

Which Quantile is the Most Informative? Maximum Entropy Quantile Regression*

Anil K. Bera[†] Antonio F. Galvao Jr.[‡] Gabriel V. Montes-Rojas[§]
Sung Y. Park[¶]

Abstract

This paper studies the connections among quantile regression, the asymmetric Laplace distribution, and the maximum entropy. We show that the maximum likelihood problem is equivalent to the solution of a maximum entropy problem where we impose moment constraints given by the joint consideration of the mean and median. Using the resulting score functions we develop a maximum entropy quantile regression estimator. This approach delivers estimates for the slope parameters together with the associated “most informative” quantile. Similarly, this method can be seen as a penalized quantile regression estimator, where the penalty is given by deviations from the median regression. We derive the asymptotic properties of this estimator by showing consistency and asymptotic normality under certain regularity conditions. Finally, an application to the U.S. wage data to evaluate the effect of training on wages illustrates the usefulness and implementation of our methodology.

Keywords: Quantile Regression; Treatment Effects; Asymmetric Laplace Distribution

JEL classification: C14; C31

*We are grateful to the participants in the South Asian and Far Eastern Meeting of the Econometrics Society, Singapore, July 2008, and specially Zhijie Xiao. However, we retain the responsibility for any remaining errors.

[†]Department of Economics, University of Illinois, 1407 W. Gregory Drive, Urbana, IL 61801. Tel.: +1-217-3334596; fax: +1-217-2446678. E-mail: abera@illinois.edu

[‡]Department of Economics, University of Wisconsin-Milwaukee, Bolton Hall 852, 3210 N. Maryland Ave., Milwaukee, WI 53201. E-mail: agalvao@uwm.edu

[§]Department of Economics, City University of London, 10 Northampton Square, London EC1V 0HB, U.K. E-mail: Gabriel.Montes-Rohas.1@city.ac.uk

[¶]Department of Economics, University of Illinois, 1407 W. Gregory Drive, Urbana, IL 61801, and The Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen, Fujian 361005, China. E-mail: sungpark@sungpark.net

1 Introduction

Different choices of loss functions determine different ways of defining the location of a random variable Y . For example, squared, absolute value, and step function lead to mean, median and mode, respectively (see Manski, 1991, for a general discussion). For a given quantile $\tau \in (0, 1)$, consider the loss function in a standard quantile estimation problem,

$$L_{1,n}(\mu; \tau) = \sum_{i=1}^n \rho_{\tau}(y_i - \mu) = \sum_{i=1}^n (y_i - \mu) (\tau - I(y_i \leq \mu)), \quad (1)$$

as proposed by Koenker and Bassett (1978). Minimizing $L_{1,n}$ with respect to the location parameter μ is identical to maximizing the likelihood based on the asymmetric Laplace probability density (ALPD):

$$f(y; \mu, \tau, \sigma) = \frac{\tau(1-\tau)}{\sigma} \exp\left(-\frac{\rho_{\tau}(y-\mu)}{\sigma}\right), \quad (2)$$

for given τ . The well known symmetric Laplace (double exponential) distribution is a special case of (2) when $\tau=1/2$.

Recent studies developed the properties of the maximum likelihood (ML) estimators based on ALPD (see for instance Kotz, Kozubowski and Podgórski, 2002, and Yu and Zhang, 2005). Machado (1993) used the ALPD to derive a Schwartz information criterion for model selection for quantile regression (QR) models, and Koenker and Machado (1999) introduced a goodness-of-fit measure for QR and related inference processes. Yu and Moyeed (2001) and Geraci and Botai (2007) used a Bayesian QR approach based on the ALPD. Komunjer (2005) constructed a new class of estimators for conditional quantiles in possibly misspecified nonlinear models with time series data. The proposed estimators belong to the family of quasi-maximum likelihood estimators (QMLEs) and are based on a family of ‘tick-exponential’ densities. Under the asymmetric Laplace density, the corresponding QMLE reduces to the Koenker and Bassett (1978) nonlinear quantile regression estimator. In addition, Komunjer (2007) developed a parametric estimator for the risk of financial time series expected short-fall based on the asymmetric power distribution, derived the asymptotic distribution of the asymmetric power distribution maximum likelihood estimator, and constructed a consistent estimator for its asymptotic covariance matrix.

Interestingly, the parameter μ in functions (1) and (2) is at the same time the location parameter, the τ -th quantile, and the mode of the ALPD. For the simple (unconditional) case, the minimization of (1) returns different order-statistics. For example, if we set $\tau = \{0.1, 0.2, \dots, 0.9\}$, the solutions are, respectively, the nine deciles of Y . In order to extract important information from the data a good summary statistic would be to choose *one* order statistics accordingly the most likely value. For a symmetric distribution one would choose the median. Using the ALPD, for given τ , maximization of the corresponding likelihood function gives that particular order statistics. Thus, the main idea of this paper is to *jointly* estimate τ and the corresponding order statistic of Y which can be taken as a good summary statistic of the data. The above notion can be easily extended to modeling the “conditional location” of Y given covariates X , as we do in Section 2.3. In this case, the ALPD model provides a twist to the QR problem, as now τ becomes the *most likely informative quantile* in a regression set-up.

We show in this paper that the score functions implied by the ALPD-ML estimation are not restricted to the true data generating process being ALPD, but they arise as the solution of a maximum entropy (ME) problem where we impose moment constraints given by the joint consideration of the mean and median. By so doing, the ALPD-ML estimator combines the information in the mean and the median to summarize the asymmetry underlying the empirical distribution (see Park and Bera, 2009, for a related discussion). Thus, we propose a novel Z-estimator that is based on the maximum entropy framework and the resulting estimating functions. In this sense, our estimator is the maximum entropy quantile regression (MEQR). This approach delivers estimates for the slope parameters together with the associated ME quantile. We derive the asymptotic properties of this estimator by showing consistency and asymptotic normality under certain regularity conditions.

The intuition behind this estimator works as follows. Suppose that the mean is larger than the median, and therefore the distribution is right skewed. Thus, taking into consideration the empirical distribution, there is more probability mass to the left of the distribution. As a result if we move towards the left of the median we have a point estimate in a place with more probability mass. The selected τ quantile does not necessarily corresponds to the mode, but

to a point estimate that maximizes the entropy. This provides a new interpretation of QR and frames it within the ME paradigm.

The proposed estimator has an interesting interpretation from policy perspective. The QR analysis gives a full range of estimators that account for heterogeneity in the response variable to certain covariates. However, by selecting the ME estimator, MEQR answers the question: of all the heterogeneity in the conditional regression model, which one is more likely to be observed? In general, the entire QR process is of interest because we would like to either test global hypotheses about conditional distributions or make comparisons across different quantiles (for a discussion about inference in QR models see Koenker and Xiao, 2002). But selecting the most informative quantile provides an estimator as parsimonious as OLS or the median estimators. The MEQR methodology is, therefore, a complement to the QR analysis rather than a competing alternative.

This set-up also allows for a different interpretation of the QR analysis. Consider the standard conditional regression set-up, $y = x'\beta + u$, and let β be partitioned into $\beta = (\beta_1, \beta_2)$. For a given value of $\beta_1 = \bar{\beta}_1$, we may be interested in finding the representative quantile of the unobservables distribution that corresponds to this level of β_1 . As an example, consider the standard treatment effects problem where $\bar{\beta}_1$ is the targeted treatment effect, and we would be interested in finding the quantile of the unobservables where this effect is more likely to occur. If β_1 is the return to education, the policy maker may be interested in the characterization of the unobservable characteristics of individuals with $\beta_1 = \bar{\beta}_1$. For such a case, instead of assuming a given quantile τ we would like to estimate it. In other words, the QR process provides us with the graph $\beta_1(\tau)$, but we argue that the graph $\tau(\beta_1)$ could be of interest too.

In order to illustrate the implementation of the MEQR, we apply the proposed estimator to the estimation of quantile treatment effects of subsidized training on wages under the Job Training Partnership Act (JTPA). We discuss the relationship between OLS, median regression and MEQR estimates of the JTPA treatment effect. We show that each estimator provides different treatment effect estimates. Moreover, we extend our MEQR estimator to Chernozhukov and Hansen (2006, 2008) instrumental variables strategy in quantile regres-

sion.

The rest of the paper is organized as follows. Section 2 develops the ML and ME frameworks of the problem. Section 3 derives the asymptotic distribution of the estimators. Section 4 deals with an empirical application to the effect of training on wages. Finally, conclusions and suggestions for future research are in the last section.

2 Maximum Likelihood and Maximum Entropy

2.1 Maximum Likelihood

Using (2), consider the maximization of the log-likelihood function of an ALPD:

$$L_{2,n}(\mu, \tau, \sigma) = n \ln \left(\frac{1}{\sigma} \tau (1 - \tau) \right) - \sum_{i=1}^n \frac{1}{\sigma} \rho_{\tau}(y_i - \mu) = n \ln \left(\frac{1}{\sigma} \tau (1 - \tau) \right) - \frac{1}{\sigma} L_{1,n}(\mu; \tau), \quad (3)$$

with respect to μ , τ and σ . The first order conditions lead to the following estimating equations (EE):

$$\sum_{i=1}^n \left(\frac{1}{2} \text{sign}(y_i - \mu) + \tau - \frac{1}{2} \right) = 0, \quad (4)$$

$$\sum_{i=1}^n \left(\frac{1 - 2\tau}{\tau(1 - \tau)} - \frac{(y_i - \mu)}{\sigma} \right) = 0, \quad (5)$$

$$\sum_{i=1}^n \left(-\frac{1}{\sigma} + \frac{1}{\sigma^2} \rho_{\tau}(y_i - \mu) \right) = 0. \quad (6)$$

Let $(\hat{\mu}, \hat{\tau}, \hat{\sigma})$ denote the solution to this system of equations. The first equation leads to the most probable order statistic. Once we have $\hat{\tau}$, $(1 - 2\hat{\tau})$ will provide a measure of asymmetry of the distribution. Equation (6) provides a straightforward measure of dispersion, namely,

$$\hat{\sigma} = \frac{1}{n} \sum_{i=1}^n \rho_{\hat{\tau}}(y_i - \hat{\mu}).$$

Then, the loss function (3) can be rewritten as a two-parameter loss function

$$-\frac{1}{n} L_{2,n}(\mu, \tau) = \ln \left(\frac{1}{n} L_{1,n}(\mu; \tau) \right) - \ln(\tau(1 - \tau)). \quad (7)$$

This determines that $L_2(\mu, \tau, \sigma)$ can be seen as a penalized quantile optimization function, where we minimize $\ln\left(\frac{1}{n}L_{1,n}(\mu; \tau)\right)$ and penalize it by $-\ln(\tau(1-\tau))$. The penalty can be interpreted as the cost of deviating from the median, i.e. for $\tau = 1/2$, $-\ln(\tau(1-\tau)) = -\ln(1/4)$ is the minimum, while for either $\tau \rightarrow 0$ or $\tau \rightarrow 1$ the penalty goes to $+\infty$.

Equation (5) selects the quantile τ and it contains the connection with the ALPD. After a little algebra, (5) becomes

$$\frac{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{\mu})}{\frac{1}{n} \sum_{i=1}^n |y_i - \hat{\mu}|} = \frac{1 - 2\hat{\tau}}{1 - 2\hat{\tau} + 2\hat{\tau}^2}.$$

Its population counterpart can be derived. If $Y \sim ALPD(\mu, \tau, \cdot)$, then

$$\frac{E[y_i - \mu]}{E[|y_i - \mu|]} = \frac{1 - 2\tau}{1 - 2\tau + 2\tau^2},$$

as can be seen in the Appendix. In this case, by using the mean and median deviations around μ , and because $1 - 2\tau$ is a measure of skewness of the distribution, it fully takes into account the asymmetric properties of the distribution.

It is important to note that the structure of the estimating functions suggests that the solution to the MLE problem can be obtained by first obtaining every quantile of the distribution, and then plugging them (with the corresponding estimator for σ) in (5) until this equation is satisfied (if the solution is unique). In other words, given all the quantiles of Y , the problem above selects the most likely quantile as if the distribution of Y were ALPD.

2.2 Maximum Entropy

The ALPD can also be characterized as a maximum entropy density obtained by maximizing Shannon's (1948) entropy measure subject to two moment constraints (see Kotz, Kozubowski and Podgórski, 2001, p.156):

$$f_{ME}(y) \equiv \arg \max_f - \int f(y) \ln f(y) dy \tag{8}$$

subject to

$$E|y - \mu| = c_1, \tag{9}$$

$$E(y - \mu) = c_2, \tag{10}$$

and the normalization constraint, $\int f(y)dy = 1$. The solution of the above optimization problem has the familiar exponential form

$$f_{ME}(y : \mu, \lambda_1, \lambda_2) = \frac{1}{\Omega(\theta)} \exp [-\lambda_1|y - \mu| - \lambda_2(y - \mu)], \quad (11)$$

where λ_1 and λ_2 are the Lagrange multipliers corresponding to the constraints (9) and (10), respectively, $\theta = (\mu, \lambda_1, \lambda_2)'$ and $\Omega(\theta)$ is the normalizing constant. It is easy to see from (11) that (symmetric) Laplace density (LD) is a special case of ALPD. The constraint (9) along with the normalization constraint characterizes LD in the sense of ME principle. Interestingly, the constraints (9) and (10) capture, respectively, the dispersion and asymmetry for ALPD. The marginal contribution of (10) is measured by the Lagrangian multiplier λ_2 . If λ_2 is close to 0, then (10) does not have useful information for the data, and therefore, symmetric LD is the most appropriate one. In our case, μ is median when the density follows LD. In this sense, the moment function $|y - \mu|$ helps to extract the most informative measure of dispersion of the data.

Letting c_1 and c_2 be the sample counterpart of the right hand side of the moment constraints, we can write (9) and (10) as

$$\int \psi_1(y, \mu) f_{ME}(y : \mu, \lambda_1, \lambda_2) dy = 0 \quad \text{and} \quad \int \psi_2(y, \mu) f_{ME}(y : \mu, \lambda_1, \lambda_2) dy = 0,$$

respectively, where $\psi_1(y, \mu) = |y - \mu| - (1/n) \sum_{i=1}^n |y_i - \mu|$ and $\psi_2(y, \mu) = (y - \mu) - (1/n) \sum_{i=1}^n (y_i - \mu)$. By substituting the solution $f_{ME}(y : \mu, \lambda_1, \lambda_2)$ into the Lagrangian of the maximization problem in (8), we obtain the profiled objective function

$$h(\lambda_1, \lambda_2, \mu) = \ln \int \exp \left[- \sum_{j=1}^2 \lambda_j \psi_j(y, \mu) \right] dy. \quad (12)$$

The parameters λ_1 , λ_2 and μ can be estimated by solving the following saddle point problem

$$\hat{\mu}_{ME} = \arg \max_{\mu} \ln \int \exp \left[- \sum_{j=1}^2 \hat{\lambda}_{MEj} \psi_j(y, \mu) \right] dy,$$

where $\hat{\lambda}_{ME} = (\hat{\lambda}_{ME1}, \hat{\lambda}_{ME2})$ is given by

$$\hat{\lambda}_{ME}(\mu) = \arg \min_{\lambda} \ln \int \exp \left[- \sum_{j=1}^2 \lambda_j \psi_j(y, \mu) \right] dy.$$

Therefore, $\hat{\mu}_{ME}$ and $\hat{\lambda}_{ME}$ have to satisfy the following first order conditions $\partial h/\partial\lambda_1 = 0$, $\partial h/\partial\lambda_2 = 0$ and $\partial h/\partial\mu = 0$, respectively:

$$\frac{\int -|y - \mu| \exp[-\lambda_1|y - \mu| - \lambda_2(y - \mu)]dy}{\int \exp[-\lambda_1|y - \mu| - \lambda_2(y - \mu)]dy} + \frac{1}{n} \sum_{i=1}^n |y_i - \mu| = 0, \quad (13)$$

$$\frac{\int -(y - \mu) \exp[-\lambda_1|y - \mu| - \lambda_2(y - \mu)]dy}{\int \exp[-\lambda_1|y - \mu| - \lambda_2(y - \mu)]dy} + \frac{1}{n} \sum_{i=1}^n (y_i - \mu) = 0, \quad (14)$$

$$\frac{\int (\lambda_1 \text{sign}(y - \mu) + \lambda_2) \exp[-\lambda_1|y - \mu| - \lambda_2(y - \mu)]dy}{\int \exp[-\lambda_1|y - \mu| - \lambda_2(y - \mu)]dy} - \frac{\lambda_1}{n} \sum_{i=1}^n \text{sign}(y_i - \mu) - \lambda_2 = 0. \quad (15)$$

After some algebra we get to the following equations,

$$-\frac{1}{\lambda_1} \frac{\lambda_1^2 + \lambda_2^2}{\lambda_1^2 - \lambda_2^2} + \frac{1}{n} \sum_{i=1}^n |y_i - \mu| = 0, \quad (16)$$

$$\frac{2\lambda_2}{(\lambda_1 + \lambda_2)(\lambda_1 - \lambda_2)} + \frac{1}{n} \sum_{i=1}^n (y_i - \mu) = 0, \quad \text{and} \quad (17)$$

$$-\frac{\lambda_1}{n} \sum_{i=1}^n \text{sign}(y_i - \mu) - \lambda_2 = 0. \quad (18)$$

It can be easily checked that the first terms in eqs. (16) and (17) are $E_{f_{ME}}[-|y - \mu|]$ and $E_{f_{ME}}[-(y - \mu)]$, and that $E_{f_{ME}}[\text{sign}(y - \mu)] = -\lambda_2/\lambda_1$. Equations (16)-(18) are nothing but the re-parameterized version of (4)-(6). In fact, from a comparison of (2) and (11) we can easily see that $\lambda_1 = 1/(2\sigma)$, $\lambda_2 = (2\tau - 1)/(2\sigma)$ and $\Omega(\theta) = \sigma/(\tau(1 - \tau))$. Given λ_1 the degree of asymmetry is explained by λ_2 that is proportionally equal to $2\tau - 1$ in ALPD. Note that $\lambda_2 = 0$ when $\tau = 0.5$, i.e., μ is a median. Thus finding the most appropriate degree of asymmetry is equivalent to estimating τ based on the ML method.

2.3 Linear Regression Model

Now consider the conditional version of the above, by taking a linear model of the form $y = x'\beta + u$ where the parameter of interest is $\beta \in \mathfrak{R}^p$, x refers to a p -vector of exogenous covariates, and u denotes the unobservable component in the linear model. As noted in

Angrist, Chernozhukov and Fernández-Val (2006), QR provides the best linear predictor for y under the asymmetric loss function

$$L_{3,n}(\beta; \tau) = \sum_{i=1}^n \rho_{\tau}(y_i - x'_i \beta) = \sum_{i=1}^n ((y_i - x'_i \beta) (\tau - I(y_i \leq x'_i \beta))), \quad (19)$$

where β is assumed to be a function of the *fixed* quantile τ of the unobservable components, that is $\beta(\tau)$. Similarly, if u is assumed to follow an ALPD, the log-likelihood function is

$$\begin{aligned} L_{4,n}(\beta, \tau, \sigma) &= n \ln \left(\frac{1}{\sigma} \tau (1 - \tau) \right) - \sum_{i=1}^n \left(\frac{1}{\sigma} \rho_{\tau}(y_i - x'_i \beta) \right) \\ &= n \ln \left(\frac{1}{\sigma} \tau (1 - \tau) \right) - \frac{1}{\sigma} L_{3,n}(\beta; \tau). \end{aligned} \quad (20)$$

Estimating β in this framework provides the marginal effect of x on the τ ME conditional quantile of y .

Computationally, the MLE can be obtained by simulating a grid of quantiles and choosing the quantile that maximizes (20), or by solving the estimating equations, $\nabla L_{4,n}(\beta, \tau, \sigma) = 0$, i.e.,

$$\frac{\partial L_{4,n}(\beta, \tau, \sigma)}{\partial \beta} = \sum_{i=1}^n \left(\frac{1}{2} \text{sign}(y_i - x'_i \beta) + \tau - \frac{1}{2} \right) x'_i = 0, \quad (21)$$

$$\frac{\partial L_{4,n}(\beta, \tau, \sigma)}{\partial \tau} = \sum_{i=1}^n \left(\frac{1 - 2\tau}{\tau(1 - \tau)} - \frac{(y_i - x'_i \beta)}{\sigma} \right) = 0, \quad (22)$$

$$\frac{\partial L_{4,n}(\beta, \tau, \sigma)}{\partial \sigma} = \sum_{i=1}^n \left(-\frac{1}{\sigma} + \frac{1}{\sigma^2} \rho_{\tau}(y_i - x'_i \beta) \right) = 0. \quad (23)$$

As we stated before, $L_{4,n}$ can be written as a penalized QR problem loss function that depends only on (β, τ) :

$$-\frac{1}{n} L_{4,n}(\beta, \tau) = \ln \left(\frac{1}{n} L_{3,n}(\beta; \tau) \right) - \ln(\tau(1 - \tau)), \quad (24)$$

and the interpretation is the same as discussed in section 2.1.

3 Consistency and Asymptotic Normality

In this section we propose a Z-estimator using equations (21)-(23). Assume that the observed data (x_i, y_i) $i = 1, \dots, n$ is a realization of a stochastic process $y_i = x'_i \beta + u_i$, $i = 1, 2, \dots, n$,

where u_i are independent observations of a continuous random variable with distribution function G .

Let $\|\cdot\|$ be the Euclidean norm and $\theta = (\beta, \tau, \sigma)$. Moreover, define

$$\Psi_n(\theta) = \frac{1}{n} \sum_{i=1}^n \begin{pmatrix} (\tau - I(u_i < 0)) x_i \\ \frac{1-2\tau}{\tau(1-\tau)} - \frac{(y_i - x_i' \beta)}{\sigma} \\ -\frac{1}{\sigma} + \frac{1}{\sigma^2} \rho_\tau(y_i - x_i' \beta) \end{pmatrix} = \frac{1}{n} \sum_{i=1}^n \psi_\theta(u_i) = 0.$$

We impose the following regularity conditions:

Assumption 1 (Compactness). Let Θ be a compact set, with $\theta \in \Theta$, where $\beta \in \mathbb{R}^p$, $\tau \in (0, 1)$, and $\sigma > 0$.

Assumption 2. The distribution function of u_i , G , is absolutely continuous with conditional densities, g , with $0 < g(\cdot) < \infty$.

Assumption 3 (Uniqueness). The function $\theta \mapsto \Psi_n(\theta)$ has a unique zero at $\hat{\theta}_n \in \Theta$. Moreover, assume that for every $\theta \in \Theta$, $\Psi_n(\theta) \xrightarrow{p} \Psi(\theta)$, where $\Psi(\theta) = E_G[\psi_\theta(u)]$, and let $\Psi(\theta_0) = 0$ for a unique $\theta_0 \in \Theta$.

Now we discuss the asymptotic properties of the estimator. First we state consistency and later asymptotic normality.

Theorem 1 Under Assumptions 1-3, $\|\hat{\theta}_n - \theta_0\| \xrightarrow{p} 0$.

Proof: First note that, under condition A3, the function $\Psi(\theta)$ satisfies,

$$\inf_{\theta: d(\theta, \theta_0) \geq \epsilon} \|\Psi(\theta)\| > 0 = \|\Psi(\theta_0)\|.$$

Secondly, under conditions A1-A3, the class of functions $\{\Psi_{1n}, \Psi_{2n}, \Psi_{3n} : \theta \in \Theta\}$ can easily be shown to be a Donsker class by standard arguments for QR and mean regression regarding finite bracketing entropy (see, for instance, Chernozhukov and Hansen (2006)). In addition, each of these class of functions belongs to a finite-dimensional vector space and hence is a Vapnik-Chervonenkis class (VC-class). Next, the stability properties of VC-class

show that the components of Ψ also run through VC-class. Since they are uniformly bounded and pointwise separable, they are Donsker. Since Donsker classes are also Glivenko-Cantelli, we have that $\sup_{\theta \in \Theta} \|\Psi_n(\theta) - \Psi(\theta)\| \xrightarrow{p} 0$.

Finally, from problem (20) and equations (21)-(23) we have that $\|\Psi_n(\hat{\theta}_n)\| \xrightarrow{p} 0$. Thus, all the conditions in Theorem 5.9 of van der Vaart (1998) are satisfied and $\|\hat{\theta}_n - \theta_0\| \xrightarrow{p} 0$. ■

Now we move our attention to the asymptotic normality of the estimator. In order to derive the limiting distribution define

$$V_1(\theta) = E_G [\psi_\theta(u)\psi_\theta(u)'], \quad (25)$$

$$V_2(\theta) = \frac{\partial E_G [\psi_\theta(u)]}{\partial \theta'}. \quad (26)$$

Note that the order of expectation and differentiation are reversed in V_2 . Here,

$$V_1(\theta) = \begin{bmatrix} \frac{1}{\sigma^2} \tau(1-\tau) E^*[xx'] & \frac{E^*[(1-2\tau)\text{sign}(y-x'\beta)-(1-2\tau)^2]x'}{2\sigma\tau(1-\tau)} - E^*\left[\frac{1}{\sigma^2} \rho_\tau(y-x'\beta)x'\right] & \frac{1}{2\sigma^3} E^*[\rho_\tau(y-x'\beta) (\text{sign}(y-x'\beta) - (1-2\tau))x'] \\ \cdot & \frac{(1-2\tau)^2}{\tau^2(1-\tau)^2} + E^*\left[\frac{1}{\sigma^2} (y-x'\beta)^2\right] - 2\frac{(1-2\tau)}{\tau(1-\tau)} E^*\left[\frac{1}{\sigma} (y-x'\beta)\right] & \frac{1}{\sigma^2} E^*[\rho_\tau(y-x'\beta) \left(\frac{1-2\tau}{\tau(1-\tau)} - \frac{1}{\sigma} (y-x'\beta)\right)] \\ \cdot & \cdot & \frac{1}{\sigma^4} E^*[\rho_\tau^2(y-x'\beta)] + \frac{1}{\sigma^2} - \frac{1}{\sigma^3} E^*[\rho_\tau(y-x'\beta)] \end{bmatrix}$$

and

$$V_2(\theta) = \begin{bmatrix} -\frac{E^*[g(y-x'\beta)xx']}{\sigma^2} & \frac{1}{\sigma} E^*[x'] & 0 \\ \cdot & \frac{-1+2\tau-2\tau^2}{\tau^2(1-\tau)^2} & \frac{1}{\sigma^2} E^*[(y-x'\beta)] \\ \cdot & \cdot & -\frac{1}{\sigma^2} \end{bmatrix}$$

where we define $\frac{1}{n} \sum_{i=1}^n z_i \xrightarrow{p} E^*(z)$. It is easy to show that $V_1(\theta) = -V_2(\theta)$ if $u \sim ALPD(0, \tau, \sigma)$ (see the Appendix for a proof).

Assumption 4. Assume that $V_1(\theta_0)$ and $V_2(\theta_0)$ exist and are finite, and $V_2(\theta_0)$ is invertible.

Theorem 2 Under Assumptions 1-4,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow N(0, V_2(\theta_0)^{-1}V_1(\theta_0)V_2(\theta_0)^{-1}).$$

Proof: First, note that $\mathcal{F} = \{\psi_\theta, \theta \in \Theta\}$ is Donsker, such that $\sqrt{n}\Psi_n(\theta_0) \Rightarrow Z$ for $\theta_0 \in \Theta$ and some tight random element Z . Now let's analyze the first element of Ψ_n defined as Ψ_{1n} . From the QR literature, e.g. He and Shao (1996) and Ruppert and Carroll (1980), it is well known that

$$\sup_{|\hat{\theta}_n - \theta_0| < \delta} \left| \sqrt{n}(\Psi_{1n}(\hat{\theta}_n) - E\Psi_{1n}(\hat{\theta}_n)) - \sqrt{n}(\Psi_{1n}(\theta_0) - E\Psi_{1n}(\theta_0)) \right| = o_p(1).$$

In addition, the same argument is true for the second element in Ψ by simple arguments of standard linear regression and triangle inequality. Finally, under the condition A2, i.e. the error terms u_i are continuously distributed given x_i , with continuous conditional density $g(u_i|x_i)$, the third element Ψ_{3n} also satisfies,

$$\sup_{|\hat{\theta}_n - \theta_0| < \delta} \left| \sqrt{n}(\Psi_{3n}(\hat{\theta}_n) - E\Psi_{3n}(\hat{\theta}_n)) - \sqrt{n}(\Psi_{3n}(\theta_0) - E\Psi_{3n}(\theta_0)) \right| = o_p(1),$$

see, e.g., Chernozhukov and Hansen (2006) for details. This fact is a consequence that the centered $\rho_\tau(u)$ is Lipschitz,¹ and we assume that G is sufficiently regular around θ_0 such that $E\rho_\tau$ is twice differentiable. Therefore, we have that

$$\sup_{|\hat{\theta}_n - \theta_0| < \delta} \left| \sqrt{n}(\Psi_n(\hat{\theta}_n) - E\Psi_n(\hat{\theta}_n)) - \sqrt{n}(\Psi_n(\theta_0) - E\Psi_n(\theta_0)) \right| = o_p(1).$$

Now note that the true value θ_0 satisfies the corresponding population condition

$$\Psi_n(\hat{\theta}_n) = \frac{1}{n} \sum_{i=1}^n \psi_\theta(u_i) = 0.$$

Moreover, using from Assumption 3, $E[\psi_{\theta_0}(u)] = 0$, and therefore

$$E[\Psi_n(\theta_0)] = E \left[\frac{1}{n} \sum_{i=1}^n \psi_{\theta_0}(u_i) \right] = \frac{1}{n} \sum_{i=1}^n E[\psi_{\theta_0}(u_i)] = 0.$$

Thus, it follows that $\sqrt{n}(\Psi_n(\theta_0) - E[\Psi_n(\hat{\theta}_n)]) \xrightarrow{p} 0$ uniformly on Θ . A mean-value expansion of $E[\Psi_n(\hat{\theta}_n)]$ around $\hat{\theta} = \theta_0$ yields

$$E[\Psi_n(\hat{\theta}_n)] = E[\Psi_n(\theta_0)] + \left[\frac{\partial E[\Psi_n(\hat{\theta}_n)]}{\partial \theta'} \right] (\hat{\theta}_n - \theta_0) + o_p(\|\hat{\theta}_n - \theta_0\|)$$

¹Note that $\rho(x+y) - \rho(x) < 2|y|$.

Substituting

$$\begin{aligned}\sqrt{n}\Psi_n(\theta_0) - \sqrt{n}E[\Psi_n(\hat{\theta}_n)] &= o_p(1), \\ \sqrt{n}\Psi_n(\theta_0) - \sqrt{n}\left[\frac{\partial E[\Psi_n(\hat{\theta}_n)]}{\partial \theta'}\right] (\hat{\theta}_n - \theta_0) &= o_p(1), \\ \sqrt{n}(\hat{\theta}_n - \theta_0) &= \left[\frac{\partial E[\Psi_n(\hat{\theta}_n)]}{\partial \theta'}\right]_{\theta=\theta_0}^{-1} \sqrt{n}\Psi_n(\theta_0) + o_p(1).\end{aligned}$$

We have that under assumption A2 and A4, $\left[\frac{\partial E[\Psi_n(\hat{\theta}_n)]}{\partial \theta'}\right]_{\theta=\theta_0}^{-1} \xrightarrow{p} \left[\frac{\partial E_G[\psi_\theta(u)]}{\partial \theta'}\right]_{\theta=\theta_0}^{-1}$. Finally as argued in the previous theorem Ψ is Donsker and this yields the asymptotic distribution of the $(\hat{\theta}_n - \theta_0)$ estimator as

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow N(0, V_2(\theta_0)^{-1}V_1(\theta_0)V_2(\theta_0)^{-1}).$$

■

4 Monte Carlo Simulations

In this section we provide a very brief glimpse into the finite sample behavior of the proposed maximum entropy quantile regression estimator. Two simple versions of our basic model are considered in the simulation experiments. In the first, reported in Table 1, the scalar covariate, x_i , exerts a pure location shift effect. In the second, reported in Table 2, x_i has a both a location and scale shift effect. In the former case the response, y_i , is generated by the model,

$$y_i = \alpha + \beta x_i + u_i,$$

while in the latter case,

$$y_i = \alpha + \beta x_i + (1 + \gamma)u_i,$$

with u_i *i.i.d.* and generated according to a standard normal distribution, t_3 distribution and χ_3^2 centered at the mean, Laplace distribution (i.e. $\tau = 0.5$), and ALPD with $\tau = 0.25$.² In the location shift model x_i follows a standard normal distribution; in the location-scale shift model, it follows a χ_3^2 .³ We also set $\alpha = \beta = 1$ and $\gamma = 0.5$.

Our interest is on the effect of the covariates in terms of bias and root mean squared error (RMSE). Sample size is fixed at $n = 200$ in all versions of the model. In all cases the reported entries are based on 5,000 replications of the simulations. Three estimators are considered: maximum entropy quantile regression (MEQR), quantile regression at the median (QR), and ordinary least squares (OLS). We pay special attention to the estimated quantile of interest τ in the MEQR. Bias is computed in each case with respect to the true slope parameter (see below for the location-scale model).

Table 1 reports the results of the location shift simulations. In all cases we compute the bias and RMSE with respect to $\beta = 1$. Bias is close to zero in all cases. In the Gaussian setting we see roughly the anticipated efficiency loss due to estimating MEQR and QR rather than the OLS. As expected, under symmetric distributions, normal, t_3 , and Laplace, the estimated quantile of interest τ in the MEQR is remarkably close to the median. Nevertheless, MEQR has RMSE close to QR, and therefore, it has a better performance in the t_3 case than OLS. In the χ_3^2 case, the MEQR estimator does better than the QR procedure, but both outperform the OLS. Note that the estimated quantile for the χ_3^2 is 0.08, consistent with the fact that the location parameter is more informative at low quantiles, and therefore a smaller RMSE can be obtained for the location. Finally, in the ALPD(0.25) case, MEQR also has a smaller RMSE than both QR and OLS estimators. Overall, Table 1 shows that the MEQR estimator retains the properties of the QR robust estimator (i.e. RMSE in the t_3 and χ_3^2 cases is smaller than OLS) but we do not need to specify a quantile of interest, since the estimator delivers the τ associated with the maximum entropy.

[Table 1]

²Although not reported, similar results were obtained for ALPD with $\tau = 0.75$.

³To avoid crossings in the quantile processes.

In the location-scale version of the model we adopt the same distributions for generating the u_i 's. However, in this case it is important that the resulting linear quantile functions do not cross, an eventuality we avoid by now taking the x_i 's as χ_3^2 instead of Gaussian, thus ensuring that the scale parameter will be positive. In the heteroscedastic case the effect of the covariate x_i on quantile of interest response in QR is given by $\beta(\tau) = \beta + \gamma Q_u(\tau)$. In MEQR we construct β used for computation of the bias and RMSE using the average estimated maximum entropy quantile.

In Table 2, the results for the normal, t_3 and Laplace distributions are similar to those in the location model, showing that all point estimates are approximately unbiased and OLS outperforms MEQR and QR in the normal case, but the contrary occurs in the t_3 and Laplace cases. However, the median QR estimator shows a better performance than MEQR. In the χ_3^2 case, the estimated quantile is $\tau = 0.085$, but in this case, MEQR does worse than both median and OLS estimators. Finally, in the ALPD(0.25) distribution, the best performance is obtained for the MEQR estimator.

[Table 2]

5 Empirical Application: The Effect of Subsidized Training, the JTPA case

As we argued in the Introduction, a very useful application of MEQR is in the quantile treatment effects (QTE) estimation. We apply the proposed estimators to the QTE of training offers on wages under the Job Training Partnership Act (JTPA). The JTPA was a large publicly-funded training program that began funding in October 1983 and continued until late 1990's. We focus on the Title II subprogram, which was offered only to individuals with "barriers to employment" (long-term use of welfare, being a high-school drop-out, 15 or more recent weeks of unemployment, limited English proficiency, physical or mental disability, reading proficiency below 7th grade level or an arrest record).

We use the database in Abadie, Angrist and Imbens (2002) that contains information

about adult male and female JTPA participants and non-participants. Individuals in the randomly assigned JTPA treatment group were offered training, while those in the control group were excluded for a period of 18 months. Let Z denote the indicator variable for those receiving a JTPA offer. Of those offered, 60% did training; of those in the control group, less than 2% did training. We refer the reader to Abadie et al. (2002) for an extensive discussion of the database and descriptive statistics.

For our purposes of illustrating the use of MEQR, we first study the effect of receiving a JTPA offer on log wages, and later we pursue their instrumental variable estimation in our MEQR context. Therefore, our first exercise is to value the option of training. In this case we consider the following regression model:

$$Y = Z\gamma + X\beta + U,$$

where the dependent variable Y is the logarithm of 30 month accumulated earnings (we exclude individuals without earnings), Z is a dummy variable for the JTPA offer, X is a set of exogenous covariates containing individual characteristics, and U is an unobservable component. The parameter of interest is γ that provides the effect of the JTPA training offer on wages.

[Table 3 and Figures 1, 2, and 3]

First, we compute the entire QR process, that is, for all $\tau \in (0.05, 0.95)$ we run a standard quantile regression of the model given in Table 3. The JTPA effect estimates for QR and OLS appear in Figure 1. Interestingly, the effect of JTPA is decreasing in τ (except for very low quantiles), which determines that those individuals with conditionally less unobserved ability benefited the most from the JTPA training program offer. Second, by solving the linear equation corresponding to the selection of τ , (22), determines that the most likely quantile is $\hat{\tau} = 0.84$, as shown in Figure 2, which implies that the distribution of unobservables is negatively skewed. This value is denoted by a vertical solid line, together with the usual 95% confidence interval given by the vertical parallel dotted lines. Note that the OLS and median regression estimators are, respectively, 0.075 (0.032) and 0.100

(0.033) which are rather similar, but they both are much higher than the MEQR estimate of 0.045 (0.022).⁴ A very interesting result is that both OLS and QR overestimate the training effect relative to MEQR. The results show that for the estimate of the most likely quantile, $\hat{\tau} = 0.84$, the most likely effect of training is much smaller when compared to the mean and median effects.

As argued in the Introduction, the MEQR framework allows for a different interpretation of the QR analysis. Suppose that we are interested in a targeted treatment effect of $\bar{\gamma} = 0.1$, and we would like to get the representative quantile of the unobservables distribution that will most likely have this effect. This corresponds to estimating the MEQR parameters for $Y - Z\bar{\gamma} = X\beta + U$. In this case, we obtain an estimated most likely quantile of $\tau(\widehat{\bar{\gamma}}) = 0.85$, which is very close to the one obtained above. In fact, the graph $\tau(\gamma)$ is flat, i.e. we obtain values of $\hat{\tau}$ close to $\widehat{\tau(\bar{\gamma})}$. Therefore, the MEQR solution does not depend on conditioning on Z .

As a result, a policy maker interested in computing a parsimonious effect of the option to training would get different answers. The nature of the unobservables would reveal that the upper quantiles are more informative, and the MEQR estimator would be more appropriate to describe the effect of JTPA on earnings. Finally, we also plot the standard errors for the treatment effects of JTPA (see Figure 3) for all QR estimates (dashed line), OLS (dotted line) and MEQR (solid line). Note that QR estimates in the upper tail of the distribution get smaller standard errors, which suggests that by choosing the most likely quantile the model implicitly solves for the smallest standard error QR estimator.

To value the option of treatment is an interesting exercise in itself, but policymakers may be more interested in the effect of actual training rather than the possibility of training. In this case the mode of interest is

$$Y = D\alpha + X\beta + U$$

where D is a dummy variable indicating if the individual actually completed the JTPA training. We have strong reasons to believe that $cov(D, U) \neq 0$ and therefore OLS and QR

⁴The numbers in parenthesis are the corresponding standard errors.

estimates will be biased. In this case, while the JTPA offer is random, those individuals that decide to undertake training do not constitute a random sample of the population. Rather, they are likely to be more motivated individuals or those that value training the most. However, the exact nature of this bias is unknown in terms of quantiles. Figure 4 reports the entire quantile process and OLS for the above equation. Interestingly the effect of training on wages is monotonically decreasing in τ . The selection of the most likely quantile determines that as in the previous case $\hat{\tau} = 0.84$.

[Figure 4]

In order to solve for the potential endogeneity, and following Abadie et al. (2002), Z can be used as a valid instrument for D . The reason is that it is exogenous as it was a randomized experiment, and it is correlated with D (see above on the percentage of individuals that undertook training given they offer status). The IV strategy is based on Chernozhukov and Hansen (2006, 2008) by considering the model

$$Y = D\alpha + X\beta + Z\gamma + U.$$

The IV method in QR proceeds as follows. Note that Z does not belong to the model, as conditional on D , undertaking training, the offer has no effect on wages. Then, we construct a grid in $\alpha \in \mathcal{A}$, which is indexed by j for each $\tau \in (0, 1)$ and we estimate the quantile regression model for fixed τ

$$Y - D\alpha_j(\tau) = X\beta + Z\gamma + U.$$

This gives $\{\hat{\beta}_j(\alpha_j(\tau), \tau), \hat{\gamma}_j(\alpha_j(\tau), \tau)\}$, the set of conditional quantile regression estimates for the new model. Next, we choose α by minimizing a given norm of γ ,

$$\hat{\alpha}(\tau) = \underset{\alpha \in \mathcal{A}}{\operatorname{argmin}} \|\hat{\gamma}(\alpha(\tau), \tau)\|.$$

Figures 5 and 6 show the values of γ^2 for the grids of α and τ . As a result we obtain the map $\tau \mapsto \{\hat{\alpha}(\tau), \hat{\beta}(\hat{\alpha}(\tau), \tau) \equiv \hat{\beta}(\tau), \hat{\gamma}(\hat{\alpha}(\tau), \tau) \equiv \hat{\gamma}(\tau)\}$.

[Figures 5 and 6]

Finally, we select the most likely quantile as in the previous ALPD set-up by using the first order condition corresponding the the selection of τ :

$$\hat{\tau} = \underset{\tau \in (0,1)}{\operatorname{argmin}} \left| \frac{1 - 2\tau}{\tau(1 - \tau)} - \frac{\sum_{i=1}^n \hat{u}_i(\tau)}{\sum_{i=1}^n \rho_{\tau}(\hat{u}_i(\tau))} \right|$$

where $\hat{u}_i(\tau) = y_i - d_i \hat{\alpha}(\tau) - x_i' \hat{\beta}(\tau) - z_i \hat{\gamma}(\tau)$. Figure 7 reports the IV estimates together with the most likely quantile. Surprisingly, the results are very much alike those of the value of the JTPA training offer. The IV least squares estimator for the effect of JTPA training gives a value of 0.116 (0.045) while IV median regression gives a much higher value of 0.142 (0.047). The most likely quantile continues to be 0.84, which has an associated training effect of 0.072 (0.033). Note that the MEQR effect continues to be smaller than the OLS and median estimates. As before, the upper quantiles are more informative to compute the effect of JTPA training.

[Figure 7]

6 Conclusions and Extensions

The maximum entropy quantile regression estimation method proposed in this paper provides a parsimonious estimator that complements the quantile process. We show that this problem can also be found as the solution of a maximum entropy problem where we impose moment constraints given by the joint consideration of the mean and the median as in Kotz, Kozubowski and Podgórsk (2001). This provides a nice interpretation of quantile regression and frames it within the maximum entropy paradigm. Potential estimates from this method has important applications. As an example, we apply the proposed estimator to a well-known dataset where extensive quantile regression inference has been made.

Many issues remain to be investigated. One interesting extension is the development of tests for asymmetry based on the estimated maximum entropy τ -quantile and the implied

measure of skewness in the data generating process. Another is further research on developing a broader class of penalized quantile regression loss functions using the maximum entropy paradigm.

Appendix

Consider the log likelihood function in (20)

$$L(\beta, \tau, \sigma) = n \ln \left(\frac{1}{\sigma} \tau (1 - \tau) \right) - \sum_{i=1}^n \left(\frac{1}{\sigma} \rho_{\tau}(y_i - x'_i \beta) \right),$$

where $\rho_{\tau}(u) = u \cdot (\tau - I(u \leq 0)) = \frac{|u|}{2} + \frac{(2\tau-1)u}{2}$.

The score functions for this problem are obtained by taking first order derivatives. First note that,

$$\frac{\partial L}{\partial \beta} = -\frac{1}{\sigma} \sum_{i=1}^n \psi_{\tau}(y_i - x'_i \beta) x'_i,$$

where $\psi_{\tau}(u) = (\tau - I(u \leq 0))$. Other way of expressing this score function is

$$\begin{aligned} \frac{\partial L}{\partial \beta} &= -\frac{1}{\sigma} \sum_{i=1}^n \left(\frac{1}{2} \frac{\partial |y_i - x'_i \beta|}{\partial \beta} + \frac{2\tau - 1}{2} (-x'_i) \right) \\ &= -\frac{1}{\sigma} \sum_{i=1}^n \left(\frac{1}{2} \text{sign}(y_i - x'_i \beta) \cdot (-x'_i) + \frac{2\tau - 1}{2} (-x'_i) \right) \\ &= \frac{1}{\sigma} \sum_{i=1}^n \left(\frac{\text{sign}(y_i - x'_i \beta)}{2} + \frac{2\tau - 1}{2} \right) x'_i \end{aligned}$$

Moreover,

$$\begin{aligned} \frac{\partial L}{\partial \tau} &= n \frac{1 - 2\tau}{\tau(1 - \tau)} - \frac{1}{\sigma} \sum_{i=1}^n (y_i - x'_i \beta) \\ \frac{\partial L}{\partial \sigma} &= -\frac{n}{\sigma} + \frac{1}{\sigma^2} \sum_{i=1}^n \rho_{\tau}(y_i - x'_i \beta) \end{aligned}$$

If $w \sim ALPD(0, \tau, 1)$, the following expectations are true:

$$\begin{aligned}
E[\text{sign}(w)] &= 1 - 2\tau \\
E[w] &= \frac{1 - 2\tau}{\tau(1 - \tau)} \\
E[w^2] &= 2\tau(1 - \tau) \left(\frac{1}{\tau^3} + \frac{1}{(1 - \tau)^3} \right) \\
E[|w|] &= \frac{1 - 2\tau + 2\tau^2}{\tau(1 - \tau)} \\
E[\rho_\tau(w)] &= \frac{1 - 2\tau + 2\tau^2}{2\tau(1 - \tau)} + \frac{(2\tau - 1)^2}{2\tau(1 - \tau)} = 1 \\
E[w^2 \cdot I(w < 0)] &= \frac{2\tau}{(1 - \tau)^2} \\
E[w^2 \cdot I(w > 0)] &= \frac{2(1 - \tau)}{\tau^2}; \\
E[w \cdot |w|] &= E[w^2 \cdot I(w > 0)] - E[w^2 \cdot I(w < 0)] \\
&= \frac{2(1 - \tau)}{\tau^2} - \frac{2\tau}{(1 - \tau)^2} = 2 \frac{(1 - \tau)^3 - \tau^3}{\tau^2(1 - \tau)^2}
\end{aligned}$$

Now consider the following cross products of the score functions, assuming $u \sim ALPD(0, \tau, \sigma)$:

$$\begin{aligned}
E \left[\frac{\partial L}{\partial \beta} \frac{\partial L}{\partial \beta'} \right] &= \frac{1}{4\sigma^2} \sum_{i=1}^n (1 + 2(2\tau - 1)E[\text{sign}(y_i - x'_i\beta)] + (2\tau - 1)^2) x_i x'_i \\
&= \frac{1}{4\sigma^2} (-4\tau^2 + 4\tau) \sum_{i=1}^n x_i x'_i = \frac{\tau(1 - \tau)}{\sigma^2} \sum_{i=1}^n x_i x'_i \\
E \left[\frac{\partial L}{\partial \beta} \frac{\partial L}{\partial \tau'} \right] &= \frac{1 - 2\tau}{2\sigma\tau(1 - \tau)} \sum_{i=1}^n E[\text{sign}(y_i - x'_i\beta)] x'_i - \frac{1}{2\sigma^2} \sum_{i=1}^n E[|y_i - x'_i\beta|] x'_i \\
&\quad - \frac{(2\tau - 1)^2}{2\sigma\tau(1 - \tau)} \sum_{i=1}^n x'_i - \frac{2\tau - 1}{2\sigma^2} \sum_{i=1}^n E[(y_i - x'_i\beta)] x'_i \\
&= \frac{-1 + 2\tau - 2\tau^2 + 4\tau^2 - 4\tau + 1}{2\sigma\tau(1 - \tau)} \sum_{i=1}^n x'_i = -\frac{1}{\sigma} \sum_{i=1}^n x'_i
\end{aligned}$$

$$\begin{aligned}
E \left[\frac{\partial L}{\partial \beta} \frac{\partial L}{\partial \sigma'} \right] &= \frac{1}{2\sigma^3} \sum_{i=1}^n (E[\text{sign}(y_i - x'_i\beta)\rho_\tau(y_i - x'_i\beta)] + (2\tau - 1)E[\rho_\tau(y_i - x'_i\beta)]) x'_i \\
&= \frac{1}{2\sigma^3} \sum_{i=1}^n (E[\text{sign}(y_i - x'_i\beta)\rho_\tau(y_i - x'_i\beta)] + \sigma(2\tau - 1)) x'_i \\
&= \frac{1}{2\sigma^2} \sum_{i=1}^n \left(\frac{1}{2}E[w + (2\tau - 1)|w|] + (2\tau - 1) \right) x'_i \\
&= \frac{1}{2\sigma^2} \left(\frac{1}{2} \left(\frac{1 - 2\tau}{\tau(1 - \tau)} + (2\tau - 1) \frac{1 - 2\tau + 2\tau^2}{\tau(1 - \tau)} \right) + (2\tau - 1) \right) \sum_{i=1}^n x'_i \\
&= \frac{1}{2\sigma^2} \left(\frac{1}{2} \left(\frac{4\tau^3 - 6\tau^2 + 2\tau}{\tau(1 - \tau)} \right) + (2\tau - 1) \right) \sum_{i=1}^n x'_i = 0
\end{aligned}$$

$$\begin{aligned}
E \left[\frac{\partial L}{\partial \tau} \frac{\partial L}{\partial \tau'} \right] &= n \frac{(1 - 2\tau)^2}{\tau^2(1 - \tau)^2} + \frac{1}{\sigma^2} \sum_{i=1}^n E[(y_i - x'_i\beta)^2] - \frac{2(1 - 2\tau)}{\sigma\tau(1 - \tau)} \sum_{i=1}^n E[y_i - x'_i\beta] \\
&= n \frac{(1 - 2\tau)^2}{\tau^2(1 - \tau)^2} + n2\tau(1 - \tau) \left(\frac{1}{\tau^3} + \frac{1}{(1 - \tau)^3} \right) - n \frac{2(1 - 2\tau)^2}{\tau^2(1 - \tau)^2} \\
&= n \frac{2 - 6\tau + 6\tau^2 - (1 - 2\tau)^2}{\tau^2(1 - \tau^2)} = n \frac{1 - 2\tau + 2\tau^2}{\tau^2(1 - \tau^2)}
\end{aligned}$$

$$\begin{aligned}
E \left[\frac{\partial L}{\partial \tau} \frac{\partial L}{\partial \sigma'} \right] &= \frac{1 - 2\tau}{\sigma^2\tau(1 - \tau)} \sum_{i=1}^n E[\rho_\tau(y_i - x'_i\beta)] - \frac{1}{\sigma^3} \sum_{i=1}^n E[\rho_\tau(y_i - x'_i\beta) \cdot (y_i - x'_i\beta)] \\
&\quad - \frac{1 - 2\tau}{\sigma\tau(1 - \tau)} + \frac{1}{\sigma^2} \sum_{i=1}^n E[(y_i - x'_i\beta)] \\
&= n \frac{1 - 2\tau}{\sigma\tau(1 - \tau)} - n \frac{1}{\sigma} E[\rho_\tau(w) \cdot w] \\
&= n \frac{1 - 2\tau}{\sigma\tau(1 - \tau)} - n \frac{1}{\sigma} \left(\frac{(1 - \tau)^3 - \tau^3}{\tau^2(1 - \tau)^2} + (2\tau - 1) \frac{(1 - \tau)^3 + \tau^3}{\tau^2(1 - \tau)^2} \right) \\
&= n \frac{1 - 2\tau}{\sigma\tau(1 - \tau)} - n \frac{2(1 - \tau)^2 - \tau^2}{\sigma\tau(1 - \tau)} \\
&= -n \frac{1 - 2\tau}{\sigma\tau(1 - \tau)}
\end{aligned}$$

$$\begin{aligned}
E \left[\frac{\partial L}{\partial \sigma} \frac{\partial L}{\partial \sigma'} \right] &= \frac{n}{\sigma^2} + \frac{1}{\sigma^4} \sum_{i=1}^n E[\rho_\tau^2(y_i - x'_i\beta)] - \frac{2}{\sigma^3} \sum_{i=1}^n E[\rho_\tau(y_i - x'_i\beta)] \\
&= \frac{n}{\sigma^2} + \frac{n}{\sigma^2} E[\rho_\tau(w)^2] - \frac{2n}{\sigma^2} E[\rho_\tau(w)] = \frac{n}{\sigma^2}
\end{aligned}$$

References

- Abadie, Alberto, Angrist, Joshua, and Imbens, Guido (2002) “Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings,” *Econometrica*, 70, 91–117.
- Angrist, Joshua, Chernozhukov, Victor and Fernández-Val, Iván (2006) “Quantile Regression under Misspecification, with an Application to the U.S. Wage Structure,” *Econometrica*, 74, 539–563.
- Bera,, Anil K., Biliyas, Y. and Simlai, P. (2006) “Estimating functions and equations: An essay on historical developments with applications to econometrics,” in Mills, T.C, Patterson, K. (Eds.), *Palgrave Handbook of Econometrics*, vol.1 , pp.427–476.
- Abadie, Alberto, Angrist, Joshua, and Imbens, Guido (2002) “Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings,” *Econometrica*, 70, 91–117.
- Chernozhukov, Victor and Hansen, Christian (2006) “Instrumental quantile regression inference for structural and treatment effect models,” *Journal of Econometrics*, 132, 491–525.
- Chernozhukov, Victor and Hansen, Christian (2008) “Instrumental variable quantile regression: A robust inference approach,” *Journal of Econometrics*, 142, 379–198.
- Geraci, Marco and Botai, Mateo (2007) “Quantile regression for longitudinal data using the asymmetric Laplace distribution,” *Biostatistics*, 8, 140-154
- He, Xuming and Shao, Qi-Man (1996) “A General Bahadur Representation of M-Estimators and its Applications to Linear Regressions with Nonstochastic Designs,” *Annals of Statistics*, 24, 2608–2630.
- Koenker, Roger and Bassett, Gilbert (1978) “Regression Quantiles,” *Econometrica*, 46, 33–50.
- Koenker, Roger and Machado, Jose A.F. (1999) “Goodness of Fit and Related Inference Processes for Quantile Regression,” *Journal of the American Statistical Association*, 94,

1296–1310.

- Koenker, Roger and Xiao, Zhijie (2002) “Inference on the Quantile Regression Process,” *Econometrica*, 70, 1583–11612.
- Komunjer, Ivana (2005) “Quasi-maximum likelihood estimation for conditional quantiles,” *Journal of Econometrics*, 128, 137–164.
- Komunjer, Ivana (2007) “Asymmetric Power Distribution: Theory and Applications to Risk Measurement,” *Journal of Applied Econometrics*, 22, 891–921.
- Kotz, Samuel, Kozubowski, Tomasz J. and Podgórski, Krzysztof (2002) “Maximum Likelihood Estimation of Asymmetric Laplace Distributions,” *Ann. Inst. Statist. Math* 54 , 816–826.
- Machado, Jose A. F. (1993) “Robust Model Selection and M -Estimation,” *Econometric Theory*, 9, 478–493.
- Manski, Charles F. (1991) “Regression,” *Journal of Economic Literature*, 29, 34–50.
- Park, Sung Y. and Bera, Anil K. (2009) “Maximum entropy autoregressive conditional heteroskedasticity model,” *Journal of Econometrics*, 150, 219–230.
- Ruppert, David and Carroll, Raymond J. (1980) “Trimmed least squares estimation in the linear model,” *Journal of the American Statistical Association*, 75, 828–838.
- Shannon, C.E., (1948). The Mathematical Theory of Communication. Bell System Technical Journal (July-Oct), 3-91. Reprinted in: C.E. Shannon and W. Weaver, *The Mathematical Theory of Communication*. University of Illinois Press. Urbana, IL.
- van der Vaart, A.W. (1998) *Asymptotic Statistics*, Cambridge University Press, Cambridge.
- Yu, Keming and Moyeed, Rana A. (2001) “Bayesian quantile regression,” *Statistics & Probability Letters*, 54, 437-447.
- Yu, Keming and Zhang, Jin (2005) “A Three-Parameter Asymmetric Laplace Distribution and Its Extension,” *Communications in Statistics - Theory and Methods*, 34, 1867–1879.

Table 1: Location-Shift Model: Bias and RMSE

		MEQR	QR	OLS
$N(0, 1)$	Bias	0.0018	0.0014	0.0011
	RMSE	0.0929	0.0899	0.0706
	τ	0.501	0.500	—
t_3	Bias	0.0039	0.0013	0.0018
	RMSE	0.1258	0.0949	0.1207
	τ	0.498	0.500	—
χ_3^2	Bias	-0.0001	0.0037	0.0024
	RMSE	0.1311	0.1715	0.1840
	τ	0.081	0.500	—
ALPD ($\tau = 0.5$)	Bias	0.0019	0.0014	0.0011
	RMSE	0.0581	0.0549	0.0710
	τ	0.498	0.500	—
ALPD ($\tau = 0.25$)	Bias	-0.0008	0.0008	0.0016
	RMSE	0.0780	0.0858	0.0907
	τ	0.247	0.500	—

Table 2: Location-Scale-Shift Model: Bias and RMSE

		MEQR	QR	OLS
$N(0, 1)$	Bias	-0.0002	0.0036	0.0004
	RMSE	0.2287	0.1468	0.1363
	τ	0.498	0.500	-
t_3	Bias	-0.0087	-0.0039	-0.0111
	RMSE	0.2401	0.1440	0.2682
	τ	0.501	0.500	-
χ_3^2	Bias	0.0081	0.0086	0.0069
	RMSE	0.5191	0.2816	0.3633
	τ	0.085	0.500	-
ALPD ($\tau = 0.5$)	Bias	-0.0005	-0.0008	0.0003
	RMSE	0.1439	0.0861	0.1461
	τ	0.499	0.500	-
ALPD ($\tau = 0.25$)	Bias	0.0055	0.0076	0.4107
	RMSE	0.1362	0.1474	0.4505
	τ	0.248	0.500	-

Table 3: JTPA offer

	MEQR ($\hat{\tau} = 0.84$)		OLS		Median regression	
Intercept	9.894	(0.059)	8.814	(0.088)	9.188	(0.086)
JTPA offer	0.045	(0.022)	0.075	(0.032)	0.100	(0.033)
FEMALE	0.301	(0.023)	0.259	(0.030)	0.260	(0.031)
HSORGED	0.201	(0.025)	0.267	(0.034)	0.297	(0.037)
BLACK	-0.102	(0.026)	-0.121	(0.036)	-0.175	(0.039)
HISPANIC	-0.032	(0.034)	-0.034	(0.050)	-0.025	(0.051)
MARRIED	0.129	(0.025)	0.242	(0.036)	0.265	(0.034)
WKLESS13	-0.255	(0.023)	-0.598	(0.032)	-0.556	(0.036)
AGE2225	0.229	(0.057)	0.175	(0.084)	0.125	(0.080)
AGE2629	0.285	(0.058)	0.192	(0.085)	0.131	(0.081)
AGE3035	0.298	(0.057)	0.191	(0.084)	0.176	(0.080)
AGE3644	0.320	(0.058)	0.130	(0.085)	0.173	(0.081)
AGE4554	0.267	(0.064)	0.110	(0.094)	0.080	(0.092)
τ	0.840	(0.051)			0.500	
σ	0.249	(0.060)			0.538	(0.0055)

Notes: 9872 observations. JTPA offer: dummy variable for individuals that received a JTPA offer; FEMALE: Female dummy variable; HSORGED: dummy variable for individuals with completed high school or GSE; BLACK: race dummy variable; HISPANIC: dummy variable for hispanic; MARRIED: dummy variable for married individuals; WKLESS13: dummy variable for individuals working less than 13 weeks in the past year; AGE2225, AGE2629, AGE3035, AGE3644 and AGE4554 age range indicator variables.

Figure 1: JTPA offer: Quantile regression process and OLS

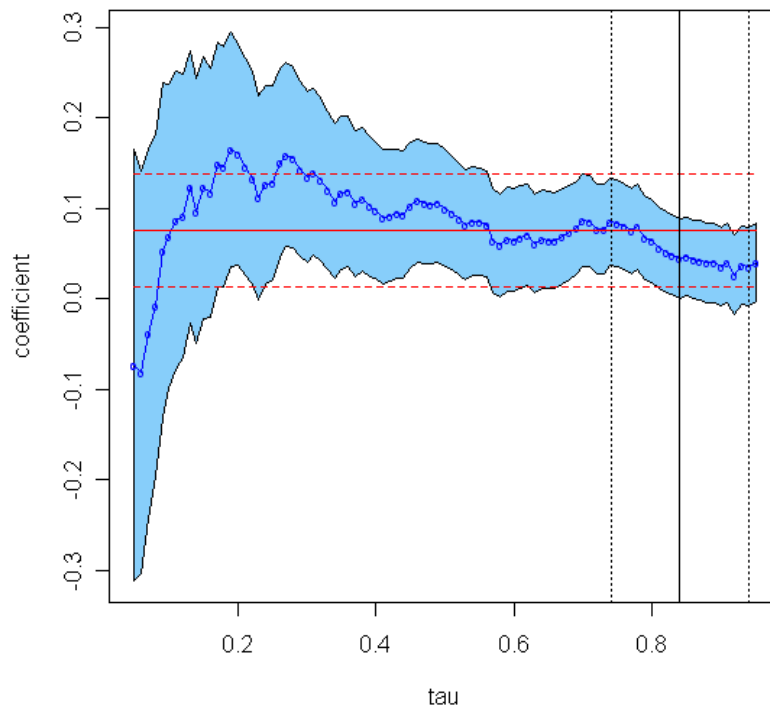


Figure 2: JTPA offer: τ -score function

$$\frac{1 - 2\tau}{\tau(1 - \tau)} - \frac{\sum_{i=1}^n (y_i - x_i' \hat{\beta}(\tau))}{n \hat{\sigma}}$$

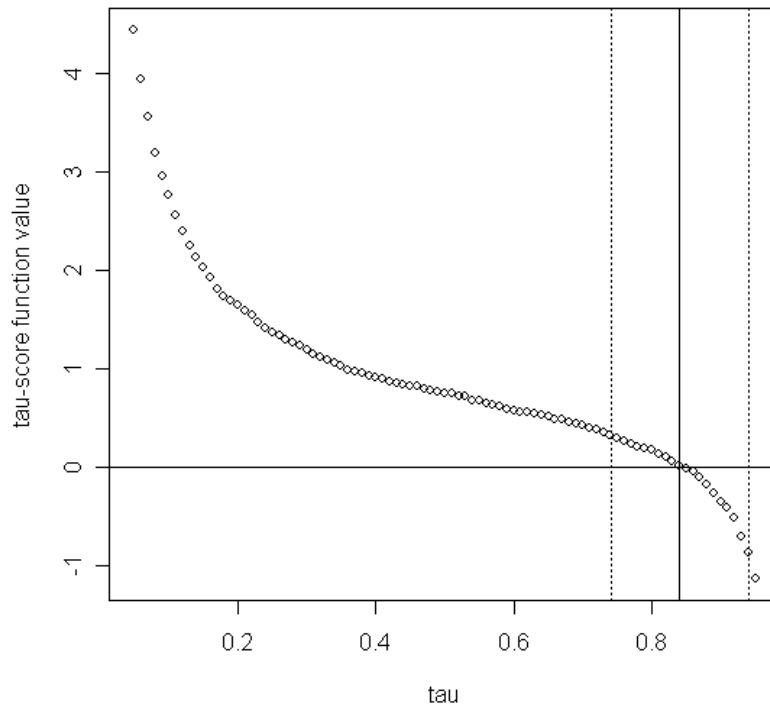


Figure 3: JTPA offer: Standard errors

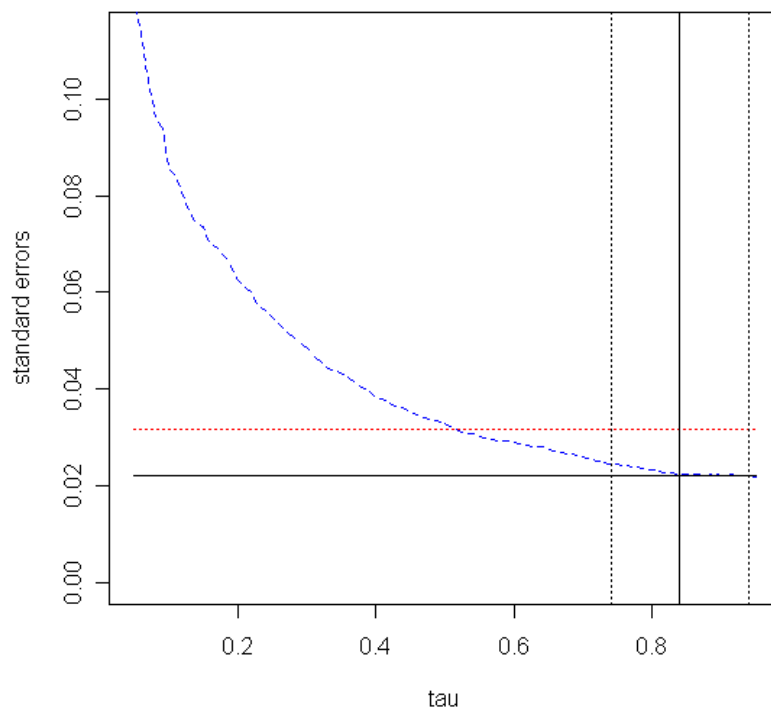


Figure 4: JTPA: Quantile regression process and OLS

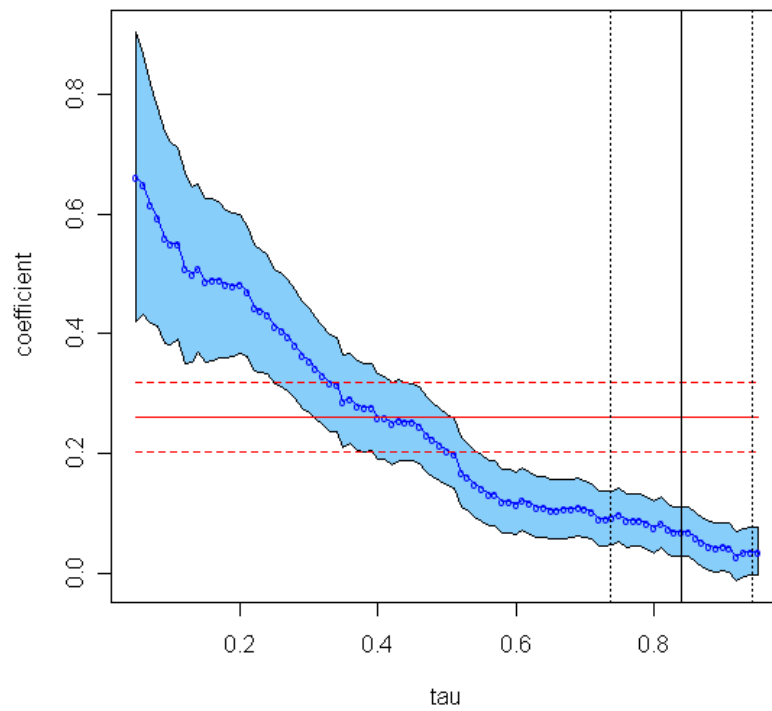


Figure 5: JTPA: Minimisation of $\|\gamma^2(\tau, \alpha)\|$

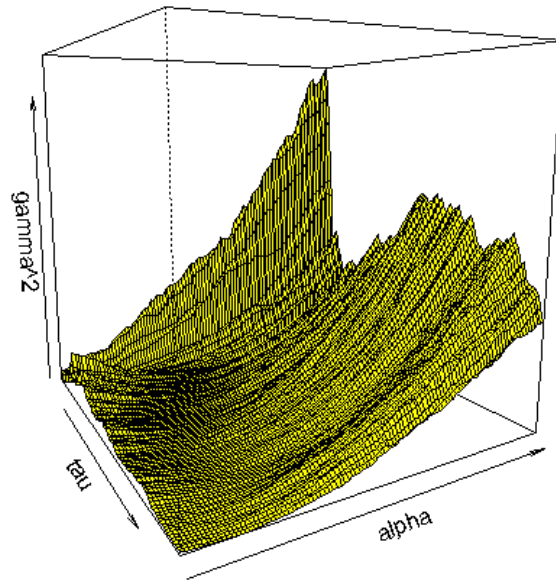


Figure 6: JTPA: Minimisation of $\|\gamma^2(\tau, \alpha)\|$

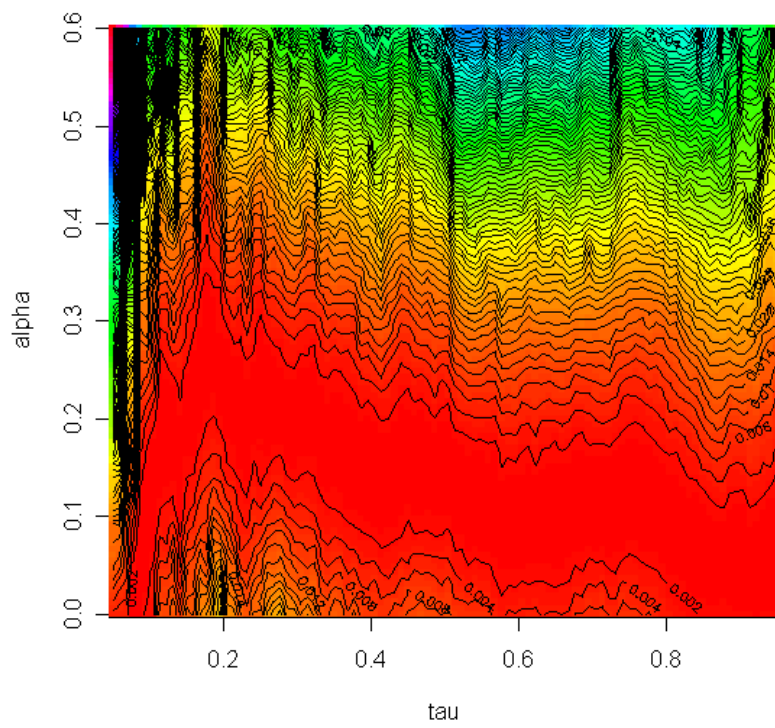


Figure 7: JTPA: IV Quantile regression process and IV OLS

