NBER WORKING PAPER SERIES

ERRORS IN THE DEPENDENT VARIABLE OF QUANTILE REGRESSION MODELS

Jerry A. Hausman
Haoyang Liu
Ye Luo
Christopher Palmer

Working Paper 25819
http://www.nber.org/papers/w25819

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2019

We thank Isaiah Andrews, Colin Cameron, Victor Chernozhukov, Denis Chetverikov, Kirill Evdokimov, Hank Farber, Brigham Frandsen, Larry Katz, Brad Larsen, Rosa Matzkin, James McDonald, Ulrich Muller, Shu Shen, and Steven A. Snell for helpful feedback and discussions, as well as seminar participants at Cornell, Harvard, MIT, Princeton, UC Davis, UCL, and UCLA. Lei Ma, Yuqi Song, and Jacob Ornelas provided outstanding research assistance. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research, the Federal Reserve Bank of New York or the Federal Reserve System.

Errors in the Dependent Variable of Quantile Regression Models
Jerry A. Hausman, Haoyang Liu, Ye Luo, and Christopher Palmer
NBER Working Paper No. 25819
May 2019
JEL No. C19,C21,C31,I24,I26,J30

ABSTRACT

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

Jerry A. Hausman
Department of Economics, E52-518
MIT
50 Memorial Drive
Cambridge, MA 02142
and NBER
jhausman@mit.edu

Haoyang Liu
Federal Reserve Bank of New York
liuhy@berkeley.edu

Ye Luo
The University of Hong Kong
Faculty of Business and Economics
kurtluo@hku.hk

Christopher Palmer
MIT Sloan School of Management
100 Main Street, E62-639
Cambridge, MA 02142
and NBER
cjpalmer@mit.edu

1. Introduction

Economists are aware of problems arising from errors in variables (EIV) in regressors but generally ignore measurement error in the dependent variable. The EIV problem has received its most 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, however, the linear model results do not hold in nonlinear models.[1] In this paper, we study left-hand-side EIV in random-coefficients models, where even an additive disturbance uncorrelated with the regressors can create bias in estimating the outcome's conditional distribution. We focus on the consequences of measurement error in the dependent variable of linear conditional quantile models, a setting where we can achieve nonparametric identification even with some discrete covariates (in contrast to the generic random-coefficients model).[2] We propose a maximum-likelihood approach to consistently estimate the distributional effects of covariates in such a setting. While EIV in regressors usually require instrumental variables, we provide sufficient conditions for our estimator to identify the conditional distribution of the outcome without instrumenting. Our estimator has fractional polynomial of n convergence speed and asymptotic normality, permitting inference by bootstrapping.

Quantile regression (Koenker and Bassett, 1978) is the most widely used special case of heterogenous-effects random-coefficients models and has become a popular tool for applied microeconomists to consider the impact of covariates on the distribution of the dependent variable. As noted, a key benefit of the restrictions imposed by quantile regression on the general linear random-coefficients model is to accommodate non-continuous covariates, which cause the general random-coefficients model to become unidentified. However, in part because left-hand side variables in microeconometrics often come from self-reported survey data, the sensitivity of traditional quantile regression to dependent variable measurement error poses a serious problem to its validity.[3] Put another way, while omitted variables are problematic in the linear model insofar as they are correlated with the regressors, in quantile regression even unobserved heterogeneity independent of included covariates causes bias. In this sense, our results are applicable to settings with many covariates and unobserved

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

[2] Hausman (2001) observes that EIV in the dependent variable in quantile regression models generally leads to significant bias in contrast to the linear model intuition.

[3] For overviews of the econometric issues associated with measurement error in survey data, see Bound et al. (2001) and Meyer et al. (2015).

heterogeneity, such as the nonparametric estimation of a panel-data models with unobserved heterogeneity studied by Evdokimov (2010).

Intuitively, the estimated quantile regression line $x_i^T\widehat{\beta}(\tau)$ 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 estimates the likelihood that a given quantile-specific residual ($\varepsilon_{ij} := y_i - x_i^T\beta(\tau_j)$) 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 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 it is more likely that $u_i \approx \tau_j$. We show in simulations that a mixture of normals can accommodate a wide set of EIV distributions.[4] 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 that observation's residuals across all quantiles.

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

The rest of the paper proceeds as follows. In Section 2, we introduce our model specification and identification conditions. In Section 3, we introduce our estimator and characterize its properties. We present Monte Carlo simulation results in Section 4, and Section 5 contains our empirical application. Section 6 concludes.

We adopt the following notation. Define x to have dimension d_x and support $\mathcal{X}$. Let x_k denote the k^{th} dimension of x, and let x_{-k} denote the subvector of x corresponding to all but the k^{th} dimension of x. Define the space of y as $\mathcal{Y}$. Let $\xrightarrow{p}$ stand for convergence in probability. Let $f(\varepsilon|\sigma)$ be the p.d.f. of the EIV ε parametrized by σ where σ has dimension d_σ and domain Σ. We denote the true coefficient and measurement error distributional parameters as $\beta_0(\cdot)$ and σ_0, respectively. Define $||(\beta,\sigma)|| := \sqrt{\int_0^1 ||\beta(\tau)||_2^2 d\tau + ||\sigma||_2^2}$ as the

[4]See Burda et al. (2008, 2012) for other applications demonstrating the flexibility of a finite mixture of normals.

L^2 norm of (β_0,σ_0), where $||\cdot||_2$ is the usual Euclidean norm. Finally, we use the notation $x \precsim y$ for $x = O(y)$ and $x \precsim_p y$ for $x = O_p(y)$.

2. Model and Identification

Consider the general linear random-coefficients model as a framework to characterize unobserved heterogeneity in marginal effects

$$y_i = x_i^T \beta_i + \varepsilon_i, \tag{2.1}$$

where the covariates vector x_i is independent of the random coefficient vector β_i. This model is nonparametrically identified even in the presence of additive unobserved heterogeneity ε_i such that additional measurement error is isomorphic to any other form of independent unobserved heterogeneity and poses no immediate problem for bias. However, identification requires x_i to be continuously distributed and practical computation requires the dimension of x_i to be low to avoid the curse of dimensionality.

When at least some covariates are discrete (the most common situation when estimating treatment effects), a special case of (2.1) that permits nonparametric identification of heterogenous treatment effects is linear conditional quantile regression, which takes the form

$$y_i^* = x_i^T \beta_0(u_i), \tag{2.2}$$

where the unobserved heterogeneity in treatment effects enters as the scalar $u_i \sim U[0,1]$ representing the unobserved quantile of y_i in the conditional distribution of $y_i|x_i$.[5] In this model, the τ^{th} conditional quantile of the dependent variable y^* is a linear function of x

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

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

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

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

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

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

[5]Here we study the linear conditional quantile model, as is ubiquitous in practice. While the conditional quantile model is identified for linear and many nonlinear specifications, it is not nonparametrically identified (Horowitz and Lee, 2005). Note that our results will allow for polynomials in x_i, somewhat relaxing the linearity assumption.

the minimization problem in the usual quantile regression

$$\beta(\tau) \in \arg\min_b E[\rho_\tau(y - x^T b)], \tag{2.3}$$

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, y^* is observed and the FOC is

$$E[x(\tau - 1(y^* < x^T\beta(\tau)))] = 0, \tag{2.4}$$

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

$$E[x(\tau - 1(y^* + \varepsilon < x^T\beta(\tau)))] = 0. \tag{2.5}$$

In Appendix A, we demonstrate the bias of bivariate quantile regression, showing that coefficient estimates are biased inwards from their minimum and maximum levels over τ, which we refer to as compression bias.[6] For intuition, note that 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 (2.4) even in the case where ε is symmetrically distributed because of the asymmetry of the check function. Observations for which $y^* \geq x^T\beta(\tau)$ and should therefore be weighted by τ in the minimization problem may end up on the left-hand side of the check function, receiving a weight of $(1-\tau)$ such that equal-sized differences on either side of zero do not cancel each other out. Note that for median regression, $\rho_{.5}(\cdot)$ 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 (2.5) holds for the true $\beta_0(0.5)$.

A Monte-Carlo simulation shows the degree of bias in a two-factor model with random disturbances in the dependent variable y to illustrate the direction and magnitude of measurement error bias.

Example 1. We consider a data-generating process

$$y_i = \beta_1(u_i) + x_{2i}\beta_2(u_i) + x_{3i}\beta_3(u_i) + \varepsilon_i$$

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 $\beta_1(\tau) = \tau$, $\beta_2(\tau) = \exp(\tau)$, and $\beta_3(\tau) = \sqrt{\tau}$. The variables x_2 and x_3 are drawn from independent lognormal distributions $LN(0, 1)$. The number of observations is 100,000.

[6] See also Arellano and Weidner (2016), who find that estimation error in the fixed effects can create bias in the quantile-regression estimate of the slope coefficients, essentially understating the degree of heterogeneity by smoothing across quantiles. However, their setup does not permit them to characterize the direction of the bias.

TABLE 1. Monte-Carlo Results: Mean Bias

Parameter	EIV Distribution	Quantile (τ) 0.1	0.25	0.5	0.75	0.9
$\beta_2(\tau) = e^\tau$	$\varepsilon = 0$	0.000	0.000	-0.000	-0.000	0.000
	$\varepsilon \sim \mathcal{N}(0, 4)$	0.156	0.126	0.027	-0.117	-0.215
	$\varepsilon \sim \mathcal{N}(0, 16)$	0.262	0.214	0.042	-0.200	-0.353
	True parameter:	1.105	1.284	1.649	2.117	2.460
$\beta_3(\tau) = \sqrt{\tau}$	$\varepsilon = 0$	0.000	-0.000	0.000	0.000	0.000
	$\varepsilon \sim \mathcal{N}(0, 4)$	0.125	0.053	-0.021	-0.069	-0.086
	$\varepsilon \sim \mathcal{N}(0, 16)$	0.196	0.091	-0.030	-0.112	-0.141
	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_2, and x_3 where $y = \beta_1(\tau) + x_2\beta_2(\tau) + x_3\beta_3(\tau) + \varepsilon$ 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_2 and x_3 are i.i.d. draws from $LN(0, 1)$. $N = 100,000$.

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 as the variance of the EIV becomes larger.

2.1. **Identification and Regularity Conditions.** To establish the nonparametric identification of our model, we require the following two assumptions.

Assumption 1 (Properties of $\beta(\cdot)$)**.** *We assume the following properties on the coefficient vectors $\beta(\tau)$.*

(1) *$\beta(\tau)$ is in the space $M[B_1 \times B_2 \times B_3 \times ... \times B_{d_x}]$ where the functional space M is defined as the collection of all functions $b = (b_1, ..., b_{d_x}) : [0, 1] \to [B_1 \times ... \times B_{d_x}]$ with $B_k \subset \mathbb{R}$ being a closed interval $\forall\, k \in \{1, ..., d_x\}$ such that $x^T b(\tau) : [0, 1] \to \mathbb{R}$ is monotonically increasing in τ for all $x \in \mathcal{X}$.*
(2) *The true parameter β_0 is a vector of C^2 functions with first-order derivatives bounded from above by a positive constant.*

Monotonicity of $x^T\beta(\cdot)$ is a key assumption in quantile regression and important for identification because in the log-likelihood function, $f(y|x) = \int_0^1 f(y - x^T\beta(u))du$ is invariant to a rearrangement of the function $\beta(u)$.[7] The function $\beta(\cdot)$ is therefore unidentified if we do not impose further restrictions. However, 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 rearrangement if $x^T\beta(\cdot)$ is monotonic in τ.

Assumption 2 (Properties of x)**.** *We assume the following properties of the vectors x that comprise the design matrix X.*

(1) *$E[x'x]$ is non-singular.*
(2) *There is at least one dimension x_1 of x such that $x_1|x_{-1}$ is continuously distributed, and the element of $\beta_0(\cdot)$ corresponding to x_1, denoted as $\beta_{0,1}(\tau)$, does not have any point mass in its probability distribution.*

The above conditions on the parameters and covariate matrix allow us to state our main nonparametric identification result.

Theorem 1 (Nonparametric Global Identification)**.** *Assume that Assumptions 1 and 2 hold and that the PDFs of ε, $f(\cdot)$ and $f_0(\cdot)$, are continuously differentiable functions such that*

(1) *$\int_{-\infty}^{\infty} \varepsilon f(\varepsilon)d\varepsilon = 0$, $\int_{-\infty}^{\infty} \varepsilon f_0(\varepsilon)d\varepsilon = 0$, and*
(2) *$\int_{-\infty}^{\infty} \varepsilon^2 f(\varepsilon)d\varepsilon < C$, $\int_{-\infty}^{\infty} \varepsilon^2 f_0(\varepsilon)d\varepsilon < C$ for some constant C.*

Then, for any $\beta(\cdot)$ and $f(\cdot)$ which generate the same density of $y|x$ almost everywhere as the true function $\beta_0(\cdot)$ and $f_0(\cdot)$, it must be that $\beta(\tau) = \beta_0(\tau)$ almost everywhere for all $\tau \in [0,1]$ and $f(\varepsilon) = f_0(\varepsilon)$ almost everywhere for all $\varepsilon \in \mathbb{R}$.

Proof. See Appendix G.1. □

Although the above identification result allows x_{-1} to enter into $x^T\beta(\cdot)$ in an unrestricted fashion, Theorem 1 holds under the presence of a continuously distributed x_1 that enters x linearly. To illustrate that nonlinear functions of x_1 are admissible, the following lemma establishes nonparametric identification when polynomials of arbitrarily high degree of x_1 are also included in x. Whenever the contribution of x_1 can be approximated by finite polynomials, nonparametric identification holds. Before stating the lemma, we restate Assumption 2 to allow for polynomials of x_1.

Assumption 3 (Properties of x allowing for polynomials of x_1)**.** *We assume the following properties of the vectors x that comprise the design matrix X.*

(1) *$E[x'x]$ is non-singular.*

[7] Note that the monotonicity assumption in Assumption 1 also requires that if $x \in \mathcal{X}$ then $-x \notin \mathcal{X}$. In practice, many quantile models assume that $x \geq 0$.

(2) *We can partition* $x = (W(x_1), x_{-w})^T$ *where* x_1 *is one dimensional,* $W(x_1) = (x_1, x_1^2, \cdots, x_1^p)^T$ *for some* p*,* $x_1|x_{-w}$ *is continuously distributed, and the element of* $\beta_0(\cdot)$ *corresponding to* x_1^p *, denoted as* $\beta_{0,x_1^p}(\tau)$*, does not have any point mass in its probability distribution.*

(3) $\beta_{0,x_1^p}(\tau)$ *has continuous and bounded derivatives with respect to* τ *for all* $\tau \in (0,1)$*.*

Lemma 1 (Nonparametric Identification with Higher-order Polynomials)**.** *Assume that Assumptions 1 and 3 hold and that the PDFs of* ε*,* $f(\cdot)$ *and* $f_0(\cdot)$*, are continuously differentiable functions such that*

(1) $\int_{-\infty}^{\infty} \varepsilon f(\varepsilon) d\varepsilon = 0$*,* $\int_{-\infty}^{\infty} \varepsilon f_0(\varepsilon) d\varepsilon = 0$*, and*

(2) $\int_{-\infty}^{\infty} \varepsilon^2 f(\varepsilon) d\varepsilon < C$*,* $\int_{-\infty}^{\infty} \varepsilon^2 f_0(\varepsilon) d\varepsilon < C$ *for some constant* C*.*

Then, for any $\beta(\cdot)$ *and* $f(\cdot)$ *which generate the same density of* $y|x$ *almost everywhere as the true function* $\beta_0(\cdot)$ *and* $f_0(\cdot)$*, it must be that* $\beta(\tau) = \beta_0(\tau)$ *almost everywhere for all* $\tau \in [0,1]$ *and* $f(\varepsilon) = f_0(\varepsilon)$ *almost everywhere for all* $\varepsilon \in \mathbb{R}$*.*

Proof. See Appendix G.1. □

3. Estimation

In this section, we first demonstrate the consistency of the ML estimator, which we then operationalize with a sieve-ML estimator, establishing its consistency and asymptotic normality. In addition, we extend Chen and Pouzo (2013) to provide sufficient conditions for inference by bootstrapping in our setting. While Theorem 1 and Lemma 1 establish identification even when the distribution of ε is nonparametric, for estimation, we require the following assumptions on the properties of the measurement error ε.[8]

Assumption 4 (Properties of EIV)**.** *The probability density function of the EIV is parametrized as* $f(\varepsilon|\sigma)$*, and the true density is abbreviated* $f_0(\varepsilon) := f(\varepsilon|\sigma_0)$*.*

(1) *The domain of the parameter* σ *is a compact space* Σ*, and the true value* σ_0 *is in the interior of* Σ*.*

(2) $f(\varepsilon|\sigma)$ *is twice differentiable in* ε *and* σ *with bounded derivatives up to the second order.*

(3) *For all* $\sigma \in \Sigma$*, there exists a uniform constant* $\bar{C} > 0$ *such that* $E[|\log f(\varepsilon|\sigma)|] < \bar{C}$*. Moreover,* $f(\cdot|\sigma)$ *is non-zero all over the entire space* $\mathbb{R}$ *and bounded from above uniformly.*

(4) $E[\varepsilon] = \int_{-\infty}^{\infty} \varepsilon f(\varepsilon|\sigma) = 0$*.*

(5) *For any* $\sigma \in \Sigma$*,* $l > 0$*, and some constant* $C_l > 0$*,* $\int_{-l}^{l} |\phi_\varepsilon(s) - \phi_{\varepsilon_0}(s)|^2 ds \geq C_l ||\sigma - \sigma_0||_2^2$*, where* $\phi_\varepsilon(s) := \int_{-\infty}^{\infty} \exp(is\varepsilon) f(\varepsilon|\sigma) d\varepsilon$ *is the characteristic function of* ε *given PDF* $f(\varepsilon|\sigma)$*.*

[8]While Assumption 4 requires knowing the distribution of the EIV up to a finite set of parameters, we show in simulations below that when the distribution of the EIV is unknown, a mixture of normals is sufficiently flexible to approximate a wide range of potential distributions.

Note that Assumption 4 holds for all mean-zero distributions in the exponential family.

Given this parameterization of $f(\cdot|\sigma)$, we define our log likelihood function as follows. Denote $\theta := (\beta(\cdot), \sigma) \in \Theta$. For any θ, define the expected log-likelihood function $L(\theta)$ as

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

with the empirical log likelihood being denoted

$$L_n(\theta) = E_n[\log g(y|x, \theta)], \tag{3.2}$$

where E_n is the empirical average operator $E_n h(x) := \frac{1}{n}\sum_{i=1}^n h(x_i)$.

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

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

Then the ML estimator is defined as

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

where $g(\cdot|\cdot, \cdot, \cdot)$ is the conditional density of y given x and parameters, as defined in equation (3.3). The following theorem states the consistency property of the ML estimator.

Lemma 2 (MLE Consistency)**.** *Under Assumptions 1, 3, and 4, the maximum-likelihood estimator defined by* (3.4) *exists and converges in probability to the true parameter* $(\beta_0(\cdot), \sigma_0)$ *under the* L^2 *norm in the functional space* M *and Euclidean norm in* Σ*.*

Proof. See Appendix G.2. □

The consistency theorem is a special version of a general MLE consistency theorem (Van der Vaart, 2000). Two conditions play critical roles here: the monotonicity of $x^T\beta(\cdot)$ for all $x \in \mathcal{X}$ and the local continuity of at least one right-hand side variable. If monotonicity fails, we lose compactness of the parameter space Θ and the consistency argument will fail.

3.1. **Sieve Maximum Likelihood Estimation.** While we have demonstrated that the maximum likelihood estimator restricted to parameter space Θ converges to the true parameter with probability approaching 1, the estimator still lives in a large space with $\beta(\cdot)$ being d_x-dimensional functions such that $x^T\beta(\cdot)$ is monotonic 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. Here, we consider a spline estimator of $\beta(\cdot)$ for their computational advantages in calculating the sieve estimator. The estimator below is easily adapted to the reader's preferred estimator. We use a piecewise-spline sieve space, which we define as follows.

Definition 1 (Sieve Space)**.** Define $\Theta_J^r := \Omega_J^r \times \Sigma$ to denote the sieve-ML parameter space, where Ω_J^r stands for the space of r^{th}-order spline functions with J knots on $[0,1]$ such that $x^T\beta(\tau)$ is monotonically increasing in $x \in \mathcal{X}$ for all $\beta(\cdot) \in \Omega_J^r$ and elements in Ω_J^r are bounded above as in Assumption 1.

For example, for any $\beta(\cdot) \in \Omega_J^1$, $\beta_k(\cdot)$ is a piecewise linear function on a set of intervals covering $[0,1]$ and $k = 1, \ldots, d_x$. Such a definition allows Ω_J^r to cover a dense set in $M[B_1 \times B_2 \times B_3 \times ... \times B_{d_x}]$ as J grows to infinity with sample size.

The space Ω_J^r can therefore be written as the collection of functions $\beta(\tau)$ such that $\beta(\tau) := \sum_{l=1}^r b_l\tau^l + \sum_{j=1}^J b_{j+r}\left(\max\{\tau - t_j, 0\}\right)^r = \sum_{l=1}^{r+J} b_l S_l(\tau)$ where t_j is the j^{th} knot, $S_l(\tau)$ and b_l with $l = 1, 2, ..., r+J$ are the spline functions and their coefficients.. In general, the L^2 distance of the space Θ_J^r to the true parameter θ_0 satisfies $d_2(\theta_0, \Theta_J^r) \leq CJ_n^{-r-1}$ for some generic constant C (Chen, 2007). It is easy to see that $\Theta_J^r \subset \Theta$.

The sieve estimator is defined as follows.

Definition 2 (Sieve Estimator)**.**

$$\widehat{\theta}_J = (\widehat{\beta}_J(\cdot), \widehat{\sigma}) = \arg\max_{\theta \in \Theta_{J_n}^r} E_n[\log g(y|x, \beta, \sigma)] \tag{3.5}$$

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

The following lemma establishes the consistency of the sieve estimator.

Lemma 3 (Sieve Estimator Consistency)**.** *If Assumptions 1, 3, and 4 hold, $J_n \to \infty$, and $J_n/n \to 0$, then the sieve estimator defined in* (3.5) *is consistent.*

Proof. See Appendix G.2. □

Our objective is to show that $\widehat{\beta}_J$ will converge to β_0 with certain speed. Doing so requires a definition of the parametric score evaluated at a functional $\beta(\cdot)$. Let the Hadamard derivative of g with respect to β in the directions of $S_1(\tau), ..., S_{J+r}(\tau)$ and evaluated at $\tilde{\beta}$ and $\tilde{\sigma}$ be defined as

$$\left.\frac{\partial g}{\partial \beta}\right|_{\tilde{\beta},\tilde{\sigma}} := \left(\int_0^1 f'(y - x^T\beta(\tau)|\sigma)S_1(\tau)d\tau, ..., \int_0^1 f'(y - x^T\beta(\tau)|\sigma)S_{J+r}(\tau)d\tau\right).$$

Note that for a $(\beta_J, \sigma) \in \Theta_J^r$, $\left.\frac{\partial g}{\partial \beta}\right|_{\beta_J,\sigma} = \left[\frac{\partial g}{\partial b_1}, \ldots, \frac{\partial g}{\partial b_{J+r}}\right]$, where $b_1 \ldots, b_{J+r}$ are the coefficients for $S_1(\tau), ..., S_{J+r}(\tau)$ in $\beta_J(\tau)$. We also define the information matrix evaluated at $(\widetilde{\beta}, \widetilde{\sigma})$ as

$$\begin{aligned} I_{\tilde{\beta},\tilde{\sigma}} &:= E\left[\left(\frac{\partial \log(g)}{\partial \beta}, \frac{\partial \log(g)}{\partial \sigma}\right)\left(\frac{\partial \log(g)}{\partial \beta}, \frac{\partial \log(g)}{\partial \sigma}\right)'\right]\Bigg|_{\tilde{\beta},\tilde{\sigma}} \\ &= E\left[\left(\frac{\frac{\partial g}{\partial \beta}, \frac{\partial g}{\partial \sigma}}{g}\right)\left(\frac{\frac{\partial g}{\partial \beta}, \frac{\partial g}{\partial \sigma}}{g}\right)'\right]\Bigg|_{\tilde{\beta},\tilde{\sigma}} \end{aligned}$$

When J goes to infinity, the smallest eigenvalue of $I(\beta_0, \sigma_0)$ goes to 0, leading to an ill-posedness problem. Intuitively, as we are trying to estimate $\beta(\cdot)$ and σ via MLE from the mixture distribution of $y = x^T\beta(\tau) + \varepsilon$, where $\tau \sim U[0,1]$ and $\varepsilon \sim f(\cdot|\sigma_0)$, the estimation of $\beta(\cdot)$ is ill-posed. However, the curse of dimensionality in β is not at play because $x^T\beta(\cdot)$ is a monotone function of a single random variable τ. We will adopt the following measure of ill-posedness.

Assumption 5 (Ill-posed Measure)**.** *Define mineigen(I) as the minimum eigenvalue for a given matrix I. Let one of the following two assumptions on the degree of ill-posedness hold*

(1) *Mild ill-posedness: mineigen$(I_{\beta,\sigma}) \geq C/J^\lambda$ for some $\lambda > 0$ and constant $C > 0$, for all $(\beta_0, \sigma_0) \in \Theta$.*
(2) *Severe ill-posedness: mineigen$(I_{\beta,\sigma}) \geq C\exp(-\lambda J)$ for some $\lambda > 0$ and constant $C > 0$, and all $(\beta_0, \sigma_0) \in \Theta$.*

These ill-posed measures are closely related to the smoothness of the PDF of the EIV (Fan, 1991). The normal distribution is severely ill-posed with $\lambda = 2$, and the Laplace distribution is mildly ill-posed with $\lambda = 1$. Unlike the usual sieve estimation problem, our problem is ill-posed with minimum eigenvalue decaying at speed J^λ under mild ill-posedness of degree λ. When the PDF of the EIV is super smooth, the problem becomes severely ill-posed. While convergence to normality will be too slow for our bootstrap results to hold, consistency still holds under super smoothness. However, we note that mild ill-posedness will be satisfied under even minor perturbations from super smoothness. In such a case, we could use a sieve mixture of non-smooth PDFs to approximate a smooth PDF and reduce the ill-posedness of the problem, a point we leave to future research. We establish consistency and the convergence rate under severe ill-posedness in Theorem 3 below.

A sufficient condition for mild ill-posedness is the following discontinuity assumption on f—see also An and Hu (2012).[9]

Assumption 6 (Discontinuity of f)**.** *There exists a positive integer λ such that $f \in C^{\lambda-1}(\mathbb{R})$, and the λ^{th} order derivative of f equals*

$$f^{(\lambda)}(x) = h(x) + c_\delta \delta(x - a), \tag{3.6}$$

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

[9]See Lemma 8 in Appendix G.2 for a formal statement and proof of this result for the special case of a piecewise-constant sieve, showing that if a function is of the class C^λ, the minimum eigenvalue of I is of order $O(J^{-\lambda})$ as $J \to \infty$ for $\lambda \in \mathbb{Z}^+$. In general, for smooth functions $f(\cdot)$, the minimum eigenvalue of I will decay with speed $O(\exp(-J^{-a}))$ for some $a > 0$.

The following final assumption on the characteristic function significantly simplifies our proof of the convergence rate of the distributional parameters. It holds whenever there exists enough variation in x such that the characteristic function is non-constant around x.

Assumption 7 (Variation on Characteristic Function)**.** *Let $\phi_{x\beta}(s|x)$ denote the characteristic function of $x^T\beta$ conditional on x. Suppose there exists a local neighborhood $N \subset \mathcal{X}$ such that there exists a constant $c > 0$ and for any $(\beta, \sigma) \in \Theta$ and any $s \in [-l, l]$,*

$$Var_{x\in N}\left(\frac{\phi_{x\beta}(s|x)}{\phi_{x\beta_0}(s|x)}\right) \geq cE_{x\in N}\left[\left|\frac{\phi_{x\beta}(s|x) - \phi_{x\beta_0}(s|x)}{\phi_{x\beta_0}(s|x)}\right|^2\right]$$

where $Var_{x\in N}$ and $E_{x\in N}$ denote the variance and expectation operators evaluated over all x in a neighborhood N.

In the lemma below, we use the stochastic equicontinuity of the log likelihood function to establish key facts about the convergence rate of $\widehat{\sigma}$, including that it converges to σ_0 at rate $n^{-\frac{1}{4}}$.

Lemma 4 (Convergence Rate of $\widehat{\sigma}$)**.** *If Assumptions 1, 3, and 7 hold and $J_n^{2r+2}/n \to \infty$, the sieve estimator $(\widehat{\beta}_J(\cdot), \widehat{\sigma})$ has the following property:*

$$\widehat{\sigma} - \sigma_0 = o_p(n^{-\frac{1}{4}}). \tag{3.7}$$

Moreover, defining $\delta := ||\widehat{\beta}_J - \beta_J^*||$, then

$$||\widehat{\sigma} - \sigma_0||^2 = O_p\left(\max\left(\frac{\log n}{n}, \frac{\delta\sqrt{-\log\delta}}{\sqrt{n}}\right)\right) \tag{3.8}$$

Proof. See Appendix G.2. □

For EIV distributions that are mildly ill-posed, we require that the sieve grid J_n grow quickly enough to overcome the bias but slowly enough to overcome the ill-posed problem, as we formalize in the following theorem.

Theorem 2 (Sieve Estimator Asymptotic Normality)**.** *Let Assumptions 1, 3, 5.1 (the mild ill-posed case), and 7 hold. Further, let the number of knots J_n satisfy $J_n^{4\lambda^2+6\lambda}\log(n)/n \to 0$ and $J_n^{2r+2}/n \to \infty$ as $n \to \infty$. Then*

$$\left|\left|\widehat{\beta}_J - \beta_0, \widehat{\sigma} - \sigma_0\right|\right| = O_p\left(\frac{1}{J_n^{r+1}}\right) = J_n^{\lambda}O_p\left(\frac{1}{J_n^{r+1}}, \frac{1}{\sqrt{n}}\right).$$

Moreover, there exists a sequence $\kappa_J \geq \frac{C}{J_n^{\lambda}}$ for some generic constant $C > 0$ such that for any fixed τ

$$\sqrt{n\kappa_J}\big(\widehat{\beta}_J(\tau) - \beta_0(\tau)\big) \xrightarrow{d} \mathcal{N}(0, \Omega_{J,\tau}),$$

where $\Omega_{J,\tau}$ is a sequence of positive definite matrices with the largest eigenvalue bounded by a constant, and

$$\sqrt{n\kappa_J}\big(\widehat{\sigma}_J - \sigma_0\big) \xrightarrow{d} \mathcal{N}(0, \Omega_{J,\sigma}),$$

where $\Omega_{J,\sigma}$ is a sequence of positive definite matrices with the largest eigenvalue bounded by a constant.

Proof. See Appendix G.2. □

The smoothness of the mapping from the data to the estimator $\beta(\cdot)$ helps with robustness to mild forms of misspecification. Following same proof as for Theorem 2 above, misspecification would produce a second residual term in addition to the stochastic term. Using this smoothness along with the capacity of our estimator to accommodate additional polynomial terms, the approximation provided by the sieve estimator would still approach the truth asymptotically.

As discussed above, while asymptotic normality need not hold under the severe ill-posed case, the following theorem establishes the convergence rate of the sieve estimator under severe ill-posedness.

Theorem 3 (Severe Ill-posedness Sieve Estimator Convergence Rate)**.** *Let J_n be a sequence of positive numbers such that $\frac{\exp(\lambda J_n)}{\sqrt{n}} = \frac{1}{J_n}$. Then under Assumptions 1, 3, 4, 5.2 (the severe ill-posed case), and 7, the sieve estimator β_{J_n} satisfies*

$$||\widehat{\beta}_{J_n} - \beta_0|| \precsim_p \frac{1}{\log(n)}.$$

Proof. See Appendix G.2. □

3.2. **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 pairs 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

$$\theta^b = \arg\max_{\theta \in \Theta_J} E_n^b[\log g^b(y_i^b | x_i^b, \theta)], \tag{3.9}$$

where 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 of each coefficients $\left\{\beta_k^b(\tau_j)\right\}_{b=1}^B$ as $\widehat{\beta}_k(\tau_j) \pm z_{1-\alpha/2} \cdot \widehat{\sigma}_{jk}$ where the critical value $z_{1-\alpha/2} \approx 1.96$ for significance level of $\alpha = .05$ and $\widehat{\sigma}_{jk}$ is the standard deviation of the bootstrapped estimates of $\beta_k(\tau_j)$.

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

Lemma 5 (Validity of the Bootstrap)**.** *As in Theorem 2, let Assumptions 1, 3, 5.1 (the mild ill-posed case), and 7 hold, and let the number of knots J_n satisfy $J_n^{4\lambda^2+6\lambda}\log(n)/n \to 0$ and $J_n^{2r+2}/n \to \infty$ as $n \to \infty$. Then $\sqrt{n\kappa_J}(\widehat{\beta}_J^b - \widehat{\beta}_J)$ and $\sqrt{n\kappa_J}(\widehat{\sigma}^b - \widehat{\sigma})$ have the same distribution as $\sqrt{n\kappa_J}(\widehat{\beta}_J - \beta_0)$ and $\sqrt{n\kappa_J}(\widehat{\sigma} - \sigma_0)$, respectively.*

Proof. See Appendix G.2. □

4. Monte-Carlo Simulations

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

$$y_i = \beta_1(u_i) + x_{2i}\beta_2(u_i) + x_{3i}\beta_3(u_i) + \varepsilon_i$$

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_{2i}, $x_{3i} \sim LN(0,1)$. The coefficient vector is a function of the conditional quantile u_i of individual i

$$\begin{pmatrix} \beta_1(u) \\ \beta_2(u) \\ \beta_3(u) \end{pmatrix} = \begin{pmatrix} 1 + 3u - u^2 \\ \exp(u) \\ \sqrt{u} \end{pmatrix}.$$

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

$$\varepsilon_i \sim \begin{cases} \mathcal{N}(-3,1) & \text{with probability } 0.5 \\ \mathcal{N}(2,1) & \text{with probability } 0.25 \\ \mathcal{N}(4,1) & \text{with probability } 0.25. \end{cases} \tag{4.1}$$

To simulate robustness to real-world settings in which the econometrician does not know the true distribution of the residuals, we also present results simulating measurement error from alternative distributions and test how well quasi-MLE modeling the error distribution as a Gaussian mixture accommodates misspecification in F_ε.[10] We use a genetic-algorithm optimizer to find the maximizer of the log-likelihood function defined in Section 3 with

[10]In Appendix B, we examine the performance of a weighted least squares EM algorithm when the measurement error is normally distributed. While estimating a mixture model allows for an arbitrary amount of measurement-error distributional flexibility by increasing the number of mixture components, we are also interested in more parsimonious specifications that may be computationally attractive to applied researchers willing to make parametric assumptions on the data-generating process. As discussed in Appendix B, if the measurement error is normally distributed, the estimand reduces to a weighted-least squares objective function, similar to how linear MLE is equivalent to OLS in the case of normally distributed stochastic terms.

start values provided by a gradient-based constrained optimizer. 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.

In Figures 1-3, we plot the true coefficient function defined above, average coefficients from quantile regression, and our MLE estimator and associated 95% confidence intervals using a sieve for $\beta(\cdot)$ consisting of 15 knots. The plotted confidence intervals are calculated pointwise as $\widehat{\beta}(\tau_j) \pm 1.96\widehat{\sigma}_j$, where $\widehat{\sigma}_j$ is the standard deviation across simulations of parameter estimates $\widehat{\beta}(\tau_j)$.[11] To test the validity of our bootstrapped confidence intervals, we further calculated bootstrap confidence intervals for each simulation using the procedure described in section 3.2 and calculated the fraction of simulations for which the true parameter lied within the bootstrapped confidence interval. We found that our confidence intervals had a coverage of 98%, suggesting them to be slightly conservative on average.[12] Focusing on Figures 2 and 3 that plot estimates of the slope coefficients $\beta_2(\cdot)$ and $\beta_3(\cdot)$, quantile regression estimates are badly biased, with lower quantiles biased upwards and upper quantiles biased downwards. In contrast, the ML estimates fall almost directly on top of the true parameter functional, and the bias of the ML estimator is statistically indistinguishable from zero at all quantiles. The average absolute bias for the ML estimates is 0.6% and 1.5% of the true coefficients for $\beta_2(\cdot)$ and $\beta_3(\cdot)$ respectively, and always less than 4% of the true magnitude. By contrast, the mean bias of the quantile regression coefficients is 12% and 22% for the two slope coefficients and exceeds 100% for some quantiles. Appendix Table C1 confirms that the quantile-regression average absolute bias is 26 and 16 times larger than the MLE bias for $\beta_2(\cdot)$ and $\beta_3(\cdot)$, respectively. Appendix Table C1 further reports MSE results, showing that the average MSE is an order of magnitude smaller for the ML estimates than the quantile-regression estimates.

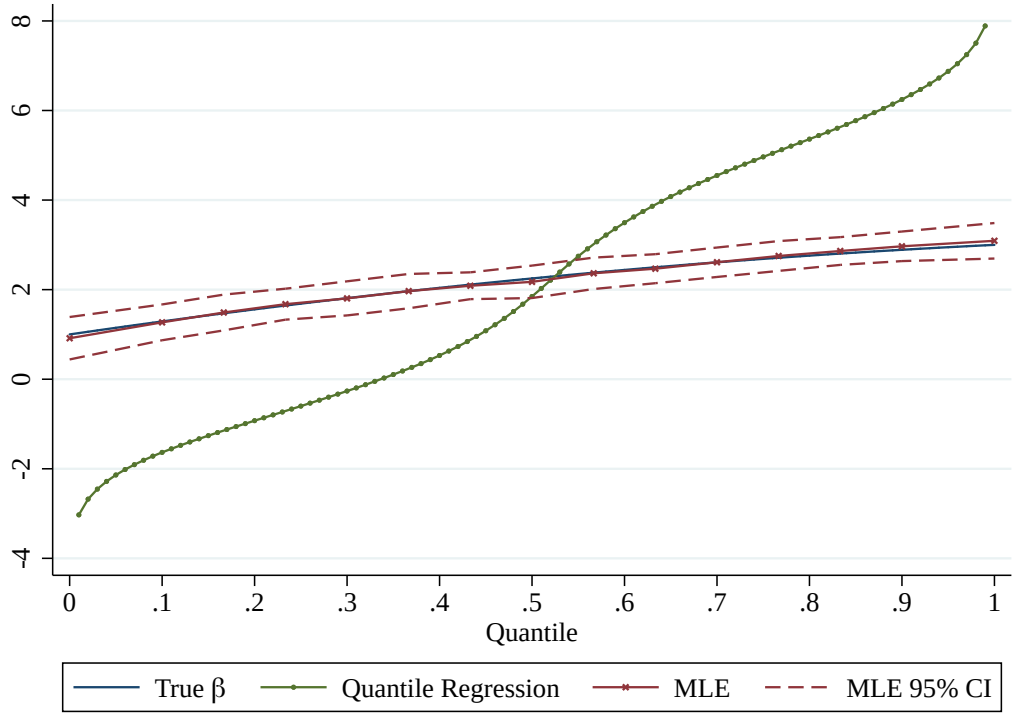
Figure 1 plots estimates of the intercept term, showing that quantile-regression estimates of $\beta_1(\cdot)$ are badly biased. Given that quantile-regression estimated intercepts ensure that the τ^{th} conditional quantile of the residuals $Q_{\hat{\varepsilon}}(\tau) = 0$, when the slope coefficients are biased, this exacerbates the bias in the constant function. Whereas the mean absolute bias of the ML estimates of $\beta_1(\cdot)$ is 2% of the true magnitude, quantile regression has a mean absolute bias of 116% of the true $\beta_1(\cdot)$ functional.

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

[11] We estimate the critical value for simultaneous confidence intervals to be 2.92, roughly 50% wider than pointwise confidence intervals.

[12] To further demonstrate the performance of the bootstrapped confidence intervals (and in particular the asymptotic results in Theorem 2), we varied the sample size in the simulations, holding the number of knots fixed, and calculated how the width of the pointwise 95% confidence intervals changed. Decreasing the sample size from 50,000 to 10,000 observations—a decrease in $\sqrt{n}$ by a factor of 2.24—increased the width of the confidence intervals for both β_1 and β_2 (averaged across quantiles) by a factor of 2.25.

FIGURE 1. Monte Carlo Simulation Results: $\widehat{\beta}_1(\tau)$

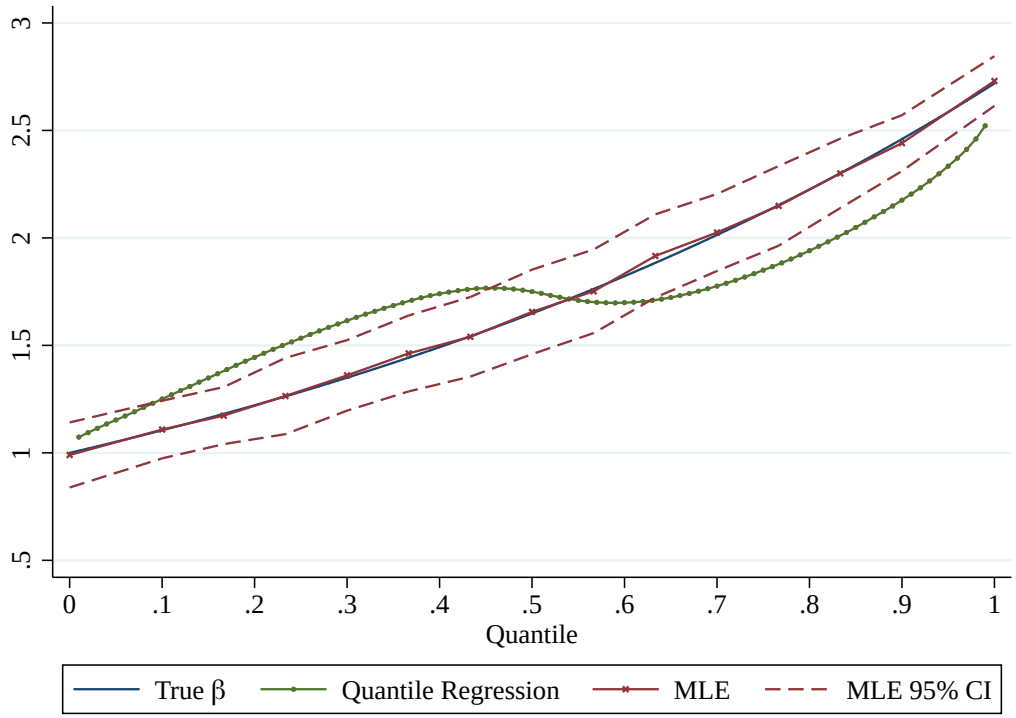


Notes: Figure plots the true $\beta_1(\tau) = 1 + 3\tau - \tau^2$ (blue line) against quantile-regression estimates (green circles), bias-corrected MLE (red xs), and 95% confidence intervals for the MLE estimates (dashed red lines) from 100 MC simulations using the data-generating process described in the text with the measurement error generated as a mixture of three normals.

from the average estimated distributional parameters across all MC simulations (solid red line). The 95% confidence interval of the estimated density (dotted green line) are estimated pointwise as the 2.5th and 97.5th percentiles of EIV densities across all simulations. Despite the bimodal nature of the true measurement error distribution, our algorithm captures the overall features of true distribution well, with the true density always within the 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 as a t distribution with three degrees of freedom (normalized to have the same variance as in (4.1)) in panel I or as a Laplace (with $\lambda = 2.29$ to again have the same variance across ε DGPs) in panel II, the ML estimates that model the EIV distribution as a mixture of three normals still significantly outperform quantile regression. As expected,

FIGURE 2. Monte Carlo Simulation Results: $\widehat{\beta}_2(\tau)$



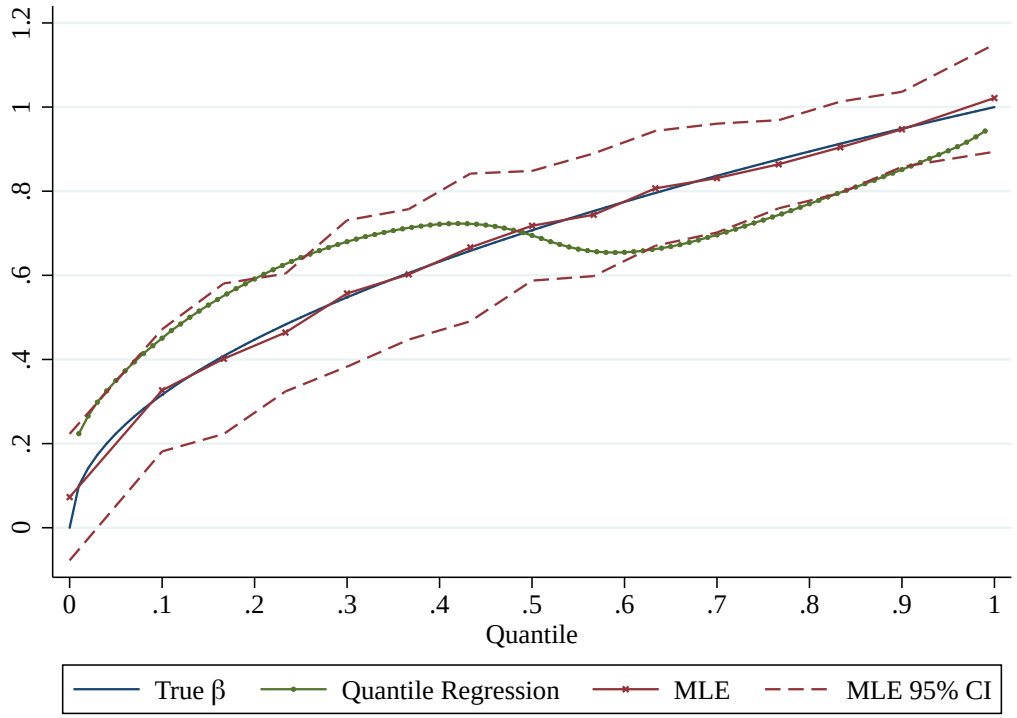
Notes: Figure plots the true $\beta_2(\tau) = \exp(\tau)$ (blue line) against quantile-regression estimates (green circles), bias-corrected MLE (red xs), and 95% confidence intervals for the MLE estimates (dashed red lines) from 100 MC simulations using the data-generating process described in the text with the measurement error generated as a mixture of three normals.

quantile regression is again biased 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. Appendix C provides additional simulation results—including both mean bias and MSE—for alternative measurement error distributions, when x_2 is binary, and when $\beta(\cdot)$ is estimated using a finer sieve space (99 knots).

5. Empirical Application

To illustrate the use of our estimator in practice, we examine distributional heterogeneity in the wage returns to education. First, we estimate the quantile-regression analog of a

FIGURE 3. Monte Carlo Simulation Results: $\widehat{\beta}_3(\tau)$



Notes: Figure plots the true $\beta_3(\tau) = \sqrt{\tau}$ (blue line) against quantile-regression estimates (green circles), bias-corrected MLE (red xs), and 95% confidence intervals for the MLE estimates (dashed red lines) from 100 MC simulations using the data-generating process described in the text with the measurement error generated as a mixture of three normals.

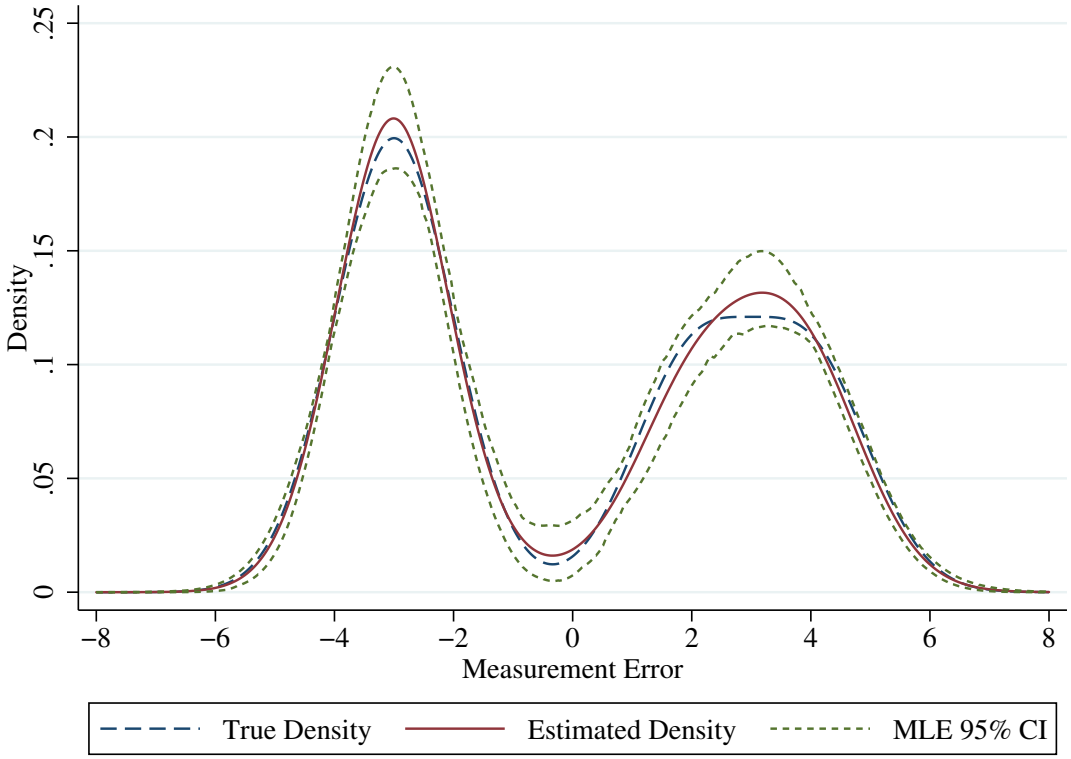
Mincer regression, replicating and extending results from Angrist et al. (2006)

$$Q_{y|x}(\tau) = \beta_1(\tau) + \beta_2(\tau)education_i + \beta_3(\tau)experience_i + \beta(\tau)experience_i^2 \tag{5.1}$$

where $Q_{y|x}(\tau)$ is the τ^{th} quantile of the conditional (on the covariates x) log-wage distribution, and the education and experience variables are measured in years. 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 $\beta(\cdot)$.

Appendix Figure E1 plots quantile-regression estimates of equation (5.1) using census microdata samples from four decennial census years: 1980, 1990, 2000, and 2010, along with

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 100 Monte Carlo simulations (solid red line). For each grid point, the dotted green line plots the 2.5th and 97.5th percentile of the EIV density function across all MC simulations.

pointwise confidence intervals.[13] Consistent with Angrist et al. (2006), we find quantile-regression evidence that heterogeneity in the returns to education across the conditional wage distribution has increased over time. Adding 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 than 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

[13]For further details on the data including summary statistics, see Appendix D. For comparability with Angrist et al. (2006) and to have a sufficient number of observations to run our estimator, we focus on prime age white males (aged 40-49). In Appendix D, we provide evidence that other demographic groups have markedly different patterns of heterogeneity in the education-wage gradient across the conditional income distribution, motivating further study on treatment effect heterogeneity.

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

	I. $\varepsilon \sim t$				*II.* $\varepsilon \sim$ *Laplace*			
	β_2		β_3		β_2		β_3	
Quantile	QR	MLE	QR	MLE	QR	MLE	QR	MLE
0.1	0.14	0.04	0.10	0.03	0.17	0.01	0.13	-0.01
0.2	0.11	0.00	0.05	-0.01	0.12	0.01	0.05	-0.01
0.3	0.09	-0.04	0.02	0.00	0.09	0.00	0.02	0.05
0.4	0.06	0.03	0.00	0.01	0.06	0.02	0.00	0.00
0.5	0.03	0.02	-0.02	-0.02	0.03	-0.03	-0.02	-0.03
0.6	0.00	-0.02	-0.03	0.00	-0.01	0.00	-0.03	-0.01
0.7	-0.05	-0.04	-0.05	-0.01	-0.05	-0.04	-0.05	0.00
0.8	-0.11	-0.01	-0.06	-0.01	-0.13	-0.01	-0.07	0.00
0.9	-0.20	-0.02	-0.08	-0.02	-0.24	0.00	-0.10	-0.03
$\overline{\lvert\text{Bias}\rvert}$	0.09	0.02	0.05	0.01	0.10	0.01	0.05	0.02

Notes: Table reports mean bias of slope coefficients for estimates from classical quantile regression and bias-corrected MLE modeling the error term as a mixture of three normals across 100 MC simulations of $N = 100,000$ observations each. The data are simulated from the data-generating process described in the text and the measurement error generated by either a Student's t distribution (panel I) with three degrees of freedom (normalized by $\sqrt{3.5}$ or a Laplace distribution with $\lambda = 2.29$ such that both data-generating processes result in measurement errors with the same variance (10.5) as in the original data-generating process in (4.1). The last row reports the mean absolute bias over the nine quantiles listed above.

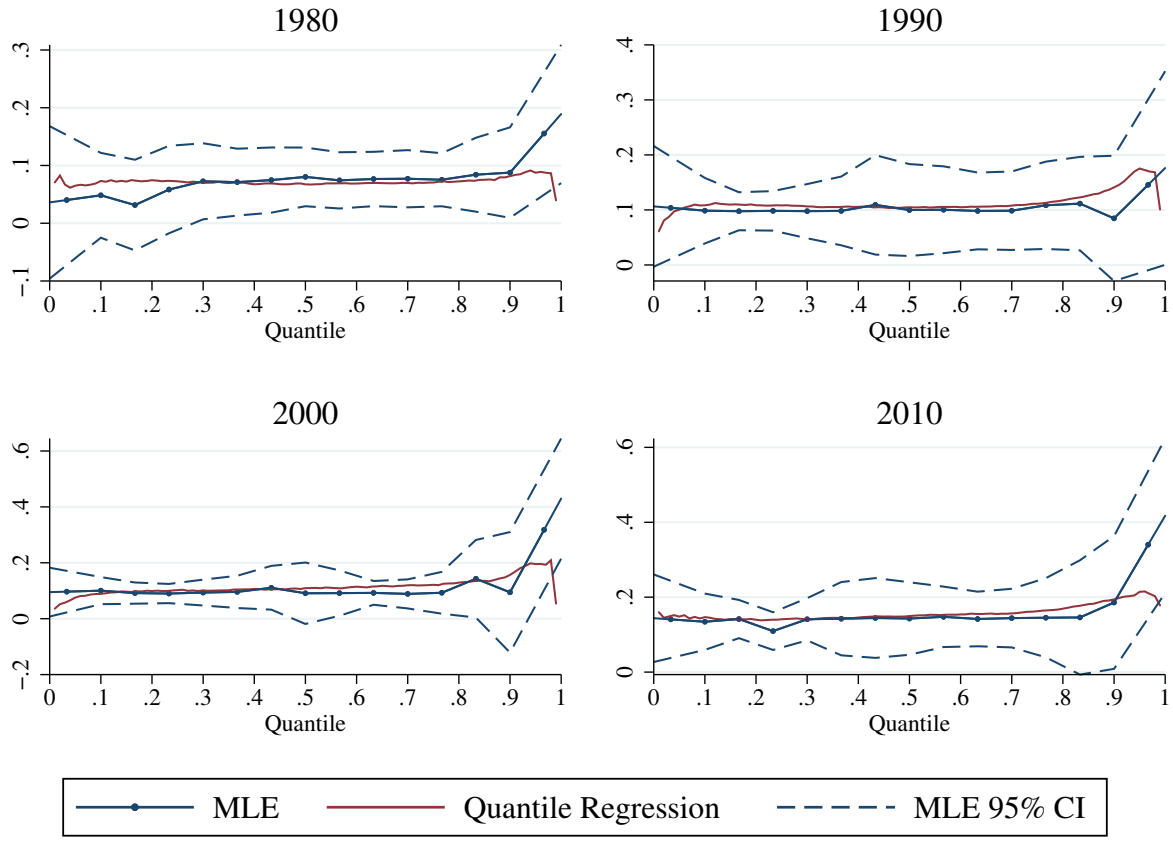
of the distribution.[14] In 2010, the education coefficient for the 95th percentile percentile was six log points higher than the education coefficient for the 5th percentile. Note, too, that traditional quantile regression estimates become quite unstable at the highest wage quantiles, characterized as the extremal quantiles problem by Chernozhukov (2005).

We observe a different pattern when we correct for measurement-error bias in the self-reported wages in the census data. Figure 5 plots the education coefficient $\widehat{\beta}_2(\tau)$ from estimating equation (5.1) by MLE and quantile regression. We approximate $\beta(\cdot)$ with a piecewise linear function consisting of 15 knots using our sieve-ML estimator developed in Section 3. We construct 95% bootstrapped confidence intervals pointwise as $\widehat{\beta}_2(\tau_j) \pm 1.96\widehat{\sigma}_j$ where $\widehat{\sigma}_j$ is the empirical standard deviation of bootstrapped estimates of $\widehat{\beta}_2(\tau_j)$.

In each year, quantile regression estimates understate the returns to education at the top of the conditional wage distribution relative to ML estimates. A formal test of the joint equality across the grid of 15 knots of QR and ML coefficients rejects equality of the education coefficient function for each year except 1990. For 1980, the quantile-regression

[14]That the wage-education gradient varies significantly with 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.

FIGURE 5. Returns to Education Correcting for LHS Measurement Error

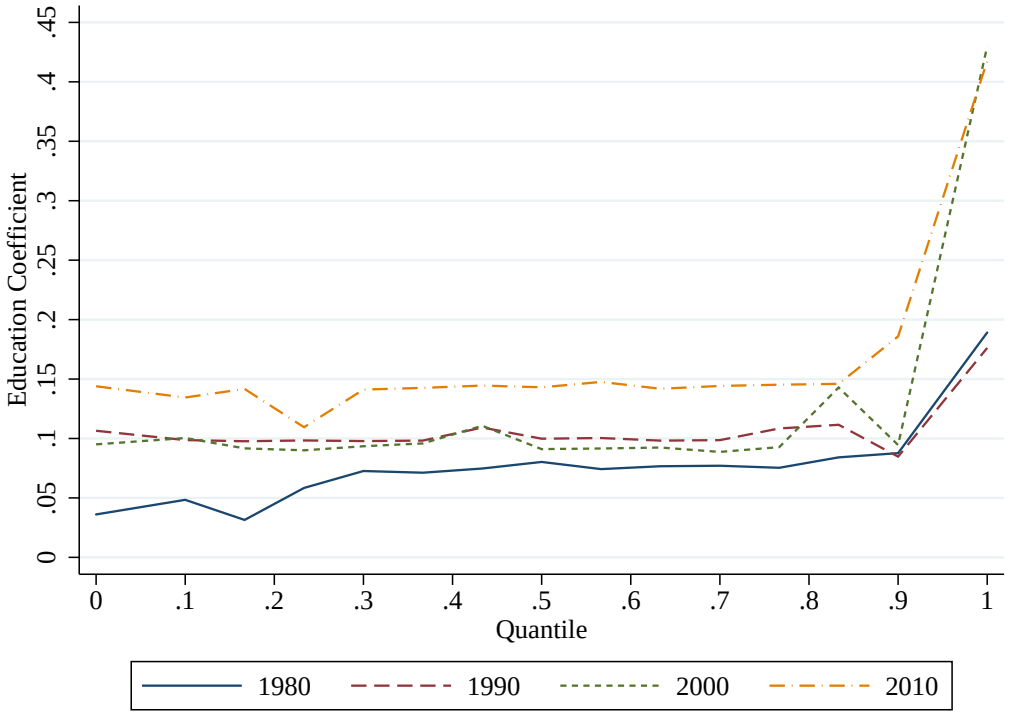


Notes: Figure reports quantile regression (red lines) and maximum likelihood estimates (dotted blue lines) of (self-reported) log weekly wages on education and a quadratic in experience. Dashed blue lines plot 95% pointwise confidence intervals from 500 bootstrap iterations. The data comes from the indicated decennial census year and consist of 40-49 year old white men with positive wages born in America. The number of observations in each sample is 60,051, 80,115, 90,201, and 98,292 in 1980, 1990, 2000, and 2010, respectively.

estimates show relatively constant returns to education across the conditional wage distribution, with a sharp decline at the very top characteristic of quantile-regression estimates at extremal quantiles. The ML estimates feature more convexity, with the pattern of increasing returns to education for higher quantiles seen in quantile-regression estimates in later years visible in the ML estimates for 1980. In 1990, the quantile-regression estimates are less affected by measurement error in the sense that the classical quantile-regression estimates and bias-corrected ML estimates are nearly indistinguishable given the typically wide confidence intervals for extremal quantiles, and we fail to reject equality of QR and ML estimates.

In the 2000 sample, 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

FIGURE 6. ML Estimated Returns to Education Across Years



Notes: Figure overlays ML estimates of the returns to education across the conditional wage distribution from Figure 5. See notes to Figure 5 for details.

measurement error in self-reported wages, the true returns to an additional year of education for 98th percentile of the conditional wage distribution is 15 log points (17 percentage points) higher than estimated by classical quantile regression. This bias correction affects the amount of inequality estimated in the education-wage gradient, with the ML estimates implying that top wage earners gained 27 log points (31 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. The quantile regression estimates at the very top of the conditional wage distribution are again outside the 95% confidence intervals for the ML estimates.

For each year besides 1990, the quantile regression lines understate the returns to education in the top decile of the wage distribution. Correcting for measurement error in self-reported wages generally increases the estimated returns to education for the top quintile of the conditional wage distribution, a distinction that is missed because of the compression bias in the quantile regression coefficients. Figure 6 overlays each year's ML estimates to facilitate easier comparisons across years. The returns to education have varied significantly over time.

Each decade—with the exception of 1990-2000—we see an increase in the returns to education broadly enjoyed across the wage distribution. However, the increase in the education-wage gradient is relatively constant across the bottom nine deciles and very different for the top decile.

These two trends—constant, moderate increases for the bottom three quartiles and acute increases in the schooling coefficient for top earners—are consistent with the observations of Angrist et al. (2006) and other work on inequality (e.g., Autor et al., 2008) that finds significant increases in income inequality post-1980. 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 the role of education in income inequality (Goldin and Katz, 2009), the rise in top-income inequality (see, for example, Piketty and Saez, 2006), and the increasing returns to cognitive performance (Lin et al., 2016).[15]

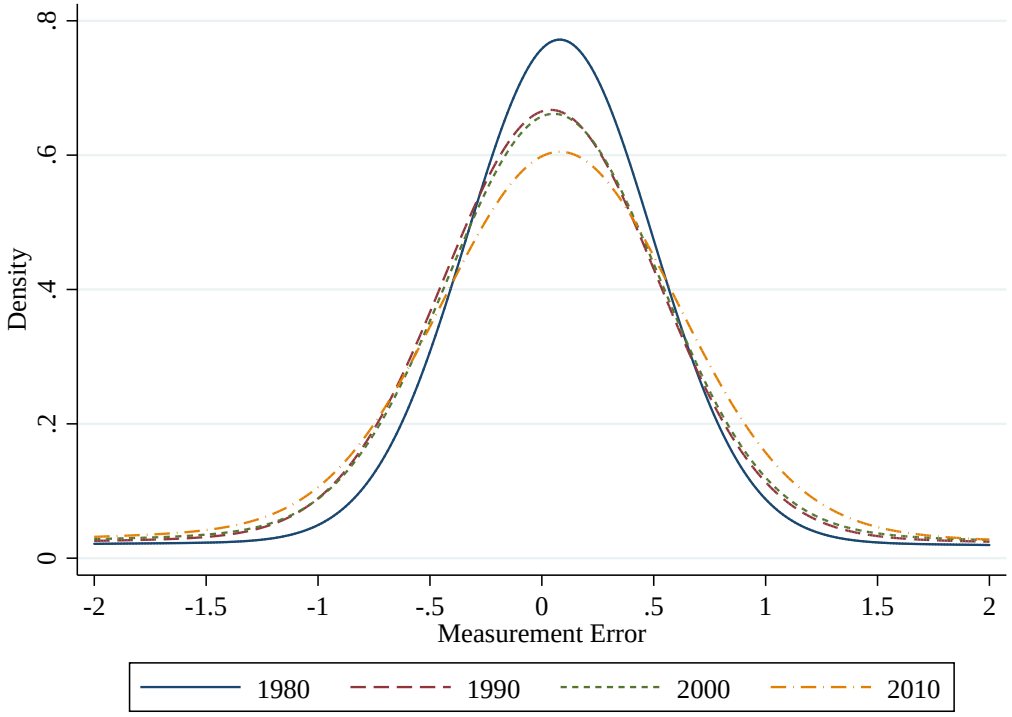
Our methodology also permits a characterization of the distribution of dependent-variable measurement error. Figure 7 plots the estimated distribution of the measurement error by census year. Despite the flexibility afforded by the mixture specification, the estimated density is unimodal but somewhat skewed with negative excess kurtosis (thinner tails) than the density of a single normal. Over time, the variance in the measurement error is increasing, consistent with recent concerns about declining response rates and a potential deterioration in the reliability of large-scale survey data (see, e.g., Bound et al., 2001; Brick and Williams, 2013; Meyer et al., 2015).

6. Conclusion

In this paper, we develop a methodology for estimating the functional parameter $\beta(\cdot)$ in quantile regression models when there is measurement error in the dependent variable. Assuming that the measurement error follows a distribution that is known up to a finite-dimensional parameter, we establish general convergence-speed results for the MLE-based approach. Under an assumption about the degree of ill-posedness of the problem (Assumption 5), we establish the convergence speed of the sieve-ML estimator. We prove the validity of bootstrapping based on asymptotic normality of our estimator and suggest using a bootstrap procedure for inference. Monte Carlo results demonstrate substantial improvements in mean bias and MSE relative to classical quantile regression when there are modest errors

[15]Our results here are not causal given that we are using observational variation in education as in Angrist et al. (2006). IV QR techniques (e.g., Chernozhukov and Hansen, 2005) could be adapted to our setting. We note that the IV literature on the returns to education has found larger effects after addressing the endogeneity of education (e.g., Griliches, 1977; Angrist and Krueger, 1991; Card, 2001).

FIGURE 7. Estimated Distribution of Wage Measurement Error



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

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 self-reported wage data, we find that top wages have grown more sensitive to education than wages in the rest of the conditional wage distribution, an important potential source of secular trends in income inequality.

REFERENCES

AN, Y., AND Y. HU (2012): "Well-posedness of measurement error models for self-reported data," *Journal of Econometrics*, 168(2), 259–269.

ANGRIST, J., V. CHERNOZHUKOV, AND I. FERNÁNDEZ-VAL (2006): "Quantile regression under misspecification, with an application to the US wage structure," *Econometrica*, 74(2), 539–563.

ANGRIST, J., AND A. KEUEGER (1991): "Does compulsory school attendance affect schooling and earnings?," *The Quarterly Journal of Economics*, 106(4), 979–1014.

ARELLANO, M., AND M. WEIDNER (2016): "Instrumental Variable Quantile Regressions in Large Panels with Fixed Effects," Working Paper.

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

BOUND, J., C. BROWN, AND N. MATHIOWETZ (2001): "Measurement error in survey data," *Handbook of Econometrics*, 5, 3705–3843.

BRICK, J. M., AND D. WILLIAMS (2013): "Explaining rising nonresponse rates in cross-sectional surveys," *The ANNALS of the American academy of political and social science*, 645(1), 36–59.

BURDA, M., M. HARDING, AND J. HAUSMAN (2008): "A Bayesian mixed logit–probit model for multinomial choice," *Journal of Econometrics*, 147(2), 232–246.

——— (2012): "A Poisson mixture model of discrete choice," *Journal of Econometrics*, 166(2), 184–203.

CARD, D. (2001): "Estimating the return to schooling: Progress on some persistent econometric problems," *Econometrica*, 69(5), 1127–1160.

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

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

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

CHERNOZHUKOV, V., AND C. HANSEN (2005): "An IV model of quantile treatment effects," *Econometrica*, 73(1), 245–261.

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

DEMPSTER, A. P., N. M. LAIRD, D. B. RUBIN, ET AL. (1977): "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, 39(1), 1–38.

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

EVDOKIMOV, K. (2010): "Identification and Estimation of a Nonparametric Panel Data Model with Unobserved Heterogeneity," Princeton University Working Paper.

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

GOLDIN, C. D., AND L. F. KATZ (2009): *The Race Between Education and Technology*. Harvard University Press.

GRILICHES, Z. (1977): "Estimating the returns to schooling: Some econometric problems," *Econometrica*, pp. 1–22.

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

HAUSMAN, J. A., J. ABREVAYA, AND F. M. SCOTT-MORTON (1998): "Misclassification of the dependent variable in a discrete-response setting," *Journal of Econometrics*, 87(2), 239–269.

HOROWITZ, J. L., AND S. LEE (2005): "Nonparametric estimation of an additive quantile regression model," *Journal of the American Statistical Association*, 100(472), 1238–1249.

KOENKER, R., AND G. BASSETT JR (1978): "Regression quantiles," *Econometrica*, pp. 33–50.

LIN, D., R. LUTTER, AND C. J. RUHM (2016): "Cognitive Performance and Labor Market Outcomes," NBER Working Paper #22470.

MEYER, B. D., W. K. MOK, AND J. X. SULLIVAN (2015): "Household surveys in crisis," *Journal of Economic Perspectives*, 29(4), 199–226.

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

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

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

RUGGLES, S., K. GENADEK, R. GOEKEN, J. GROVER, AND M. SOBEK (Minneapolis: University of Minnesota, 2015): "Integrated Public Use Microdata Series Version 6.0 [Machine-readable database]," .

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

VAN DER VAART, A. W. (2000): *Asymptotic Statistics*, vol. 3. Cambridge University Press.

VAN DER VAART, A. W., AND J. A. WELLNER (1996): "Weak convergence," in *Weak Convergence and Empirical Processes*, pp. 16–28. Springer.

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

Appendix A. Bias Characterization

In this section, we prove compression bias for the quantile regression slope coefficient. We make the following assumptions:

(1) Besides the constant, there is one covariate x, which is nonnegative and strictly positive with a positive probability.
(2) Let $\beta_1(\tau)$ and $\beta_2(\tau)$ denote the true constant and slope coefficient functions. We assume that $\beta_2(\tau)$ is not a constant, i.e. $\min_\tau \beta_2(\tau) < \max_\tau \beta_2(\tau)$. We also assume that with a positive probability, $\beta_2(\tau)$ is strictly greater than $\min_\tau \beta_2(\tau)$ and strictly smaller than $\max_\tau \beta_2(\tau)$.
(3) We assume that the true data generating process is $y = \beta_1(\tau) + \beta_2(\tau)x + \varepsilon$, where the EIV ε has a positive probability density everywhere between $-\infty$ and ∞.

Let $\widehat{\beta}_1(\tau_0)$ and $\widehat{\beta}_2(\tau_0)$ denote the estimated constant and slope coefficients at τ_0. In the following, we will show that with left-hand side measurement error ε, $\min_\tau \beta_2(\tau) < \widehat{\beta}_2(\tau_0) < \max_\tau \beta_2(\tau)$ holds for every τ_0. In other words, the quantile-regression estimated slope coefficient is always strictly bounded by the lower and upper bounds of the true slope coefficient function. We first write out the first-order conditions for $\widehat{\beta}_1(\tau_0)$ and $\widehat{\beta}_2(\tau_0)$ respectively:

$$E_{x,\tau,\varepsilon}\left[1(y - \widehat{\beta}_1(\tau_0) - \widehat{\beta}_2(\tau_0)x < 0)\right] = \tau_0$$
$$E_{x,\tau,\varepsilon}\left[x1(y - \widehat{\beta}_1(\tau_0) - \widehat{\beta}_2(\tau_0)x < 0)\right] = \tau_0 E[x]$$

where $E_{x,\tau,\varepsilon}[\cdot]$ denotes an expectation taken over the domains of x, τ, and ε. Using iterated expectations, the first-order conditions can be written as

$$E_x\left[\alpha_{\tau_0}(x)\right] = \tau_0 \tag{A.1}$$
$$E_x\left[x\alpha_{\tau_0}(x)\right] = \tau_0 E[x], \tag{A.2}$$

where

$$\begin{aligned}\alpha_{\tau_0}(x) &= E_{\tau,\varepsilon}\left[1\left(y - \widehat{\beta}_1(\tau_0) - \widehat{\beta}_2(\tau_0)x < 0\right)\right] \\ &= E_{\tau,\varepsilon}\left[1\left(\varepsilon < \widehat{\beta}_1(\tau_0) - \beta_1(\tau) + (\widehat{\beta}_2(\tau_0) - \beta_2(\tau))x\right)\right] \\ &= E_{\tau,\varepsilon}\left[1\left(\varepsilon < \widehat{\beta}_1(\tau_0) - \beta_1(\tau) + (\widehat{\beta}_2(\tau_0) - \min_\tau \beta_2(\tau))x + ((\min_\tau \beta_2(\tau)) - \beta_2(\tau))x\right)\right] \end{aligned} \tag{A.3}$$

We prove that $\widehat{\beta}_2(\tau_0) > \min_\tau \beta_2(\tau)$ by contradiction. Suppose that $\widehat{\beta}(\tau_0) \leq \min_\tau \beta_2(\tau)$. Then the slope for x inside (A.3) is nonpositive for every τ and negative for some τ by the assumption that $\beta_2(\tau)$ is not everywhere equal to its minimum. This together with the assumption that ε has a positive probability density everywhere implies that $\alpha_{\tau_0}(x)$ is a strictly decreasing function of x. However, the monotonicity of $\alpha_{\tau_0}(x)$ causes a contradiction

to (A.1) and (A.2). (A.1) claims that the mean of $\alpha_{\tau_0}(x)$ over the range of x is τ_0. The left-hand side of (A.2) is a weighted average of $\alpha_{\tau_0}(x)$ over the range of x, where the average weight is $E[x]$, and the weight increases as x increases. Since $\alpha_{\tau_0}(x)$ is strictly decreasing, the weighted average in (A.2) must be smaller than the average weight times the mean of $\alpha_{\tau_0}(x)$. In other words, the left-hand side of (A.2) must be smaller than $\tau_0 E[x]$ and cannot be equal to $\tau_0 E[x]$. This causes a contradiction to (A.2). By a similar argument, $\widehat{\beta}_2(\tau_0) < \max_\tau \beta_2(\tau)$. Therefore,

$$\min_\tau \beta_2(\tau) < \widehat{\beta}_2(\tau_0) < \max_\tau \beta_2(\tau),$$

which we refer to as compression bias because the estimated parameters strictly lie in the interior of their true maximum and minimum values over $\tau \in [0, 1]$.

Appendix B. Weighted Least Squares

Iterative weighted-least squares is a computationally attractive estimator that is theoretically equivalent to MLE when the true DGP is for ε to be distributed as a single normal random variable. WLS estimates provide useful alternative start values.

Under a normality assumption of the EIV term ε, the maximization of $Q(\cdot|\theta)$ reduces to the minimization of a weighted least squares problem.[16] Suppose the disturbance $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Then the maximization problem (3.5) becomes the following, with the parameter vector $\theta = [\beta(\cdot), \sigma]$

$$\max_{\theta'} Q(\theta'|\theta) := E\left[\log(f(y - x\beta'(\tau))|\theta')w(x, y, \theta)|\theta\right] \tag{B.1}$$

$$= E\left[\int_0^1 \frac{f(y - x^T\beta(\tau)|\sigma)}{\int_0^1 f(y - x^T\beta(u)|\sigma)du}\left(-\frac{1}{2}\log(2\pi\sigma'^2) - \frac{(y - x\beta'(\tau))^2}{2\sigma'^2}\right)d\tau\right].$$

It is apparent 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 (B.1). 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 $\widehat{W}$ of a weighting matrix W, the weighted least squares estimates of β and σ are

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

where $\widehat{W}_j$ is the diagonal matrix formed from the j^{th} column of $\widehat{W}$, which has elements $\widehat{w}_{ij}$.

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

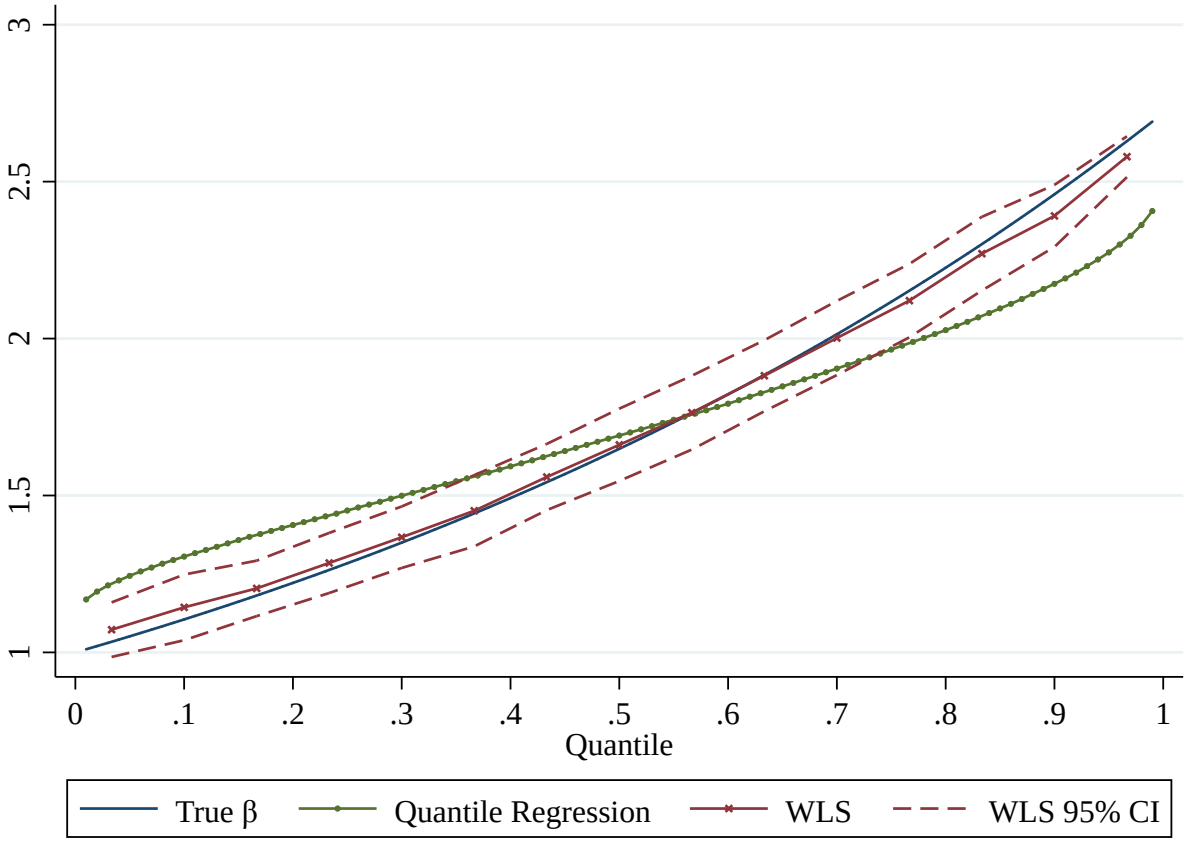
$$\widehat{w}_{ij} = \frac{\phi\left(\widehat{\varepsilon}_{ij}/\widehat{\sigma}\right)}{\frac{1}{J}\sum_j \phi\left(\widehat{\varepsilon}_{ij}/\widehat{\sigma}\right)} \tag{B.2}$$

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 $\{\tau_j\} = \{0.1, 0.2, ..., 0.9\}$.

Figures B1 and B2 show results from estimating the coefficient vectors $\beta_2(\cdot)$ and $\beta_3(\cdot)$ from iterated WLS when the DGP is taken to be $\varepsilon \sim \mathcal{N}(0, 10.5)$ to match the variance in the other simulation designs, along with confidence intervals constructed as $\pm 1.96\sqrt{\widehat{Var}(\widehat{\beta})}$ where $\widehat{Var}(\widehat{\beta})$ is the empirical variance of the coefficient estimates across MC simulations. The results confirm that WLS is a successful alternative to MLE when the data is normal.

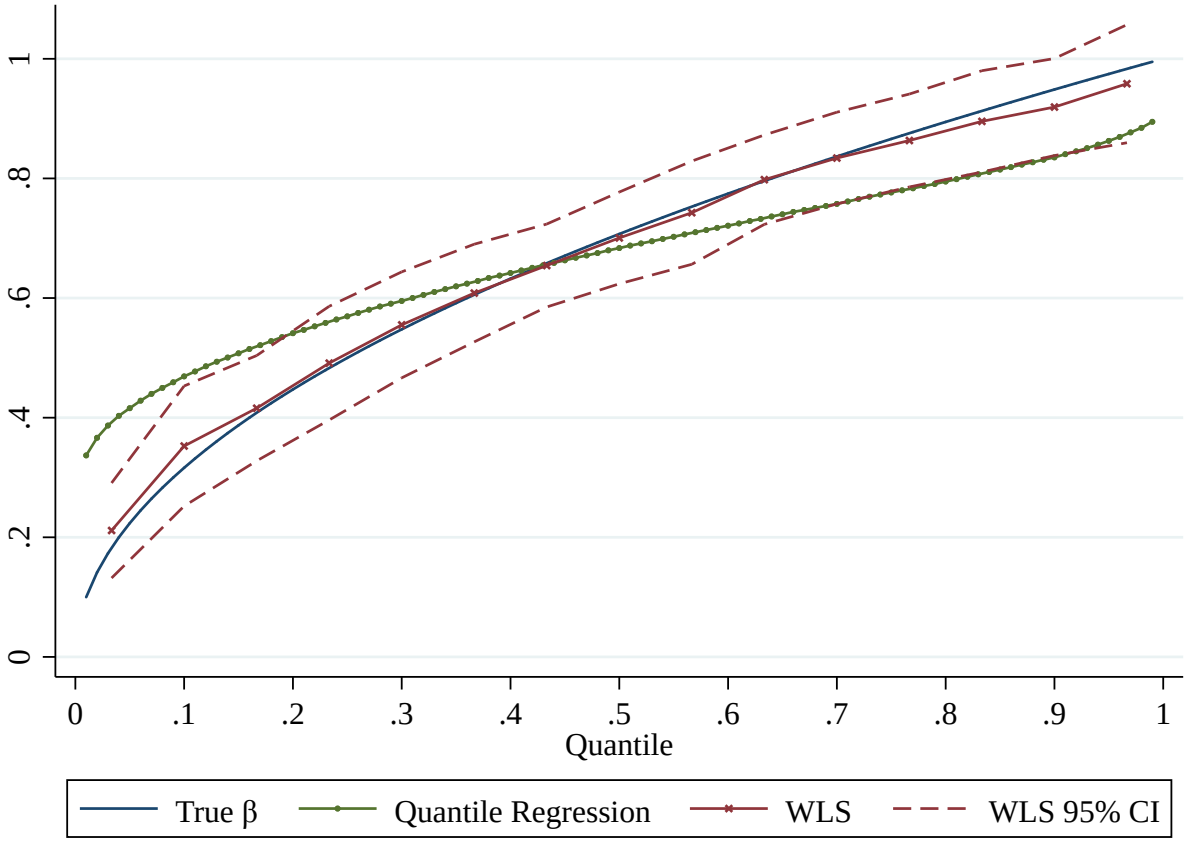
[16] See, also, the related method in Section 4.6 of Dempster et al. (1977).

FIGURE B1. Weighted Least Squares Monte Carlo Simulation Results: $\widehat{\beta}_2(\tau)$



Notes: Figure plots the true $\beta_2(\tau) = \exp(\tau)$ (blue line) against quantile-regression estimates (green circles), weighted least squares estimates (red xs), and 95% confidence intervals for the WLS estimates (dashed red lines) from 100 MC simulations with 40 WLS iterations each. The data-generating process is described in the text with the measurement error generated as a normal random variable distributed $\mathcal{N}(0, 10.5)$.

FIGURE B2. Weighted Least Squares Monte Carlo Simulation Results: $\widehat{\beta}_3(\tau)$



Notes: Figure plots the true $\beta_3(\tau) = \sqrt{\tau}$ (blue line) against quantile-regression estimates (green circles), weighted least squares (red xs), and 95% confidence intervals for the WLS estimates (dashed red lines) from 100 MC simulations with 40 WLS iterations each. The data-generating process is described in the text with the measurement error generated as a normal random variable distributed $\mathcal{N}(0, 10.5)$.

Appendix C. Additional Simulation Results

In this appendix, we present Monte Carlo simulation results (mean bias and MSE) under alternative data generating processes. For each design, quasi-ML estimation continues to treat the measurement error as a mixture of three normals. After simulating measurement error under alternative measurement error distributions (all normalized such that ε has equal variance across designs), Appendix Tables C5 and C6, respectively, present results when x_1 is binary (normalized to have equal variance across simulation designs) and when a 99-knot sieve is used to approximate $\beta(\cdot)$.

Table C1. Mean Bias and Mean Squared Error: $\varepsilon \sim 3\mathcal{N}$

	β_1		β_2		β_3	
Quantile	QR	MLE	QR	MLE	QR	MLE
			I. Mean Bias			
0.1	-2.924	-0.020	0.145	0.003	0.134	0.010
0.2	-2.486	0.022	0.223	-0.003	0.144	-0.014
0.3	-2.074	-0.005	0.265	0.011	0.132	0.009
0.4	-1.510	-0.013	0.248	0.009	0.089	0.002
0.5	-0.402	-0.075	0.101	0.007	-0.012	0.011
0.6	1.055	-0.023	-0.123	0.012	-0.120	0.001
0.7	1.939	0.002	-0.238	0.011	-0.141	-0.005
0.8	2.601	0.047	-0.285	-0.001	-0.125	-0.010
0.9	3.355	0.078	-0.284	-0.019	-0.097	-0.002
$\overline{\lvert\text{Bias}\rvert}$	2.038	0.032	0.213	0.008	0.110	0.007
			II. Mean Squared Error			
0.1	8.548	0.042	0.021	0.005	0.018	0.006
0.2	6.179	0.020	0.050	0.002	0.021	0.003
0.3	4.302	0.037	0.070	0.007	0.018	0.008
0.4	2.280	0.019	0.062	0.004	0.008	0.003
0.5	0.164	0.040	0.011	0.010	0.000	0.004
0.6	1.113	0.018	0.015	0.004	0.014	0.003
0.7	3.761	0.028	0.057	0.008	0.020	0.004
0.8	6.767	0.019	0.082	0.003	0.016	0.001
0.9	11.259	0.034	0.081	0.005	0.010	0.002
$\overline{\text{MSE}}$	4.930	0.029	0.050	0.005	0.014	0.004

Notes: Table reports mean bias (panel I) and MSE (panel II) for estimates from classical quantile regression (QR) and MLE across 100 MC simulations of $N = 100,000$ observations using data simulated from the data-generating process described in Section 4. The last row reports the mean absolute bias (panel I) and the mean MSE (panel II) over the nine quantiles listed above.

TABLE C2. Mean Bias and Mean Squared Error: $\varepsilon \sim 2\mathcal{N}$

	β_1		β_2		β_3	
Quantile	QR	MLE	QR	MLE	QR	MLE
			I. Mean Bias			
0.1	-3.894	0.049	0.237	0.001	0.205	0.014
0.2	-3.082	0.010	0.342	-0.001	0.206	-0.008
0.3	-1.663	-0.027	0.259	0.022	0.096	0.006
0.4	0.537	0.004	-0.049	0.003	-0.105	-0.002
0.5	1.184	0.012	-0.141	-0.004	-0.125	-0.007
0.6	1.536	0.007	-0.182	0.009	-0.111	0.001
0.7	1.807	-0.080	-0.196	0.022	-0.090	0.000
0.8	2.077	-0.025	-0.192	0.010	-0.070	0.002
0.9	2.457	-0.069	-0.174	0.020	-0.055	0.002
|Bias|	2.026	0.031	0.197	0.010	0.118	0.005
			II. Mean Squared Error			
0.1	15.163	0.036	0.057	0.005	0.042	0.007
0.2	9.500	0.020	0.117	0.002	0.043	0.002
0.3	2.768	0.073	0.067	0.016	0.009	0.008
0.4	0.289	0.026	0.003	0.005	0.011	0.002
0.5	1.403	0.079	0.020	0.017	0.016	0.006
0.6	2.360	0.021	0.033	0.004	0.012	0.002
0.7	3.266	0.093	0.039	0.019	0.008	0.005
0.8	4.315	0.014	0.037	0.002	0.005	0.001
0.9	6.037	0.024	0.031	0.003	0.003	0.002
MSE	5.011	0.043	0.045	0.008	0.017	0.004

Notes: Table reports mean bias (panel I) and MSE (panel II) for estimates from classical quantile regression (QR) and quasi-MLE modeling the error term as a mixture of three normals across 100 MC simulations of $N = 100,000$ observations each. The data are simulated from the data-generating process described in Section 4 and the measurement error generated as a mixture of two normals $\mathcal{N}(-4.36, 1)$ and $\mathcal{N}(2.18, 1)$ with weights 1/3 and 2/3, respectively, such that the variance of the measurement error is equal across simulation designs. The last row reports the mean absolute bias (panel I) and the mean MSE (panel II) over the nine quantiles listed above.

TABLE C3. Mean Bias and Mean Squared Error: $\varepsilon \sim t$

	β_1		β_2		β_3	
Quantile	QR	MLE	QR	MLE	QR	MLE
			I. Mean Bias			
0.1	-1.962	0.039	0.142	0.039	0.103	0.028
0.2	-1.087	0.084	0.112	0.000	0.048	-0.015
0.3	-0.649	0.018	0.088	-0.038	0.021	-0.004
0.4	-0.336	-0.099	0.062	0.028	0.000	0.007
0.5	-0.062	0.001	0.032	0.015	-0.016	-0.015
0.6	0.227	0.053	-0.004	-0.020	-0.031	-0.005
0.7	0.581	0.109	-0.052	-0.041	-0.045	-0.011
0.8	1.098	0.074	-0.115	-0.007	-0.062	-0.013
0.9	2.082	0.049	-0.200	-0.022	-0.082	-0.017
$\overline{\lvert\text{Bias}\rvert}$	0.898	0.058	0.090	0.023	0.045	0.013
			II. Mean Squared Error			
0.1	3.850	0.216	0.020	0.021	0.011	0.036
0.2	1.183	0.104	0.013	0.009	0.002	0.009
0.3	0.422	0.164	0.008	0.028	0.001	0.022
0.4	0.113	0.120	0.004	0.017	0.000	0.010
0.5	0.004	0.191	0.001	0.044	0.000	0.017
0.6	0.052	0.064	0.000	0.017	0.001	0.007
0.7	0.338	0.108	0.003	0.033	0.002	0.011
0.8	1.205	0.039	0.013	0.006	0.004	0.003
0.9	4.337	0.018	0.040	0.011	0.007	0.007
$\overline{\text{MSE}}$	1.278	0.114	0.011	0.021	0.003	0.013

Notes: Table reports mean bias (panel I) and MSE (panel II) for estimates from classical quantile regression (QR) and quasi-MLE modeling the error term as a mixture of three normals across 100 MC simulations of $N = 100,000$ observations each. The data are simulated from the data-generating process described in Section 4 but measurement error generated as a Student's t random variable with three degrees of freedom, multiplied by $\sqrt{3.5}$ to ensure the variance of the measurement error is equal across simulation designs. The last row reports the mean absolute bias (panel I) and the mean MSE (panel II) over the nine quantiles listed above.

TABLE C4. Mean Bias and Mean Squared Error: $\varepsilon \sim$ *Laplace*

	β_1		β_2		β_3	
Quantile	QR	MLE	QR	MLE	QR	MLE
			I. Mean Bias			
0.1	-2.499	0.082	0.173	0.008	0.126	-0.009
0.2	-1.325	0.026	0.124	0.009	0.052	-0.009
0.3	-0.773	-0.098	0.092	-0.001	0.019	0.049
0.4	-0.391	-0.047	0.062	0.020	-0.001	0.002
0.5	-0.058	0.099	0.031	-0.031	-0.018	-0.028
0.6	0.286	0.019	-0.007	0.000	-0.032	-0.014
0.7	0.697	0.044	-0.054	-0.037	-0.046	-0.004
0.8	1.327	0.041	-0.127	-0.012	-0.068	-0.005
0.9	2.631	0.047	-0.245	0.000	-0.100	-0.030
$\overline{\lvert\text{Bias}\rvert}$	1.110	0.056	0.102	0.013	0.051	0.017
			II. Mean Squared Error			
0.1	6.247	0.117	0.030	0.014	0.016	0.021
0.2	1.756	0.064	0.016	0.008	0.003	0.008
0.3	0.598	0.155	0.009	0.031	0.000	0.026
0.4	0.154	0.072	0.004	0.013	0.000	0.008
0.5	0.004	0.178	0.001	0.040	0.000	0.017
0.6	0.082	0.047	0.000	0.011	0.001	0.006
0.7	0.486	0.096	0.003	0.033	0.002	0.009
0.8	1.762	0.019	0.016	0.005	0.005	0.003
0.9	6.922	0.014	0.060	0.009	0.010	0.006
$\overline{\text{MSE}}$	2.001	0.085	0.015	0.018	0.004	0.011

Notes: Table reports mean bias (panel I) and MSE (panel II) for estimates from classical quantile regression (QR) and quasi-MLE modeling the error term as a mixture of three normals across 100 MC simulations of $N = 100,000$ observations each. The data are simulated from the data-generating process described in Section 4 but measurement error generated as a Laplace random variable with $\lambda = 2.29$ to ensure the variance of the measurement error is equal across simulation designs. The last row reports the mean absolute bias (panel I) and the mean MSE (panel II) over the nine quantiles listed above.

TABLE C5. Mean Bias and Mean Squared Error: Binary x_2

	β_1		β_2		β_3	
Quantile	QR	MLE	QR	MLE	QR	MLE
			I. Mean Bias			
0.1	-3.019	0.205	0.177	-0.026	0.139	0.025
0.2	-2.608	0.154	0.296	-0.022	0.138	-0.019
0.3	-2.242	0.028	0.386	0.015	0.112	-0.014
0.4	-1.759	0.054	0.377	-0.021	0.074	0.014
0.5	-0.512	0.031	0.162	-0.006	-0.018	-0.004
0.6	1.238	-0.026	-0.197	0.008	-0.100	-0.002
0.7	2.035	-0.054	-0.348	-0.016	-0.108	0.007
0.8	2.731	-0.120	-0.394	0.036	-0.118	-0.014
0.9	3.535	-0.115	-0.356	-0.001	-0.102	-0.002
|Bias|	2.186	0.087	0.299	0.017	0.101	0.011
			II. Mean Squared Error			
0.1	9.113	0.160	0.031	0.009	0.019	0.012
0.2	6.800	0.081	0.088	0.007	0.019	0.005
0.3	5.027	0.161	0.149	0.034	0.013	0.014
0.4	3.093	0.085	0.143	0.012	0.006	0.003
0.5	0.264	0.189	0.026	0.041	0.001	0.014
0.6	1.534	0.076	0.039	0.012	0.010	0.004
0.7	4.141	0.156	0.121	0.035	0.012	0.008
0.8	7.456	0.062	0.155	0.007	0.014	0.002
0.9	12.495	0.075	0.127	0.011	0.010	0.003
MSE	5.547	0.116	0.098	0.019	0.012	0.007

Notes: Table reports mean bias (panel I) and MSE (panel II) for estimates from classical quantile regression (QR) and MLE across 100 MC simulations of $N = 100,000$ observations using data simulated from the data-generating process described in Section 4 but where $x_2 \in \{0, 4.32\}$ with equal probability (to ensure that the variance of x_2 is equal across simulation designs). The last row reports the mean absolute bias (panel I) and the mean MSE (panel II) over the nine quantiles listed above.

TABLE C6. Mean Bias and Mean Squared Error: 99 Knots

	β_1		β_2		β_3	
Quantile	QR	MLE	QR	MLE	QR	MLE
			I. Mean Bias			
0.1	-2.924	-0.220	0.145	0.007	0.134	-0.018
0.2	-2.486	-0.237	0.223	-0.001	0.144	0.021
0.3	-2.074	-0.243	0.265	-0.002	0.132	0.009
0.4	-1.510	-0.196	0.248	-0.006	0.089	-0.008
0.5	-0.402	-0.184	0.101	0.007	-0.012	-0.018
0.6	1.055	-0.220	-0.123	0.008	-0.120	0.003
0.7	1.939	-0.281	-0.238	0.034	-0.141	0.006
0.8	2.601	-0.156	-0.285	-0.002	-0.125	-0.005
0.9	3.355	-0.108	-0.284	-0.015	-0.097	-0.009
$\overline{\lvert\text{Bias}\rvert}$	2.038	0.205	0.213	0.009	0.110	0.011
			II. Mean Squared Error			
0.1	8.548	0.343	0.021	0.007	0.018	0.012
0.2	6.179	0.315	0.050	0.013	0.021	0.020
0.3	4.302	0.315	0.070	0.019	0.018	0.014
0.4	2.280	0.256	0.062	0.024	0.008	0.012
0.5	0.164	0.254	0.011	0.025	0.000	0.014
0.6	1.113	0.313	0.015	0.025	0.014	0.010
0.7	3.761	0.359	0.057	0.027	0.020	0.005
0.8	6.767	0.212	0.082	0.023	0.016	0.004
0.9	11.259	0.070	0.081	0.015	0.010	0.004
$\overline{\text{MSE}}$	4.930	0.271	0.050	0.020	0.014	0.011

Notes: Table reports mean bias (panel I) and MSE (panel II) for estimates from classical quantile regression (QR) and MLE across 100 MC simulations of $N = 100,000$ observations using data simulated from the data-generating process described in Section 4 and when a sieve of $J = 99$ knots is used in estimation. The last row reports the mean absolute bias (panel I) and the mean MSE (panel II) over the nine quantiles listed above.

Appendix D. Data Appendix

Following the sample selection criteria of Angrist et al. (2006), our data comes from 1% samples of decennial census data available via IPUMS.org (Ruggles et al., 2015) from 1980–2010. From each database, we select annual wage income, education, age, and race data for prime-age (age 40-49) black and white males who have at least five years of education, were born in the United States, had positive earnings and hours worked in the reference year, and whose responses for age, education, and earnings were not imputed (which would have been an additional source of measurement error). Our dependent variable is log weekly wage, obtained as annual wage income divided by weeks worked. For 1980, we take the number of years of education to be the highest grade completed and follow the methodology of Angrist et al. (2006) to convert the categorical education variable in 1990, 2000, and 2010 into a measure of the number of years of schooling. Experience is defined as age minus years of education minus five. For 1980, 1990, and 2000, we use the exact extract of Angrist et al. (2006), and draw our own data to extend the data to include the 2010 census. Table D1 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 D1. Education and Wages Summary Statistics

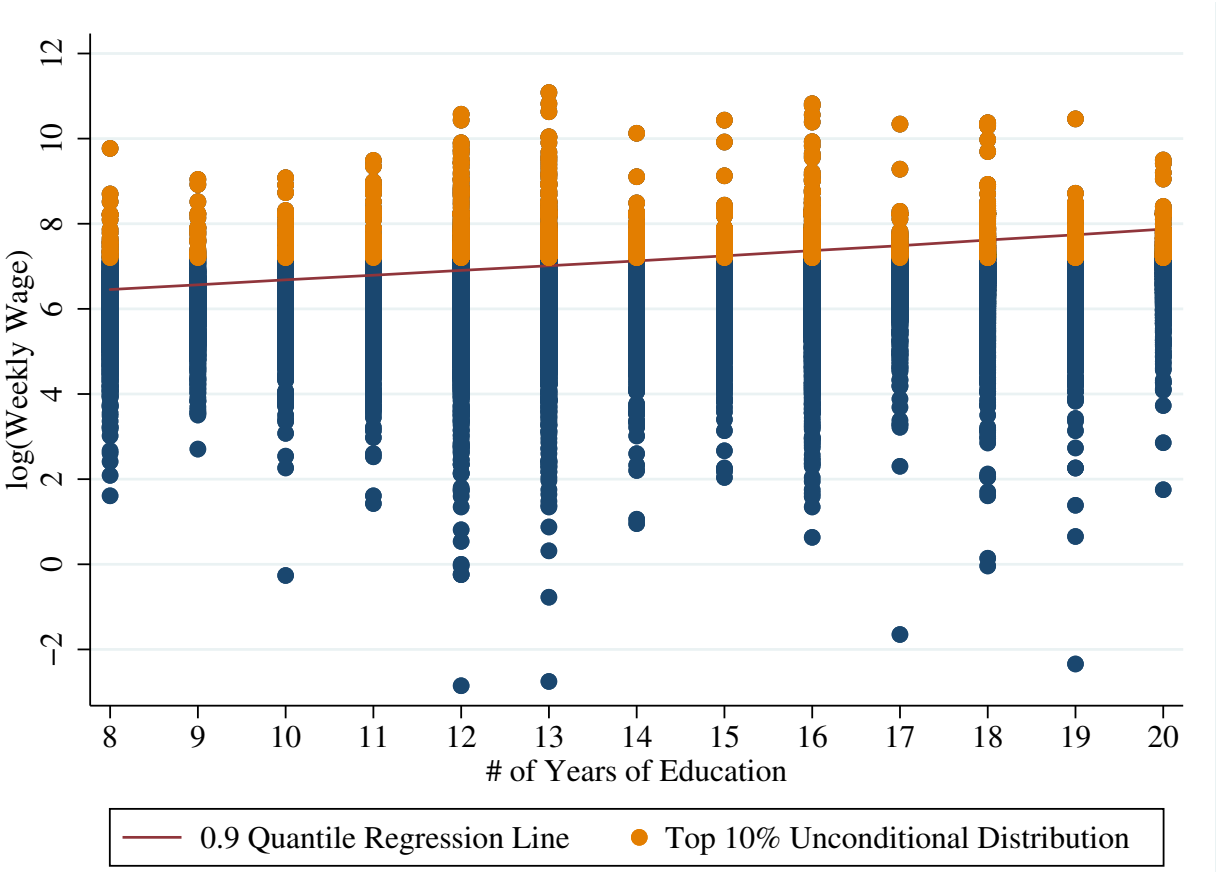
Year	1980	1990	2000	2010
Log weekly wage	6.43	6.48	6.50	8.37
	(0.66)	(0.69)	(0.74)	(0.76)
Education	12.99	13.97	13.90	14.12
	(3.08)	(2.66)	(2.41)	(2.39)
Experience	25.38	24.45	24.45	24.55
	(4.32)	(4.01)	(3.60)	(3.83)
Number of Observations	60,051	80,115	90,201	98,292

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).

Although quantile regression recovers effects on the conditional distribution of the outcome, it is worth noting that given the substantial variation in wages left unexplained by the Mincer model, the empirical difference between effects on the unconditional and conditional distributions of the dependent variable is likely small. See DiNardo et al. (1996) and Powell (2013) for further discussion and methods that recover effects on the unconditional distribution. Appendix Figure D1 illustrates this point for the 0.9 quantile estimates, showing that because of the relatively low goodness of fit of equation (5.1) (as is the case in many

cross-sectional applied microeconomics settings), over 63% of the observations in the top unconditional decile are also in the top conditional decile.

FIGURE D1. Overlap between Unconditional and Conditional Wage Distribution



Notes: Figure plots log weekly wages against years of education from the 1990 decennial Census microdata extract used by Angrist et al. (2006). The regression line plots the average predicted values by year of education from estimating equation (5.1) by classic quantile regression. Lighter colored dots indicate observations in the top 10% of the unconditional wage distribution (individuals with over $1,326 in weekly wages in 1989 dollars).

Appendix E. Quantile Regression Results across Demographic Subgroups

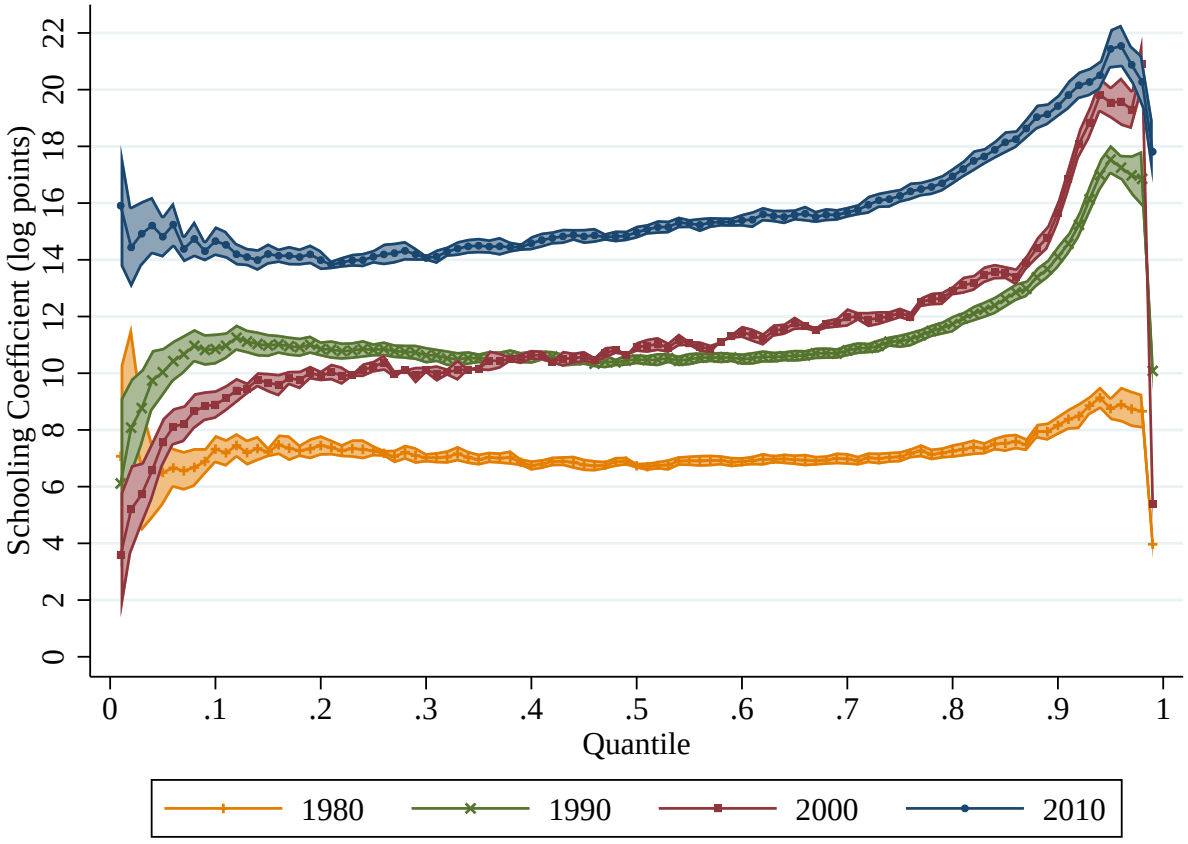
In this appendix, we plot quantile-regression estimates of the return to education for several demographic subgroups: white, black, Asian, and hispanic men and women. For comparability, we again follow Angrist et al. (2006) and restrict the sample to 40-49 year olds. Our main results focus on white males for two reasons. First, only the white-male subgroup had enough observations in the census data to have sufficient power for our semiparametric estimator to have enough precision to be useful. Second, the results below show that the level and quantile dependency of the education-wage gradient varies substantially across subgroups. Accordingly, pooling demographic subgroups to estimate a common returns to education quantile function results in misspecification that likely significantly misstates the relationship between education and wages for many demographics.

Over time, the returns to education have been increasing for all ethnic subgroups at almost all quantiles. The education coefficients are increasing across quantiles only for white men.[17] While most ethnic subgroups have similar quantile patterns for men and women, white women (Figure E5) have a hump-shaped relationship between quantile and the Mincerian education coefficient, with education consistently less valuable lower in the conditional distribution of income. For black men and women (Figures E2 and E6), the relationship is generally downward sloping such that low-income blacks appear to have the highest returns to education relative to blacks elsewhere in the conditional wage distribution. The return on education for the bottom of the conditional wage distribution for low-income blacks is higher in 2010 than for any other demographic in any year at any quantile, potentially motivating future research into the social return on increasing educational attainment among low-income black men. Hispanic men and women (Figures E3 and E7) also exhibit a downward trend in the education-wage gradient across quantiles, while the education-wage gradient is relatively constant across the conditional wage distribution for Asian men and women (Figures E4 and E8).

We leave an explanation of this heterogeneity—as well as an exploration of whether the education-wage gradient differs for other age groups—to future research.

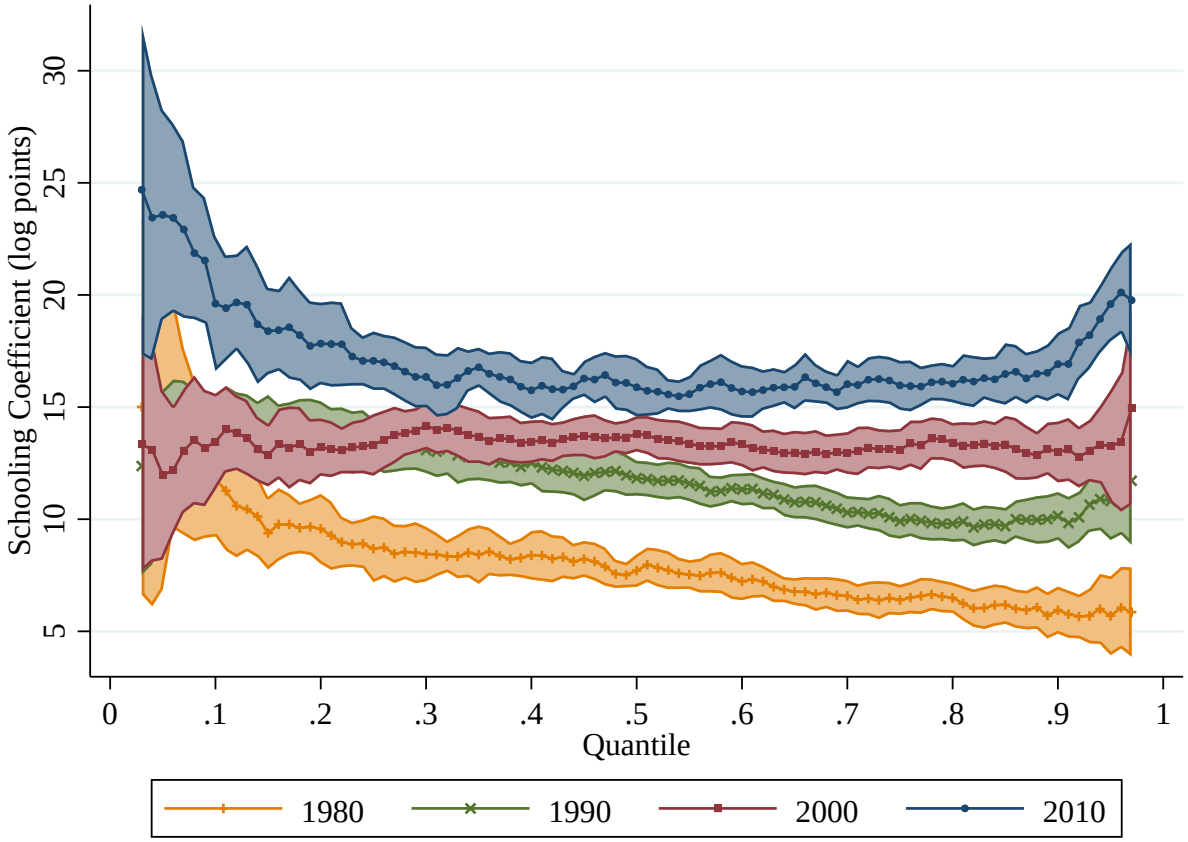
[17] We report extremal quantile estimates of the return to education only for white men to highlight their instability and extremal quantile problem discussed in Chernozhukov (2005).

FIGURE E1. QR Estimates of the Returns to Education for White Men



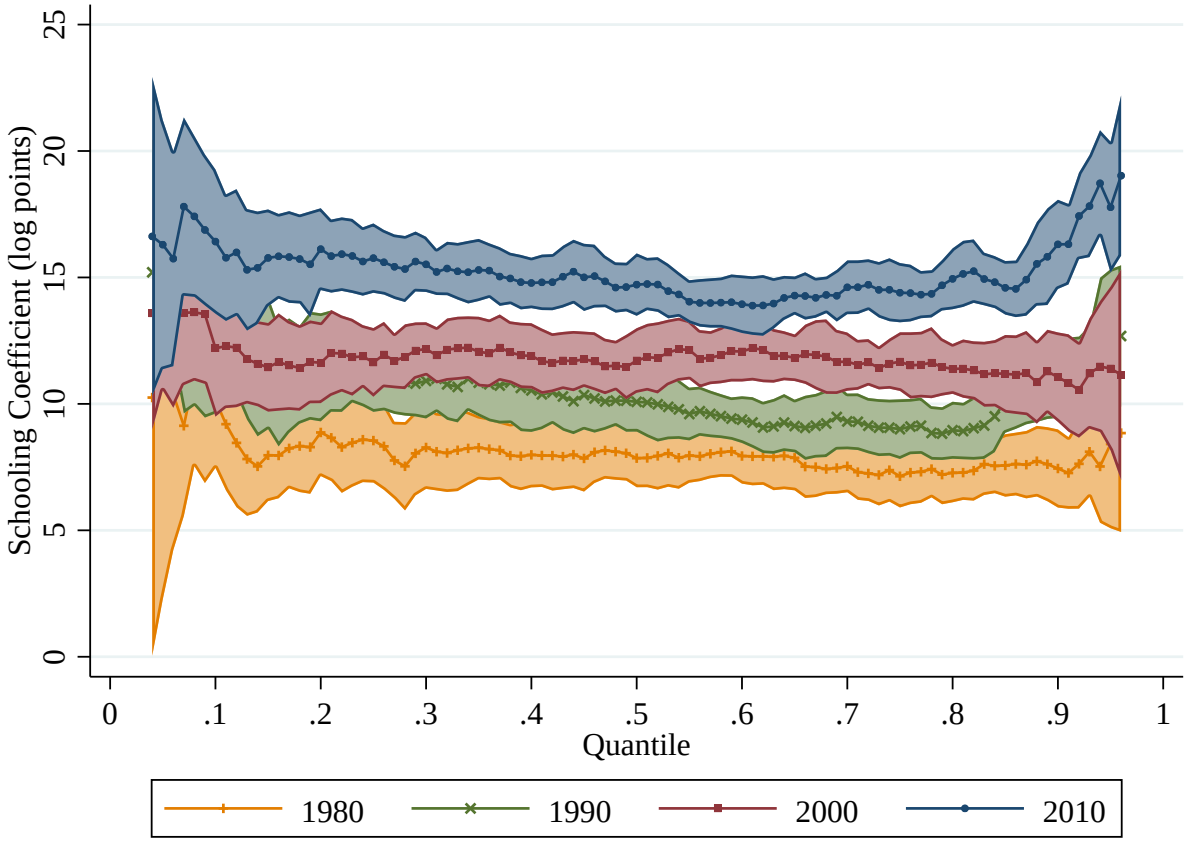
Notes: Figure reports quantile regression estimates of log weekly wages (self-reported) on education and a quadratic in experience for a grid of 99 quantiles from 0.01 to 0.99. Robust 95% pointwise confidence intervals are plotted for the quantile regression estimates for each year. The data comes from the indicated decennial census year and consist of 40-49 year-old U.S. born white men with positive reported wages. The number of observations in each sample is 60,051, 80,115, 90,201, and 98,292 in 1980, 1990, 2000, and 2010, respectively.

FIGURE E2. QR Estimates of the Returns to Education for Black Men



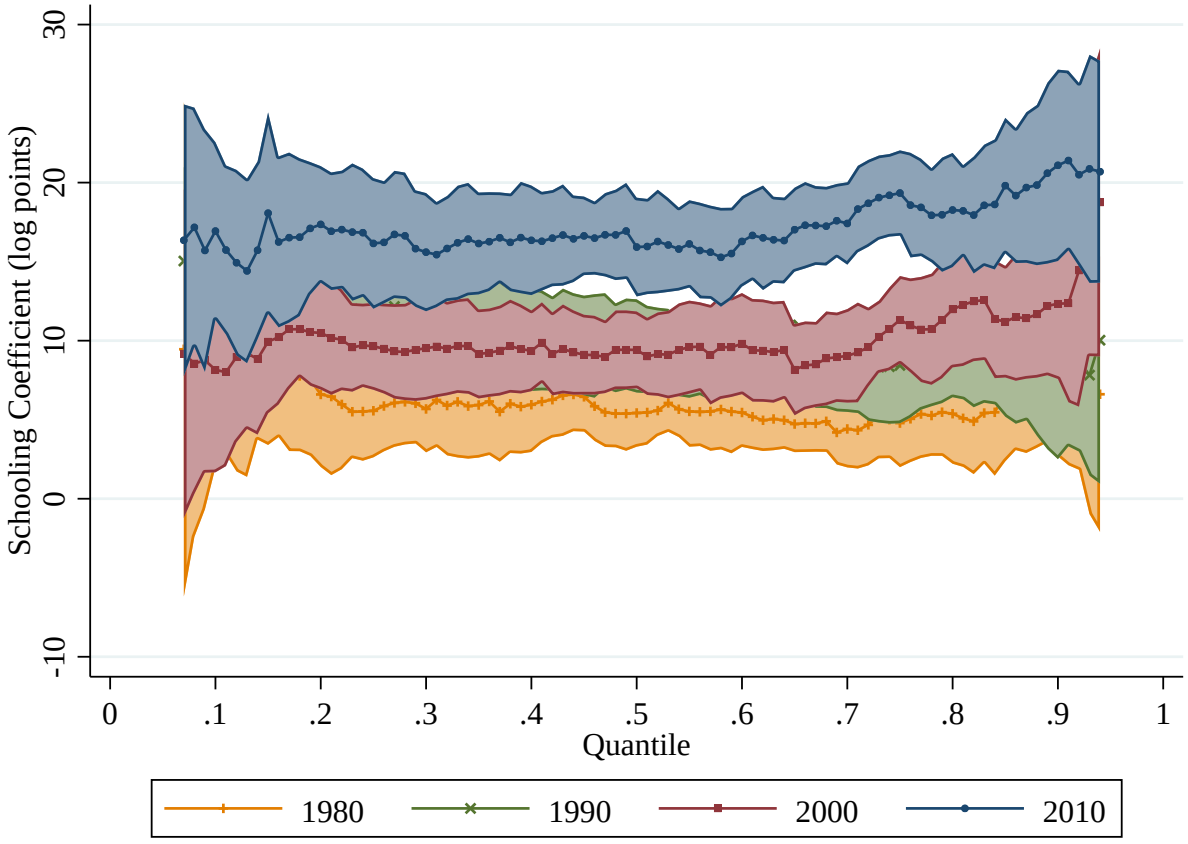
Notes: Figure reports quantile regression estimates of log weekly wages (self-reported) on education and a quadratic in experience for a grid of 99 quantiles from 0.02 to 0.98. Horizontal lines plot OLS estimates for each year, and robust 95% pointwise confidence intervals are plotted for the quantile regression estimates for each year. The data comes from the indicated decennial census year and consist of 40-49 year-old U.S. born black men with positive reported wages. The number of observations in each sample is 4,991, 6,855, 6,733, and 7,365 in 1980, 1990, 2000, and 2010, respectively.

FIGURE E3. QR Estimates of the Returns to Education for Hispanic Men



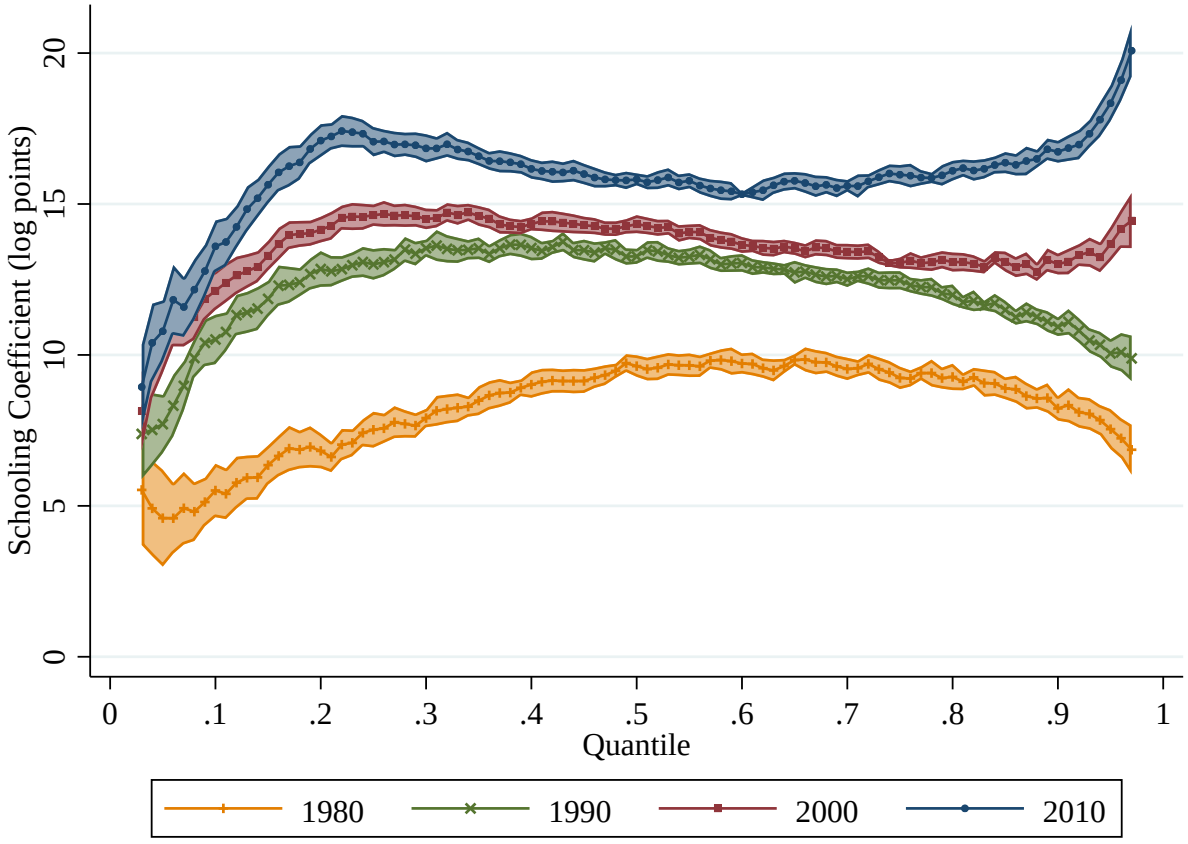
Notes: Figure reports quantile regression estimates of log weekly wages (self-reported) on education and a quadratic in experience for a grid of 99 quantiles from 0.03 to 0.97. Horizontal lines plot OLS estimates for each year, and robust 95% pointwise confidence intervals are plotted for the quantile regression estimates for each year. The data comes from the indicated decennial census year and consist of 40-49 year-old U.S. born Hispanic men with positive reported wages. The number of observations in each sample is 1,948, 3,005, 3,666, and 5,501 in 1980, 1990, 2000, and 2010, respectively.

FIGURE E4. QR Estimates of the Returns to Education for Asian Men



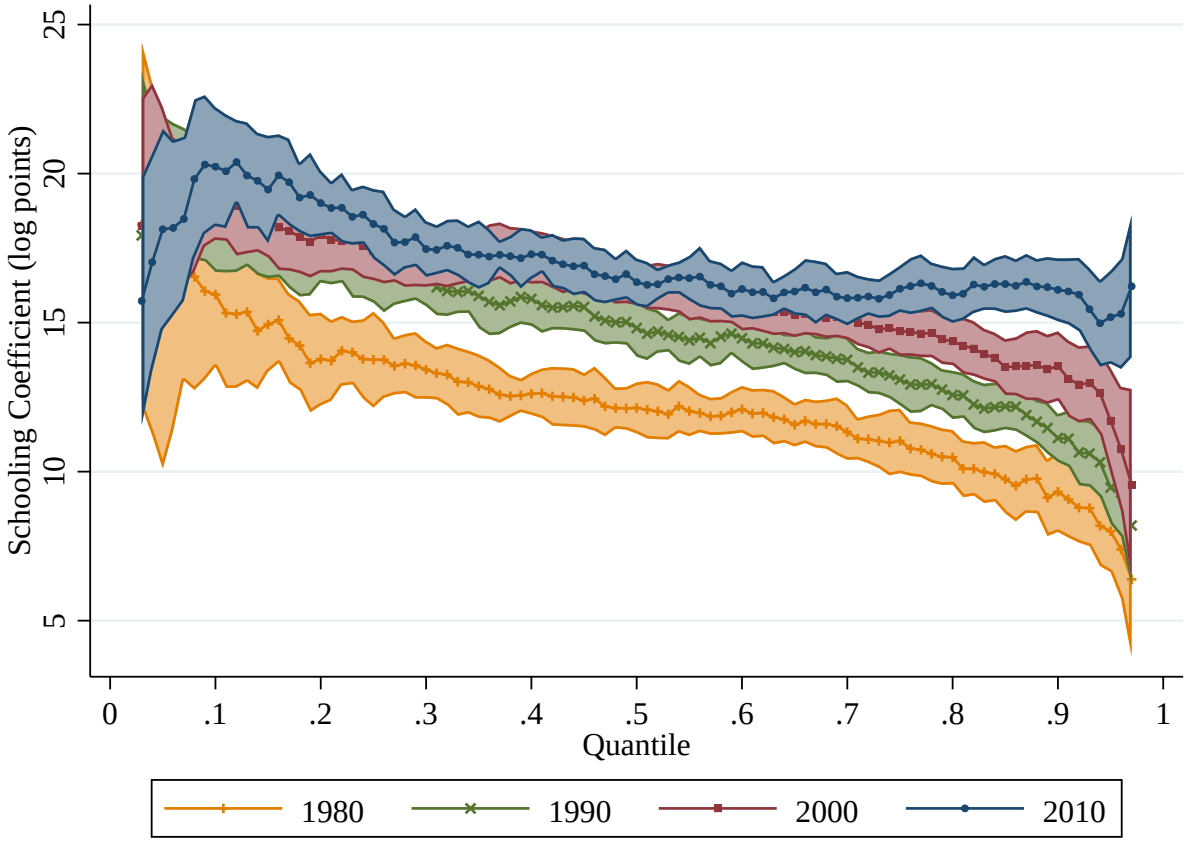
Notes: Figure reports quantile regression estimates of log weekly wages (self-reported) on education and a quadratic in experience for a grid of 99 quantiles from 0.06 to 0.94. Horizontal lines plot OLS estimates for each year, and robust 95% pointwise confidence intervals are plotted for the quantile regression estimates for each year. The data comes from the indicated decennial census year and consist of 40-49 year-old U.S. born men of Asian ethnicity with positive reported wages. The number of observations in each sample is 348, 532, 544, and 786 in 1980, 1990, 2000, and 2010, respectively.

FIGURE E5. QR Estimates of the Returns to Education for White Women



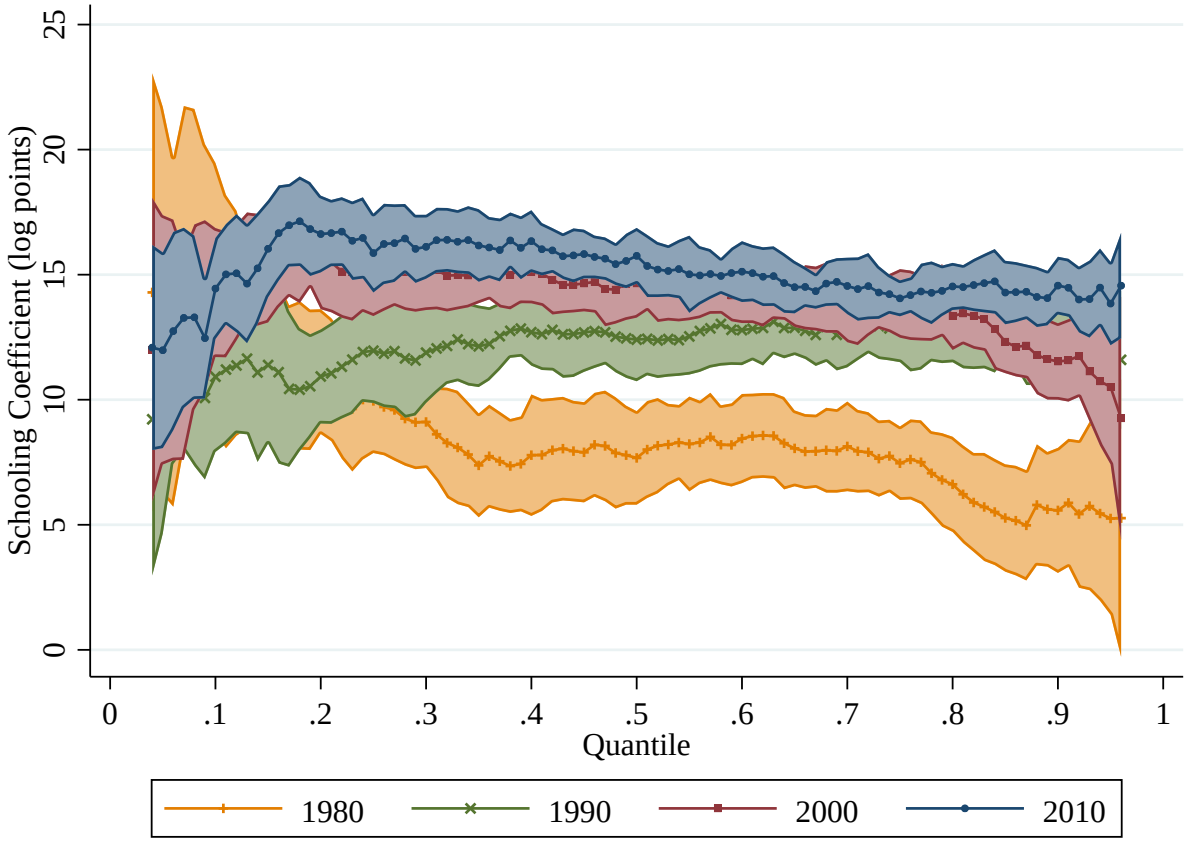
Notes: Figure reports quantile regression estimates of log weekly wages (self-reported) on education and a quadratic in experience for a grid of 99 quantiles from 0.02 to 0.98. Horizontal lines plot OLS estimates for each year, and robust 95% pointwise confidence intervals are plotted for the quantile regression estimates for each year. The data comes from the indicated decennial census year and consist of 40-49 year-old U.S. born white women with positive reported wages. The number of observations in each sample is 43,965, 69,903, 74,878, and 76,985 in 1980, 1990, 2000, and 2010, respectively.

FIGURE E6. QR Estimates of the Returns to Education for Black Women



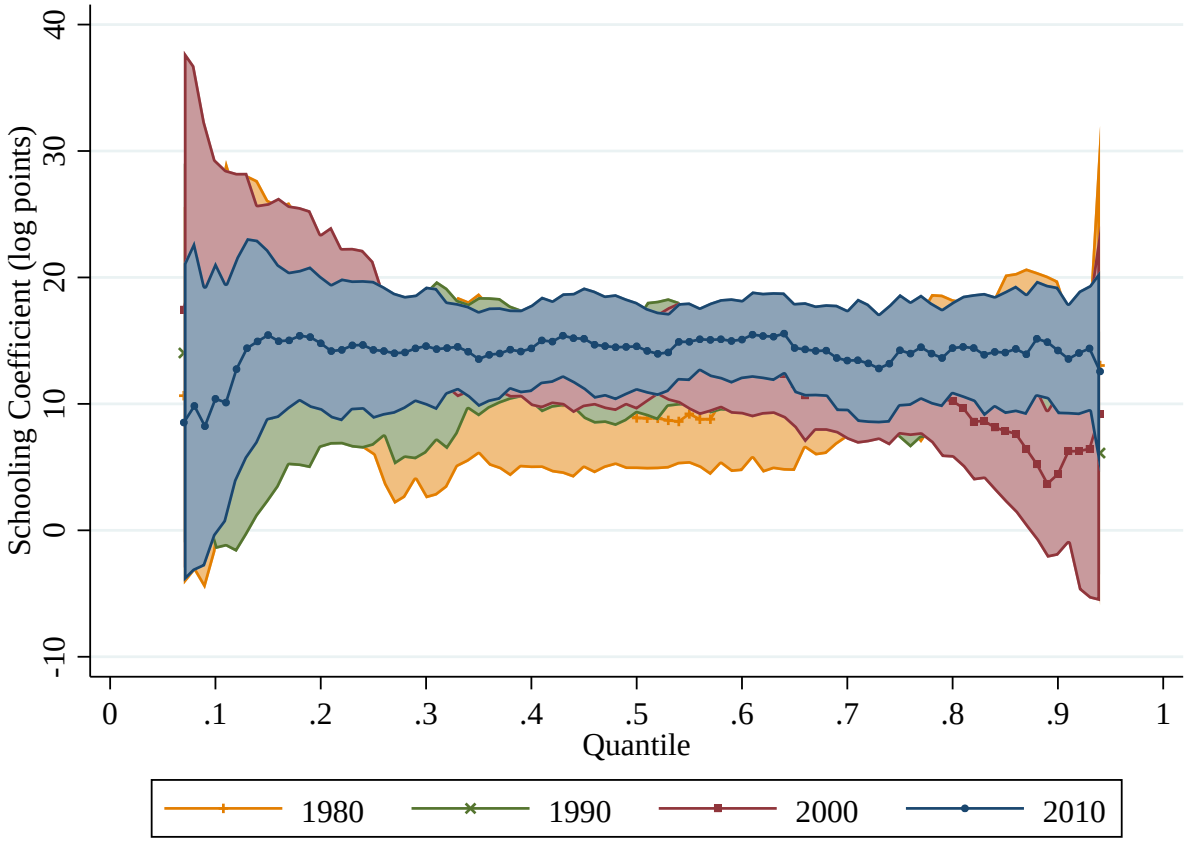
Notes: Figure reports quantile regression estimates of log weekly wages (self-reported) on education and a quadratic in experience for a grid of 99 quantiles from 0.02 to 0.98. Horizontal lines plot OLS estimates for each year, and robust 95% pointwise confidence intervals are plotted for the quantile regression estimates for each year. The data comes from the indicated decennial census year and consist of 40-49 year-old U.S. born black women with positive reported wages. The number of observations in each sample is 5,295, 8,215, 8,479, and 9,351 in 1980, 1990, 2000, and 2010, respectively.

FIGURE E7. QR Estimates of the Returns to Education for Hispanic Women



Notes: Figure reports quantile regression estimates of log weekly wages (self-reported) on education and a quadratic in experience for a grid of 99 quantiles from 0.03 to 0.97. Horizontal lines plot OLS estimates for each year, and robust 95% pointwise 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 U.S. born Hispanic women with positive reported wages. The number of observations in each sample is 1,326, 2,711, 3,645, and 5,616 in 1980, 1990, 2000, and 2010, respectively.

FIGURE E8. QR Estimates of the Returns to Education for Asian Women

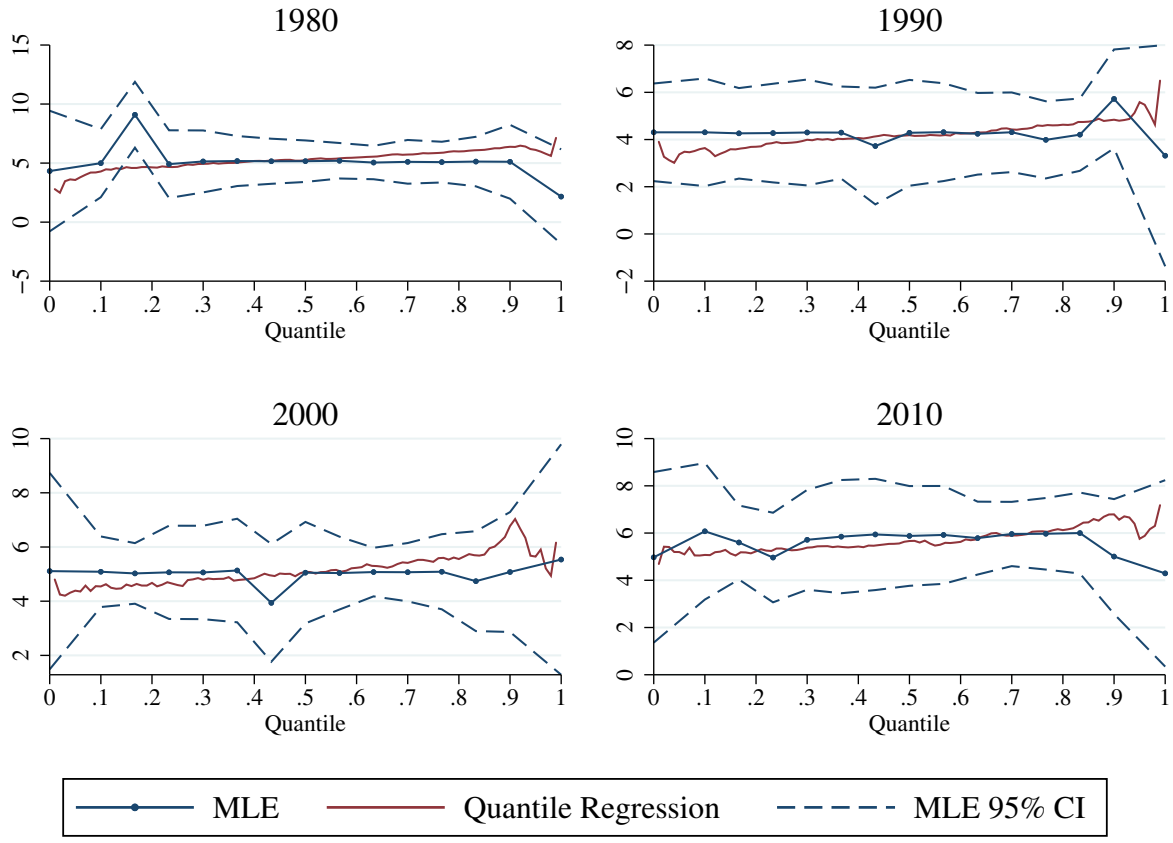


Notes: Figure reports quantile regression estimates of log weekly wages (self-reported) on education and a quadratic in experience for a grid of 99 quantiles from 0.06 to 0.94. Horizontal lines plot OLS estimates for each year, and robust 95% pointwise confidence intervals are plotted for the quantile regression estimates for each year. The data comes from the indicated decennial census year and consist of 40-49 year-old U.S. born women of Asian ethnicity with positive reported wages. The number of observations in each sample is 272, 451, 463, and 720 in 1980, 1990, 2000, and 2010, respectively.

Appendix F. Additional Maximum Likelihood Estimates of Mincer Equation

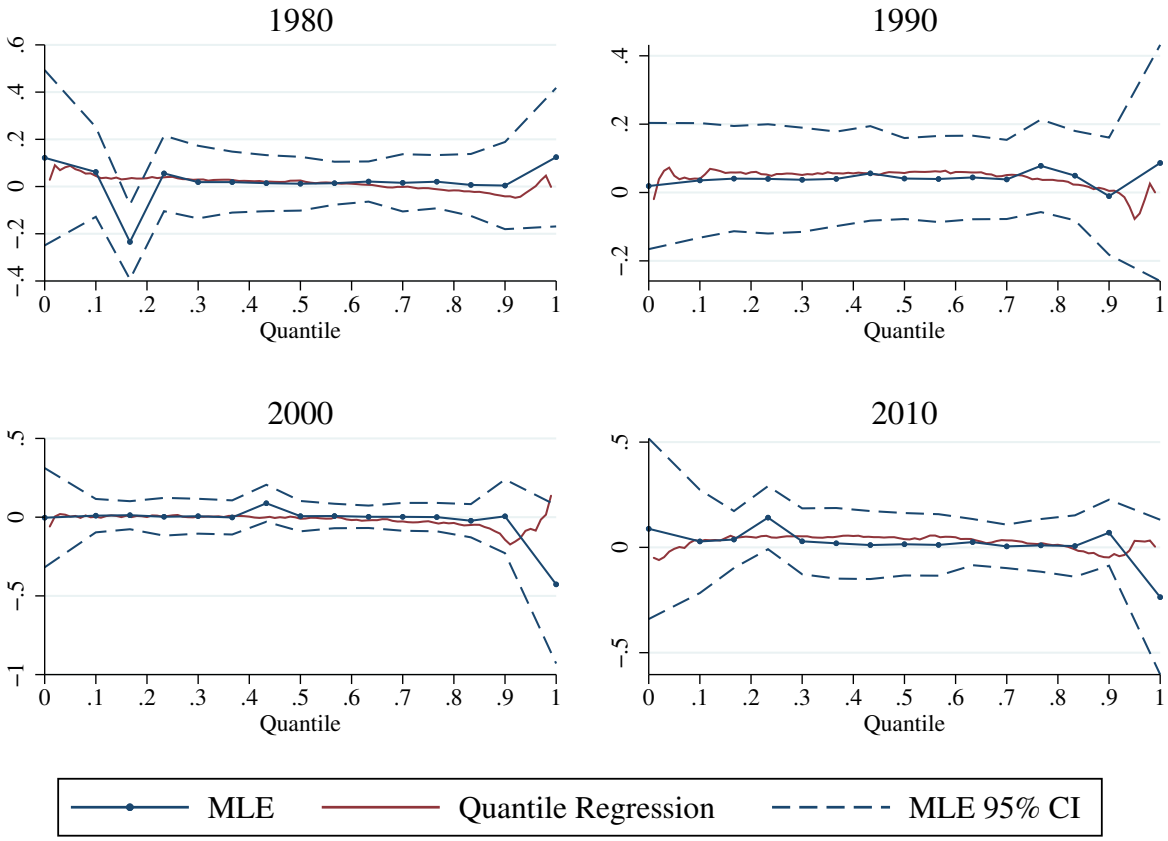
In this appendix, we plot ML estimates of equation (5.1) for the non-education terms in the model, consisting of the constant term $\beta_1(\cdot)$, the coefficient on experience $\beta_3(\cdot)$, and the coefficient on experience squared $\beta_4(\cdot)$. For each graph, we plot the quantile-regression estimate, the ML estimate, and the bootstrapped 95% confidence intervals associated with the ML estimates.

Figure F1. ML Estimates of Quantile Intercept in Log Wage Model



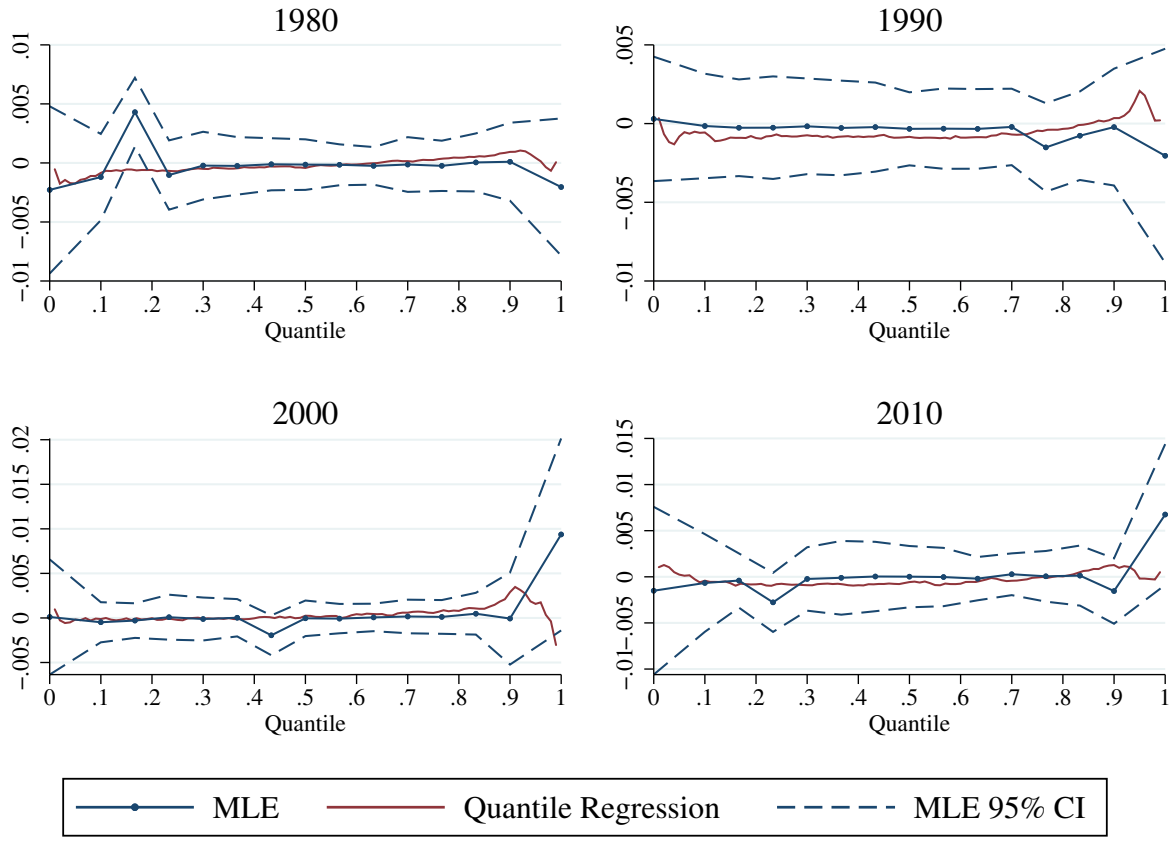
Notes: Graphs plot estimated intercept term using quantile regression (red lines) and the ML estimator described in the text (blue line). Dashed blue lines plot 95% pointwise confidence intervals. See notes to Figure E1 for further details.

FIGURE F2. ML Estimates of Experience Coefficient in Log Wage Model



Notes: Graphs plot experience coefficients estimated using quantile regression (red lines) and the ML estimator described in the text (blue line). Dashed blue lines plot 95% pointwise confidence intervals. See notes to Figure E1 for further details.

FIGURE F3. ML Estimates of Experience Quadratic Term in Log Wage Model



Notes: Graphs plot experience^2 coefficients estimated using quantile regression (red lines) and the ML estimator described in the text (blue line). Dashed blue lines plot 95% pointwise confidence intervals. See notes to Figure E1 for further details.

Appendix G. Proofs of Lemmas and Theorems

G.1. **Lemmas and Theorems in Section 2.**

Proof of Theorem 1. If there exist $\beta(\cdot)$ and $f(\cdot)$ which generate the same density $g(y|x,\beta(\cdot),f(\cdot))$ as the true parameters $\beta_0(\cdot)$ and $f_0(\cdot)$ then by applying a Fourier transformation and conditional on x,

$$\phi_\varepsilon(s)\int_0^1 \exp(isx^T\beta(\tau))d\tau = \phi_{\varepsilon 0}(s)\int_0^1 \exp(isx^T\beta_0(\tau))d\tau.$$

Denote $m(s)=\frac{\phi_{\varepsilon 0}(s)}{\phi_\varepsilon(s)}$. Without loss of generality, by Assumption 2, we can assume that x_1 is the continuous variable and the support of $x_1|x_{-1}$ contains an open neighborhood of 0. A Taylor expansion on both sides around $x_1=0$ gives us

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

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

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

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

$$\int_0^1 \beta_1(\tau)^k d\tau = \int_0^1 \beta_{0,1}(\tau)^k d\tau. \tag{G.2}$$

We now show that (G.2) implies that random variables β_1 and $\beta_{0,1}$ share the same distribution. The characteristic function of $\beta_1(\tau)$ can be written as

$$\phi_{\beta_1(\tau)}(s) = \sum_{k=0}^{\infty}\frac{(is)^k}{k!}\int_0^1 \beta_1(\tau)^k d\tau \tag{G.3}$$

Since $\beta_1(\tau)$ is bounded, $\int_0^1\beta_1(\tau)^k d\tau \leq M^k$ for some constant $M>0$. Therefore, $\left|\phi_{\beta_1(\tau)}(s)\right| \leq \sum_{k=0}^{\infty}\frac{|s|^k}{k!}M^k = \exp(M|s|)<\infty$ for any s, and the right-hand side of (G.3) is well defined. Combining (G.2) and (G.3), we have

$$\phi_{\beta_1(\tau)}(s) = \phi_{\beta_{0,1}(\tau)}(s),$$

and thus β_1 and $\beta_{0,1}$ share the same distribution almost everywhere as two random variables. Thus there exists a measurable one-to-one reordering mapping $\pi:[0,1]\mapsto[0,1]$. Then $\beta_1(\pi(\tau))=\beta_{0,1}(\tau)$ almost everywhere, and $\int h(\tau)d\tau = \int h(\pi(\tau))d\tau$ for all integrable functions $h(\cdot)$ defined on [0,1].

Now consider (G.1) again. Dividing both sides by $(is)^k/k!$, we have, for all $k\geq 0$

$$\begin{aligned}\int_0^1 \exp(isx_{-1}^T\beta_{-1}(\pi(\tau)))\beta_1(\pi(\tau))^k d\tau &= \int_0^1 \exp(isx_{-1}^T\beta_{-1}(\tau))\beta_1(\tau)^k d\tau \\ &= m(s)\int_0^1 \exp(isx_{-1}^T\beta_{0,-1}(\tau))\beta_{0,1}(\tau)^k d\tau. \end{aligned}\tag{G.4}$$

Consider the first-order terms of s in (G.4). Since both f and f_0 satisfy $\int_{-\infty}^{\infty} \varepsilon f(\varepsilon) = 0$, we have $m'(0) = 0$, and hence the coefficients for the first-order terms of s in (G.4) can be written as

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

By Assumption 2.2, $\beta_{0,1}(\tau)^k, k \geq 0$ is a functional basis of $L^2[0,1]$, therefore

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

almost everywhere and everywhere for $\tau \in [0,1]$ by invoking the continuity of $\beta_{-1}(\cdot)$ and $\beta_{0,-1}(\cdot)$. Hence $E[x_{-1}x_{-1}^T]\beta_{-1}(\pi(\tau)) = E[x_{-1}x_{-1}^T]\beta_{0,-1}(\tau)$ almost everywhere for $\tau \in [0,1]$. By Assumption 2.2, $E[xx^T]$ is non-singular. Ergo $E[x_{-1}x_{-1}^T]$ is also non-singular. Multiplying both sides of $E[x_{-1}x_{-1}^T]\beta_{-1}(\pi(\tau)) = E[x_{-1}x_{-1}^T]\beta_{0,-1}(\tau)$ by $E[x_{-1}x_{-1}^T]^{-1}$,we get $\beta_{-1}(\pi(\tau)) = \beta_{0,-1}(\tau)$ for almost all $\tau \in [0,1]$.

For any x, $x^T\beta(\pi(\tau)) = x^T\beta_0(\tau)$. Since conditional on x, $x^T\beta(\tau)$ has the same distribution as $x^T\beta(\pi(\tau))$, $x^T\beta(\tau)$ has the same distribution as $x^T\beta_0(\tau)$. By the monotonicity of $x^T\beta(\tau)$ and $x^T\beta_0(\tau)$, they must equal each other at almost all τ. Since $E[xx^T]$ is non-singular, $\beta(\tau) = \beta_0(\tau)$ almost everywhere. Consequently, $\phi_\varepsilon(s) = \phi_{\varepsilon 0}(s)$, and $f(\varepsilon) = f_0(\varepsilon)$ almost everywhere. □

Proof of Lemma 1. We first prove the case when $W(x_1) = [x_1, x_1^2]^T$ and then describe how the proof can be generalized to the case where $W(x_1)$ is a p^{th}-order polynomial. If there exist $\beta(\cdot)$ and $f(\cdot)$ which generate the same density as the true parameters $\beta_0(\cdot)$ and $f_0(\cdot)$ then by applying a Fourier transformation and conditional on x,

$$\phi_{x\beta}(s|x)\phi_\varepsilon(s) = \phi_{x\beta_0}(s|x)\phi_{\varepsilon_0}(s). \tag{G.5}$$

Then

$$\phi_{x\beta}(s|x) = m(s)\phi_{x\beta_0}(s|x), \tag{G.6}$$

where $m(s) = \frac{\phi_{\varepsilon_0}(s)}{\phi_\varepsilon(s)}$ is a function depending only on s. Let $\beta_w(\tau) = [\beta_{x_1}, \beta_{x_1^2}]^T$ and $\beta_{0,w} = [\beta_{0,x_1}, \beta_{0,x_1^2}]^T$ denote the subvectors of β and β_0 associated with $W(x_1) = [x_1, x_1^2]^T$. Expanding (G.6) around $s = 0$, we get

$$\begin{aligned}
&\sum_{k=0}^{\infty} \int_0^1 \frac{(is)^k}{k!}[(x_1, x_1^2)\beta_w(\tau) + x_{-w}^T\beta_{-w}(\tau)]^k d\tau \\
=&\sum_{k=0}^{\infty} a_k s^k \sum_{k=0}^{\infty} \int_0^1 \frac{(is)^k}{k!}[(x_1, x_1^2)\beta_{0,w}(\tau) + x_{-w}^T\beta_{0,-w}(\tau)]^k d\tau, \\
=&\sum_{k=0}^{\infty} s^k \Big[\sum_{l=0}^{k} a_{k-l} \int_0^1 \frac{i^l}{l!}[(x_1, x_1^2)\beta_{0,w}(\tau) + x_{-w}^T\beta_{0,-w}(\tau)]^l d\tau \Big]
\end{aligned} \tag{G.7}$$

where $\sum_{k=0}^{\infty} a_k s^k$ is a Taylor expansion of $m(s)$ around $s = 0$. Since both ε and ε_0 have zero mean, we have $a_0 = 1$, and $a_1 = 0$. Comparing the coefficients on both sides of (G.7) for s^k, we have

$$\int_0^1 \frac{i^k}{k!}[(x_1, x_1^2)\beta_w(\tau) + x_{-w}^T\beta_{-w}(\tau)]^k d\tau = \sum_{l=0}^{k} a_{k-l} \int_0^1 \frac{i^l}{l!}[(x_1, x_1^2)\beta_{0,w}(\tau) + x_{-w}^T\beta_{0,-w}(\tau)]^l d\tau \tag{G.8}$$

holding for any k, x_1 and x_{-w}. Comparing both sides of (G.8) for any k and the coefficients for x_1^{2k}, we have

$$\int_0^1 \frac{i^k}{k!}\big(\beta_{x_1^2}(\tau)\big)^k d\tau = \int_0^1 \frac{i^k}{k!}\big(\beta_{0,x_1^2}(\tau)\big)^k d\tau.$$

Using the same argument as in the proof for Theorem 1 through the characteristic functions, the two random variables $\beta_{x_1^2}(\tau)$ and $\beta_{0,x_1^2}(\tau)$ share the same distribution, and there exists a measurable reordering mapping $\pi : [0,1] \mapsto [0,1]$ such that $\beta_{x_1^2}(\pi(\tau)) = \beta_{0,x_1^2}(\tau)$ almost everywhere. Comparing both sides of (G.8) for any k and the coefficients for x_1^{2k-1}, we have

$$\int_0^1 \frac{i^k}{k!}\big(\beta_{x_1^2}(\tau)\big)^{k-1}\beta_{x_1}(\tau)d\tau = \int_0^1 \frac{i^k}{k!}\big(\beta_{0,x_1^2}(\tau)\big)^{k-1}\beta_{0,x_1}(\tau)d\tau = \int_0^1 \frac{i^k}{k!}\big(\beta_{x_1^2}(\pi(\tau))\big)^{k-1}\beta_{0,x_1}(\tau)d\tau,$$

where we used the fact that $\beta_{x_1^2}(\pi(\tau)) = \beta_{0,x_1^2}(\tau)$. As argued above in the proof of the previous lemma, by Assumption 3, we know that $\big(\beta_{x_1^2}(\tau)\big)^{k-1}$ for $k \geq 1$ forms a functional basis of $L^2[0,1]$, implying that

$$\beta_{x_1}(\pi(\tau)) = \beta_{0,x_1}(\tau)$$

almost everywhere. Comparing both sides of (G.8) for any k and the coefficients for x_1^{2k-2}, we have

$$\begin{aligned}&\frac{i^k}{k!}\int_0^1 \big(\beta_{x_1^2}(\tau)\big)^{k-2}\big(\beta_{x_1}(\tau)\big)^2 + \big(\beta_{x_1^2}(\tau)\big)^{k-1}x_{-w}^T\beta_{-w}(\tau)d\tau \\ =&\frac{i^k}{k!}\int_0^1 \big(\beta_{0,x_1^2}(\tau)\big)^{k-2}\big(\beta_{0,x_1}(\tau)\big)^2 + \big(\beta_{0,x_1^2}(\tau)\big)^{k-1}x_{-w}^T\beta_{0,-w}(\tau)d\tau \end{aligned} \tag{G.9}$$

where we used the fact that $a_1 = 0$ and thus for a fixed k, the only l on the right-hand side of (G.8) that can generate a nonzero coefficient for x_1^{2k-2} is $l = 0$. Since we already proved that $\beta_{x_1}(\pi(\tau)) = \beta_{0,x_1}(\tau)$ and $\beta_{x_1^2}(\pi(\tau)) = \beta_{0,x_1^2}(\tau)$ almost everywhere, (G.9) can be rewritten as

$$\begin{aligned}\int_0^1 \big(\beta_{x_1^2}(\tau)\big)^{k-1}x_{-w}^T\beta_{-w}(\tau)d\tau &= \int_0^1 \big(\beta_{0,x_1^2}(\tau)\big)^{k-1}x_{-w}^T\beta_{0,-w}(\tau)d\tau \\ &= \int_0^1 \big(\beta_{x_1^2}(\pi(\tau))\big)^{k-1}x_{-w}^T\beta_{0,-w}(\tau)d\tau.\end{aligned}$$

Again, using the fact that $\big(\beta_{x_1^2}(\tau)\big)^{k-1}, k \geq 1$ forms a functional basis of $L^2[0,1]$, we have for any x_{-w}

$$x_{-w}^T\beta_{-w}(\pi(\tau)) = x_{-w}^T\beta_{0,-w}(\tau)$$

almost everywhere in $\tau \in [0,1]$. Following the same argument as in Theorem 1, we know that there is sufficient variation in x_{-w} such that $x_{-w}^T\beta_{-w}(\pi(\tau)) = x_{-w}^T\beta_{0,-w}(\tau)$ implies $\beta_{-w}(\pi(\tau)) = \beta_{0,-w}(\tau)$ almost everywhere. By monotonicity of $x^T\beta(\tau)$ and $x\beta_0^T(\tau)$, we have $\pi(\tau) = \tau$ almost everywhere, and thus $\beta(\tau) = \beta_0(\tau)$ and $f(\varepsilon) = f_0(\varepsilon)$ almost everywhere.

The argument for the case of $W(x_1)$ being a p^{th} order polynomial is very similar to the argument above. We start from a Taylor expansion similar to (G.7). Then we compare the coefficients for each term s^k and get

$$\begin{aligned}&\int_0^1 \frac{i^k}{k!}[(x_1, \cdots, x_1^p)\beta_w(\tau) + x_{-w}^T\beta_{-w}(\tau)]^k d\tau \\ =&\sum_{l=0}^k a_{k-l}\int_0^1 \frac{i^l}{l!}[(x_1, \cdots, x_1^p)\beta_{0,w}(\tau) + x_{-w}^T\beta_{0,-w}(\tau)]^l d\tau.\end{aligned} \tag{G.10}$$

Using the fact that for each $k \geq 1$, the coefficients for x_1^{kp} on both sides of (G.10) must equal each other, we can show that there exists a reordering mapping $\pi(\tau)$ such that $\beta_{x_1^p}(\pi(\tau)) = \beta_{0,x_1^p}(\tau)$. Using the fact that for each k, the coefficients for $x_1^{kp-l}, 1 \leq l \leq k-1$ on both sides of (G.10) must equal each other, we can show that $\beta_{x_1^{p-l}}(\pi(\tau)) = \beta_{0,x_1^{p-l}}(\tau)$ almost everywhere. Because the coefficients for $x_1^{k(p-1)}$ must equal each other on both sides of (G.10), we can also show that $x_{-w}^T \beta_{-w}(\pi(\tau)) = x_{-w}^T \beta_{0,-w}(\tau)$ almost everywhere. The rest follows the same argument as in the proof for Theorem 1, and we have $\beta(\tau) = \beta_0(\tau)$ almost everywhere for all $\tau \in [0,1]$ and $f(\varepsilon) = f_0(\varepsilon)$ almost everywhere for all $\varepsilon \in \mathbb{R}$. □

G.2. **Lemmas and Theorems in Section 3.** The following lemmas are used in the proofs of Lemmas 2 and 3.

Lemma 6. *The space $M[B_1 \times B_2 \times B_3 ... \times B_{d_x}]$ is a compact and complete space under L^p for any $p \geq 1$.*

Proof of Lemma 6. For bounded monotonic functions, pointwise convergence is equivalent to uniform convergence, making a space of bounded monotonic functions compact under any L^p norm for $p \geq 1$. Hence the product space $B_1 \times B_2 \times ... \times B_{d_x}$ is compact. It is complete since the L^p functional space is complete and the limit of monotonic functions is still monotonic. □

Lemma 7 (Donskerness of Θ)**.** *The set of functions*

$$\mathcal{G} = \{h(y, x, \beta(\cdot), \sigma) := \log(g(y|x, \beta(\cdot), \sigma))|(\beta(\cdot), \sigma) \in \Theta\}$$

is μ-Donsker, where μ is the joint PDF of (y, x).

Proof. By Theorem 2.7.5 of Van der Vaart and Wellner (1996), the bracketing number $N_{[]}$ of the space of uniformly bounded monotone functions $\mathcal{F}$ satisfies

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

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

$$|q(y, x, \theta_1) - q(y, x, \theta_2)| \leq ||\theta_1 - \theta_2||_2 w(y, x). \tag{G.11}$$

$$E_Q[|w(y, x)|^2] < \infty, \tag{G.12}$$

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

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

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

In particular, let $q = \log g$ and $Q = \mu$, where μ is the joint PDF of (x, y). By Assumption 4.5, equation (G.11) holds with $w(y|x) := \int_0^1 |y - x^T \beta(\tau)|^\gamma d\tau$. Equation (G.12) is satisfied by Assumption 4.3. Hence, $\mathcal{G}$ is μ-Donsker. □

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

(1) The parameter space $\Theta = M \times \Sigma$ is compact.
(2) Global identification holds, i.e., there exists no other $\theta' = (\beta', \sigma') \in \Theta$ such that $E[\log \int_0^1 f(y - x\beta'(\tau)|\sigma')d\tau] = E[\log \int_0^1 f(y - x\beta_0(\tau)|\sigma_0)d\tau]$.
(3) The objective function $E[\log \int_0^1 f(y - x\beta'(\tau)|\sigma')d\tau]$ is continuous for all $\theta' = (\beta', \sigma') \in \Theta$.
(4) Stochastic equicontinuity of $E_n[\log \int_0^1 f(y - x^T\beta(\tau)|\sigma)]$, with $\theta \in \Theta$.

Condition 1 is established by Lemma 6. Condition 2 is provided by Lemma 1. Condition 3 holds under Assumption 4. For the proof of point 4, see Lemma 7 above. Therefore, the ML estimator defined herein is consistent. □

The following lemma establishes mild ill-posedness (Assumption 5.1) under the special case of piecewise-constant sieves.

Lemma 8 (Sufficient Condition for Mild Ill-posedness)**.** *If the function f satisfies Assumptions 1, 3, 4, and 6 with degree $\lambda > 0$ then*

(1) *The minimum eigenvalue of I, denoted as $v(I)$, satisfies $\frac{1}{J^\lambda} \precsim v(I)$.*
(2) *For any θ, $\frac{1}{J^\lambda} \precsim \sup_{p \in \Theta_J - \theta, p \neq 0} \frac{||p||_d}{||p||}$ where $||p||_d := |p'Ip|^{1/2}$.*

Proof. Suppose f satisfies the discontinuity condition in Assumption 6 with degree $\lambda > 0$, and without loss of generality, assume $c_\delta = 1$. Denote $l_J = (f_{\tau_1}, f_{\tau_2}, ..., f_{\tau_J}, g_\sigma)$. For any $p_J \in \Theta_J - \theta$, $p_J I p_J' = E[\int_{\mathbb{R}} \frac{(l_J p_J')^2}{g} dy] \geq CE[(l_J p_J')^2]$ for some constant $C > 0$ since g is bounded from above.

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

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

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

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

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

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

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

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

By the discontinuity assumption,

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

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

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

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

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

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

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

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

Proof of Lemma 3. By Lemma 7, the set of log likelihood functions indexed by $\widehat{\theta}_n\in\Theta$ is Donsker such that the sample-average log likelihood converges uniformly to its population counterpart:

$$E[-\log g(y|x,\widehat{\theta}_n)]\leq E_n[-\log g(y|x,\widehat{\theta}_n)]+o_p(1).$$

By Chen (2007), there exists a $\theta_n^*\to\theta_0$ as $J_n\to\infty$ where $\theta_n^*\in\Theta_{J_n}^r$ given that $d_2(\theta_0,\Theta_{J_n}^r)=O(J_n^{-\min(p,r)})$, denoting the degree of smoothness of $\beta_0(\cdot)$ as p. Because $\widehat{\theta}_n$ is the minimizer of the negative log likelihood,

$$E_n[-\log g(y|x,\widehat{\theta}_n)]\leq E_n[-\log g(y|x,\theta_n^*)].$$

Again, by uniform convergence,

$$\begin{aligned}E_n[-\log g(y|x,\theta_n^*)]&\leq E[-\log g(y|x,\theta_n^*)]+o_p(1)\\&\leq E[-\log g(y|x,\theta_0)]+o_p(1)\end{aligned}$$

where the last step used the continuity of the population log-likelihood function around θ_0. Since Θ is compact, by identification (i.e., Theorem 1), we have $\widehat{\theta}_n\xrightarrow{p}\theta_0$ as $J_n\to\infty$. □

Proof of Lemma 4. By Lemma 3, we know that sieve-ML estimators for β_0 and σ_0 are consistent, i.e., $||\widehat{\sigma}-\sigma_0||\xrightarrow{p}0$ and $||\widehat{\beta}_J-\beta_0||_2\xrightarrow{p}0$. MLE by definition implies that

$$E_n[\log(g(y|x,\widehat{\beta}_J,\sigma))]\geq E_n[\log(g(y|x,\beta_J^*,\sigma_0))]$$

where by construction of the sieve, there exists a β_J^* such that $||\beta_J^* - \beta_0||_2 \leq CJ_n^{-r-1}$ for some generic constant $C > 0$. Therefore, $||(\widehat{\beta}_J, \widehat{\sigma}) - (\beta_J^*, \sigma_0)|| \xrightarrow{p} 0$ as $J_n \to \infty$.

By Lemma 7, $\mathcal{G} = \{h(y, x, \beta(\cdot), \sigma) := \log(g(y|x, \beta(\cdot), \sigma))|(\beta(\cdot), \sigma) \in \Theta\}$ is Donsker. Thus by stochastic equicontinuity,

$$\begin{aligned} &E_n[\log(g(y|x, \widehat{\beta}_J, \widehat{\sigma}))] - E[\log(g(y|x, \widehat{\beta}_J, \widehat{\sigma}))] \\ &= E_n[\log(g(y|x, \beta_J^*, \sigma_0))] - E[\log(g(y|x, \beta_J^*, \sigma_0))] + o_p(1/\sqrt{n}), \end{aligned}$$

implying that

$$\begin{aligned} &E[\log(g(y|x, \widehat{\beta}_J, \widehat{\sigma}))] - E[\log(g(y|x, \beta_J^*, \sigma_0))] \\ &= E_n[\log(g(y|x, \widehat{\beta}_J, \widehat{\sigma}))] - E_n[\log(g(y|x, \beta_J^*, \sigma_0))] - o_p(1/\sqrt{n}) \geq -o_p(1/\sqrt{n}). \end{aligned}$$

Define $G_n := \sqrt{n}(E_n - E)$. By the maximal inequality, for any $\delta > 0$,

$$E\left[\max_{||\beta_J - \beta_J^*||_2 < \delta} |G_n \log g(y|x, \widehat{\beta_J}, \widehat{\sigma}) - G_n \log g(y|x, \beta_J^*, \sigma_0)|\right] \leq K \int_0^\delta \sqrt{\log N(r, M, ||\cdot||_2)} dr\Bigg],$$

where $N(r, M, ||\cdot||_2)$ is the covering number of r balls on M, the space of β, and K is a generic constant. Since M is a bounded and co-monotone space ($x\beta(\tau)$ is monotone in τ for all $x \in \mathcal{X}$), $N(r, M, ||\cdot||_2) < \delta^{d_x}$. Therefore,

$$E\left[\max_{||\beta_J - \beta_J^*||_2 < \delta} |G_n \log g(y|x, \widehat{\beta_J}, \widehat{\sigma}) - G_n \log g(y|x, \beta_J^*, \sigma_0)|\right] \leq \delta\sqrt{-\log \delta},$$

where $\delta = \max(||\widehat{\sigma} - \sigma||, ||\widehat{\beta_J} - \beta_J^*||)$. Consequently,

$$\begin{aligned} &E[\log(g(y|x, \widehat{\beta}_J, \widehat{\sigma}))] - E[\log(g(y|x, \beta_J^*, \sigma_0))] \\ &= E_n[\log(g(y|x, \widehat{\beta}_J, \widehat{\sigma}))] - E_n[\log(g(y|x, \beta_J^*, \sigma_0))] - O_p(\frac{\delta\sqrt{-\log \delta}}{\sqrt{n}}) \geq -O_p(\frac{\delta\sqrt{-\log \delta}}{\sqrt{n}}). \end{aligned}$$

By consistency of $(\widehat{\beta_J}, \widehat{\sigma}), \delta \to_p 0$.

Since $E[\log g(y|x, \beta, \sigma)]$ is maximized at (β_0, σ_0), the Hadamard derivative of $E[\log g(y|x, \beta_0, \sigma_0)]$ with respect to $\beta \in \Theta$ is 0. By Assumption 4.2, the $\log g(\cdot|\cdot, \cdot, \cdot)$ function is twice differentiable with bounded derivatives up to the second order. Therefore, for some generic constant $C_1 > 0$,

$$E[\log g(y|x, \beta_J^*, \sigma_0)] - E[\log g(y|x, \beta_0, \sigma_0)] \geq -C_1||\beta_J^* - \beta_0||_2^2 \geq -C_1C^2J_n^{-2r-2} = O(\frac{1}{n}).$$

Then

$$\begin{aligned} &E[\log g(y|x, \widehat{\beta}_J, \widehat{\sigma})] - E[\log g(y|x, \beta_0, \sigma_0)] \\ =&E[\log g(y|x, \widehat{\beta}_J, \widehat{\sigma})] - E[\log g(y|x, \beta_J^*, \sigma_0)] \\ &+ E[\log g(y|x, \beta_J^*, \sigma_0)] - E[\log g(y|x, \beta_0, \sigma_0)] \\ \geq& -O_p(\frac{\delta\sqrt{-\log \delta}}{\sqrt{n}}) - C_1C^2J_n^{-2r-2} = -O_p(\frac{\delta\sqrt{-\log \delta}}{\sqrt{n}}) \end{aligned}$$

where the last step used the assumption that $\frac{J_n^{2r+2}}{n} \to \infty$.

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

$$E_x[||z(y|x)||_1^2] \leq D(g(\cdot|\beta_0, \sigma_0)||g(\cdot|\widehat{\beta}_J, \widehat{\sigma})) \tag{G.13}$$

$$\leq 2(E[\log(g(y|x, \beta_0, \sigma_0))] - E[\log(g(y|x, \widehat{\beta}_J, \widehat{\sigma}))]) = O_p(\frac{\delta\sqrt{-\log\delta}}{\sqrt{n}}),$$

where $D(P|Q)$ is the K-L divergence between two probability distribution P and Q.

Now consider the characteristic functions of $x\widehat{\beta}_J(\tau)$ and $x\beta_0(\tau)$ conditional on x and given that $\tau \sim U[0,1]$,

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

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

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

$$\left| m(s) - \frac{\phi_{x\widehat{\beta}_J}(s|x)}{\phi_{x\beta_0}(s|x)} \right| \leq \frac{||z(y|x)||_1}{|\phi_{x\beta_0}(s|x)\phi_\varepsilon(s|\sigma_0)|}. \tag{G.14}$$

Plugging in (G.14) back into (G.13), we have

$$E_x\left[\left| m(s) - \frac{\phi_{x\widehat{\beta}_J}(s|x)}{\phi_{x\beta_0}(s|x)} \right|^2 \right] \leq E_x\left[\frac{||z(y|x)||_1^2}{|\phi_{x\beta_0}(s|x)\phi_\varepsilon(s|\sigma_0)|^2} \right] \tag{G.15}$$

$$\leq E_x\left[||z(y|x)||_1^2 \right] \frac{s^2}{C^2\phi_\varepsilon(s|\sigma_0)^2} = o_p\left(E_x[||z(y|x)||_1^2] \frac{1}{\phi_\varepsilon(s|\sigma_0)^2} \right),$$

where in the last step we require that $s \in [-l, l]$ for some $l > 0$ such that $|\phi_{x\beta_0}(s|x)|$ is bounded away from zero. Using the fact that for any random variable a and any number b, $Var(a) \leq E[(a-b)^2]$, we have that

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

Inequality (G.15) then implies that

$$Var_x\left(\frac{\phi_{x\widehat{\beta}_J}(s|x)}{\phi_{x\beta_0}(s|x)} \right) \precsim_p E_x[||z(y|x)||_1^2] \frac{1}{\phi_\varepsilon(s|\sigma_0)}. \tag{G.16}$$

Applying Assumption 7, inequality (G.16) implies that

$$E_x\left[\left| \frac{\phi_{x\widehat{\beta}_J}(s|x) - \phi_{x\beta_0}(s|x)}{\phi_{x\beta_0}(s|x)} \right|^2 \right] = O_p\left(E_x[||z(y|x)||_1^2] \frac{1}{\phi_\varepsilon(s|\sigma_0)^2} \right). \tag{G.17}$$

We can rewrite $m(s) - \frac{\phi_{x\widehat{\beta}_J}(s|x)}{\phi_{x\beta_0}(s|x)}$ as $(m(s)-1) - \frac{\phi_{x\widehat{\beta}_J}(s|x) - \phi_{x\beta_0}(s|x)}{\phi_{x\beta_0}(s|x)}$. Using that $\frac{1}{2}a^2 - b^2 \leq (a-b)^2$ for any $a, b \in \mathbb{R}$, we can bound inequality (G.17) from below such that

$$\frac{1}{2}E_x\left[|m(s)-1|^2\right] - E\left[\left|\frac{\phi_{x\widehat{\beta}_J}(s|x) - \phi_{x\beta_0}(s|x)}{\phi_{x\beta_0}(s|x)}\right|^2\right] \leq E_x\left[\left|m(s) - \frac{\phi_{x\widehat{\beta}_J}(s|x)}{\phi_{x\beta_0}(s|x)}\right|^2\right] \tag{G.18}$$

$$= O_p\left(E_x[||z(y|x)||_1^2]\frac{1}{\phi_\varepsilon(s|\sigma_0)^2}\right). \tag{G.19}$$

Combining (G.18) with (G.17),

$$E_x\left[|m(s)-1|^2\right] = O_p\left(E_x[||z(y|x)||_1^2]\frac{1}{\phi_\varepsilon(s|\sigma_0)^2}\right), \tag{G.20}$$

or, equivalently, for any $s \in [-l, l]$ where l is some fixed constant,

$$|\phi_\varepsilon(s|\sigma_0) - \phi_\varepsilon(s|\widehat{\sigma})|^2 = O_p\left(E_x[||z(y|x)||_1^2]\right). \tag{G.21}$$

Applying Assumption 4.5 along with (G.21), it follows that $||\widehat{\sigma} - \sigma_0||^2 = O_p\left(E_x[||z(y|x)||_1^2]\right) = O_p(\frac{\delta\sqrt{-\log\delta}}{\sqrt{n}})$.

If $||\widehat{\sigma} - \sigma|| > ||\widehat{\beta_J} - \beta_J^*||$, then $\delta = ||\widehat{\sigma} - \sigma||$, and it follows that $||\widehat{\sigma} - \sigma_0||^2 = O_p(\frac{\log n}{n})$.

If $||\widehat{\sigma} - \sigma|| \leq ||\widehat{\beta_J} - \beta_J^*||$, then $\delta = ||\widehat{\beta}_J - \beta_J^*||$, and $||\widehat{\sigma} - \sigma_0||^2 = O_p(\frac{\delta\sqrt{-\log\delta}}{\sqrt{n}})$.

Therefore, $||\widehat{\sigma} - \sigma_0||^2 = O_p\left(\max\left(\frac{\log n}{n}, \frac{||\widehat{\beta}_J - \beta_J^*||\sqrt{-\log||\widehat{\beta}_J - \beta_J^*||}}{\sqrt{n}}\right)\right)$. □

The following lemma will be instrumental in proving asymptotic normality.

Lemma 9. *Under Assumptions 1, 3, 4, 5.1 (the mild ill-posed case), 7, and* $\frac{J_n^{2r+2}}{n} \to \infty$, $||\widehat{\beta}_J - \beta_J^*|| = O_p\left(\left(J_n^{2\lambda}\frac{\log^{\frac{1}{2}} n}{n^{\frac{1}{2}}}\right)^{\frac{1}{2\lambda+1}}\right)$.

Proof. Our argument follows the proof of Lemma 4. Let $z(y|x) := g(y|x, \widehat{\beta}_J, \sigma_0) - g(y|x, \beta_J^*, \sigma_0)$. By the Scheffe Theorem and Pinsker's Inequality,

$$E_x[||z(y|x)||_1^2] \leq D(g(\cdot|\widehat{\beta}_J, \sigma_0)||g(\cdot|\beta_J^*, \sigma_0)) \leq 2(E[\log(g(y|x, \beta_J^*, \sigma_0))] - E[\log(g(y|x, \widehat{\beta}_J, \sigma_0))]), \tag{G.22}$$

where $D(P|Q)$ is the K-L divergence between two probability distribution P and Q.

By the maximal inequality, for any $\delta > 0$,

$$E\left[\max_{||\beta_J - \beta_J^*||_2 < \delta} |G_n \log g(y|x, \beta_J^*, \sigma_0) - G_n \log g(y|x, \beta_J, \sigma_0)|\right] \leq K\int_0^\delta \sqrt{\log N(r, M, ||\cdot||_2)}dr\Bigg],$$

where $N(r, M, ||\cdot||_2)$ is the covering number of r balls on M, the space of β. K is a generic constant. Since M is a bounded and co-monotone space ($x\beta(\tau)$ is monotone in τ for all $x \in \mathcal{X}$), $N(r, M, ||\cdot||_2) < \delta^{d_x}$. Therefore,

$$E\left[\max_{||\beta_J - \beta_J^*||_2 < \delta} |G_n \log g(y|x, \beta_J^*, \sigma_0) - G_n \log g(y|x, \beta_J, \sigma_0)|\right] \leq \delta\sqrt{-\log\delta}.$$

Hence,

$$|G_n \log g(y|x,\beta_J^*,\sigma_0) - G_n \log g(y|x,\widehat{\beta}_J,\sigma_0)| = O_p\big(\widehat{\delta}\sqrt{-\log\widehat{\delta}}\big),$$

where $\widehat{\delta} := ||\beta_J^* - \widehat{\beta}_J||$. Using a similar argument as in the proof of Lemma 4, we can show that $||\widehat{\sigma} - \sigma_0||^2 = O_p(\frac{1}{\sqrt{n}}\widehat{\delta}\sqrt{-\log\widehat{\delta}})$. Thus,

$$\begin{aligned}
&E[\log(g(y|x,\beta_J^*,\sigma_0))] - E[\log(g(y|x,\widehat{\beta}_J,\sigma_0))] \\
=&\frac{1}{\sqrt{n}}G_n[\log(g(y|x,\beta_J^*,\sigma_0))] - \frac{1}{\sqrt{n}}G_n[\log(g(y|x,\widehat{\beta}_J,\sigma_0))] \\
&+ E_n[\log(g(y|x,\beta_J^*,\sigma_0))] - E_n[\log(g(y|x,\widehat{\beta}_J,\sigma_0))] \qquad \text{(G.23)}
\end{aligned}$$

The first term on the right-hand side of (G.23), $\frac{1}{\sqrt{n}}G_n[\log(g(y|x,\beta_J^*,\sigma_0))] - \frac{1}{\sqrt{n}}G_n[\log(g(y|x,\widehat{\beta}_J,\sigma_0))]$, is $O_p(\frac{1}{\sqrt{n}}\widehat{\delta}\sqrt{-\log\widehat{\delta}})$. For the second term on the right-hand side of (G.23), we have

$$\begin{aligned}
&E_n[\log(g(y|x,\beta_J^*,\sigma_0))] - E_n[\log(g(y|x,\widehat{\beta}_J,\sigma_0))] \\
=&E_n[\log(g(y|x,\beta_J^*,\sigma_0))] - E_n[\log(g(y|x,\widehat{\beta}_J,\widehat{\sigma}))] \\
&+ E_n[\log(g(y|x,\widehat{\beta}_J,\widehat{\sigma}))] - E_n[\log(g(y|x,\widehat{\beta}_J,\sigma_0))].
\end{aligned}$$

We know that $E_n[\log(g(y|x,\beta_J^*,\sigma_0))] - E_n[\log(g(y|x,\widehat{\beta}_J,\widehat{\sigma}))] \leq 0$ by the first-order condition. We also have that

$$E_n[\log(g(y|x,\widehat{\beta}_J,\widehat{\sigma}))] - E_n[\log(g(y|x,\widehat{\beta}_J,\sigma_0))] = O_p(||\widehat{\sigma} - \sigma_0||^2) = O_p(\frac{1}{\sqrt{n}}\widehat{\delta}\sqrt{-\log\widehat{\delta}}).$$

Combining the results on different terms in (G.23), we have

$$E[\log(g(y|x,\beta_J^*,\sigma_0))] - E[\log(g(y|x,\widehat{\beta}_J,\sigma_0))] \precsim_p \frac{1}{\sqrt{n}}\widehat{\delta}\sqrt{-\log\widehat{\delta}}.$$

It follows that

$$E_x[||z(y|x)||_1^2] \leq 2E[\log(g(y|x,\beta_J^*,\sigma_0))] - E[\log(g(y|x,\widehat{\beta}_J,\sigma_0))] = O_p(\frac{1}{\sqrt{n}}\widehat{\delta}\sqrt{-\log\widehat{\delta}}).$$

Now consider the characteristic functions of $x\widehat{\beta}_J(\tau)$ and $x\beta_J^*(\tau)$ conditional on x, $\tau \sim U[0,1]$

$$\begin{aligned}
\phi_{x\widehat{\beta}_J}(s|x) &= \frac{\int_{-\infty}^{\infty} g(y|x,\widehat{\beta}_J,\sigma_0)e^{isy}dy}{\phi_\varepsilon(s|\sigma_0)} \\
\phi_{x\beta_J^*}(s|x) &= \frac{\int_{-\infty}^{\infty} g(y|x,\beta_J^*,\sigma_0)e^{isy}dy}{\phi_\varepsilon(s|\sigma_0)}
\end{aligned}$$

It follows that

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

Then

$$|\phi_{x\widehat{\beta}_J}(s|x) - \phi_{x\beta_J^*}(s|x)| \leq_p \frac{||z(y|x)||_1}{\phi_\varepsilon(s|\sigma_0)}.$$

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

$$F_{x\widehat{\beta}_J}(w) - F_{x\beta_J^*}(w) = \lim_{q\to\infty} \int_{-q}^{q} \frac{\exp(iws)}{2\pi is} (\phi_{x\widehat{\beta}_J}(s|x) - \phi_{x\beta_J^*}(s|x)) ds.$$

Then since in our sieve setting $\widehat{\beta}_J$ and β_J^* are r^{th}-order spline functions with grid interval size of order $O(1/J_n)$, we know that $\max\left(\left|\phi_{x\widehat{\beta}_J}(s|x)\right|, \left|\phi_{x\beta_J^*}(s|x)\right|\right) \leq J_n \frac{c}{s}$ for some constant $c > 0$. Therefore,

$$E_x\left[\left|F_{x\widehat{\beta}_J}(w) - F_{x\beta_J^*}(w)\right|\right]$$
$$\leq E_x\left[\int_{-q}^{q} \left|\frac{\exp(iws)}{2\pi is}\right| \frac{||z(y|x)||_1}{|\phi_\varepsilon(s|\sigma_0)|} ds\right] + E_x\left[2\int_q^{\infty} \frac{1}{2\pi s} \left|\phi_{x\widehat{\beta}_J}(s|x) - \phi_{x\beta_J^*}(s|x)\right| ds\right]. \tag{G.24}$$

The first term of the right-hand side of equation (G.24) is weakly bounded from above by

$$\frac{1}{2\pi}\int_{-q}^{q} \frac{1}{s\phi_\varepsilon(s|\sigma_0)} ds E_x[||z(y|x)||_1] = o_p\left(\frac{q^\lambda}{n^{1/4}}\sqrt{\widehat{\delta}}(-\log\widehat{\delta})^{\frac{1}{4}}\right)$$

where λ is the degree of mild ill-posedness. The second term of (G.24) is weakly bounded by $J_n\frac{4c}{q} \precsim J_n/q$. Putting these together, the right-hand side of (G.24) has an upper bound of $O_p\left(\frac{q^\lambda}{n^{1/4}}\sqrt{\widehat{\delta}}(-\log\widehat{\delta})^{\frac{1}{4}} + \frac{J_n}{q}\right)$ for an arbitrary q.

Since $\widehat{\delta} = ||\beta_J^* - \widehat{\beta}_J||_2 = O\left(E_x\left[\left|F_{x\widehat{\beta}_J}(w) - F_{x\beta_J^*}(w)\right|\right]\right)$, we have

$$\widehat{\delta} = O_p\left(\frac{q^\lambda}{n^{1/4}}\sqrt{\widehat{\delta}}(-\log\widehat{\delta})^{\frac{1}{4}} + \frac{J_n}{q}\right).$$

If $\widehat{\delta} = O_p\left(\frac{1}{n}\right)$,then the conclusion holds. If $\widehat{\delta}$ converges to 0 slower than $\frac{1}{n}$,we have

$$\widehat{\delta} = O_p\left(\frac{q^\lambda}{n^{1/4}}\sqrt{\widehat{\delta}}\log^{\frac{1}{4}} n + \frac{J_n}{q}\right),$$

which implies

$$\widehat{\delta} = O_p\left(\frac{q^{2\lambda}}{n^{1/2}}\log^{\frac{1}{2}} n + \frac{J_n}{q}\right).$$

The optimal q is $\left(\frac{J_n n^{\frac{1}{2}}}{\log^{\frac{1}{2}} n}\right)^{\frac{1}{2\lambda+1}}$. Then we have

$$\widehat{\delta} = ||\widehat{\beta}_J - \beta_J^*|| = O_p\left((J_n^{2\lambda}\frac{\log^{\frac{1}{2}} n}{n^{\frac{1}{2}}})^{\frac{1}{2\lambda+1}}\right).$$

□

Proof of Theorem 2. Suppose $\widehat{\theta}_J = (\widehat{\beta}_J(\cdot), \widehat{\sigma}) \in \Theta_J^r$ is the r^{th}-order sieve estimator. By the consistency of the sieve estimator established by Lemma 3, $||\widehat{\theta}_J - \theta_0||_2 \xrightarrow{p} 0$. It is easy to see that $\Theta_J^r \subset \Theta$. By Lemma 4, $\widehat{\sigma}$ will always converge to σ_0 at rate of at least $n^{-\frac{1}{4}}$. By construction of the sieve, there exists a set of parameters (β_J^*, σ_0) in Θ_J^r such that $||\beta_J^* - \beta_0||_2 = O_p\left(\frac{1}{J_n^{r+1}}\right)$.

Let G_n denote the operator $\sqrt{n}(E_n - E)$. Then we have

$$\begin{aligned}\frac{1}{\sqrt{n}}G_n\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{(\widehat{\beta}_J,\widehat{\sigma})} &= E_n\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{(\widehat{\beta}_J,\widehat{\sigma})} - E\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{(\widehat{\beta}_J,\widehat{\sigma})} \\ &= -E\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{(\widehat{\beta}_J,\widehat{\sigma})}, \end{aligned} \tag{G.25}$$

where we used the first-order condition $E_n\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\big|_{(\widehat{\beta}_J,\widehat{\sigma})} = 0$. For the left-hand side of (G.25), by Donskerness of $\left\{\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\big|_{(\tilde{\beta},\tilde{\sigma})} \middle| (\tilde{\beta}, \tilde{\sigma}) \in \Theta\right\}$, we have

$$\frac{1}{\sqrt{n}}G_n\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{(\widehat{\beta}_J,\widehat{\sigma})} = \frac{1}{\sqrt{n}}(1 + o_p(1))G_n\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{(\beta_0,\sigma_0)},$$

which is asymptotically Gaussian. Next, we work on the right-hand side of (G.25). It can be expanded as

$$\begin{aligned}&- E\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{(\widehat{\beta}_J,\widehat{\sigma})} \\ =& -\left[E\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{(\widehat{\beta}_J,\widehat{\sigma})} - E\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{(\beta_J^*,\sigma_0)}\right] - E\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{(\beta_J^*,\sigma_0)},\end{aligned} \tag{G.26}$$

and then a Taylor expansion of the term inside brackets of (G.26) gives

$$I_{\beta_J,\sigma_0}\left(\widehat{b}_J - b_J^*, \widehat{\sigma} - \sigma_0\right) + O_p\left(||\widehat{b}_J - b_J^*||^2 + ||\widehat{\sigma} - \sigma_0||^2\right), \tag{G.27}$$

where $\widehat{b}_J$ and b_J^* denote the coefficient vectors for the spline functions in $\widehat{\beta}_J$ and β_J^*. The second term on the right-hand side of (G.26), $-E\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\big|_{(\beta_J^*,\sigma_0)}$, equals

$$E\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{(\beta_0,\sigma_0)} - E\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{(\beta_J^*,\sigma_0)},$$

because (β_0,σ_0) is the truth and therefore $E\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Big|_{(\beta_0,\sigma_0)} = 0$.

Since $||\beta_J^* - \beta_0|| = O_p\big(\frac{1}{J_n^{r+1}}\big)$, by continuity

$$\begin{aligned}&E\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{(\beta_0,\sigma_0)} - E\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{(\beta_J^*,\sigma_0)} \\ =& O_p\big(||\beta_J^* - \beta_0||\big) = O_p\left(\frac{1}{J_n^{r+1}}\right).\end{aligned}$$

Combining for both sides of (G.25), we have

$$\begin{aligned}&\frac{1}{\sqrt{n}}(1+o_p(1))G_n\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{(\beta_0,\sigma_0)}\\ =&-I_{\beta_J^*,\sigma_0}\left(\widehat{b}_J-b_J^*,\,\widehat{\sigma}-\sigma_0\right)+O_p\left(||\widehat{b}_J-b_J^*||^2+||\widehat{\sigma}-\sigma_0||^2\right)+O_p\Big(\frac{1}{J_n^{r+1}}\Big).\end{aligned}\tag{G.28}$$

Since $||\widehat{\sigma}-\sigma_0||^2=o_p\big(\frac{1}{\sqrt{n}}\big)$, it is dominated by the Gaussian term on the left-hand side of (G.28). By Lemma 9 and the condition that $\frac{J_n^{4\lambda^2+6\lambda}\log(n)}{n}\to 0$, we know that $||\widehat{\beta}_J-\beta_J^*||^2=J_n^{-\lambda}o_p\big(||\widehat{\beta}_J-\beta_J^*||\big)$ and $||\widehat{b}_J-b_J^*||^2=J_n^{-\lambda}o_p\big(||\widehat{b}_J-b_J^*||\big)$, for J_n satisfying the growth rate conditions stated in the theorem. Therefore, (G.28) becomes

$$\begin{aligned}&-(1+o_p(1))I_{\beta_J^*,\sigma_0}\left(\widehat{b}_J-b_J^*,\,\widehat{\sigma}-\sigma_0\right)\\ =&\frac{1}{\sqrt{n}}(1+o_p(1))G_n\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{(\beta_0,\sigma_0)}+O_p\Big(\frac{1}{J_n^{r+1}}\Big).\end{aligned}\tag{G.29}$$

By continuity of the information matrix as a function of β, we know that the smallest eigenvalue of $I_{\beta_J^*,\sigma_0}$ is on the same order as the smallest eigenvalue of I_{β_0,σ_0}, i.e. bounded by $\frac{c}{J_n^\lambda}$ from below with c as a constant. Hence (G.29) implies

$$\left|\left|\widehat{b}_J-b_J^*,\,\widehat{\sigma}-\sigma_0\right|\right|=J_n^\lambda O_p\Big(\frac{1}{J_n^{r+1}},\frac{1}{\sqrt{n}}\Big),$$

or

$$\left|\left|\widehat{\beta}_J-\beta_0,\,\widehat{\sigma}-\sigma_0\right|\right|=O_p\Big(\frac{1}{J_n^{r+1}}\Big)+\left|\left|\widehat{\beta}_J-\beta_J^*,\,\widehat{\sigma}-\sigma_0\right|\right|=J_n^\lambda O_p\Big(\frac{1}{J_n^{r+1}},\frac{1}{\sqrt{n}}\Big),$$

establishing the convergence rate of the Sieve estimator. For asymptotic normality, note that if $J_n^{r+1}/\sqrt{n}\to\infty$, then the first term on the right-hand side of (G.29) dominates the second term on the right-hand side of (G.29), so we have

$$I_{\beta_J,\sigma_0}\left(\widehat{b}_J-b_J^*,\,\widehat{\sigma}-\sigma_0\right)=\frac{1}{\sqrt{n}}(1+o_p(1))G_n\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{(\beta_0,\sigma_0)}.$$

Therefore,

$$\left(\widehat{b}_J-b_J^*,\,\widehat{\sigma}-\sigma_0\right)=\frac{1}{\sqrt{n}}(1+o_p(1))I^{-1}_{\beta_J^*,\sigma_0}G_n\left(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\right)\Bigg|_{(\beta_0,\sigma_0)}.$$

Since $G_n\big(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\big)\big|_{(\beta_0,\sigma_0)}$ is asymptotically normal, $I^{-1}_{\beta_J^*,\sigma_0}G_n\big(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\big)\big|_{(\beta_0,\sigma_0)}$ is also asymptotically normal, with distribution $N(0, I^{-1}_{\beta_J^*,\sigma_0}\Omega_{G_n}I^{-1}_{\beta_J^*,\sigma_0})$, where

$$\Omega_{G_n}:=\mathrm{Var}\big(G_n\big(\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma}\big)\big|_{(\beta_0,\sigma_0)}\big)\to_p I_{\beta_0,\sigma_0}.$$

Let κ_J denote the smallest eigenvalue for I_{β_0,σ_0}. By Assumption, $\kappa_J\geq C/J^\lambda$ for some generic positive constant $C>0$. Since $||\beta_J^*-\beta_0||=O(\frac{1}{J^{r+1}})$,

$$||I_{\beta_0,\sigma_0}-I_{\beta_J^*,\sigma_0}||_\infty=O(||\beta_J^*-\beta_0||)=O(\frac{1}{J^{r+1}}).$$

By the growth conditions, we have $r+1>\lambda$, so $||I_{\beta_0,\sigma_0}-I_{\beta_J^*,\sigma_0}||_2=O(\frac{1}{J^r})=o(\kappa_J)$. Hence, $I^{-1}_{\beta_J^*,\sigma_0}=(I_{\beta_0,\sigma_0}-I_{\beta_0,\sigma_0}+I_{\beta_J^*,\sigma_0})^{-1}=(I_{\beta_0,\sigma_0}(1+o(1)))^{-1}\to I^{-1}_{\beta_0,\sigma_0}$.

Let Ω_J denote $\kappa_J I^{-1}_{\beta_J^*,\sigma_0}\Omega_{G_n}I^{-1}_{\beta_J^*,\sigma_0}\to\kappa_j I^{-1}_{\beta_0,\sigma_0}$. We have

$$\sqrt{n\kappa_J}\left(\widehat{b}_J-b_J^*,\,\widehat{\sigma}-\sigma_0\right)\xrightarrow{d}N(0,\Omega_J)=O_p(1).$$

Then $\sqrt{n\kappa_J}\big(\sigma-\sigma_0\big)$ converges to $N(0,\Omega_{J,\sigma})$ in distribution, where $\Omega_{J,\sigma}$ denotes the submatrix of Ω_J for σ. Since the largest eigenvalue of Ω_J is bounded by a constant, the largest eigenvalue of $\Omega_{J,\sigma}$ is bounded by the same constant. Similarly $\sqrt{n\kappa_J}\big(\widehat{b}_J-b_J^*\big)$ converges to $N(0,\Omega_{J,b})$ in distribution, where $\Omega_{J,b}$ is the submatrix of Ω_J for b. It follows that for each τ,

$$\sqrt{n\kappa_J}\big(\widehat{\beta}_J(\tau)-\beta_J^*(\tau)\big)\xrightarrow{d}\big(S^J(\tau)\otimes D_{d_x}\big)'N(0,\Omega_{J,b}),$$

where D_{d_x} denotes a $d_x\times d_x$ identity matrix, and $S^J(\tau)$ denotes $\big(S_1(\tau),\cdots,S_{J+r}(\tau)\big)$. By assumption that $\sum_{l=1}^{J+r}|S_k(\tau)|^2$ is bounded by a constant, the largest eigenvalue of $\big(S^J(\tau)\otimes D_{d_x}\big)'\Omega_{J,b}\big(S^J(\tau)\otimes D_{d_x}\big)$ is also bounded by a constant. Note that because $||\beta_J^*(\tau)-\beta_0(\tau)||=O_p\big(\frac{1}{J_n^{r+1}}\big)$, the limiting distribution of $\sqrt{n\kappa_J}\big(\widehat{\beta}_J(\tau)-\beta_J^*(\tau)\big)$ is the same as the limiting distribution of $\sqrt{n\kappa_J}\big(\widehat{\beta}_J(\tau)-\beta_0(\tau)\big)$. Thus we have

$$\sqrt{n\kappa_J}\big(\widehat{\beta}_J(\tau)-\beta_0(\tau)\big)\xrightarrow{d}\big(S^J(\tau)\otimes D_{d_x}\big)'N(0,\Omega_{J,b})\big(S^J(\tau)\otimes D_{d_x}\big),$$

where $\big(S^J(\tau)\otimes D_{d_x}\big)'N(0,\Omega_{J,b})\big(S^J(\tau)\otimes D_{d_x}\big)$ is a positive definite matrix with the largest eigenvalue bounded by a constant. □

Proof of Theorem 3. Following the same argument as those in the proof of Theorem 2, we have:

$$I(\widehat{\beta}_J-\beta_J^*,\widehat{\sigma}-\sigma_0)+O_p(||(\widehat{\beta}_J-\beta_J^*,\widehat{\sigma}-\sigma_0)||_2^2)=\frac{-1}{\sqrt{n}}\mathbb{G}_{n,J_n}. \tag{G.30}$$

By setting J_n such that $\frac{\exp(\lambda J_n)}{\sqrt{n}}=\frac{1}{J_n}$, we have $(\frac{1}{2\lambda}-\eta)\log(n)<J_n<\frac{1}{2\lambda}\log(n)$, for any small $\eta>0$ and n large enough.

By Assumption 5.2, the minimum eigenvalue of I is bounded by $C\exp(-\lambda J_n)$ for some $\lambda>0$ and $C>0$. It follows that $||I(\widehat{\beta}_J-\beta_J^*,\widehat{\sigma}-\sigma_0)||\geq C\exp(-\lambda J_n)||(\widehat{\beta}_J-\beta_J^*,\widehat{\sigma}-\sigma_0)||_2$.

(a) If $||(\widehat{\beta}_J-\beta_J^*,\widehat{\sigma}-\sigma_0)||_2\geq C_1/C\exp(-\lambda J_n)$, for some constant C_1 large enough, then with probability approaching 1, we have $||I(\widehat{\beta}_J-\beta_J^*,\widehat{\sigma}-\sigma_0)||_2>2O_p(||(\widehat{\beta}_J-\beta_J^*,\widehat{\sigma}-\sigma_0)||_2^2)$, where $O_p(||(\beta-\beta_J^*,\widehat{\sigma}-\sigma_0)||_2^2)$ is the higher order residual term in the equation (G.30). It follows that $||(\widehat{\beta}_J-\beta_J^*,\widehat{\sigma}-\sigma_0)||_2\precsim_p\frac{\exp(\lambda J_n)}{\sqrt{n}}$.

(b) Else we have $||(\beta-\beta_J^*,\widehat{\sigma}-\sigma_0)||_2\leq C_1/C\exp(-\lambda J_n)\leq C_1/C_n(1/2-\eta\lambda)=o(\frac{\exp(\lambda J_n)}{\sqrt{n}})$.

Combining the two situations, we have $||(\beta-\beta_J^*,\widehat{\sigma}-\sigma_0)||_2=O_p(\frac{\exp(\lambda J_n)}{\sqrt{n}})$.

By construction of the sieve, $||(\beta-\beta_J^*,\widehat{\sigma}-\sigma_0)||_2=O(\frac{1}{J_n})$. Hence, $||(\beta-\beta_J^*,\widehat{\sigma}-\sigma_0)||_2=O(\max(\frac{\exp(\lambda J_n)}{\sqrt{n}},\frac{1}{J_n}))$. By assumption, we set J_n such that $\frac{\exp(\lambda J_n)}{\sqrt{n}}=\frac{1}{J_n}=O(\frac{1}{\log(n)})$. Therefore, the sieve estimator satisfies: $||(\beta-\beta_J^*,\widehat{\sigma}-\sigma_0)||_2=O_p(\frac{1}{\log(n)})$. □

Proof of Lemma 5. A bootstrap process can be considered as putting non-negative weights $w_{i,n}$ on the i^{th} observation. We require $E[w_{i,n}]=1$, and $E[w_{i,n}]^2=\sigma_{w,n}^2<\infty$. One example is to

let $(w_{i,1},...,w_{i,n}) \sim Multinomial(n, \frac{1}{n},...,\frac{1}{n})$, which is the nonparametric pairs bootstrap recommended in the text and used in the simulation and empirical results. The bootstrapped estimator $(\widehat{\beta}_J^b, \widehat{\sigma}^b)$ should satisfy the first-order condition

$$E_n^b\left[\frac{\partial \log g^b}{\partial \beta}\Big|_{\widehat{\beta}_J^b,\widehat{\sigma}^b}, \frac{\partial \log g^b}{\partial \sigma}\Big|_{\widehat{\beta}_J^b,\widehat{\sigma}^b}\right] = 0.$$

By Assumption 4.3, $E[\sup_{\beta,\sigma}|w_{i,n}\log g(y_i|x_i,\beta,\sigma)|] < \infty$. Moreover, by Assumption 4.2, $w_{i,n}\log g(y_i|x_i,\beta,\sigma)$ satisfies that

$$E[|w\log g(y_i|x_i,\beta,\sigma) - w'\log g(y_i|x_i,\beta',\sigma')|] \leq C_1(|w-w'| + ||(\beta,\sigma)-(\beta',\sigma')||)$$

for some generic constant $C_1 > 0$. By the ULLN for any $(\beta,\sigma) \in M \times \Sigma$, $E_n^b[\log g(\beta,\sigma)] \to_p E_n[\log g(\beta,\sigma)]$, which converges to $E[\log g(\beta,\sigma)]$ with probability approaching 1. Since $M \times \Sigma$ is compact and identification holds, it must be that $(\widehat{\beta}_J^b, \widehat{\sigma}^b) \to_p (\beta_0, \sigma_0)$. Therefore,

$$||(\widehat{\beta}_J^b, \widehat{\sigma}^b) - (\widehat{\beta}_J, \widehat{\sigma})|| \to_p 0.$$

Denote $G(\beta,\sigma) = (\frac{\partial \log g}{\partial \beta}, \frac{\partial \log g}{\partial \sigma})$. By stochastic equicontinuity,

$$E_n^b\Big[G\big(\widehat{\beta}_J^b, \widehat{\sigma}^b\big)\Big] - E_n\Big[G\big(\widehat{\beta}_J^b, \widehat{\sigma}^b\big)\Big] = E_n^b\Big[G\big(\widehat{\beta}_J, \widehat{\sigma}\big)\Big] - E_n\Big[G\big(\widehat{\beta}_J, \widehat{\sigma}\big)\Big] + o_p\big(\frac{1}{\sqrt{n}}\big).$$

In the above equation, $E_n^b\Big[G\big(\widehat{\beta}_J^b, \widehat{\sigma}^b\big)\Big] = E_n\Big[G\big(\widehat{\beta}_J, \widehat{\sigma}\big)\Big] = 0$ by the first-order condition. Thus we have

$$E_n^b\Big[G\big(\widehat{\beta}_J, \widehat{\sigma}\big)\Big] - E_n\Big[G\big(\widehat{\beta}_J, \widehat{\sigma}\big)\Big] + o_p\big(\frac{1}{\sqrt{n}}\big) = -\Big(E_n\Big[G\big(\widehat{\beta}_J^b, \widehat{\sigma}^b\big)\Big] - E_n\Big[G\big(\widehat{\beta}_J, \widehat{\sigma}\big)\Big]\Big). \tag{G.31}$$

Next we will show that the left-hand side of (G.31) is asymptotically normal and the right-hand side of (G.31) can be written as a matrix multiplied by $\big(\widehat{b}_J^b, \widehat{\sigma}^b\big) - \big(\widehat{b}_J, \widehat{\sigma}\big)$, where $\widehat{b}_J$ and $\widehat{b}_J^b$ are the coefficients for the sieve functions in $\widehat{\beta}_J$ and $\widehat{\beta}_J^b$.

For the left-hand side of (G.31),

$$E_n^b\Big[G\big(\widehat{\beta}_J, \widehat{\sigma}\big)\Big] - E_n\Big[G\big(\widehat{\beta}_J, \widehat{\sigma}\big)\Big] = E_n^b\Big[\big(w_{i,n}-1\big)G\big(\widehat{\beta}_J, \widehat{\sigma}\big)\Big]$$

Note that

$$\sqrt{n}\frac{n}{n-1}E_n^b\Big[\big(w_{i,n}-1\big)G\big(\widehat{\beta}_J, \widehat{\sigma}\big)\Big] \overset{a}{\sim} \mathcal{N}\Big(0, E\Big[G\big(\widehat{\beta}_J, \widehat{\sigma}\big)G\big(\widehat{\beta}_J, \widehat{\sigma}\big)'\Big]\Big).$$

By Theorem 2, $||(\widehat{\beta}_J, \widehat{\sigma}_J) - (\beta_0,\sigma_0)|| = O_p(\frac{J_n^\lambda}{\sqrt{n}})$, therefore, $E\Big[G\big(\widehat{\beta}_J, \widehat{\sigma}\big)G\big(\widehat{\beta}_J, \widehat{\sigma}\big)'\Big] = I_{\beta_0,\sigma_0} + O_p(\frac{J_n^\lambda}{\sqrt{n}})$, where $I_{\beta_0,\sigma_0} := E\Big[G\big(\beta_0,\sigma_0\big)G\big(\beta_0,\sigma_0\big)'\Big]$. By assumption, $\frac{J_n^\lambda}{\sqrt{n}}/\text{mineigen}(I_{\beta_0,\sigma_0}) \to 0$ as $n \to \infty$, thus I_{β_0,σ_0} dominates $O_p(\frac{J_n^\lambda}{\sqrt{n}})$.

For the right-hand side of (G.31), by Donskerness of $M \times \Sigma$ and stochastic equicontinuity, we have

$$E_n\Big[G\big(\widehat{\beta}_J^b, \widehat{\sigma}^b\big)\Big] - E_n\Big[G\big(\widehat{\beta}_J, \widehat{\sigma}\big)\Big] = E\Big[G\big(\widehat{\beta}_J^b, \widehat{\sigma}^b\big)\Big] - E\Big[G\big(\widehat{\beta}_J, \widehat{\sigma}\big)\Big] + o_p(\frac{1}{\sqrt{n}}),$$

where the remainder term $o_p(\frac{1}{\sqrt{n}})$ does not affect the derivation further and is dropped.

By Taylor expansion,

$$\begin{aligned} E\Big[G(\widehat{\beta}_J^b,\widehat{\sigma}^b)\Big] - E\Big[G(\widehat{\beta}_J,\widehat{\sigma})\Big] &= I_{\widehat{\beta}_J,\widehat{\sigma}}(\widehat{b}_J^b-\widehat{b}_J,\widehat{\sigma}^b-\widehat{\sigma}) + O\Big(\big|\big|(\widehat{b}_J^b-\widehat{b}_J,\widehat{\sigma}^b-\widehat{\sigma})\big|\big|^2\Big) \\ &= \big(I_{\beta_0,\sigma_0} + O\big(\big|\big|(\widehat{b}_J-b_0,\widehat{\sigma}-\sigma_0)\big|\big|\big)\big)(\widehat{b}_J^b-\widehat{b}_J,\widehat{\sigma}^b-\widehat{\sigma}) \\ &\quad + O\Big(\big|\big|(\widehat{b}_J^b-\widehat{b}_J,\widehat{\sigma}^b-\widehat{\sigma})\big|\big|^2\Big) \\ &= \big(I_{\beta_0,\sigma_0} + o_p(1)\big)(\widehat{b}_J^b-\widehat{b}_J,\widehat{\sigma}^b-\widehat{\sigma}) + O\Big(\big|\big|(\widehat{b}_J^b-\widehat{b}_J,\widehat{\sigma}^b-\widehat{\sigma})\big|\big|^2\Big) \end{aligned}$$

Combining different terms in (G.31), we have

$$(1+o_p(1))I_{\beta_0,\sigma_0}(\widehat{b}_J^b-\widehat{b}_J,\widehat{\sigma}^b-\widehat{\sigma}) + O\Big(\big|\big|(\widehat{b}_J^b-\widehat{b}_J,\widehat{\sigma}^b-\widehat{\sigma})\big|\big|^2\Big) = \frac{1}{\sqrt{n}}N(0,(I_{\beta_0,\sigma_0}+O_p(\frac{J_n^\lambda}{\sqrt{n}}))). \quad \text{(G.32)}$$

Similar to Theorem 2, we need to show that $\big|\big|(\widehat{b}_J^b-\widehat{b}_J,\widehat{\sigma}^b-\widehat{\sigma})\big|\big|^2$ is dominated by $(1+o_p(1))I_{\beta_0,\sigma_0}(\widehat{b}_J^b-\widehat{b}_J,\widehat{\sigma}^b-\widehat{\sigma})$. Our strategy is to use stochastic equicontinuity, which implies that

$$E_n^b\Big[\log g\big(\widehat{\beta}_J^b,\widehat{\sigma}^b\big)\Big] - E_n\Big[\log g\big(\widehat{\beta}_J^b,\widehat{\sigma}^b\big)\Big] = E_n^b\Big[\log g\big(\widehat{\beta}_J,\widehat{\sigma}\big)\Big] - E_n\Big[\log g\big(\widehat{\beta}_J,\widehat{\sigma}\big)\Big] + o_p\big(\frac{1}{\sqrt{n}}\big)$$

or

$$E_n\Big[\log g\big(\widehat{\beta}_J,\widehat{\sigma}\big)\Big] - E_n\Big[\log g\big(\widehat{\beta}_J^b,\widehat{\sigma}^b\big)\Big] = E_n^b\Big[\log g\big(\widehat{\beta}_J,\widehat{\sigma}\big)\Big] - E_n^b\Big[\log g\big(\widehat{\beta}_J^b,\widehat{\sigma}^b\big)\Big] + o_p\big(\frac{1}{\sqrt{n}}\big),$$

where $E_n^b\Big[\log g\big(\widehat{\beta}_J,\widehat{\sigma}\big)\Big] - E_n^b\Big[\log g\big(\widehat{\beta}_J^b,\widehat{\sigma}^b\big)\Big] \leq 0$ by optimality of $\big(\widehat{\beta}_J^b,\widehat{\sigma}^b\big)$. Thus we have

$$E_n\Big[\log g\big(\widehat{\beta}_J,\widehat{\sigma}\big)\Big] - E_n\Big[\log g\big(\widehat{\beta}_J^b,\widehat{\sigma}^b\big)\Big] \leq o_p\big(\frac{1}{\sqrt{n}}\big).$$

On the other hand, we know that

$$E_n\Big[\log g\big(\widehat{\beta}_J,\widehat{\sigma}\big)\Big] - E_n\Big[\log g\big(\widehat{\beta}_J^b,\widehat{\sigma}^b\big)\Big] \geq 0$$

by optimality of $\big(\widehat{\beta}_J,\widehat{\sigma}\big)$. Hence,

$$\Big|E_n\big[\log g(\widehat{\beta}_J^b,\widehat{\sigma}^b)\big] - E_n\big[\log g(\widehat{\beta}_J,\widehat{\sigma})\big]\Big| = o_p\big(\frac{1}{\sqrt{n}}\big).$$

With this, we can apply similar arguments as in Lemma 4 and Lemma 9 to show that

$$\begin{aligned} ||\widehat{\sigma}^b-\widehat{\sigma}|| &= o_p\big(n^{-\frac{1}{4}}\big) \\ ||\widehat{\beta}_J^b-\widehat{\beta}_J|| &= O_p\left(\left(J_n^{2\lambda}\frac{\log^{\frac{1}{2}}n}{n^{\frac{1}{2}}}\right)^{\frac{1}{2\lambda+1}}\right) \end{aligned}$$

By an argument similar to the one in Theorem 2, (G.32) implies that $\sqrt{n\kappa_J}(\widehat{\beta}_J^b-\widehat{\beta}_J)$ and $\sqrt{n\kappa_J}(\widehat{\sigma}^b-\widehat{\sigma})$ have the same distributions as $\sqrt{n\kappa_J}(\widehat{\beta}_J-\beta_0)$ and $\sqrt{n\kappa_J}(\widehat{\sigma}-\sigma_0)$. □