应用概率统计 第 33 卷
第 5 期 2017 年 10 月

Chinese Journal of Applied Probability and Statistics
Oct., 2017, Vol. 33, No. 5, pp. 475-486
doi: 10.3969/j.issn.1001-4268.2017.05.004

Instrumental Variable Estimation in Linear Quantile Regression Models with Measurement Error *

GUAN Jing
(*School of Mathematics, Tianjin University, Tianjin 300350, China*)
WANG LiQun*
(*Department of Statistics, University of Manitoba, Winnipeg, Manitoba, Canada R3T 2N2*)

Abstract: We extend the instrumental variable method for the mean regression models to linear quantile regression models with errors-in-variables. The proposed estimator is consistent and asymptotically normally distributed under some fairly general conditions. Moreover, this approach is practical and easy to implement. Simulation studies show that the finite sample performance of the estimator is satisfactory. The method is applied to a real data study of education and wages.

Keywords: errors in variables; instrumental variables; least absolute deviation; measurement error; quantile regression

2010 Mathematics Subject Classification: 62F10; 62F12; 62J99

Citation: Guan J, Wang L Q. Instrumental variable estimation in linear quantile regression models with measurement error [J]. Chinese J. Appl. Probab. Statist., 2017, 33(5): 475–486.

§1. Introduction

Quantile regression has drawn much attention in the literature. It is a useful tool for estimating conditional quantiles of a response variable given a set of predictors [1]. While the mean regression describes the effect of predictors on the average response, the quantile regression describes the effect on its entire distribution. It has been applied in many areas such as biology, ecology, economics, finance and environmental science.

In practice, often some predictors are not directly observable or are measured with substantial errors. It is known that simple substitution of the surrogate data for the latent variables will result in attenuated and inconsistent estimators in either mean or quantile regression models [2–4].

*The project was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).
*Corresponding author, E-mail: Liqun.Wang@ad.umanitoba.ca.
Received December 11, 2015. Revised July 8, 2016.

Quantile regression with mismeasured covariates has been investigated before, however, the publications in this area are sparse due to the difficult nature of the problem. Brown [5] examined a median regression model and described the difficulty of associated parameter estimation. He and Liang [6] studied the orthogonal distance quantile regression in a linear measurement error model where the measurement and regression errors have a joint spherically symmetric distribution. Wei and Carroll [4] proposed consistent estimators for a linear quantile regression model using modified score function and replicate data. Ma and Yin [7] used a composite weighting estimation method to deal with censored linear quantile regression with spherically distributed measurement error. On the other hand, Hu and Schennach [8] and Schennach [9] used instrumental variable method to establish the identifiability and propose consistent estimators of a nonparametric quantile regression model.

While the instrumental variables (IV) method has been widely used in mean regression with errors in variables (e.g., [2]; [3]; [10]; [11]; [12]; [13]; [14]; [15]), its use in parametric quantile regression is rare. This paper attempts to explore this possibility. Generally speaking, the measurement error problem is difficult to deal with in quantile regression setup, mainly due to the lack of additive property of probability quantiles. For example, unlike the mean function the quantile of the sum of two random variables does not equal to the sum of two quantiles. This prevents the extension of usual manipulations of additive measurement errors in mean regression setups. Due to these difficulties, in this paper we consider a simple linear quantile regression model and demonstrate that the usual IV procedure in the mean regression setup can be carried over to certain quantile regression models.

To illustrate the IV method in the mean regression setup, let us consider the linear model

$$y = \alpha_1 + \alpha_2' x + \epsilon, \tag{1}$$

where $y \in I\!R$ is the response variable, $x \in I\!R^k$ is the vector of predictor variables, ϵ is a random error with $\mathsf{E}(\epsilon \,|\, x) = 0$, and $\alpha = (\alpha_1, \alpha_2')'$ is an unknown parameter vector. Here it is assumed that x is unobservable and the observed predictor is

$$w = x + u, \tag{2}$$

where u is the random measurement error satisfying $\mathsf{E}(u \,|\, x) = 0$. Usually the variance-covariance matrix of u, Σ_u, can be singular to allow some components of x to be measured without error. It is well-known that simply substituting w for x in equation (1) will give inconsistent estimator for α because w and u are dependent. One possible way to overcome

this problem is to use the instrumental variables which, by definition, are correlated with x but uncorrelated with ϵ and u [2]. Following [12,13], we assume that there exists a vector of instrumental variables $z \in I\!R^{\ell}$ that is related to x through

$$x = \beta_1 + \beta_2' z + \delta, \tag{3}$$

where the random error δ is uncorrelated with u and ϵ and satisfies $\mathsf{E}(\delta \mid z) = 0$.

The IV method for the mean regression consists of the following steps. First, substituting (3) into (1) yields a usual mean regression equation

$$y = \gamma_1 + \gamma_2' z + \nu, \tag{4}$$

where $\nu = \alpha_2'\delta + \epsilon$ and

$$\gamma_1 = \alpha_1 + \alpha_2'\beta_1, \tag{5}$$

$$\gamma_2 = \beta_2\alpha_2. \tag{6}$$

Since z is uncorrelated with δ and ϵ, γ can be consistently estimated by least squares fitting of y on z. Further, substituting (3) into (2) yields

$$w = \beta_1 + \beta_2' z + (\delta + u) \tag{7}$$

which can be used to obtain consistent estimators for β_1 and β_2. Finally, consistent estimators for α are obtained by solving equations (5)–(6), which yields $\alpha_2 = (\beta_2'\beta_2)^{-1}\beta_2'\gamma_2$ and $\alpha_1 = \gamma_1 - \alpha_2'\beta_1$. In subsequent sections, we adapt this procedure to least absolute deviation (LAD) estimation and more general quantile regression setups.

The paper is organized as follows. Section 2 introduces the IV method to LAD estimation in the linear median regression model with errors in variables. Section 3 extends this method to more general quantile regression models. Simulation studies are presented in Section 4, while an empirical example is given in Section 5. Finally, conclusions and discussion are given in Section 6.

§2. LAD Estimation

In this section, we extend the IV method described in Section 1 to the LAD estimation of model (1)–(3). To this end we make the following assumptions.

Assumption 1 ϵ, δ and u are symmetrically distributed about zero. Further, ν has a continuous density satisfying $f_\nu(0) > 0$.

Assumption 2 The data (y_i, w_i, z_i), $i = 1, 2, \ldots, n$ are independent and identically distributed.

Assumption 3 $M_z = \mathsf{E}(zz')$ is positive definite and $\max\limits_{1\leqslant i\leqslant n} \|z_i\| = o_p(\sqrt{n}\,)$.

Then under Assumption 1, $\nu = \alpha_2'\delta + \epsilon$ in equation (4) is symmetric about zero. Therefore the LAD estimator $\widehat{\gamma}$ of $\gamma = (\gamma_1, \gamma_2')'$ can be obtained by minimizing the objective function

$$\sum_{i=1}^{n} |y_i - \gamma_1 - \gamma_2' z_i|. \tag{8}$$

According to [16], Assumptions 1 and 3 ensure that $\sqrt{n}(\widehat{\gamma} - \gamma) \to \mathrm{N}(0, \Sigma_{\widehat{\gamma}})$, where $\Sigma_{\widehat{\gamma}} = M_z^{-1}/[4f_\nu^2(0)]$. Similarly, from equation (7) the LAD estimator $\widehat{\beta}$ of $\beta = (\beta_1, \beta_2')'$ can be obtained by minimizing the objective function

$$\sum_{i=1}^{n} |w_i - \beta_1 - \beta_2' z_i|. \tag{9}$$

Now, write (5) and (6) jointly as

$$\gamma = B\alpha, \tag{10}$$

where $\alpha = (\alpha_1, \alpha_2')'$ and

$$B = \begin{pmatrix} 1 & \beta_1' \\ 0 & \beta_2' \end{pmatrix}.$$

Then consistent estimator of α can be obtained by solving (10) provided consistent estimators of γ and $\beta = (\beta_1, \beta_2')'$ are available. Specifically, given the consistent estimators $\widehat{\gamma}$ and $\widehat{B}$, consistent estimator for α can be obtained by minimizing $(\widehat{\gamma} - \widehat{B}\alpha)' A_n (\widehat{\gamma} - \widehat{B}\alpha)$, where A_n is a non-negative definite weight matrix which may depend on the data. The resulting minimum distance estimator is given by

$$\widehat{\alpha} = (\widehat{B}' A_n \widehat{B})^{-1} \widehat{B}' A_n \widehat{\gamma}. \tag{11}$$

Further, by delta-method we have

$$\sqrt{n}(\widehat{\alpha} - \alpha) \to \mathrm{N}(0, (B'AB)^{-1} B' A \Sigma_{\widehat{\gamma}} AB (B'AB)^{-1}), \tag{12}$$

where $A = \mathrm{plim}(A_n/n)$. The above asymptotic variance-covariance matrix has a lower bound $(B'\Sigma_{\widehat{\gamma}}^{-1} B)^{-1}$ which is attained when $A = \Sigma_{\widehat{\gamma}}^{-1}$. Therefore the most efficient choice of wight is $A_n = \widehat{\Sigma}_{\widehat{\gamma}}$ which is a consistent estimator for $\Sigma_{\widehat{\gamma}}$. Another less efficient but practical choice is $A_n = Z'Z$, which gives $\widehat{\alpha} = (W'W)^{-1} W' Z \widehat{\gamma}$, where $W = Z\widehat{B}$ and

$$Z' = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ z_1 & z_2 & \cdots & z_n \end{pmatrix}.$$

§3. General Quantile Regression

In this section we consider more general quantile regression model

$$Q_\tau(y \mid x) = \alpha_1 + \alpha_2' x + F_\epsilon^{-1}(\tau), \qquad \tau \in (0,1), \tag{13}$$

where F_ϵ is the distribution of ϵ. If x is observed, then the quantile regression estimator of $\alpha(\tau) = (\alpha_1 + F_\epsilon^{-1}(\tau), \alpha_2)$ can be obtained by minimizing the objective function

$$\sum_{i=1}^{n} \rho_\tau(y_i - \alpha_1 - \alpha_2' x_i), \tag{14}$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$ and $I(\cdot)$ is the indicator function. For any given τ, the above quantile regression produces consistent estimator for $\alpha(\tau)$. In particular, the estimator corresponding to $\tau = 0.5$ is the least absolute deviation (LAD) estimator.

In the case of unobserved x, the instrumental variable z can be used instead. Unfortunately, the above IV procedure for mean regression cannot be directly applied here because the quantiles $F_\epsilon^{-1}(\tau)$ and $F_\nu^{-1}(\tau)$ are not the same. In order to derive consistent quantile regression estimators, we modify the procedure as follows. First, rewrite (4) as

$$y = \alpha_1 + \alpha_2' t + \nu, \tag{15}$$

where $t = \beta_1 + \beta_2' z$. For the sake of clarity, we first assume that the true value of β is known. The case of estimated β will be discussed later. Then the corresponding quantile regression model is

$$Q_{\tau^*}(y \mid t) = \alpha_1 + \alpha_2' t + F_\nu^{-1}(\tau^*), \qquad \tau^* \in (0,1), \tag{16}$$

and therefore, under some regularity conditions, the estimator given by

$$\widehat{\widetilde{\alpha}}(\tau^*) = \underset{\widetilde{\alpha} \in I\!R^k}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_{\tau^*}(y_i - \alpha_1 - \alpha_2' t_i) \tag{17}$$

converges to $\widetilde{\alpha}(\tau^*) = (\alpha_1 + F_\nu^{-1}(\tau^*), \alpha_2)$. Thus consistent estimator of $\alpha(\tau)$ in model (13) can be obtained by choosing τ^* such that $F_\nu^{-1}(\tau^*) = F_\epsilon^{-1}(\tau)$. Note that $F_\nu^{-1}(\tau^*)$ can be estimated by using the empirical distribution of the residuals from the quantile regression of (16) at $\tau^* = 0.5$. However, it is generally difficult to determine $F_\epsilon^{-1}(\tau)$ unless either the true distribution of ϵ is known or the corresponding "residuals" are available. Therefore, we make the following assumption.

Assumption 4 The distributions of ϵ and ν belong to the same family of location-scale distributions.

Under this assumption, ϵ/σ_ϵ and ν/σ_ν have the same distribution and hence $F_\epsilon^{-1}(\tau)/\sigma_\epsilon = F_\nu^{-1}(\tau)/\sigma_\nu$. Therefore τ^* can be determined by $F_\nu^{-1}(\tau^*) = F_\nu^{-1}(\tau)\sigma_\epsilon/\sigma_\nu$. Further, since $F_\nu^{-1}(\tau)$ and σ_ν can be estimated using the empirical residuals from the quantile regression of (16) at $\tau^* = 0.5$, it only remains to estimate σ_ϵ, which can be achieved using the method of moments as follows.

From (1) we have

$$\begin{aligned}\mathsf{E}(y^2 \,|\, z) &= \mathsf{E}[(\alpha_1 + \alpha_2' x)^2 \,|\, z] + \mathsf{E}(\epsilon^2 \,|\, z) \\ &= \sigma_\epsilon^2 + \mathsf{E}[(\gamma_1 + \gamma_2' z + \alpha_2'\delta)^2 \,|\, z)] \\ &= \sigma_\epsilon^2 + \alpha_2'\Sigma_\delta\alpha_2 + (\gamma_1 + \gamma_2' z)^2 \\ &= \sigma_\epsilon^2 + \alpha_2'\Sigma_\delta\alpha_2 + \gamma_1^2 + 2\gamma_1\gamma_2' z + \gamma_2' z z'\gamma_2. \end{aligned} \tag{18}$$

Hence $\varphi = \sigma_\epsilon^2 + \alpha_2'\Sigma_\delta\alpha_2 + \gamma_1^2$ can be consistently estimated by regressing y^2 on z and the elements of zz'. Therefore, σ_ϵ^2 can be obtained as long as $\Sigma_\delta\alpha_2$ can be estimated. To obtain the estimate of $\Sigma_\delta\alpha_2$, we consider

$$\begin{aligned}\mathsf{E}(wy \,|\, z) &= \mathsf{E}[x(\alpha_1 + \alpha_2' x) \,|\, z] \\ &= \mathsf{E}[(\beta_1 + \beta_2' z + \delta)(\gamma_1 + \gamma_2' z + \alpha_2'\delta) \,|\, z)] \\ &= (\beta_1 + \beta_2' z)(\gamma_1 + \gamma_2' z) + \Sigma_\delta\alpha_2 \\ &= \Sigma_\delta\alpha_2 + \gamma_1\beta_1 + (\gamma_1\beta_2' + \beta_1\gamma_2')z + \beta_2' z z'\gamma_2. \end{aligned} \tag{19}$$

It follows that $\psi = \Sigma_\delta\alpha_2 + \gamma_1\beta_1$ can be estimated by regressing wy on z and the elements of zz'. Finally, we have $\Sigma_\delta\alpha_2 = \psi - \gamma_1\beta_1$ and thus $\sigma_\epsilon^2 = \varphi - \gamma_1^2 - \alpha_2'\Sigma_\delta\alpha_2$.

Now we derive the asymptotic property of the estimator $\widehat{\widehat{\alpha}}(\tau^*)$. First, by Assumption 3 we have

$$\max_{1\leqslant i\leqslant n} \|t_i\| \leqslant \Big(\|\beta_1\| + \|\beta_2\| \max_{1\leqslant i\leqslant n} \|z_i\|\Big) = o_p(\sqrt{n}\,).$$

Further,

$$\begin{aligned}\operatorname*{plim}_{n\to\infty} n^{-1} \sum_{i=1}^n t_i t_i' &= \operatorname*{plim}_{n\to\infty} n^{-1} \sum_{i=1}^n (\beta_1\beta_1' + \beta_2' z_i\beta_1' + \beta_1 z_i'\beta_2 + \beta_2' z_i z_i'\beta_2) \\ &= \beta_1\beta_1' + \beta_1\mathsf{E}(z')\beta_2 + (\beta_1\mathsf{E}(z')\beta_2)' + \beta_2'\mathsf{E}(zz')\beta_2 \\ &= (\beta_1 + \beta_2'\mathsf{E}(z))(\beta_1 + \beta_2'\mathsf{E}(z))' + \beta_2'\mathsf{E}(z)\beta_2 \\ &\equiv M_t \end{aligned}$$

which is clearly positive definite. It follows from Theorem 4.1 of [17] that, for any $\tau \in (0,1)$ and $\tau^* = F_\nu(F_\epsilon^{-1}(\tau))$, as $n \to \infty$, $\sqrt{n}(\widehat{\widehat{\alpha}}(\tau^*) - \alpha(\tau)) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \omega^2 M_t^{-1})$, where $\omega^2 = \tau^*(1-\tau^*)/f^2(F_\nu^{-1}(\tau^*))$, provided $f_\nu(F_\nu^{-1}(\tau^*)) > 0$.

In summary, the practical implementation of the proposed IV-QR method cosists of the following steps:

(1) use a subset of the sample to calculate the LAD estimate $\widehat{\beta}$;

(2) use the rest of the sample to calculate $t_i = \widehat{\beta}_1 + \widehat{\beta}_2' z_i$ and obtain the residuals ν_i from the median regression of (16) (with $\tau^* = 0.5$);

(3) estimate σ_ϵ^2 through mean regressions (18) and (19);

(4) calculate the empirical CDF of the residuals ν_i and determine τ^* such that $\widehat{F}_\nu^{-1}(\tau^*) = \widehat{F}_\nu^{-1}(\tau)\sigma_\epsilon/\sigma_\nu$; and

(5) calculate the estimate $\widehat{\widetilde{\alpha}}(\tau^*)$ by solving the optimization problem (17).

Since we estimate β and α using separate samples, it is easy to see that the final estimator $\widehat{\widetilde{\alpha}}(\tau^*)$ is consistent and asymptotically normally distributed conditional on the pre-estimator $\widehat{\beta}$. However, the unconditional asymptotic covariance matrix of $\widehat{\widetilde{\alpha}}(\tau^*)$ should be larger than $\omega^2 M_t^{-1}$ and can be calculated using the delta-method.

§4. Simulation Studies

In this section, we illuminate our method and examine its finite sample performance through some simulation studies. For simplicity we consider models with $k = 1$ predictor and $\ell = 1$ instrumental variable. In particular, we compare our proposed IV estimator with the "gold standard" estimator computed using the true observations (y_i, x_i). We also include the orthogonal regression estimator of [6] which is consistent for $(\alpha_\tau^*, \alpha_2(\tau)) = (\alpha_1 + q_\tau\sqrt{1+|\alpha_2|^2}, \alpha_2(\tau))$ and is calculated by minimizing the objective function

$$\sum_{i=1}^{n} \rho_\tau\Big(\frac{y_i - a - w_i b}{\sqrt{1+|b|^2}}\Big),$$

where q_τ is the solution of $\mathsf{E}\rho_\tau(\epsilon_i - q) = 0$. However, for comparison purpose in the following we will deduce $\widehat{\alpha}_1(\tau) = \alpha_1 + (\widehat{\alpha}_\tau^* - \alpha_1)/\sqrt{1+|\widehat{\alpha}_2(\tau)|^2}$ from α_τ^* for the H-L estimator. Note that this method requires that the regression error and measurement error to have equal variances.

In all studies, we consider model (13) where variables and parameters are generated as follows: $z_i \sim \mathrm{U}(-6, 6)$; $x_i = \beta_1 + \beta_2 z_i + \delta_i$ with $\beta_1 = 1$, $\beta_2 = 0.5$ and $\delta_i \sim \mathrm{N}(0, \sigma_\delta^2)$; $w_i = x_i + u_i$, where $u_i \sim \mathrm{N}(0, 1)$; and $y_i = \alpha_1 + \alpha_2 x_i + \epsilon_i$ with $\alpha_1 = 2$, $\alpha_2 = 1$ and $\epsilon_i \sim \mathrm{N}(0, \sigma_\epsilon^2)$. All variables are generated independently and the sample size $n = 500$. In each simulation, 1 000 Monte Carlo replications are carried out. We consider three scenarios.

We first consider the case where the regression and measurement errors have equal variances $\sigma_\epsilon^2 = \sigma_u^2 = 1$. For this case, the results of selected quantile estimates are given in Table 1, where the corresponding simulation standard errors (SSE) are given in parentheses. Compared with the "gold standard" estimator, both the IV and H-L estimators perform fairly well, with the later more efficient.

Table 1 The case of $\alpha_1 = 2$, $\alpha_2 = 1$, $\epsilon \sim \mathrm{N}(0,1)$, $u \sim \mathrm{N}(0,1)$, $\delta \sim \mathrm{N}(0,1)$

	Gold standard		IV method		H-L method	
τ	$\widehat{\alpha}_1$	$\widehat{\alpha}_2$	$\widehat{\alpha}_1$	$\widehat{\alpha}_2$	$\widehat{\alpha}_1$	$\widehat{\alpha}_2$
0.10	0.7228	0.9996	0.7209	0.9866	0.7200	1.0023
	(0.0027)	(0.0013)	(0.0168)	(0.0081)	(0.0028)	(0.0018)
0.25	1.3271	1.0002	1.3297	0.9987	1.3269	1.0003
	(0.0022)	(0.0010)	(0.0094)	(0.0036)	(0.0021)	(0.0015)
0.50	1.9985	1.0010	1.9959	1.0015	2.0024	0.9996
	(0.0021)	(0.0009)	(0.0087)	(0.0028)	(0.0020)	(0.0013)
0.75	2.6712	1.0003	2.6596	0.9987	2.6723	1.0006
	(0.0021)	(0.0010)	(0.0093)	(0.0031)	(0.0022)	(0.0014)
0.90	3.2785	0.9985	3.2262	0.9947	3.2789	1.0010
	(0.0027)	(0.0012)	(0.0214)	(0.0070)	(0.0027)	(0.0018)

The second case we consider is where the regression and measurement errors have unequal variances $\sigma_\epsilon^2 = 2$ while $\sigma_u^2 = 1$. This represents a scenario where the assumption of equal variance for the H-L estimator is violated. However, we still include the results of this estimator to demonstrate the consequence of model misspecification. The results are given in Table 2. Again the IV estimates are similar to the "gold standard" estimates, while the H-L estimates are significantly different. In this sense the IV method is more robust to model misspecification than the H-L method. This experiment was repeated with slightly different value $\sigma_\delta^2 = 1.5$ and the corresponding results are reported in Table 3, which show the similar pattern as in Table 2.

To further investigate the robustness property of the IV estimator, in the rest of this section we simulate the scenario where the Assumption 4 does not hold and the distribution of ϵ is unknown. In particular, we generated data using $\epsilon \sim \mathrm{F}(2,2)$ but do not use this distributional information in the estimation procedure. Instead, we use the simulated results of ϵ to estimate $F_\epsilon^{-1}(\tau)$. Since β_1, β_2 and σ_ϵ^2 are not of main interest and can be easily estimated using the ordinary least squares regression, we use their true values here to eliminate the effects of estimating nuisance parameters. Since the H-L method does not apply to this scenario, we only compare the results of the IV estimator with the "gold

Table 2 The case of $\alpha_1 = 2$, $\alpha_2 = 1$, $\epsilon \sim \mathrm{N}(0, 2)$, $u \sim \mathrm{N}(0, 1)$, $\delta \sim \mathrm{N}(0, 1)$

	Gold standard		IV method		H-L method	
τ	$\widehat{\alpha}_1$	$\widehat{\alpha}_2$	$\widehat{\alpha}_1$	$\widehat{\alpha}_2$	$\widehat{\alpha}_1$	$\widehat{\alpha}_2$
0.10	-0.5432	0.9960	-0.5124	1.0006	-0.1534	-5.5928
	(0.0054)	(0.0025)	(0.0108)	(0.0046)	(0.0042)	(7.0428)
0.25	0.6572	0.9984	0.6642	1.0013	0.7433	1.4417
	(0.0044)	(0.0019)	(0.0086)	(0.0039)	(0.0033)	(0.0034)
0.50	2.0011	0.9997	2.0040	1.0062	1.7490	1.4418
	(0.0039)	(0.0017)	(0.0073)	(0.0037)	(0.0031)	(0.0030)
0.75	3.3522	1.0002	3.3512	0.9978	2.7539	1.4455
	(0.0043)	(0.0019)	(0.0083)	(0.0040)	(0.0033)	(0.0034)
0.90	4.5622	1.0015	4.5516	1.0032	3.6606	1.4571
	(0.0055)	(0.0025)	(0.0113)	(0.0045)	(0.0041)	(0.0045)

Table 3 The case of $\alpha_1 = 2$, $\alpha_2 = 1$, $\epsilon \sim \mathrm{N}(0, 2)$, $u \sim \mathrm{N}(0, 1)$, $\delta \sim \mathrm{N}(0, 1.5)$

	Gold standard		IV method		H-L method	
τ	$\widehat{\alpha}_1$	$\widehat{\alpha}_2$	$\widehat{\alpha}_1$	$\widehat{\alpha}_2$	$\widehat{\alpha}_1$	$\widehat{\alpha}_2$
0.10	-0.5599	0.9994	-0.5301	1.0049	-0.1289	1.3301
	(0.0054)	(0.0021)	(0.0125)	(0.0050)	(0.0041)	(0.0033)
0.25	0.6511	1.0004	0.6672	1.0059	0.7856	1.3277
	(0.0043)	(0.0017)	(0.0096)	(0.0045)	(0.0033)	(0.0026)
0.50	1.9953	0.9997	2.0087	0.9961	1.8041	1.3221
	(0.0039)	(0.0015)	(0.0083)	(0.0041)	(0.0030)	(0.0025)
0.75	3.3504	0.9988	3.3588	1.0000	2.8286	1.3234
	(0.0043)	(0.0017)	(0.0093)	(0.0046)	(0.0031)	(0.0026)
0.90	4.5588	1.0036	4.5419	1.0034	3.7384	1.3362
	(0.0052)	(0.0021)	(0.0122)	(0.0048)	(0.0042)	(0.0035)

standard" estimates. The simulation results are shown in Table 4. We can see that the proposed IV method performs very well in this scenario.

§5. Application

In this section we apply the proposed method to the quantile regression analysis of wages and schooling for the twins. Ashehfelter and Krueger [18] used mean regression model to study a data set of 149 pairs of wins from the US Bureau of the Census for the Current Population Survey. The objective of the study was to evaluate the effects of education on

Table 4 The case of $\alpha_1 = 2$, $\alpha_2 = 1$, $\epsilon \sim \mathrm{F}(2,2)$, $u \sim \mathrm{N}(0,1)$, $\delta \sim \mathrm{N}(0,1)$

	Gold standard		IV method	
τ	$\widehat{\alpha}_1$	$\widehat{\alpha}_2$	$\widehat{\alpha}_1$	$\widehat{\alpha}_2$
0.10	2.1142	0.9996	2.1208	0.9985
	(0.0006)	(0.0003)	(0.0044)	(0.0028)
0.25	2.3388	0.9993	2.3352	1.0053
	(0.0012)	(0.0005)	(0.0046)	(0.0029)
0.50	3.0094	1.0001	3.0149	1.0022
	(0.0032)	(0.0014)	(0.0063)	(0.0031)
0.75	5.0269	1.0040	5.0634	1.0075
	(0.0112)	(0.0050)	(0.0172)	(0.0075)
0.90	11.2262	0.9926	11.5267	0.9776
	(0.0508)	(0.0218)	(0.0779)	(0.0356)

wages. The response variable y is the intrapair difference in log wages of twins, and the observed covariate is the intrapair difference in years of schooling while the true covariate is the education. In this study the observations on the intrapair difference in cross-sibling reports are available which can be used as instrumental variable. In particular, we use the data of the first 49 pairs of twins to estimate β and the rest of sample points to estimate α.

Table 5 Empirical example: wage and schooling data

	Naive method		IV method	
τ	$\widehat{\alpha}_1$	$\widehat{\alpha}_2$	$\widehat{\alpha}_1$	$\widehat{\alpha}_2$
0.10	-0.5939	0.0454	-0.5562	0.0708
0.25	-0.2656	0.1069	-0.2231	0.1175
0.50	0.0297	0.0816	0.0297	0.1038
0.75	0.2894	0.0572	0.2624	0.0853
0.90	0.6043	0.1230	0.5469	0.1117
mean	0.0672	0.0770	0.0560	0.0959

Table 5 gives the parameter estimates for different quantile levels by using the IV and naive method that ignores the measurement error, where the last row contains the mean regression estimates. The results show clearly that the IV estimates correct the naive estimates which are attenuated by the presence of measurement error. Furthermore, the IV estimates show a clear linear increasing trend between wages and years of schooling for different quantile levels. For example, when the intrapair difference in years of

schooling is 2, which means one sibling has two more years schooling than the other one, the corresponding increase in log wages at quantile levels 10%, 25%, 50%, 75% and 90% respectively will be not higher than −41.46%, 1.19%, 23.73%, 43.30% and 77.03%. In contrast, the mean regression only indicates an average increase of the wages by 24.78%. Therefore the quantile regression results can provide much more information than the classical mean regression. Moreover, the numerical estimates indicate that education has different effects for different income groups, namely it has larger effects in higher income group, and smaller effects in lower income groups.

§6. Conclusions and Discussion

The problem of measurement error arises in many real data analyses in social and natural sciences. The treatment of the problem is notoriously difficult, especially in quantile regression models. In this paper, we demonstrated that the usual instrumental variable (IV) approach in mean regression can be extended to quantile regression with measurement error in the linear setup. This method is intuitive and easy to implement. The advantage of IV approach is that it is more flexible and general than the other approaches dealing with measurement errors. For example, any second independent measurement of the unobserved covariate can be used as an instrument. In this paper, we discussed a special case where the regression and measurement error distributions belong to the same location-scale family. The simulation studies show that our method performs fairly well under the model assumptions. Compared with the method of [6], this method does not require the regression error to have the same variance as the measurement error. In this sense it is more robust against model misspecification. Moreover, the limited simulation study shows that this method may even work well when the error term ϵ has asymmetric distribution, e.g. $\epsilon_i \sim \mathrm{F}(2,2)$.

References

[1] Koenker R, Bassett Jr G. Regression quantiles [J]. *Econometrica*, 1978, **46(1)**: 33–50.

[2] Fuller W A. *Measurement Error Models* [M]. New York: Wiley, 1987.

[3] Carroll R J, Ruppert D, Stefanski L A, et al. *Measurement Error in Nonlinear Models: A Modern Perspective* [M]. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC, 2006.

[4] Wei Y, Carroll R J. Quantile regression with measurement error [J]. *J. Amer. Statist. Assoc.*, 2009, **104(487)**: 1129–1143.

[5] Brown M L. Robust line estimation with errors in both variables [J]. *J. Amer. Statist. Assoc.*, 1982, **77(377)**: 71–79.

[6] He X M, Liang H. Quantile regression estimates for a class of linear and partially linear errors-in-variables models [J]. *Statist. Sinica*, 2000, **10(1)**: 129–140.

[7] Ma Y Y, Yin G S. Censored quantile regression with covariate measurement errors [J]. *Statist. Sinica*, 2011, **21(2)**: 949–971.

[8] Hu Y Y, Schennach S M. Instrumental variable treatment of nonclassical measurement error models [J]. *Econometrica*, 2008, **76(1)**: 195–216.

[9] Schennach S M. Quantile regression with mismeasured covariates [J]. *Econometric Theory*, 2008, **24(4)**: 1010–1043.

[10] Carroll R J, Ruppert D, Crainiceanu C M, et al. Nonlinear and nonparametric regression and instrumental variables [J]. *J. Amer. Statist. Assoc.*, 2004, **99(467)**: 736–750.

[11] Schennach S M. Instrumental variable estimation of nonlinear errors-in-variables models [J]. *Econometrica*, 2007, **75(1)**: 201–239.

[12] Wang L Q, Hsiao C. Two-stage estimation of limited dependent variable models with errors-in-variables [J]. *Econom. J.*, 2007, **10(2)**: 426–438.

[13] Wang L Q, Hsiao C. Method of moments estimation and identifiability of semiparametric nonlinear errors-in-variables models [J]. *J. Econometrics*, 2011, **165(1)**: 30–44.

[14] Abarin T, Wang L Q. Instrumental variable approach to covariate measurement error in generalized linear models [J]. *Ann. Inst. Statist. Math.*, 2012, **64(3)**: 475–493.

[15] Xu K, Ma Y Y, Wang L Q. Instrument assisted regression for errors in variables models with binary response [J]. *Scand. J. Stat.*, 2015, **42(1)**: 104–117.

[16] Polland D. Asymptotics for least absolute deviation regression estimators [J]. *Econometric Theory*, 1991, **7(2)**: 186–199.

[17] Koenker R. *Quantile Regression* [M]. Cambridge, UK: Cambridge University Press, 2005.

[18] Ashenfelter O, Krueger A. Estimates of the economic return to schooling from a new sample of twins [J]. *Am. Econ. Rev.*, 1994, **84(5)**: 1157–1173.

带测量误差的线性分位数回归模型的工具变量估计

关 静 王力群

(天津大学数学学院, 天津, 300350) (曼尼托巴大学统计系, 加拿大, R3T 2N2)

摘 要: 本文将工具变量法由研究带变量误差的均值回归模型推广到研究带变量误差的线性分位数回归模型. 所得到的估计量是一致的且在一般条件下具有渐近正态分布. 这种方法可行且易于操作. 模拟研究表明该估计量在有限样本下性质表现非常好. 最后这种方法被应用到实际问题, 研究工资与教育程度之间的关系.

关键词: 变量误差; 工具变量; 最小绝对偏差; 测量误差; 分位数回归

中图分类号: O212.1