ERRORS IN THE DEPENDENT VARIABLE OF QUANTILE REGRESSION MODELS

JERRY HAUSMAN, HAOYANG LIU, YE LUO, AND CHRISTOPHER PALMER

Abstract. We study the consequences of measurement error in the dependent variable of random-coefficients models, focusing on the particular case of quantile regression. 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. Monte Carlo evidence verifies our method outperforms quantile regression in mean bias and MSE. Finally, we 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.

Keywords: Measurement Error, Quantile Regression, Functional Analysis

---

Date: October 2020.

Hausman: MIT and NBER; jhausman@mit.edu.
Liu: Federal Reserve Bank of New York; haoyang.liu@ny.frb.org.
Luo: HKU Business School; kurtluo@hku.hk.
Palmer: MIT Sloan and NBER; cjpalmer@mit.edu.

Code to implement the estimator in this paper can be found at https://github.com/palmercj/EIV-QR.

We thank three anonymous referees, Isaiah Andrews, Colin Cameron, Victor Chernozhukov, Denis Chetverikov, Kirill Evdokimov, Hank Farber, Brigham Frandsen, Larry Katz, Brad Larsen, Rosa Matzkin, James McDonald, 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, Jacob Ornelas, and John Wilson provided outstanding research assistance. Luo acknowledges support from HK SAR RGC Grant T35/710/20R. The views expressed herein are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of New York or the Federal Reserve System.

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 bias estimates of an 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 under the standard assumptions of the linear conditional quantile model. 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.[3] We show that under certain assumptions on the degree of ill-posedness, 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.[4] Put another way, while omitted variables are problematic in the linear model insofar as they are correlated with the regressors, in quantile

[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] See Chetverikov et al. (2016) for a quantile-regression framework that can accommodate measurement error in group-level covariates without instruments.

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

regression additive unobserved heterogeneity causes bias even when independent of included covariates. In this sense, our results are applicable to linear quantile regression models with many covariates and additive unobserved 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. Jointly estimating the conditional quantiles and the distribution of 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.[5] 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 $f(\cdot)$ be the p.d.f. of the EIV ε. We denote the true coefficient and

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

measurement error distributional parameters as $\beta_0(\cdot)$ and σ_0, respectively. 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

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

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 ξ_i such that additional measurement error in y 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.

In practice, improving upon the generic random coefficient model in (2.1) requires relatively strong assumptions. 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

$$y_i^* = x_i^T \beta_0(u_i),$$

where all unobserved heterogeneity in enters through the treatment effects as the scalar $u_i \sim U[0,1]$ representing the unobserved quantile of y_i in the conditional distribution of $y_i|x_i$.[6] In this model, the τ^{th} conditional quantile of the dependent variable y^* is a linear function of x

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

implying that $x^T \beta_0(\cdot)$ is monotonic. 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. Our ability to separately identify the conditional quantile coefficient function $\beta_0(\cdot)$ and the measurement error distribution in this parsimonious model relies on the structure afforded by the two assumptions embedded in the quantile-regression model: univariate unobserved heterogeneity u_i and the

[6]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.

monotonicity of $x^T\beta_0(\cdot)$. Moreover, the estimator we propose below cannot ultimately differentiate between whether independent and additively separable unobserved heterogeneity ε in observed outcomes y consists of pure noise or omitted independent covariates whose treatment effects do not vary with u_i. In either case, however, independent and additively separable unobserved heterogeneity in observed outcomes will bias estimated treatment effects.

Unlike the linear-regression case where EIV in the left-hand side variable does not inhibit consistency or 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

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

the minimization problem in the usual quantile regression

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

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

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

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

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

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. 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 (2.2) 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 and receive a weight of $(1-\tau)$. Such asymmetry implies that equal-sized differences on either side of zero do not cancel each other out as they do for estimators with symmetric loss functions. Note that for median regression, $\rho_{.5}(\cdot)$ is symmetric around zero. Accordingly, if ε is symmetrically distributed and $\beta(\tau)$ symmetrically distributed around $\tau = .5$ (as would be the case, for example, if $\beta(\tau)$ were linear in τ), the expectation in equation (2.3) holds for the true $\beta_0(0.5)$.

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 bounded 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 vector of observables x.*

(1) *$E[xx^T]$ is non-singular.*
(2) *There is at least one dimension x_1 of x such that for each value of x_{-1} with strictly positive probability density or mass, there exists an open neighborhood of x_1 within which $x_1|x_{-1}$ is continuously distributed with strictly positive probability density.*
(3) *The element of $\beta_0(\cdot)$ corresponding to x_1, denoted as $\beta_{0,1}(\cdot)$, is strictly monotonic.*

Assumption 2.2 ensures that there is at least one dimension x_1 of x such that for every reachable value of x_{-1}, x_{-1} cannot fully explain x_1. Finally, we assume that the measurement error is mean zero and has finite moments.

Assumption 3 (Properties of EIV)**.** *We assume the following properties of any PDF $f(\cdot)$ of the measurement error ε.*

(1) *$f(\cdot)$ is continuously differentiable.*
(2) *$\int_{-\infty}^{\infty} \varepsilon f(\varepsilon)d\varepsilon = 0$.*
(3) *There exists a constant $C > 0$ such that for all $k > 0$, $\int_{-\infty}^{\infty} \left|\varepsilon^k\right| f(\varepsilon)d\varepsilon < k! \cdot C^k$ where $k!$ denotes k factorial.*

[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 $x \geq 0$.

The above conditions on the parameters, covariates, and measurement error distribution 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 $f_0(\cdot)$, both satisfy the conditions of Assumption 3. 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 B. □

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 more flexible functions of x_1 are admissible, the following lemma establishes nonparametric identification when finite polynomials of x_1 are also included in x. Before stating the lemma, we restate Assumption 2 to allow for polynomials of x_1.

Assumption 4 (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) *$E[xx^T]$ is non-singular.*
(2) *We can partition $x = (W(x_1)^T, x_{-w}^T)^T$ where x_1 is one dimensional, $W(x_1) = (x_1, x_1^2, ..., x_1^p)^T$ for some p, and for each value of x_{-w} with strictly positive marginal probability density or mass, there exists an open neighborhood of x_1 within which $x_1|x_{-w}$ is continuously distributed with strictly positive probability density.*
(3) *The element of $\beta_0(\cdot)$ corresponding to x_1^p , denoted as $\beta_{0,x_1^p}(\cdot)$, is strictly monotonic and 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 4 hold and that the PDFs of ε, $f(\cdot)$ and $f_0(\cdot)$, both satisfy the conditions of Assumption 3. Then, for any $\beta(\cdot)$ and $f(\cdot)$ which generate the same density of $y|x$ almost everywhere as the true functions $\beta_0(\cdot)$ 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 B. □

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 provide sufficient conditions for inference by pairs bootstrapping in our sieve-ML setting, paralleling the residual bootstrapping procedure in Chen and Pouzo's (2013) sieve-GMM 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 5 (Parametric 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 with dimension d_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) *There exists a constant $C > 0$ such that for all $k > 0$, $\int_{-\infty}^{\infty} \left|\varepsilon^k\right| f(\varepsilon|\sigma)d\varepsilon \leq k! \cdot C^k$.*
(6) *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 $f(\varepsilon|\sigma)$.*
(7) *There exists a constant $C > 0$ such that both $E\left[\left\| x \frac{f'(y - x^T\beta_0(\tau)|\sigma_0)}{\int_0^1 f(y - x^T\beta_0(\tau)|\sigma_0)d\tau} \right\|^4\right] < C$ and $E\left[\left\| \frac{\int_0^1 f_\sigma(y - x^T\beta_0(\tau)|\sigma_0)d\tau}{\int_0^1 f(y - x^T\beta_0(\tau)|\sigma_0)d\tau} \right\|^4\right] < C$.*

Note that Assumption 5 holds for all mean-zero distributions in the exponential family. Assumption 5.7 is a mild condition required by the Triangular Central Limit Theorem and guarantees that the information matrix exists and its empirical analog is well behaved.

Given this parameterization of $f(\cdot|\sigma)$, we define our log likelihood function as follows. Denote $\theta := (\beta(\cdot), \sigma) \in \Theta$. Define $||(\beta,\sigma)|| := \sqrt{\int_0^1 ||\beta(\tau)||_2^2 d\tau + ||\sigma||_2^2}$ as the L^2 norm of (β_0,σ_0), where $||\cdot||_2$ is the usual Euclidean norm. For any θ, define the expected log-likelihood function $L(\theta)$ as

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

with the empirical log likelihood being denoted

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

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

[8]While Assumption 5 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.

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

$$g(y|x,\theta) = \int_0^1 f(y - x^T\beta(u)|\sigma)du. \tag{3.1}$$

Then the ML estimator is defined as

$$\widehat{\theta} = (\widehat{\beta}(\cdot), \widehat{\sigma}) \in \arg \max_{(\beta(\cdot),\sigma)\in\Theta} E_n[\log g(y|x, \beta(\cdot), \sigma)], \tag{3.2}$$

where $g(\cdot|\cdot,\cdot,\cdot)$ is the conditional density of y given x and parameters, as defined in equation (3.1). The following theorem states the consistency property of the ML estimator.

Lemma 2 (MLE Consistency)**.** *Under Assumptions 1, 4, and 5, the maximum-likelihood estimator defined by* (3.2) *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 Online Appendix D.1. □

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

(1) $x^T\beta(\tau)$ is monotonically increasing in $\tau \in [0,1]$ for all $x \in \mathcal{X}$ for all $\beta(\cdot) \in \Omega_J^r$ and
(2) 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)**.**

$$\widehat{\theta}_J = (\widehat{\beta}_J(\cdot), \widehat{\sigma}) = \arg \max_{\theta \in \Theta_{J_n}^r} E_n[\log g(y|x, \beta, \sigma)] \tag{3.3}$$

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, 4, and 5 hold, $J_n \to \infty$, and $J_n/n \to 0$, then the sieve estimator defined in* (3.3) *is consistent.*

Proof. See Online Appendix D.1. □

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 $(\beta_J, \sigma) \in \Theta_J^r$, $\frac{\partial g}{\partial \beta}\big|_{\beta_J,\sigma} = \left[\frac{\partial g}{\partial b_1}, \ldots, \frac{\partial g}{\partial b_{J+r}}\right]$, 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

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

When J goes to infinity, the smallest eigenvalue of $\mathcal{I}_{\beta_0,\sigma_0}$ goes to 0, leading to an ill-posedness problem. Intuitively, as we are trying to estimate $\beta(\cdot)$ and σ via sieve 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 6 (Ill-posed Measure)**.** *Define mineigen(A) as the minimum eigenvalue for a given matrix A. Let one of the following two assumptions on the degree of ill-posedness hold*

(1) *Mild ill-posedness: mineigen*$(\mathcal{I}_{\beta,\sigma}) \geq C/J^{\lambda}$ *for some* $\lambda > 0$ *and constant* $C > 0$*, for all* $(\beta_0, \sigma_0) \in \Theta$*.*
(2) *Severe ill-posedness: mineigen*$(\mathcal{I}_{\beta,\sigma}) \geq C\exp(-\lambda J^{\zeta})$ *for some* $\lambda > 0$*,* $\zeta > 0$*, 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 $\zeta = 1$, 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 7 (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*

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

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*.*

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 8 (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* $s \in [-l, l]$*,*

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

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

[9]See Lemma 8 in Online Appendix D.1 for a formal statement and proof of this result for sieves of splines.

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, 4, 5, and 8 hold and $J_n^{2r+2}/n \to \infty$, the sieve estimator $(\widehat{\beta}_J(\cdot), \widehat{\sigma})$ satisfies*

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

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

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

Proof. See Online Appendix D.1. □

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, 4, 5, 6.1 (the mildly ill-posed case), and 8 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 $n \to \infty$ and let $r+1 > \lambda$. 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

$$\sqrt{n\kappa_J}\Omega_{J,\tau}^{-1/2}\left(\widehat{\beta}_J(\tau) - \beta_0(\tau)\right) \xrightarrow{d} \mathcal{N}(0, I_{d_x}),$$

where $\Omega_{J,\tau}$ is a sequence of positive definite matrices with the largest eigenvalue bounded by a constant and I_{d_x} is an identity matrix of dimension $d_x \times d_x$, and

$$\sqrt{n\kappa_J}\Omega_{J,\sigma}^{-1/2}\left(\widehat{\sigma}_J - \sigma_0\right) \xrightarrow{d} \mathcal{N}(0, I_{d_\sigma}),$$

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

Proof. See Online Appendix D.1. □

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 $\frac{\exp(\lambda J_n^{\zeta})}{\sqrt{n}} = \frac{1}{J_n}$. Then under Assumptions 1, 4, 5, 6.2 (the severe ill-posed case), and 8, the sieve estimator β_{J_n} satisfies*

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

Proof. See Online Appendix D.1. □

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

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

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 confidence intervals 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 $z_{1-\alpha/2} = 1.96$ 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, 4, 5, 6.1 (the mildly ill-posed case), and 8 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 $n \to \infty$ and let $r + 1 > \lambda$. Then there exists a sequence $\kappa_J \geq \frac{C}{J_n^{\lambda}}$ for some generic constant $C > 0$ such that for any fixed τ,*

$$\sqrt{n\kappa_J}\Omega_{J,\tau}^{-1/2}\big(\widehat{\beta}_J^b(\tau) - \widehat{\beta}_J(\tau)\big) \xrightarrow{d} \mathcal{N}(0, I_{d_x})$$

and

$$\sqrt{n\kappa_J}\Omega_{J,\sigma}^{-1/2}\big(\widehat{\sigma}^b - \widehat{\sigma}\big) \xrightarrow{d} \mathcal{N}(0, I_{d_\sigma}),$$

where $\Omega_{J,\tau}$ and $\Omega_{J,\sigma}$ are the same as in Theorem 2.

Proof. See Online Appendix D.1. □

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

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

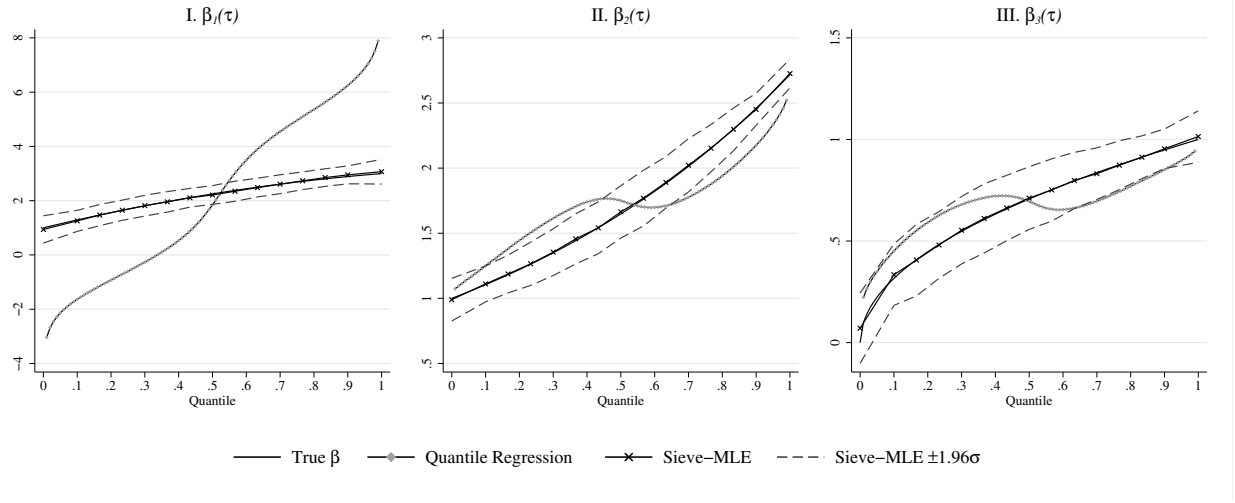
To simulate robustness to real-world settings in which the econometrician does not know the true distribution of the residuals, we also present results simulating measurement error from alternative distributions and test how well quasi-MLE via our sieve-ML estimator modeling the error distribution as 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 start values provided by a gradient-based constrained optimizer (see Online Appendix A for details on the implementation of our estimator). 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 Figure 1, we plot the true coefficient function defined above, average coefficients from quantile regression, and our sieve-ML estimator using a sieve for $\beta(\cdot)$ consisting of 15 knots. While Online Appendix B provides MSE results, to give a visual sense of the variability in the ML estimates across simulations, for each knot τ_j, we also plot pointwise error bands equal to $\widehat{\beta}(\tau_j) \pm 1.96\widehat{\sigma}_j$, where $\widehat{\sigma}_j$ is the standard deviation across simulations of parameter estimates $\widehat{\beta}(\tau_j)$.[11] To test whether our recommended bootstrapped confidence intervals

[10] We note that our asymptotic normality results (Theorem 2) require the EIV distribution to be mildly ill-posed, while the mixture of normals we consider here technically may be severely ill-posed. However, empirically, a mixture of normals can produce relatively thick density tails similar to a mildly ill-posed distribution, and we show below that our estimator based on a mixture of normals is well behaved.

[11] If we assume asymptotic normality, we estimate the critical value for simultaneous confidence intervals to be 2.92, roughly 50% wider than the 1.96 used in these pointwise error bands.

FIGURE 1. Monte Carlo Simulation Results

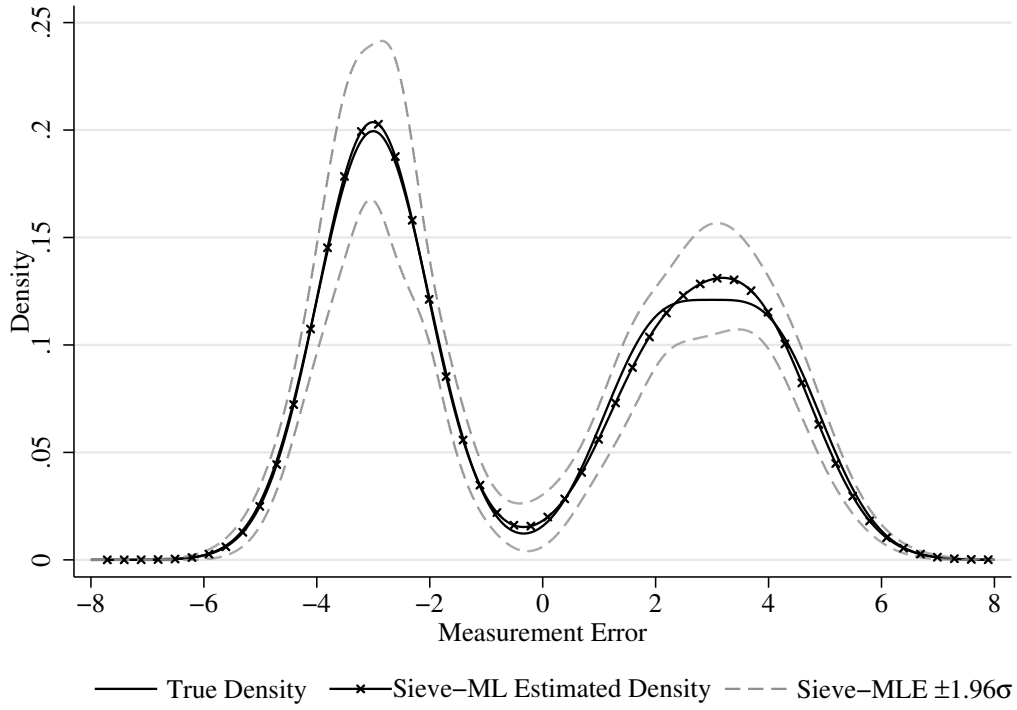


Notes: Figure plots the true coefficient vectors (lines) against quantile-regression estimates (circles), sieve MLE (×s), and $\pm 1.96\sigma$ error bands for the ML estimates (dashed lines) from 500 Monte Carlo simulations using the data-generating process described in the text with the measurement error generated as a mixture of three normals.

have desirable coverage when using a mixture of normals, we further calculate bootstrap confidence intervals for each simulation using the procedure described in section 3.2. We then calculate the fraction of simulations for which the true parameter lies within the bootstrapped confidence interval. Averaging the coverage across all $\beta(\cdot)$ and σ parameters, our bootstrapped confidence intervals have a coverage of 98%, suggesting them to be slightly conservative on average—see Online Appendix Table B5.

Focusing on Panels II and III of Figure 1 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 sieve-ML estimates fall almost directly on top of the true parameter function, and the bias of the sieve-ML estimator is nearly indistinguishable from zero at all quantiles. The average absolute bias for the sieve-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. Online Appendix Table B1 confirms that the quantile-regression average absolute bias is 26 and 16 times larger than the sieve MLE bias for $\beta_2(\cdot)$ and $\beta_3(\cdot)$, respectively. Online Appendix Table B1 further reports MSE results, showing that the average MSE is an order of magnitude smaller for the sieve-ML estimates than the quantile-regression estimates. Figure 1 also shows that quantile-regression estimates of the intercept term $\beta_1(\cdot)$ are badly biased. Given that quantile-regression estimated intercepts ensure that the τ^{th} conditional

FIGURE 2. Monte Carlo Simulation Results: Distribution of Measurement Error



Notes: Figure reports the true measurement error density (solid line) and the average sieve-ML estimated density from 500 Monte Carlo simulations (×s). The true measurement error distribution is 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. For each grid point, the dashed lines plot $\pm 1.96\sigma$ error bands for the ML estimates of the EIV density function, where σ is the standard deviation of the estimated density at that grid point across all Monte Carlo simulations.

quantile of the residuals $Q_{\widehat{\varepsilon}}(\tau) = 0$, when the slope coefficients are biased, this exacerbates the bias in the constant function. Whereas the mean absolute bias of the sieve-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 2 shows the true mixed-normal distribution of the measurement error ε as defined above (solid line) plotted with the estimated distribution of the measurement error from the average estimated distributional parameters across all Monte Carlo simulations (line with ×s). Monte-Carlo confidence intervals for the estimated density (dashed lines) are estimated pointwise as $\pm 1.96\sigma$ where σ is the standard deviation of the density estimates for each grid point 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 1 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 sieve-ML estimates that model the EIV distribution as a mixture of three normals still significantly outperform quantile regression.[12] As expected, 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, sieve-ML estimates are generally much less biased than quantile regression for both data-generating processes. Our sieve-ML framework easily accommodates mixtures of more than three normal components for additional distributional flexibility in a quasi-MLE approach. Online Appendix B provides additional simulation results—including both mean bias and MSE—for alternative measurement error distributions 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 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, the Skorohod representation of 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 $\beta(\cdot)$.

Figure 3 plots quantile-regression estimates of equation (5.1) using census microdata samples from four decennial census years: 1980, 1990, 2000, and 2010.[13] Consistent with Angrist

[12]Notably, the Laplace distribution is mildly ill-posed and our estimator using a mixture of normals accommodates such a DGP quite well.

[13]For further details on the data including summary statistics, see Online Appendix C. For comparability with Angrist et al. (2006) and to have a sufficient observations to run our estimator, we focus on prime age white males (aged 40-49). In Hausman et al. (2019), we provide evidence that other demographic groups have

TABLE 1. 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	SMLE	QR	SMLE	QR	SMLE	QR	SMLE
0.1	0.14	0.02	0.10	0.00	0.18	0.02	0.13	0.00
0.2	0.11	-0.01	0.05	-0.01	0.13	0.00	0.05	0.00
0.3	0.09	-0.02	0.02	-0.01	0.09	0.01	0.02	0.01
0.4	0.06	0.02	0.00	0.00	0.06	0.02	0.00	0.00
0.5	0.03	0.02	-0.02	0.01	0.03	-0.01	-0.02	-0.01
0.6	0.00	0.00	-0.03	0.00	0.00	-0.01	-0.03	-0.01
0.7	-0.05	0.00	-0.04	-0.01	-0.05	-0.02	-0.05	0.00
0.8	-0.11	-0.02	-0.06	-0.01	-0.13	-0.01	-0.07	-0.01
0.9	-0.20	-0.02	-0.08	-0.01	-0.24	-0.01	-0.10	-0.01
$\overline{\lvert\text{Bias}\rvert}$	0.09	0.01	0.05	0.01	0.10	0.01	0.05	0.01

Notes: Table reports mean bias of slope coefficients for estimates from classical quantile regression and sieve MLE modeling the error term as a mixture of three normals across 500 Monte Carlo 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.

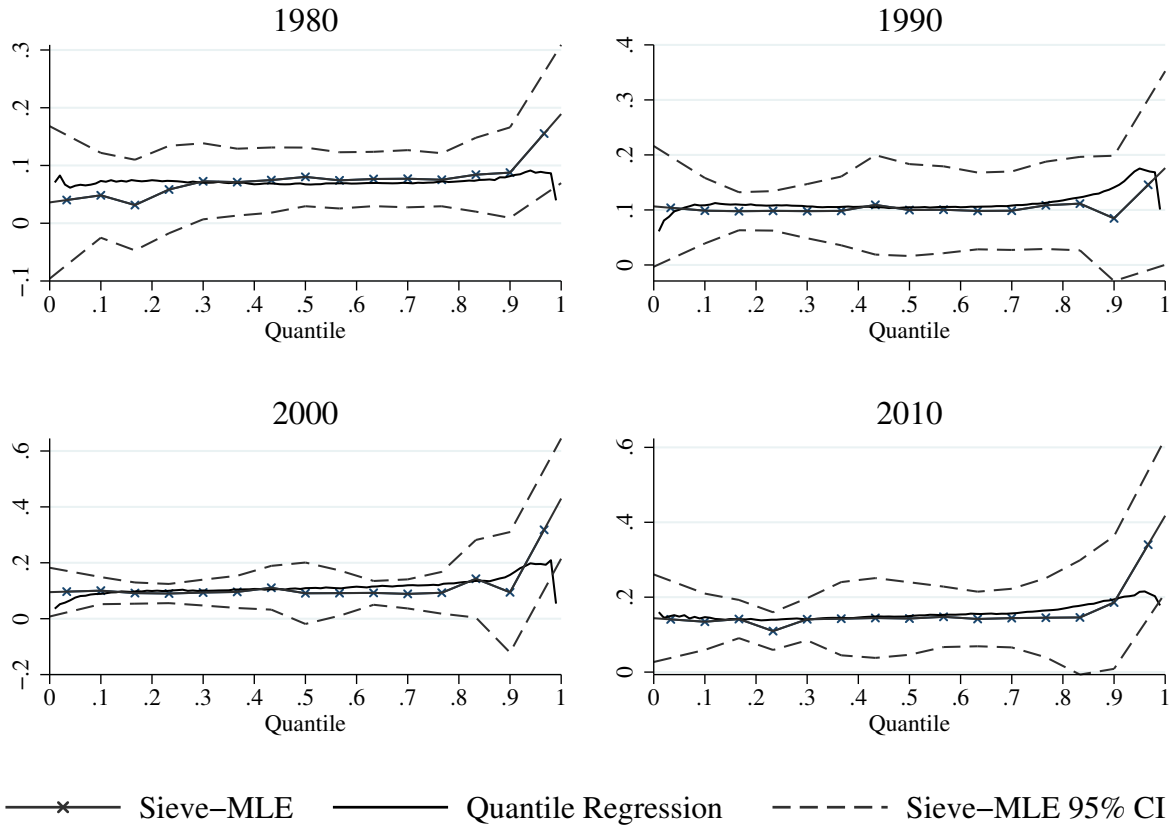
et al. (2006), we find quantile-regression evidence that heterogeneity in the returns to education across the conditional wage distribution has increased over time. Adding data from 2010 shows a large jump in the returns to education for the entire distribution, with top conditional incomes increasing 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 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 3 also plots the education coefficient $\widehat{\beta}_2(\tau)$ from

markedly different patterns of heterogeneity in the education-wage gradient across the conditional income distribution, motivating further study on treatment effect heterogeneity.

[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 3. Returns to Education Correcting for LHS Measurement Error



Notes: Figure reports quantile regression (solid lines) and sieve-ML estimates (lines with circles) of (self-reported) log weekly wages on education and a quadratic in experience. Dashed 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.

estimating equation (5.1) by sieve MLE. We approximate $\beta(\cdot)$ with a piecewise linear function consisting of 15 knots using our sieve-ML estimator developed in Section 3 (see Online Appendix A for implementation details). 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 sieve-ML estimates. A formal test of the joint equality across the grid of 15 knots of QR and sieve-ML coefficients rejects equality of the education coefficient function for each year except 1990. For 1980, the quantile-regression estimates show relatively constant returns to education across the conditional

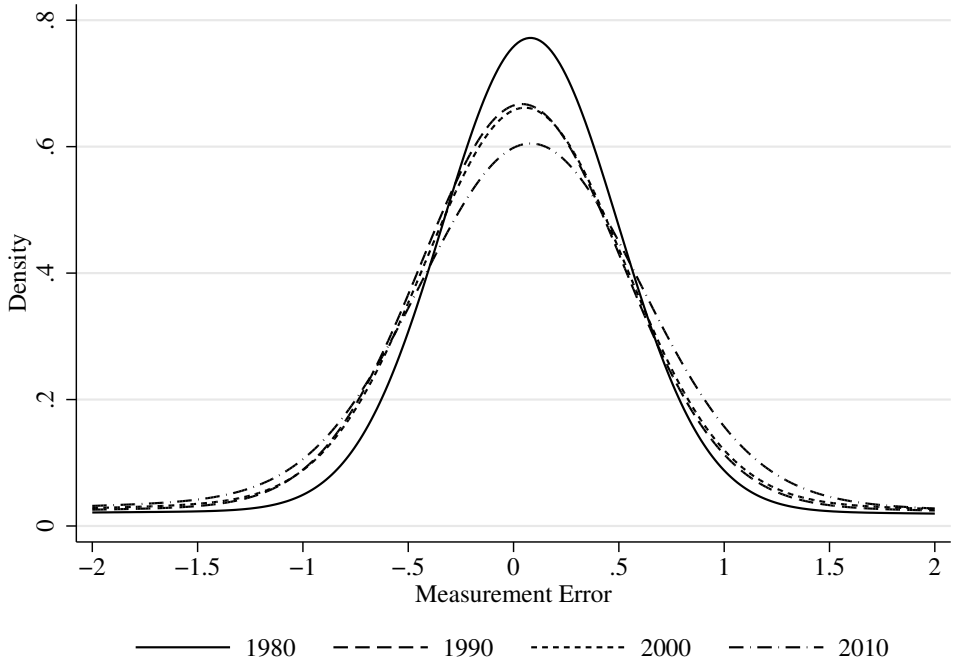
wage distribution, with a sharp decline at the very top characteristic of quantile-regression estimates at extremal quantiles. The sieve-ML estimates feature more convexity, with the pattern of increasing returns to education for higher quantiles seen in quantile-regression estimates in later years visible in the sieve-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 sieve-ML estimates are nearly indistinguishable given the typically wide confidence intervals for extremal quantiles, and we fail to reject equality of QR and sieve-ML estimates.

In the 2000 sample, the quantile-regression and sieve-ML estimates of the returns to education again diverge for top incomes, with the point estimate suggesting that after correcting for measurement error in self-reported wages, the true returns to an additional year of education for 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 sieve-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 sieve-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 sieve-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. 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,

FIGURE 4. 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.

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 4 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. Notably, Figure 4 implies larger average errors in self-reported income than found by Bollinger's (1998) analysis of the Current Population Survey using validation data. This suggests a less literal interpretation of measurement error that includes other forms of additively separable independent unobserved heterogeneity in wages not captured by our covariates. As mentioned above, our approach cannot distinguish between additively separable measurement error and additively separable independent unobserved heterogeneity, although the presence of either significantly changes estimates of the coefficients of interest in a quantile-regression model. Over time, the variance in the measurement

[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).

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 sieve-ML-based approach. Under an assumption about the degree of ill-posedness of the problem (Assumption 6), 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 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.

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.

Bollinger, C. R. (1998): "Measurement Error in the Current Population Survey: A Nonparametric Look," *Journal of Labor Economics*, 16(3), 576–594.

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

Chernozhukov, V., and C. Hansen (2005): "An IV model of quantile treatment effects," *Econometrica*, 73(1), 245–261.

Chetverikov, D., B. Larsen, and C. Palmer (2016): "IV quantile regression for group-level treatments, with an application to the distributional effects of trade," *Econometrica*, 84(2), 809–833.

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

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

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

Hausman, J. A., H. Liu, Y. Luo, and C. Palmer (2019): "Errors in the dependent variable of quantile regression models," NBER Working Paper #25819.

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 appendix, 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. Proofs of Theorem 1 and Lemma 1

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 $m(s) = \frac{\phi_{\varepsilon 0}(s)}{\phi_\varepsilon(s)}$. 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

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

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,

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

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

$$\int_0^1 \beta_1(\tau)^k d\tau = \int_0^1 \beta_{0,1}(\tau)^k d\tau. \quad \text{(B.2)}$$

Note that the denominator of $m(s)$ is well behaved as $s \to 0$. Since $\phi_\varepsilon(0) = 1$, $f(\cdot|\sigma)$ is continuous in σ, and Σ is compact, $\phi_\varepsilon(\cdot)$ is uniformly continuous. Then there is an open neighborhood of $s = 0$ such that $C_1 < |\phi_\varepsilon(s)| < C_2$ and $C_1 < |\phi_{\varepsilon 0}(s)| < C_2$ for some positive constants C_1 and C_2, implying that $m(s)$ exists and is bounded from above and below in an open neighborhood of $s = 0$.

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

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

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 (B.3) is well defined. Combining (B.2) and (B.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. 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 (B.1) again. Dividing both sides by $(is)^k/k!$, we have, for all $k \geq 0$

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

Consider the first-order terms of s in (B.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 (B.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 2.3, $\beta_{0,1}(\tau)^k, k \geq 0$ 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 2.1, $E[xx^T]$ 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 have $\beta_{-1}(\pi(\tau)) = \beta_{0,-1}(\tau)$ 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, $\phi_{x\beta}(s|x)\phi_\varepsilon(s) = \phi_{x\beta_0}(s|x)\phi_{\varepsilon_0}(s)$. Then $\phi_{x\beta}(s|x) = m(s)\phi_{x\beta_0}(s|x)$, 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 $\phi_{x\beta}(s|x)$ around $s = 0$,

$$\begin{aligned}
&\sum_{k=0}^{\infty} \int_0^1 \frac{(is)^k}{k!} [(x_1, x_1^2)\beta_w(\tau) + x_{-w}^T \beta_{-w}(\tau)]^k d\tau \\
=&\left(\sum_{k=0}^{\infty} a_k s^k \right) \left(\sum_{k=0}^{\infty} \int_0^1 \frac{(is)^k}{k!} [(x_1, x_1^2)\beta_{0,w}(\tau) + x_{-w}^T \beta_{0,-w}(\tau)]^k d\tau \right), \\
=&\sum_{k=0}^{\infty} s^k \left[\sum_{l=0}^{k} a_{k-l} \int_0^1 \frac{i^l}{l!} \left[(x_1, x_1^2)\beta_{0,w}(\tau) + x_{-w}^T \beta_{0,-w}(\tau) \right]^l d\tau \right] \qquad \text{(B.5)}
\end{aligned}$$

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 (B.5) for s^k, we have

$$\int_0^1 \frac{i^k}{k!} [(x_1, x_1^2)\beta_w(\tau) + x_{-w}^T \beta_{-w}(\tau)]^k d\tau = \sum_{l=0}^{k} a_{k-l} \int_0^1 \frac{i^l}{l!} [(x_1, x_1^2)\beta_{0,w}(\tau) + x_{-w}^T \beta_{0,-w}(\tau)]^l d\tau \qquad \text{(B.6)}$$

holding for any k, x_1 and x_{-w}. Comparing both sides of (B.6) 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, $\beta_{x_1^2}(\tau)$ 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 (B.6) 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 4, 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 (B.6) for any k and the coefficients for x_1^{2k-2}, we have

$$\begin{aligned}
&\frac{i^k}{k!} \int_0^1 \big(\beta_{x_1^2}(\tau)\big)^{k-2} \big(\beta_{x_1}(\tau)\big)^2 + \big(\beta_{x_1^2}(\tau)\big)^{k-1} x_{-w}^T \beta_{-w}(\tau) d\tau \\
=&\frac{i^k}{k!} \int_0^1 \big(\beta_{0,x_1^2}(\tau)\big)^{k-2} \big(\beta_{0,x_1}(\tau)\big)^2 + \big(\beta_{0,x_1^2}(\tau)\big)^{k-1} x_{-w}^T \beta_{0,-w}(\tau) d\tau \qquad \text{(B.7)}
\end{aligned}$$

where we used the fact that $a_1 = 0$ and thus for a fixed k, the only l on the right-hand side of (B.6) 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, (B.7) can be rewritten as

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

Again, using the fact that $\left(\beta_{x_1^2}(\tau)\right)^{k-1}, k \geq 1$ forms a functional basis of $L^2[0,1]$, we have for any x_{-w} that $x_{-w}^T \beta_{-w}(\pi(\tau)) = x_{-w}^T \beta_{0,-w}(\tau)$ 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 $x^T\beta_0(\tau)$, 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 (B.5). Then we compare the coefficients for each term s^k and get

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

Using the fact that for each $k \geq 1$, the coefficients for x_1^{kp} on both sides of (B.8) 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 (B.8) 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 (B.8), 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 $\varepsilon \in \mathbb{R}$. □