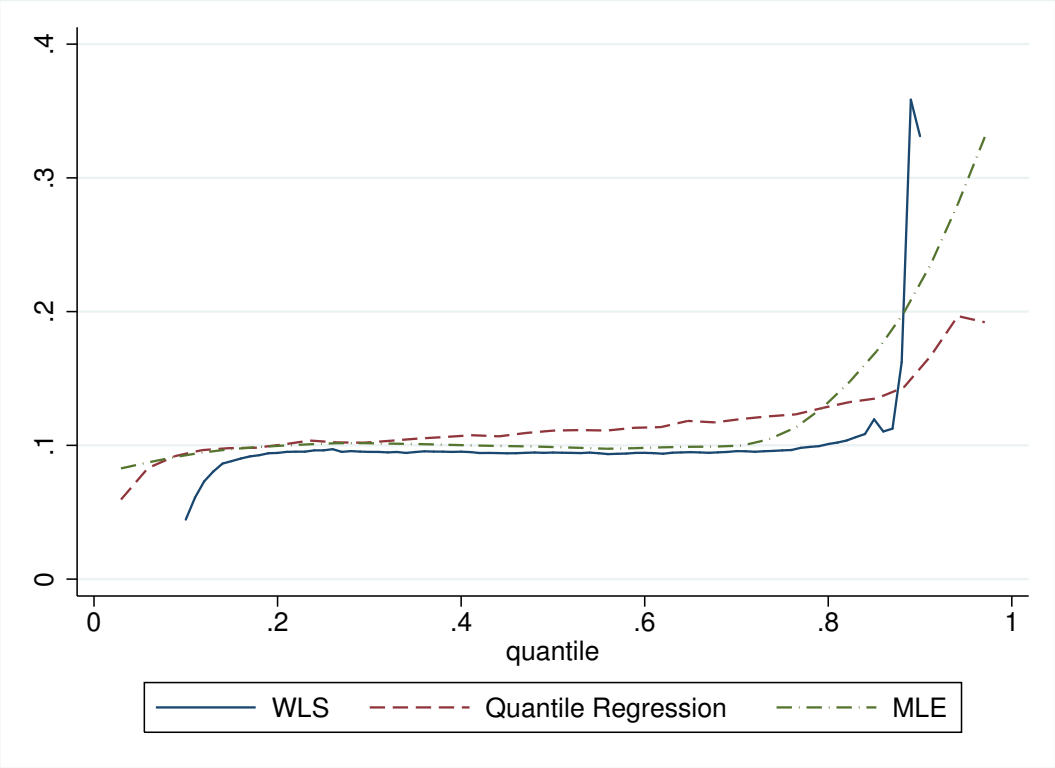


Note: Graph plots the estimated probability density function of the measurement error in 1990 when specified as a mixture of three normal distributions.

and Saez, 2006). 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 inequity in the returns to education than estimates using 2010 data. By 2010, not only had the overall returns to education increased across nearly the entire wage distribution, but the particularly high education gradient enjoyed by the top quartile seems to have been smoothed out and shared by the top half of the 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 (solid blue line) in the 1990 data. Despite the flexibility afforded by the mixture specification, the

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



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). 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 MLE estimates revise the right-tail estimates of the relationship between education and 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 MLE 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 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 a discontinuity assumption (C8), 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

[1] Joshua Angrist, Victor Chernozhukov, and Iván Fernández-Val. Quantile regression under misspecification, with an application to the us wage structure. *Econometrica*, 74(2):539–563, 2006.

[2] David H Autor, Lawrence F Katz, and Melissa S Kearney. Trends in us wage inequality: Revising the revisionists. *The Review of economics and statistics*, 90(2):300–323, 2008.

[3] Songnian Chen. Rank estimation of transformation models. *Econometrica*, 70(4):1683–1697, 2002.

[4] Xiaohong Chen. Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, 6:5549–5632, 2007.

[5] Xiaohong Chen and Demian Pouzo. Sieve quasi likelihood ratio inference on semi/nonparametric conditional moment models1. 2013.

[6] Victor Chernozhukov. Extremal quantile regression. *Annals of Statistics*, pages 806–839, 2005.

[7] Victor Chernozhukov, Ivan Fernandez-Val, and Alfred Galichon. Improving point and interval estimators of monotone functions by rearrangement. *Biometrika*, 96(3):559–575, 2009.

[8] Stephen R Cosslett. Efficient semiparametric estimation of censored and truncated regressions via a smoothed self-consistency equation. *Econometrica*, 72(4):1277–1293, 2004.

[9] Stephen R Cosslett. Efficient estimation of semiparametric models by smoothed maximum likelihood*. *International Economic Review*, 48(4):1245–1272, 2007.

[10] Arthur P Dempster, Nan M Laird, Donald B Rubin, et al. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal statistical Society*, 39(1):1–38, 1977.

[11] Jianqing Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, pages 1257–1272, 1991.

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

[13] Jerry A Hausman, Jason Abrevaya, and Fiona M Scott-Morton. Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics*, 87(2):239–269, 1998.

[14] Jerry A Hausman, Andrew W Lo, and A Craig MacKinlay. An ordered probit analysis of transaction stock prices. *Journal of financial economics*, 31(3):319–379, 1992.

[15] Joel L Horowitz. Applied nonparametric instrumental variables estimation. *Econometrica*, 79(2):347–394, 2011.

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

[17] Michael A Newton and Adrian E Raftery. Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–48, 1994.

[18] Thomas Piketty and Emmanuel Saez. The evolution of top incomes: A historical and international perspective. *The American Economic Review*, pages 200–205, 2006.

[19] Steven Ruggles, Katie Genadek, Ronald Goeken, Josiah Grover, and Matthew Sobek. Integrated public use microdata series version 6.0 [machine-readable database], Minneapolis: University of Minnesota, 2015.

[20] Susanne M Schennach. Quantile regression with mismeasured covariates. *Econometric Theory*, 24(04):1010–1043, 2008.

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

[22] Ying Wei and Raymond J Carroll. Quantile regression with measurement error. *Journal of the American Statistical Association*, 104(487), 2009.

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

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

where ε 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 π_ℓ on each component ℓ must sum to unity. Similarly, for the measurement error to be mean zero, $\sum_\ell \mu_\ell\pi_\ell = 0$, where μ_ℓ 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 ℓ.

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 β_j and distributional parameter are

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

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

Incorporating the constraints on the final L^{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 $\partial l/\partial \mu_\ell$:

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

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. 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 3 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 3. 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).

Appendix C. Deconvolution Method

Although the direct MLE method is used to obtain identification and some asymptotic properties of the estimator, the main asymptotic property of the maximum likelihood estimator depends on discontinuity assumption C.9. Evodokimov (2011) establishes a method using deconvolution to solve a similar problem under panel data condition. He adopts a nonparametric setting and therefore needs kernel estimation method.

Instead, in the special case of quantile regression setting, we can further explore the linear structure without using the nonparametric method. Deconvolution offers a straight forward view of the estimation procedure.

The CDF function of a random variable w can be computed from the characteristic function $\phi_w(s)$.

$$F(w) = \frac{1}{2} - \lim_{q\to\infty} \int_{-q}^{q} \frac{e^{-iws}}{2\pi i s} \phi_w(s) ds. \tag{C.1}$$

$\beta_0(\tau)$ satisfying the following conditional moment condition:

$$E[F(x\beta_0(\tau)|x)] = \tau. \tag{C.2}$$

Since we only observe $y = x\beta(\tau) + \varepsilon$, the $\phi_{x\beta}(s) = \frac{E[\exp(isy)]}{\phi_\varepsilon(s)}$. Therefore $\beta_0(\tau)$ satisfies:

$$E[\frac{1}{2} - \tau - \lim_{q\to\infty} \int_{-q}^{q} \frac{E[e^{-i(y-x\beta_0(\tau))s}|x]}{2\pi i s \phi_\varepsilon(s)} ds|x] = 0 \tag{C.3}$$

In the above expression, the limit symbol before the integral can not be exchanged with the followed conditional expectation symbol. But in practice, they can be exchanged if the integral is truncated.

A simple way to explore information in the above conditional moment equation is to consider the following moment conditions:

$$E\left[x\left(\frac{1}{2} - \tau + \mathrm{Im}(\lim_{q\to\infty} \int_{-q}^{q} \frac{E[e^{-i(y-x\beta_0(\tau))s}|x]}{2\pi s \phi_\varepsilon(s)} ds)\right)\right] = 0 \tag{C.4}$$

Let $F(x\beta, y, \tau, q) = \frac{1}{2} - \tau - \mathrm{Im}\left\{\int_{-q}^{q} \frac{e^{-i(y-x\beta)s|x}}{2\pi s \phi_\varepsilon(s)} ds\right\}$.

Given fixed truncation threshold q, the optimal estimator of the kind $E_n[f(x)F(x\beta, y, \tau, q)]$ is:

$$E_n[E[\frac{\partial F}{\partial \beta}|x]/\sigma(x)^2 F(x\beta, y, \tau, q)] = 0, \tag{C.5}$$

where $\sigma(x)^2 := E[F(x\beta, y, \tau)^2|x]$.

Another convenient estimator is:

$$\arg\min_{\beta} E_n[F(x\beta, y, \tau, q)^2]. \tag{C.6}$$

These estimators have similar behaviors in term of convergence speed, so we will only discuss the optimal estimator. Although it can not be computed directly because the weight $\frac{\partial F}{\partial \beta}/\sigma(x)^2$ depends on β, it can be achieved by an iterative method.

Theorem 3. *(a) Under condition OS with $\lambda \geq 1$, the estimator β satisfying equation (3.15) with truncation threshold $q_n = CN^{\frac{1}{2\lambda}}$ satisfies:*

$$\beta - \beta_0 = O(N^{\frac{-1}{N^{2\lambda}}}), \tag{C.7}$$

(b) Under condition SS with, the estimator $\widehat{\beta}$ satisfies equation (3.15) with truncation threshold $q_n = C\log(n)^{\frac{1}{\lambda}}$ satisfies:

$$\beta - \beta_0 = O(\log(n)^{\frac{-1}{\lambda}}). \tag{C.8}$$

Appendix D. Proofs of Lemmas and Theorems

D.1. **Lemmas and Theorems in Section 2.** In this section we prove the Lemmas and Theorems in section 2.

Proof of Lemma 1.

Proof. For any sequence of monotone functions $f_1, f_2, ..., f_n, ...$ with each one mapping $[a, b]$ into some closed interval $[c, d]$. For bounded monotonic functions, point convergence means uniform convergence, therefore this space is compact. Hence the product space $B_1 \times B_2 \times ... \times B_k$ is compact. It is complete since L^2 functional space is complete and limit of monotone functions is still monotone. □

Proof of Lemma 2.

Proof. WLOG, under condition C.1-C.3, we can assume the variable x_1 is continuous. If there exists $\beta(\cdot)$ and $f(\cdot)$ which generates the same density $g(y|x, \beta(\cdot), f)$ as the true paramter $\beta_0(\cdot)$ and $f_0(\cdot)$, then by applying Fourier transformation,

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

Denote $m(s) = \frac{\phi(s)}{\phi_0(s)} = 1 + \sum_{k=2}^{\infty} a_k(is)^k$ around a 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,

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

Divide both sides of the above equation by $(is)^k/k!$, as let s approaching 0, we get:

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

By assumption C.2, $\beta_1(\cdot)$ 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. divide 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 $k \geq 0$.

Since $\beta_{0,1}(\tau)^k, k \geq 1$ 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 at 0 (notice that $m'(0) = 0$), we get:

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

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

By assumption C.2, $E[x'x]$ is non-singular. Therefore $E[x'_{-1}x_{-1}]$ is also non-singular. Hence, the above equation suggests $\beta_{-1}(\tau) = \beta_{0,-1}(\tau)$, for all $\tau \in [0,1]$. Therefore, $\beta(\tau) = \beta_0(\tau)$.This implies that $\phi(s) = \phi_0(s)$. Hence $f(\varepsilon) = f_0(\varepsilon)$.

□

Proof of Lemma 3.

Proof. Proof: (a) If the equation stated in (a) holds a.e., then we can apply Fourier transformation on both sides, i.e., conditional on x,

$$\int e^{isy}\int_0^1 f_y(y - x\beta(\tau))xt(\tau)d\tau dy = \int e^{isy}\int_0^1\int_0^1 f_\sigma(y - x\beta(\tau))d\tau dy. \tag{D.1}$$

Let $q(\tau) = x\beta(\tau)$, which is a strictly increasing function. We can use change of variables to change τ to q.

We get:

$$\int_0^1 e^{isx\beta(\tau)}xt(\tau)\int e^{is(y-x\beta(\tau))}f'(y-x\beta(\tau))dyd\tau = \int_0^1 e^{isx\beta(\tau)}\int e^{is(y-x\beta(\tau))}f_\sigma(y-x\beta(\tau))dyd\tau. \tag{D.2}$$

or, equivalently,

$$\int_0^1 e^{isx\beta(\tau)}xt(\tau)d\tau(-is)\phi_\varepsilon(s) = \int_0^1 e^{isx\beta(\tau)}d\tau\frac{\partial\phi_\varepsilon(s)}{\partial\sigma}, \tag{D.3}$$

given that $\int f'(y)e^{isy}dy = -is\int f(y)e^{isy}dy$.

Denote $\int_0^1 e^{isx\beta(\tau)}xt(\tau)d\tau = q(x,s)$, and $\int_0^1 e^{isx\beta(\tau)}d\tau = r(x,s)$. Therefore,

$$-isq(s,x)\phi(s) = r(s,x)\frac{\partial\phi_\varepsilon(s)}{\partial\sigma}.$$

First we consider normal distribution. $\phi_\varepsilon(s) = \exp(-\sigma^2s^2/2)$. So $\frac{\partial\phi_\varepsilon(s)}{\partial\sigma} = -\sigma s^2\phi_\varepsilon(s)$.

Therefore, $q(s,x) = is\sigma r(s,x)$, for all s and x with positive probability. Since $x\beta(\tau)$ is strictly increasing, let $z = x\beta(\tau)$, and $I(y) = 1(x\beta(0) \le y \le x\beta(1))$.

So $q(s,x) = \int_{-\infty}^{\infty} e^{isw}\{\frac{xt(z^{-1}(w))}{xz'(z^{-1}(w))}I(w)\}dw$, and $r(s,x) = \int_{-\infty}^{\infty} e^{isw}\{\frac{1}{xz'(z^{-1}(w))}I(w)\}dw$.

$q(s,x) = is\sigma r(s,x)$ means that there is a relationship between the integrands. More specifically, it must be that

$$\frac{xt(z^{-1}(w))}{xz'(z^{-1}(w))}I(w) = d\frac{1}{xz'(z^{-1}(w))}I(w)/dw. \tag{D.4}$$

Obviously they don't equal to each other for a continuous interval of x_i, because the latter function has a δ function component.

For more general distribution function f, $-isq(s,x)\phi_\varepsilon(s) = r(s,x)\frac{\partial\phi_\varepsilon(s)}{\partial\sigma}$, with $\frac{\partial\phi_\varepsilon(s)}{\partial\sigma} = \phi_\varepsilon(s)s^2m(s)$, for some $m(s) := \sum_{j=0}^{\infty} a_j(is)^j$.

Therefore,

$$-ism(s)r(s,x) = q(s,x).$$

Consider local expansion of s in both sides, we get:

$$-is(\sum_{j=0}^{\infty} a_j(is)^j)(\sum_{k=0}^{\infty}\int_0^1 (x\beta(\tau))^k d\tau\frac{(is)^k}{k!}) = \sum_{j=0}^{\infty}\int_0^1 (x\beta(\tau))^k xt(\tau)d\tau\frac{(is)^k}{k!}.$$

Compare the k^{th} polynomial of s, we get:

$\int_0^1(x\beta(\tau))^k xt(\tau)d\tau = 0$, if $k = 0$.

$\int_0^1(x\beta(\tau))^k xt(\tau)d\tau = -\sum_{1\le j<k} a_j(is)^j(\int_0^1 x\beta(\tau))^{j-1}d\tau\frac{(is)^{j-1}}{(j-1)!}$.

Let x_1 be the continuous variable, and x_{-1} be the variables left. Let β_1 be the coefficient corresponds to x_1 and let β_{-1} be the coefficient corresponds to x_{-1}. Let t_1 be the component in t that corresponds to x_1, and let t_{-1} be the component in t that corresponds to x_{-1}.

Then, consider the $(k+1)^{th}$ order polynomial of x_1, we get:

$$\int_0^1 \beta_1(\tau)^k t_1(\tau)d\tau = 0.$$

since $\beta_1(\tau)^k$, $k \ge 0$ form a functional basis in $L^2[0,1]$, $t_1(\tau)$ must equal to 0.

Consider the k^{th} order polynomial of x_1, we get:

$$\int_0^1 \beta_1(\tau)^k (x_{-1}t_{-1}(\tau))d\tau = 0.$$

So $x_{-1}t_{-1}(\tau) = 0$ for all $\tau \in [0,1]$. Since $E[x_{-1}x'_{-1}]$ has full rank, it must be that $t_{-1}(\tau) = 0$. Hence local identification condition holds. □

D.2. **Lemmas and Theorems in Section 3.** In this subsection, we prove the Lemmas and Theorems in section 3.

Lemma 9 (Donskerness of Θ)**.** *The parameter space Θ is Donsker.*

Proof. By theorem 2.7.5 of Van der Vaart and Wellner (2000), the space of bounded monotone function, denoted as $\mathcal{F}$, satisfies:

$$logN_{[]}(\varepsilon, \mathcal{F}, L_r(Q)) \le K\frac{1}{\varepsilon},$$

for every probability measure Q and every $r \ge 1$, and a constant K which depends only on r.

Since Θ is a product space of bounded monotone functions M and a finite dimensional bounded compact set Σ, therefore the bracketing number of Θ given any measure Q is also bounded by:

$$logN_{[]}(\varepsilon, \times, L_r(Q)) \leq Kd_x \frac{1}{\varepsilon},$$

Therefore, $\int_0^\delta \sqrt{\log N_{[]}(\varepsilon, \Theta, Q)} < \infty$, i.e., Θ is Donsker. □

Proof of Lemma 4.

Proof. $L_n(\theta) \geq L_n(\theta_0)$ by definition, and we know that $L(\theta_0) \geq L(\theta)$ by information inequality.

By Lemma 10, since Θ is Donsker, then the following stochastic equicontinuity condition holds:

$$\sqrt{n}(L_n(\theta) - L_n(\theta_1) - (L(\theta) - L(\theta_1))) \to_p r_n,$$

for some generic term $r_n = o_p(1)$, if $|\theta - \theta_1| \leq \eta_n$ for small enough η_n. Therefore uniform convergence of $L_n(\theta) - L(\theta)$ holds in Θ. In addition, by Lemma 2, the parameter θ is globally identified. Also, it is obvious that L is continuous in θ.

Therefore, the usual consistency result of MLE estimator applies here. □

Proof of Lemma 5.

Proof of Lemma 6.

Proof. By definition,

$$\mathbb{E}_n[\log(g(y|x, \beta, \sigma))] \geq \mathbb{E}_n[\log(g(y|x, \beta_0, \sigma_0))]. \tag{D.5}$$

And we know that by information inequality, $\mathbb{E}[\log(g(y|x, \beta, \sigma))] \leq \mathbb{E}[\log(g(y|x, \beta_0, \sigma_0))]$.

Therefore, $\mathbb{E}[\log(g(y|x, \beta, \sigma))] - \mathbb{E}[\log(g(y|x, \beta_0, \sigma_0))] \geq \frac{1}{\sqrt{n}}\mathbb{G}_n[\log(g(y|x, \beta, \sigma)) - g(y|x, \beta_0, \sigma_0))]$.

By Donskerness of parameter space Θ, the following equicontinuity condition holds:

Hence,

$$0 \leq \mathbb{E}[\log(g(y|x, \beta_0, \sigma_0))] - \mathbb{E}[\log(g(y|x, \beta, \sigma))] \precsim_p \frac{1}{\sqrt{n}}. \tag{D.6}$$

Because σ is converging to the true parameter in $\sqrt{n}$ speed, we will simply use σ_0 here in the argument because it won't affect our analysis later. In the notation, I will abbreviate σ_0.

Let $z(y, x) = g(y|\beta, x) - g(y|\beta_0, x)$. In equation 3.19, we know that the KL-discrepancy of $g(y|x, \beta_0)$ and $g(y|x, \beta)$, noted as $D(g(\cdot|\beta_0)||g(\cdot|\beta))$, is bounded by $C\frac{1}{\sqrt{n}}$ with $p \to 1$.

So $E_x[||z(y|x)||_1]^2 \leq D(g(\cdot|\beta_0)||g(\cdot|\beta)) \leq 2(E[\log(g(y|x, \beta_0, \sigma_0))] - E[\log(g(y|x, \beta, \sigma))]) \precsim_p \frac{1}{\sqrt{n}}$, where $|z(y|x)|_1 = \int_{-\infty}^{\infty} |z(y|x)|dy$.

Now consider the characteristic function of y conditional on x:

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

and

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

Therefore

$$\phi_{x\beta}(s) - \phi_{x\beta_0}(s) = \frac{\int_{-\infty}^{\infty} z(y|x)e^{isy}dy}{\phi_\varepsilon(s)}, \tag{D.7}$$

and $|\phi_{x\beta}(s) - \phi_{x\beta_0}(s)| \leq \frac{||z(y|x)||_1}{|\phi_\varepsilon(s)|}$.

Let $F_\eta(w) = \frac{1}{2} - \lim_{q\to\infty} \int_{-q}^{q} \frac{e^{-iws}}{2\pi s}\phi_\eta(s)ds$. Then $F(w)$ is the CDF of the random variable η.

Let $F_\eta(w, q) = \frac{1}{2} - \int_{-q}^{q} \frac{e^{-iws}}{2\pi s}\phi_\eta(s)ds$. If the p.d.f of η is a C^1 function and η has finite second order moment, then $F_\eta(w, q) = F_\eta(w) + O(\frac{1}{q^2})$. Since x and β are bounded, let $\eta = x\beta(\tau)$, then there exists a common C such that $|F_{x\beta}(w, q) - F_{x\beta}(w)| \leq \frac{C}{q^2}$ for all $x \in \mathcal{X}$ and $w \in \mathbb{R}$ for q being large enough.

Therefore $|F_{x\beta}(w, q) - F_{x\beta_0}(w, q)| = |\int_{-q}^{q} \frac{e^{isw}}{2\pi s}(\phi_{x\beta}(s) - \phi_{x\beta_0}(s))ds| \leq |\int_{-q}^{q} \frac{\frac{||z(y|x)||_1}{2\pi s}}{|\phi_\varepsilon(s)|}ds| \precsim ||z(y|x)||_1\chi(q)$.

Consider the follow moment condition:

$$E[x(F_{x\beta_0}(x\beta_0) - \tau)] = 0 \tag{D.8}$$

Let F_0 denote $F_{x\beta_0}$ and F denote $F_{x\beta}$.

We will replace the moment condition stated in (J.8) by the truncated function $F(w, q)$. So by the truncation property, for any fixed $\delta \in \mathbb{R}^{d_x}$:

$$\delta E[x(F_0(x\beta) - F_0(x\beta_0))] = \delta E[x(F_0(x\beta) - F(x\beta))] = \delta E[x(F_0(x\beta, q) - F(x\beta, q))] + O(\frac{1}{q^2}). \tag{D.9}$$

The first term of the right hand side is bounded by:

$$E[|\delta x|\chi(q)||z(y|x)||_1] \precsim E[||z(y|x)||_1]\chi(q).$$

And $\delta E[x(F_0(x\beta) - F_0(x\beta_0))] = \delta E[x' x f(x\beta_0(\tau))](\beta(\tau) - \beta_0(\tau))(1 + o(1))$, where $f(x\beta_0(\tau))$ is the p.d.f of the random variable$x\beta_0(\cdot)$ evaluated at τ.

Hence, $\beta(\tau) - \beta_0(\tau) = O_p(\frac{\chi(q)}{n^{\frac{1}{4}}}) + O(\frac{1}{q^2})$.

Under OS condition, $\chi(q) \precsim q^\lambda$, so by choosing $q = Cn^{\frac{1}{4(\lambda+2)}}$, $||\beta(\tau) - \beta_0(\tau)||_2 \precsim_p n^{-\frac{1}{2(\lambda+2)}}$.

Under SS condition, $\chi(q) \precsim C_1(1 + |q|^{C_2}) \exp(|q|^\lambda/C_3)$. By choosing $q = C\log(n)^{\frac{1}{4\lambda}}$, $||\beta(\tau) - \beta_0(\tau)||_2 \precsim_p \log(n)^{-\frac{1}{2\lambda}}$. □

D.3. **Lemmas and Theorems in Sections 4 and 5.** In this subsection we prove the Lemmas and Theorems in section 4.

Proof of Lemma 8.

Proof. Suppose f satisfies discontinuity condition C8 with degree $\lambda > 0$.

Denote $l_k = (f_{\tau_1}, f_{\tau_2}, ..., f_{\tau_k}, g_\sigma)$.

For any $(p_k, \delta) \in \mathcal{P}_k$, $p_k I_k p_k' = E[\int_{\mathbb{R}} \frac{(l_k p_k')^2}{g} dy] \geq E[(l_k p_k')^2]$, since g is bounded from above.

Assume $c := \inf_{x \in \S, \tau \in [0,1]} (x\beta'(\tau)) > 0$. Now consider an interval $[y, y + \frac{1}{2\lambda kc}]$.

Let $S(\lambda) := \sum_{i=0}^{\lambda} (\lambda C_\lambda^i)^2$, where C_a^b stands for the combinatorial number choosing b elements from a set with size a.

So,

$$S(\lambda) E[\int_{\mathbb{R}} \frac{(l_k p_k')^2}{d} y] = E[\sum_{j=0}^{\lambda} (C_\lambda^j)^2 \int_{\mathbb{R}} (l_k(y + j/(2\lambda ukc)) p_k'(y + j/(2\lambda kc)))^2 dy].$$

And by Cauchy-Schwarz inequality,

$$E[\sum_{j=0}^{\lambda} (C_\lambda^j)^2 \int_{\mathbb{R}} (l_k(y + j/(2\lambda ukc)) p_k'(y + j/(2\lambda ukc)))^2 dy]$$

$$\geq E[\int_{\mathbb{R}} (\sum_{j=0}^{\lambda} (-1)^j C_\lambda^j l_k(y + j/(2\lambda ukc)) p_k'(y + j/(2\lambda ukc)))^2 dy].$$

let $J_k^i := [a + x\beta(\frac{i+1/2}{k}) - \frac{1}{2\lambda uk}, a + x\beta(\frac{i+1/2}{k}) + \frac{1}{2\lambda uk}]$, so

$$E[\sum_{j=0}^{\lambda} (C_\lambda^j)^2 \int_{\mathbb{R}} (l_k(y + j/(2\lambda ukc)) p_k'(y + j/(2\lambda ukc)))^2 dy]$$

$$\geq E[\int_{J_k^i} (\sum_{j=0}^{\lambda} (-1)^j C_\lambda^j l_k(y + j/(2\lambda ukc)) p_k'(y + j/(2\lambda ukc)))^2 dy].$$

By discontinuity assumption, ,

$$\sum_{j=0}^{\lambda} (-1)^j C_\lambda^j l_k(y + j/(2\lambda ukc)) p_k'(y + j/(2\lambda ukc))$$

$$= \frac{1}{(uk)^{\lambda-1}} (\frac{c}{(\lambda-1)!} x_i p_i' + \frac{c_j}{uk} \sum_{j=1, j \neq i}^{k} x_j p_j),$$

with c_j uniformly bounded from above, since $f^{(\lambda)}$ is L^1 Lipschitz except at a.

Notice that J_k^i does not intersect each other, so $S(\lambda) E[\int_{\mathbb{R}} (l_k p_k')^2 dy] \geq S(\lambda) \sum_{i=1}^{k} E[\int_{J_k^i} (l_k p_k')^2 dy]$
$\geq E[\frac{c'}{2uk} \sum_{i=1}^{k} \{\frac{1}{(uk)^{\lambda-1}} (\frac{c}{(\lambda-1)!} x_i p_i' + \frac{c_{jk}}{uk} \sum_{j=1, j \neq i}^{k} x_j p_j)\}^2]$

For constant u being large enough (only depend on λ, $\sup c_{jk}$ and c),

$E[\sum_{i=1}^k\{\frac{1}{(uk)^{\lambda-1}}(\frac{c}{(\lambda-1)!}x_ip_i' + \frac{c_{jk}}{uk}\sum_{j=1,j\neq i}^k x_jp_j)\}^2] \geq c(\lambda)\frac{1}{k^{2\lambda-1}}\sum_{j=1}^k E[x_j^2p_j^2] \asymp \frac{1}{k^{2\lambda}}||p||_2^2$, with some constant $c(\lambda)$ only depends on λ and u.

In addition, from condition C5, we know that $\theta I_k\theta \gtrsim ||\delta||_2^2$.

Therefore $\theta I_k\theta \gtrsim ||\delta||_2^2 + \frac{1}{k^{2\lambda}}||p||_2^2 \gtrsim \frac{1}{k^{2\lambda}}||\theta||_2^2$. Hence, the smallest eigenvalue $r(I_k)$ of I_k is bounded by $\frac{c}{k^\lambda}$, with some generic constant c depending on λ, x, the the L^1 Lipshitz coefficient of $f^{(k)}$ at set $\mathbb{R} - [a-\eta, a+\eta]$. □

Proof of Lemma 9.

Proof. The proof of this Lemma is based on Chen (2008) results of sieve extremum estimator consistency. Below I recall the five Conditions listed in Theorem 3.1 of Chen (2008). Theorem 3.1 in Chen (2008) shows that as long as the following conditions are satisfied, the sieve estimator is consistent.

*Condition C*11. [Identification] (1) $L(\theta_0) < \infty$, and if $L(\theta_0) = -\infty$ then $L(\theta) > -\infty$ for all $\theta \in \Theta_k \setminus \{\theta_0\}$ for all $k \geq 1$.

(2)there are a nonincreasing positive function $\delta(\cdot)$ and a positive function $m(\cdot)$ such that for all $\varepsilon > 0$ and $k \geq 1$,

$$L(\theta_0) - \sup_{\theta\in\Theta_k: d(\theta,\theta_0)L(\theta_k)\geq\varepsilon} \geq \delta(k)m(\varepsilon) > 0.$$

*Condition C*12. [Sieve Space] $\Theta_k \subset \Theta_{k+1} \subset \Theta$ for all $k \geq 1$; and there exists a sequence $\pi_k\theta_0 \in \Theta_k$ such that $d(\theta_0, \pi_k\theta_0) \to 0$ as $k \to 0$.

*Condition C*13. [Continuity] (1)$L(\theta)$ is upper semicontinuous on Θ_k under metric $d(\cdot,\cdot)$.

(2) $|L(\theta_0) - L(\pi_{k(n)}\theta_0)| = o(\delta(k(n)))$.

*Condition C*14. [compact Sieve space] Θ_k is compact under $d(\cdot,\cdot)$.

*Condition C*15. [uniform convergence] 1) For all $k \geq 1$, $\operatorname{plim}\sup_{\theta\in\Theta_k}|L_n(\theta) - L(\theta)| = 0$. 2) $\widehat{c}(k(n)) = o_p(\delta(k(n)))$ where $\widehat{c}(k(n)) := \sup_{\theta\in\Theta_k}|L_n(\theta) - L(\theta)|$. 3) $\eta_{k(n)} = o(\delta(k(n)))$.

These assumptions are verified below:

For Identification: Let d be the metric induced by theL^2 norm $||\cdot||_2$defined on Θ. By assumption $C4$ and compactness of Θ, $L(\theta_0) - \sup_{\theta\in\Theta:d(\theta,\theta_0)} L(\theta_k) > 0$ for any $\varepsilon > 0$. So let $\delta(k) = 1$, Identification condition holds.

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

For condition Continuity: Since we assume f is continuous and L^1 Lipshitz, (1) is satisfied. (2) is statisfied under the construction of our Sieve space.

Condition Compact Sieve Space is trivial.

The condition Uniform Convergence is also easy to verify since the entropy metric of space Θ_k is finite and uniformly bounded by the entropy metric of Θ. Therefore we have stochastic equi-continuity, i.e., $\sup_{\theta\in\Theta_k}|L_n(\theta)-L(\theta)| = O_p(\frac{1}{\sqrt{n}})$.

In (3), η_k is defined as the error in the maximization procedure. Since δ is constant, as soon as our maximizer $\widehat{\theta_k}$ satisfies $L_n(\widehat{\theta_k}) - \sup_{\theta\in\Theta_k} L_n(\theta_k) \to 0$, (3) is verified.

Hence, the Sieve MLE estimator is consistent.

□

Proof of Lemma 10.

Proof. This theorem follows directly from Lemma 8. If the neighborhood $\cdot_j$ of θ_0 is chosen as $\Delta_j := \{\theta\in\Theta_k : |\theta-\theta_0|_2 \le \frac{\mu_k}{k^\lambda}\}$,

with μ_k being a sequence converging to 0, then the non-linear term in $E_x[\int_{\mathcal{Y}} \frac{(g(y|x,\theta_0)-g(y|x,\theta))^2}{g(y|x,\theta_0)}dy]$ is dominated by the first order term $\theta I^k\theta'$.

□

Proof of Theorem 1.

Proof. Suppose $\theta = (\beta(\cdot),\sigma) \in \Theta_k$ is the Sieve estimator. By Lemma 8, the consistency of Sieve estimator shows that $||\theta-\theta_0||_2 \to_p 0$.

Denote $\tau_i = \frac{i-1}{k}$, $i = 1,2,...,k$

First order condition gives:

$$E_n[-xf'(y-x\beta(\tau_i)|\sigma)] = 0. \tag{D.10}$$

and

$$E_n[g_\sigma(y|x,\beta,\sigma)] = 0. \tag{D.11}$$

Assume that we choose $k = k(n)$ such that:

(1)$k^{\lambda+1}||\theta-\theta_0||_2 \to_p 0$.

(2)$k^r||\theta-\theta_0||_2 \to \infty$ with $p\to 1$.

And assume that $r > \lambda+1$. Such sequence $k(n)$ exists because Lemma 7 shows an upper bound for $||\theta-\theta_0||$, i.e., $||\theta_0-\theta||_2 \precsim_p n^{-\frac{1}{2(2+\lambda)}}$.

From the first order condition, we know that:

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

Denote $f_{\theta_k} = (f_{\tau_1}(y|x,\theta), f_{\tau_2}(y|x,\theta),...,f_{\tau_k}(y|x,\theta))$.

Therefore,

$$0 = E_n[(\frac{f_{\theta_k}}{g}, \frac{g_\sigma}{g})] = (E_n[(\frac{f_{\theta_k}}{g}, \frac{g_\sigma}{g})] - E_n[(\frac{f_{\theta_k,0}}{g_0}, \frac{g_{\sigma,0}}{g_0})]) + E_n[(\frac{f_{\theta_k,0}}{g_0}, \frac{g_{\sigma,0}}{g_0})].$$

The first term can be decomposed as:

$$E_n[(\frac{f_{\theta_k}}{g},\frac{g_\sigma}{g})]-E_n[(\frac{f_{\theta_k}}{g_0},\frac{g_\sigma}{g_0})]+E_n[(\frac{f_{\theta_k}}{g_0},\frac{g_\sigma}{g_0})-(\frac{f_{\theta_k,0}}{g_0},\frac{g_{\sigma,0}}{g_0})].$$

By C10, we know that CLT holds pointwise in the space Θ_k for $E_n[(\frac{f_{\theta_k}}{g_0},\frac{g_\sigma}{g_0})]$.

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 Θ and $U[0,1]$, Σ is Donsker. By C10, we have pointwise CLT for $H((u,\theta))$.

By condition C2, the Lipshitz 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 γ small enough,

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

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

Notice $E[(\frac{f_{\theta_k}}{g_0},\frac{g_\sigma}{g_0})-(\frac{f_{\theta_k,0}}{g_0},\frac{g_{\sigma,0}}{g_0})]=0$, so $E_n[(\frac{f_{\theta_k}}{g_0},\frac{g_\sigma}{g_0})-(\frac{f_{\theta_k,0}}{g_0},\frac{g_{\sigma,0}}{g_0})]=o_p(\frac{1}{\sqrt{n}})+E[(\frac{f_{\theta_k}}{g_0},\frac{g_\sigma}{g_0})-(\frac{f_{\theta_k,0}}{g_0},\frac{g_{\sigma,0}}{g_0})]=o_p(\frac{1}{\sqrt{n}})$.

Therefore, the first term $=E_n[(\frac{f_{\theta_k}}{g},\frac{g_\sigma}{g})]-E_n[(\frac{f_{\theta_k}}{g_0},\frac{g_\sigma}{g_0})]+o_p(1)$.

Similarly, for the second term, $E_n[(\frac{f_{\theta_{k,0}}}{g_0},\frac{g_{\sigma,0}}{g_0})]$, we can establishes a functional CLT:

$$E_n[(\frac{f_{\theta_{k,0}}}{g_0},\frac{g_{\sigma,0}}{g_0})]=\frac{1}{\sqrt{n}}\mathbb{G}_n(\tau_1,\tau_2,...,\tau_k)+o_p(\frac{1}{\sqrt{n}})\in\mathbb{R}^{krd_x+d_p},$$

where $\mathbb{G}_n(\cdot):=\sqrt{n}E_n[(\frac{f_\tau}{g_0}|_{0\le\tau\le1},\frac{g_{\sigma,0}}{g})]$ is weakly converging to a tight Gaussian process.

Therefore,

$$E_n[(\frac{f_{\theta_k}}{g},\frac{g_\sigma}{g})]-E_n[(\frac{f_{\theta_k}}{g_0},\frac{g_\sigma}{g_0})]]=-\frac{1}{\sqrt{n}}\mathbb{G}_n(\tau_1,\tau_2,...,\tau_k)(1+o_p(1)).$$

Similarly by stochastic equicontinuity, we can replace f_{θ_k} with f_{θ_0} and g_σ with σ_0 on the left hand side of the above equation. And now we get:

$$E[(f_{\theta_0},g_{\sigma_0})(\frac{g-g_0}{g_0^2})]=\frac{1}{\sqrt{n}}\mathbb{G}_n(\tau_1,\tau_2,...,\tau_k)(1+o_p(1)),\tag{D.12}$$

where $g-g_0=g(y|x,\theta)-g(y|x,\theta_0):=g(y|x,\theta)-g(y|x,\pi_k\theta_0)+(g(y|x,\pi_k\theta_0)-g(y|x,\theta_0))$.

By the construction of Sieve, $g(y|x,\pi_k\theta_0)-g(y|x,\theta_0)=O(\frac{1}{k^{r+1}})$, which depends only on the order of Spline employed in the estimation process.

The first term $g(y|x,\theta)-g(y|x,\pi_k\theta_0)=(\theta-\pi_k\theta_0)(f_{\widetilde{\theta}},g_{\widetilde{\sigma}})'$,

where $\widetilde{\theta}$ and $\widetilde{\sigma}$ are some value between θ and $\pi_k\theta_0$.

Denote $\widehat{I}_k=E[(f_{\theta_0},g_{\sigma_0})'(f_{\widetilde{\theta}},g_{\widetilde{\sigma}})]$.

The eigenvalue norm $||\widehat{I}_k-I_k||_2=||E[(f_{\theta_0},g_{\sigma_0})'(f_{\widetilde{\theta}}-f_{\theta_0},g_{\widetilde{\sigma}}-g_{\sigma_0})]||_2$.

While $||f_{\widetilde{\theta}}-f_{\theta_0},g_{\widetilde{\sigma}}-g_{\sigma_0}||_2\precsim||\theta-\pi_k\theta_0||_2+||\pi_k\theta_0-\theta_0||_2\precsim||\theta-\pi_k\theta_0||_2$ (By the choice of k, $||\theta-\pi_k\theta_0||_2\ge||\theta-\theta_0||_2-||\pi_k\theta_0-\theta_0||_2$. The first term dominates the second term because $O(\frac{1}{k^r})\succsim||\pi_k\theta_0-\theta_0||_2$ and $k^r||\theta-\theta_0||_2\to_p\infty$.)

Since $\sup_{x,y} ||(f_{\theta_0}, g_{\sigma_0})||_2 \precsim k$, $||\widehat{I}_k - I_k||_2 \precsim k||\theta - \pi_k\theta_0||_2$.

Denote $H := \widehat{I}_k - I_k$, so

$$||H||_2 \precsim k||\theta - \pi\theta_0||_2 \precsim ||\theta - \theta_0||_2 \precsim_p \frac{1}{k^\lambda}. \tag{D.13}$$

$\widehat{I}_k^{-1} = (I_k + H)^{-1} = I_k^{-1}(I + I_k^{-1}H)^{-1}$.

By (J.13), $||I_k^{-1}H||_2 \leq k^\lambda \frac{o(1)}{k^\lambda} = o_p(1)$. Therefore, $I + I_k^{-1}H$ is invertible and has eigenvalues bounded from above and below uniformly in k. Hence the matrix $\widehat{I}_k^{-1}$'s largest eigenvalue is bounded by Ck^λ for some generic constant C with $p \to 1$. Furthermore, $\widehat{I}_k^{-1} = I_k^{-1}(1 + o_p(1))$.

Apply our result in (J.12), we get: $\theta - \pi\theta_0 = I_k^{-1}\frac{1}{\sqrt{n}}\mathbb{G}_n(1 + o_p(1))$. So $||\theta - \theta_0||_2 = O_p(\frac{k^\lambda}{\sqrt{n}})$. Moreover, under condition C.11,

$||\theta - \theta_0||_2 \succsim_p$

Since at the beginning we assume: (1) $k^{\lambda+1}||\theta - \theta_0||_2 \to 0$. (2) $k^r||\theta - \theta_0||_2 \to \infty$ with $p \to 1$, and $||\theta_0 - \pi\theta_0||_2 = O_p(\frac{1}{k^r})$.

Now we prove the condition that $||\theta - \theta_0||_2 = O_p(\frac{k^\lambda}{\sqrt{n}})$

the regularity condition for k is:

(a)$\frac{k^{2\lambda+1}}{\sqrt{n}} \to 0$.

(b)$\frac{k^{\lambda+r-\alpha}}{\sqrt{n}} \to \infty$.

If k satisfies growth conditions (1) and (2), the Sieve estimator θ_k satisfies:

$$\theta_k - \theta_0 = \theta_k - \pi\theta_0 + \pi\theta_0 - \theta_0 = O_p(\frac{k^\lambda}{\sqrt{n}}) + O(\frac{1}{k^r}) = O_p(\frac{k^\lambda}{\sqrt{n}}).$$

Under condition C.11,

$$\theta - \pi\theta_0 = I_k^{-1}\frac{1}{\sqrt{n}}\mathbb{G}_n(1 + o_p(1)) \succsim_p \frac{k^{\lambda-\alpha}}{\sqrt{n}} \succsim \frac{1}{k^r}.$$

Therefore,

$$\theta - \theta_0 = I_k^{-1}\frac{1}{\sqrt{n}}\mathbb{G}_n(1 + o_p(1)).$$

And pointwise asymptotic normality holds.

For parameter σ, we would like to recall Chen's (2008) result. In fact, we use a result slightly stronger than that in Chen (2008), but the reasoning are similar. For k satisfying growth condition in (1), we know that $L(\theta) - L(\theta_0) \asymp ||\theta - \theta_0||_k$, where $\theta_k \in \Theta_k \cap \mathcal{N}_k$ and $||\cdot||_k$ is defined as $< \cdot, I^k \cdot >$.

For a functional of interest, $h : \Theta :\to \mathbb{R}$, define the partial derivative:

$$\frac{\partial h}{\partial \theta} := \lim_{t\to 0} \frac{h(\theta_0 + t[\theta - \theta_0]) - h(\theta_0)}{t}.$$

Define $\bar{V}$ be the space spanned by $\Theta - \theta_0$.

By Riesz representation theory, there exists a $v^* \in \bar{V}$ such that $\frac{\partial h}{\partial \theta}[\theta - \theta_0] =< \theta - \theta_0, v^* >$, where $< \cdot, \cdot >$ is induced by $||\cdot||$.

Below are five conditions needed for asymptotic normality of $h(\theta)$.

*Condition C*16. (i)There is $\omega > 0$ such that $|h(\theta-\theta_0-\frac{\partial h(\theta_0)}{\partial\theta}[\theta-\theta_0])| = O(||\theta-\theta_0||^{\omega})$ uniformly in $\theta \in \Theta_n$ with $||\theta-\theta_0|| = o(1)$;
(ii)$||\frac{\partial h(\theta_0)}{\partial\theta}|| < \infty$.
(iii)there is $\pi_n v^* \in \Theta_k$ such that $||\pi_n v^* - v^*|| \times ||\widehat{\theta}_n - \theta_0|| = o_p(n^{\frac{-1}{2}})$.

This assumption is easy to verify because our function h is a linear function in σ: $h(\theta) = \delta'\sigma$. Therefore the first and third condition is automatically verified. The second condition is true because of assumption C.6. So $v^* = (0, E[g_{\sigma,0}g'_{\sigma,0}]^{-1}\delta)$. Denote $I_\sigma := E[g_{\sigma,0}g'_{\sigma,0}]$

Let ε_n denote any sequence satisfying: $\varepsilon_n = o(n^{-1/2})$. Suppose the Sieve estimator converges to θ_0 faster than δ_n.

Define the functional operator $\mu_n = E_n - E$.

*Condition C*17.

$$\sup_{\theta\in\Theta_{k(n)};||\theta-\theta_0||\leq\delta_n} \mu_n(l(\theta,Z) - l(\theta \pm \varepsilon_n\pi_n v^*, Z) - \frac{\partial l(\theta_0,Z)}{\partial\theta}[\pm\varepsilon_n\pi_n v^*]) = O_p(\varepsilon_n^2).$$

In fact this assumption is the stochastic continuity assumption. It can be verified by P-Donskerness of Θ and L^1 Lipshitz property that we assume on the function f.

Denote $K(\theta_1,\theta_2) = L(\theta_1) - L(\theta_2)$.

*Condition C*18.

$$E[\frac{\partial l(\widehat{\theta}_n,Z)}{\partial\theta}[\pi_n v^*]] =< \widehat{\theta}_n - \theta_0, \pi_n v^* > +o(n^{-1/2}).$$

We know that $E[\frac{\partial l(\widehat{\theta}_n,Z)}{\partial\theta}[\pi_n v^*]] = E[\frac{g_\sigma(y|x,\widehat{\theta}_n)}{g}' I_\sigma^{-1}\delta] = E[\frac{g_\sigma(y|x,\widehat{\theta}_n)}{g}']I_\sigma^{-1}\delta$.

We know that $|\widehat{\theta}_n - \theta_0|_2 = O_p(\frac{k^\lambda}{\sqrt{n}})$ and $\frac{k^{2\lambda+1}}{\sqrt{n}} \to 0$, so $|\theta-\theta_0|_2^2 = o_p(n^{-(\lambda+1)/(2\lambda+1)}) = o(n^{-1/2})$.

And we know that $E[\frac{g_\sigma(y|x,\widehat{\theta}_n)}{g_0}] = 0$. So

$$E[\frac{g_\sigma(y|x,\widehat{\theta}_n)}{g}] = E[g_\sigma(y|x,\widehat{\theta}_n)(\frac{1}{g} - \frac{1}{g_0})].$$

By L^1 Lipshitz condition $|g-g_0| \leq C_1(x,y)|\theta-\theta_0|_2$, and $|g_\sigma(|\widehat{\theta}_n)-g\sigma(|\theta_0)| \leq C_2(x,y)|\theta-\theta_0|_2$. And such coefficients has nice tail properties. Therefore,

$$E[g_\sigma(y|x,\widehat{\theta}_n)(\frac{1}{g} - \frac{1}{g_0})] = E[g_\sigma(y|x,\theta_0)(\frac{\partial g/\partial\theta}{g^2})(\widehat{\theta}_n - \theta_0)] + O(|\widehat{\theta}_n - \theta_0|_2^2).$$

Then $E[\frac{\partial l(\widehat{\theta}_n,Z)}{\partial\theta}[\pi_n v^*]] = E[(\widehat{\theta}_n - \theta_0)'\frac{\partial g/\partial\theta}{g^2}g_\sigma(y|x,\theta_0)'\delta] + o_p(n^{-1/2})$.

The first part of the equation above is $< \widehat{\theta}_n - \theta_0, \pi_n v^* >$.

*Condition C*19. (i) $\mu_n(\frac{\partial l(\theta_0,Z)}{\partial\theta}[\pi_n v^* - v^*]) = o_p(n^{-1/2})$.

(ii) $E[\frac{\partial l(\theta_0,Z)}{\partial \theta}[\pi_n v^*]] = o(n^{-1/2})$.

(i) is obvious since $v^* = \pi_n v^*$. (ii) is also obvious since it equals to 0.

*Condition C*20.

$$\mathbb{G}_n\{\frac{\partial l(\theta_0, Z)}{\partial \theta}[v^*]\} \to_d \mathcal{N}(0, \sigma_v^2),$$

with$\sigma_v^2 > 0$.

This assumption is also not hard to verify because $\frac{\partial l(\theta_0,Z)}{\partial \theta}[v^*]$ is a random variable with variance bounded from above. It is not degenerate because $E[g_\sigma g_\sigma']$ is positive definite.

Then by theorem 4.3 in Chen (2008), for any σ, our sieve estimator θ_n have the property:

$$\sqrt{n}\delta'(\sigma_n - \sigma_0) \to_d \mathcal{N}(0, \sigma_v^2).$$

Since the above conclusion is true for arbitrary δ, $\sqrt{n}(\sigma_n - \sigma_0)$ must be jointly asymptotic normal, i.e., there exists a positive definite matrix V such that $\sqrt{n}(\sigma_n - \sigma_0) \to_d \mathcal{N}(0, V)$.

□

D.4. **Lemmas and Theorems in Appendix I.** In this subsection we prove the Lemmas and Theorems stated in section 5.

Proof of Theorem 2.

Proof. For simplicity, we assume the $\phi_\varepsilon(s)$ is a real function. The proof is similar when it is complex. Then,

$$F(x\beta, y, q) = \frac{1}{2} - \tau + \int_0^q \frac{\sin((y - x\beta)s)}{\pi s} \frac{1}{\phi_\varepsilon(s)} ds.$$

Under condition OS,

$|F(x\beta, y, q)| \leq C(q^{\lambda-1} + 1)$ if $\lambda \geq 1$, for some fixed C and q large enough.

Under condition SS, $|F(x\beta, y, q)| \leq C_1 \exp(q^\lambda / C_2)$ for some fixed C_1 and C_2.

Let $h(x, y) := E[\int_0^{q_n} \frac{cos((y-x\beta)s)}{\pi} \frac{ds}{\phi_\varepsilon(s)}]$.

By change of integration, $E[\int_0^{q_n} \frac{cos((y-x\beta)s)}{\pi} \frac{ds}{\phi_\varepsilon(s)}] \leq C$.

Let $\sigma(x)^2 = E[F(x\beta, y, \tau, q)^2|x]$.

Under condition OS, by similar computation, $C_1 q^{2(\lambda-1)} \leq \sigma(x)^2 \leq C_2 q^{2(\lambda-1)}$, and under condition SS, $C_1 \exp(q^\lambda/C_3) \leq \sigma(x)^2 \leq C_2 \exp(q^\lambda/C_4)$, for some fixed C_1, C_2, C_3, C_4.

For simplicity, let $m(\lambda) = q^{\lambda-1}$ if OS condition holds, and let $m(\lambda) = \exp(q^\lambda/C)$ if SS holds with constant C.

β is solving:

$$E_n[x \frac{h(x)}{\sigma(x)^2} F(x\beta, y, \tau, q)] = 0. \tag{D.14}$$

Adopting the usual local expansion:

$$E_n[\frac{h(x,y)}{\sigma(x)^2}(F(x\beta,y,\tau,q)-F(x\beta_0,y,\tau,q))] = -\{E_n[\frac{h(x,y)}{\sigma(x)^2}(F(x\beta_0,y,\tau,q)]-E[\frac{h(x)}{\sigma(x)^2}(F(x\beta_0,y,\tau,q)]\} \\ -E[\frac{h(x,y)}{\sigma(x)^2}(F(x\beta_0,y,\tau,q)]. \tag{D.15}$$

The left hand side equals:

$$(\beta-\beta_0)E[x'x\frac{h(x,y)^2}{\sigma(x)^2}](1+o_p(1)) = \frac{q^2}{m(\lambda)^2}E[l(x,q)x'x](\beta-\beta_0)(1+o_p(1)).$$

$l(x,q)$ is a bounded function converging uniformly to $l(x,\infty)$.

The first term of the RHS is:

$$-\frac{1}{\sqrt{n}}G_n[x\frac{h(x,y)}{\sigma(x)}\frac{f(x\beta,y,\tau,q)}{\sigma(x)}] = O_p(\frac{1}{\sqrt{n}m(\lambda)}).$$

The second term of the RHS is:

$$E[\frac{h(x)}{\sigma(x)^2}(F(x\beta_0,y,\tau,q)] = O(\frac{q}{m(\lambda)^2}).$$

Therefore, $\beta-\beta_0 = O_p(\frac{m(\lambda)}{\sqrt{n}}) + O(\frac{1}{q})$.

Under SS, the optimal q level is $\log(n)^{\frac{1}{\lambda}}$, and the optimal convergence speed is therefore $\log(n)^{\frac{-1}{\lambda}}$. Under OS condition, the optimal $q = n^{\frac{1}{2\lambda}}$, and the optimal convergence speed is $\frac{1}{n^{\frac{1}{2\lambda}}}$. □