# GMM Quantile Regression*

Sergio Firpo† Antonio Galvao‡ Cristine Pinto§ Alexandre Poirier¶
Graciela Sanroman‖

June 24, 2019

**Abstract**

This paper develops generalized method of moments (GMM) estimation and inference procedures for quantile regression models when allowing for general parametric restrictions on the parameters of interest over a set of quantiles. First, we suggest a GMM estimator for simultaneous estimation across multiple quantiles. This estimator exploits a partition of the quantile space, which induces a weighting matrix that is independent of the parameters of interest and the number of partitions itself. The GMM estimator is designed to estimate a fixed number of quantiles simultaneously, is flexible since it allows for imposing restrictions on the parameters of interest over a set of moments indexed by the quantiles, and accounts for information across quantiles to improve efficiency. Second, we study the properties of the GMM estimator when the number of partitions diverge to infinity, and derive its efficiency bound. Third, we suggest an alternative smooth GMM estimation procedure for large number moments. We establish the asymptotic properties of both GMM estimators. These methods have the advantage of being simple to implement in practice. Monte Carlo simulations show numerical evidence of the finite sample properties of the methods. Finally, we apply the proposed methods to estimate the effects of various covariates on birthweight of live infants at the extreme bottom of the conditional distribution.

Keywords: Quantile Regression, Generalized Method of Moments, Extreme Quantile
JEL Codes: C21, C31

*We would like to thank Matias Cattaneo and seminar participants at Warwick University, University of Michigan, Insper, New York Camp Econometrics XIV, 2018 Midwest Econometrics Group, and 2017 Rio-Sao Paulo Conference for helpful comments and discussions. Luis Fantozzi and Alejandra Marroig provided excellent research assistance. Computer programs to replicate all the numerical analyses are available from the authors. All the remaining errors are ours.

†Insper Institute of Education and Research: firpo@insper.edu.br
‡Department of Economics, University of Arizona: agalvao@email.arizona.edu
§Sao Paulo School of Economics - FGV : cristine.pinto@fgv.br
¶Department of Economics, Georgetown University: alexandre.poirier@georgetown.edu
‖UDELAR: gsanroman@decon.edu.uy

# 1 Introduction

In a seminal paper, Koenker and Bassett (1978) introduced quantile regression (QR) models. By estimating conditional quantile functions, QR models have provided a valuable tool in economics, finance, and statistics as a way of capturing heterogeneous effects that policy variables may have on the outcome of interest, exposing a wide variety of forms of conditional heterogeneity under weak distributional assumptions. This is especially valuable for program evaluation studies, where these methods help analyze how treatments or social programs affect the outcome's distribution.

This paper develops generalized method of moments (GMM) estimation and inference procedures for QR models when allowing for general parametric restrictions on the parameters of interest over a set of quantiles. In many empirical applications, researchers are interested in modeling a specific part of the distribution, or the entire distribution, of an outcome of interest. For example, in extremal quantile models, the researcher is interested in modeling the tails of the conditional distribution. These models have become popular with an increasing number of economic, financial, and statistical applications such as value-at-risk, analysis of tail risk over time, production frontiers, determinants of low infant birth weights, and auction models (see Chernozhukov et al. (2017) for an overview of quantile extremal models).[1] This paper provides a useful alternative to extremal quantile regression since the imposition of parametric restrictions on the quantile coefficients allows us to use GMM estimators we describe below. These estimators follow standard (Gaussian) asymptotic distributions as opposed to extremal quantile regression estimators, which are nonstandard.

There are several other examples where the proposed methods may be useful. For instance, Donald and Paarsch (1993) use a parametric model to test whether the predictions of game theory models for actions are compatible with observable data; Koenker and Geling (2001) use parametric restrictions to estimate a QR survival analysis model. In these examples, the researcher does not need to specify the entire conditional distribution of outcomes, but only the relationship for a subset of quantiles. Another set of examples include specification of quantiles over the entire conditional distribution as, among others, constant treatment effects, stochastic dominance, existence of threshold, and changes in distribution (see, e.g.,

[1] Chernozhukov (2005) and Zhang (2016) derive the asymptotic theory for extreme quantile estimators.

Qu and Yoon (2017), Chiang and Sasaki (2017), Firpo and Pinto (2015), and Galvao et al. (2011)).

The first main contribution of this paper to the literature is to develop a GMM-QR estimator for simultaneous estimation across quantiles. Concurrent estimation of multiple quantiles has not received much attention in the literature.[2] In many applications one simply estimates a standard QR model for a given quantile, and then varies the quantile $\tau$ in a given grid set (see, e.g., Koenker (2005)). The proposed GMM-QR estimator, however, is designed to estimate a fixed number of quantiles simultaneously by using moment conditions indexed by different quantiles $\tau \in \mathcal{T} \subset (0,1)$. This estimator is flexible since it allows for imposing restrictions on the parameters of interest over the set of moments. Importantly, the GMM-QR estimator accounts for information across quantiles to improve efficiency. In order to make the estimation procedure feasible, we exploit an equally spaced partition of the quantile interval using a finite number of grid points, $L-1$ in $(0,1)$, which form a partition of the unit interval. Based on this partition, we construct moment conditions determined by the exogenous variables and quantiles. This partitioning technique has the advantage of producing a weighting matrix that does not depend on the parameters of interest. We establish the asymptotic properties of the GMM-QR estimator and show that it is consistent and asymptotically normal. By exploiting the estimation of multiple quantiles simultaneously, under the correct restriction at the coefficients of interest, the GMM-QR estimator is more efficient than the simple quantile-by-quantile estimation.

We then discuss the optimal GMM-QR estimator by using the optimal weighting matrix from the standard GMM theory (see, e.g., Hall (2005)). The optimal GMM-QR also uses the new partition matrix, which together with the optimal moment conditions for QR produce the optimal weighting matrix. As expected, we show that the optimal GMM-QR estimator is asymptotically more efficient than the GMM-QR estimator.

Next we study the efficiency properties of the GMM-QR estimator when the number of partitions of the quantile interval, $L$, diverge to infinity. First, for large $L$, we derive the limiting GMM objective function, as well as the efficiency bound from the calculation of the

[2]Koenker (2004) and Koenker (1984) proposed estimation for several quantiles simultaneously by considering a weighted QR estimation in which the quantiles are also predetermined. Nevertheless, these methods do not account for information across quantiles. To the best of our knowledge, the paper that is closest to ours is Yang and He (2012). In this paper, the authors use prior information about the relationship across quantile and Bayesian Empirical Likelihood to estimate several quantiles jointly.

asymptotic variance-covariance matrix of the Maximum Likelihood estimator (MLE). We then show that even when $L$ grows to infinity, the GMM-QR is not efficient. Nevertheless, we show that when the number of partitions, $L$, diverges to infinity the optimal GMM-QR estimator's asymptotic variance converges to the semiparametric efficiency bound.

We also suggest GMM-QR estimation using an alternative smooth GMM (SGMM-QR) estimator for large number of moment conditions. This SGMM-QR is based on a smooth basis functions that uses all the information embedded in the identifying restriction. This method has been used in the literature to obtain efficient estimators under conditional moment restrictions (See Newey (1990), Newey (1993), Donald et al. (2003)) and also under independence restrictions (Poirier (2017)). We establish the limiting properties of the SGMM-QR estimator, consistency and asymptotic normality, and show that it achieves the lower bound as well. We note that, regarding the SGMM-QR estimator, a restriction on the vector of parameters, which is a vector function of $\tau$, needs to be satisfied for all $\tau$ in its support, $\mathcal{T}$. When we increase the number of partitions in the support, the number of moments restrictions increases. The main technical difficulty with this statistical model is that we have an infinite number of moment conditions.

Compared to the existing procedures for estimation and inference of QR models, our approach has several distinctive advantages. First, an important application of the proposed methods is to allow researchers to estimate models under restrictions on the coefficient functions. Second, a direct implication of imposing restrictions on coefficients is to test hypotheses on the shape of the quantile curve. Third, the estimator is flexible and does not necessarily require modeling the entire conditional distribution of the variable of interest. In particular, our methods can be useful as an alternative to estimation of extreme quantiles. Fourth, we derive the optimal GMM-QR, which is efficient. Finally, our algorithm is computationally simple and easy to implement in practice. The proposed procedures should be useful for those empirical settings based on QR models in which estimation of the entire or part of the conditional quantile function is a concern. For example, in the survival analysis, the restrictions across quantiles come from the parametric assumption about the survival function. In the tail risk application, there is a particular relationship among the extreme quantiles.

To illustrate the proposed methods, we consider an empirical application to a birthweight study using data from the National Center for Health Statistics. We estimate the effects

of various covariates (marital status, gender, smoke, number of cigarettes a day, and lack of prenatal care) on birthweight of live infants at the extreme bottom of the conditional distribution. Given the difficulty to perform inference at the extremes of the distribution, the empirical results document important statistically significant effects of smoking at the bottom tail. Although intuitive, these findings complement the existing results in the literature.

We now briefly review the literature related to this paper. The paper that is closest to ours is Yang and He (2012). The authors impose a prior information on the quantile coefficients across several values of quantiles. They estimate the quantiles all together using a Bayesian Empirical Likelihood estimator. They establish the asymptotic distribution of the posterior and compute efficiency gains from informative priors. As in our method, they can impose restrictions in specific subsets of the quantile interval. There is also a literature using moment conditions to estimate QR models, among many others, Xu et al. (2017), Chen and Lee (2017), Kaplan and Sun (2017), Chen et al. (2015), Chen and Liao (2015), Chen and Pouzo (2012, 2009), Chernozhukov and Hong (2003), and Buchinsky (1998). In this paper, we extend the literature on GMM for conditional average models as well as that on GMM for QR models by adapting the GMM methods for estimation and inference of QR models allowing for imposing parametric restriction on the parameters of interest as well as simultaneous estimation across quantiles. There is also a small literature combining information from QR. One may consider combining information over different quantiles via the criterion or loss function. For example, Zou and Yuan (2008) and Bradic et al. (2011) proposed the composite QR for parameter estimation and variable selection in the classical linear regression models. Zhao and Xiao (2014) construct efficient estimators of regression models via QR. The paper is also related to the literature about adaptive estimators. Newey (1988) use an adaptive GMM estimator for linear regression model. He shows that when the number of moments increase with the sample size, the GMM estimator approaches the MLE for that problem, assuming that the error term has a known distribution function. Portnoy and Koenker (1989) also develop adaptive L-estimator for linear regression models by combining information based on estimators at different quantiles for the linear regression model. The adaptive estimator proposed by the authors is a weighted integral of linear functions of estimators for the slope and intercept of the linear regression quantile models. The integral is taken over the unit interval $(0, 1)$ of possible quantile values. This adaptive

L-estimator is similar to our GMM estimator since it uses a infinite combination of quantile regressions for different values of quantiles in the interval $(0, 1)$. The big different is that in our case we use use information across quantiles and estimate all the quantiles regressions simultaneous. As Newey (1988), we show that our estimator attains the efficient bound (the variance of the MLE), when the number of moments increases with the sample size.

The remaining of the paper is organized as follows. Section 2 presents the statistical model. The GMM-QR estimator is described in Section 3. Section 4 describes the MLE and establishes the relationship between the GMM and the MLE. It also discusses the SGMM-QR. Section 5 illustrates the methods with an application to the estimation various covariates on birthweight. Finally, Section 6 concludes the paper. All proofs and numerical Monte Carlo exercises are provided in a Supplemental Online Appendix.

# 2 Statistical Model

Many statistical models can be written as follows

$$Y = \Gamma(X, U), \tag{2.1}$$

where $Y$ is a dependent variable, $X$ is a $K$-vector of regressors, and $U$ is a scalar disturbance that is independent of $X$. The potentially non-separable function $\Gamma(\cdot)$ allows for the effect of $X$ to depend on the unobserved component $U$.

Our main model is a version of the standard linear quantile regression (QR) model. Recall the baseline model for the linear QR representation,

$$Y = X^\top \beta(U), \tag{2.2}$$

where $U$ represents the heterogeneity in responses. Imposing that $u \longmapsto X^\top \beta(u)$ is strictly increasing in $u$ with probability 1, from equation (2.2) we can write the conditional quantile of $Y$ given $X$ as

$$Q_Y(\tau|X) = X^\top \beta(Q_U(\tau|X)),$$

where $\tau \in (0, 1)$ is the quantile of interest and and $Q_Y(\tau|X)$ is the conditional $\tau$-quantile of $Y$ given $X$. Under the exogeneity assumption, i.e., the unobserved heterogeneity $U$ is inde-

pendent from $X$ and without loss of generality normalizing $U$ to have a uniform distribution on $[0,1]$, $U \sim \text{Unif}[0,1]$, we obtain the standard representation for the linear QR model,

$$Q_Y(\tau|X) = X^\top \beta(\tau), \tag{2.3}$$

where $X \in \mathcal{X} \subset \mathbb{R}^K$. In this model, $X^\top \beta(\tau)$ is an increasing function of $\tau$.

It is important to note that even though $X^\top \beta(\tau)$ is a general way to model the conditional quantile of $Y$ as a linear function of $X$, it creates an asymmetry between the effects of $\tau$ and the effects of $X$ on the conditional quantile function of $Y$. For a fixed $\tau$, we know that the marginal effect of changes in $X$ on $Y$ is $\beta(\tau)$, which does not depend on $X$. But the marginal effect of changes in $\tau$, for a fixed $X = x$, is $x^\top \partial\beta(\tau)/\partial\tau$, which generally depends on $\tau$. Thus, in standard QR models, the main parametric restrictions imposed on the model are captured by how the observable components enter function (2.3), while the format of how $\tau$ affects $Y$ is left unspecified.

In this paper, we add to the linear QR model a flexible assumption about how $\tau$ affects the conditional quantile function. We assume that the slope function can be represented as follows

$$\beta(\tau) = g(\theta, \tau), \tag{2.4}$$

where $g(\cdot)$ is a $K$-vector of known functions and $\theta$ is a vector of unknown parameters.

Under the restriction (2.4), the model in equation (2.2) relates the dependent variable $Y$ as a function of observables $X$ and the unobservable random variable $U$ as,

$$Y = X^\top g(\theta, U),$$

and the QR model in (2.3) is then parametrically specified in both $X$ and $\tau$ as,

$$Q_Y(\tau|X) = X^\top \beta(\tau) = X^\top g(\theta, \tau).$$

The restriction in (2.4) can be motivated by different reasons. Theoretical models often assume that there is some specified unobservable factor entering into the outcome equation. Observable factors may not be entirely modeled and are in many cases treated as some 'nuisance' parameters or functions. Another important 'non-structural' reason for imposing parametric restriction on the conditional quantile process is for inference at the tails. Moreover, the restrictions in (2.4) on how the quantiles $\tau$ enter the conditional quantile function

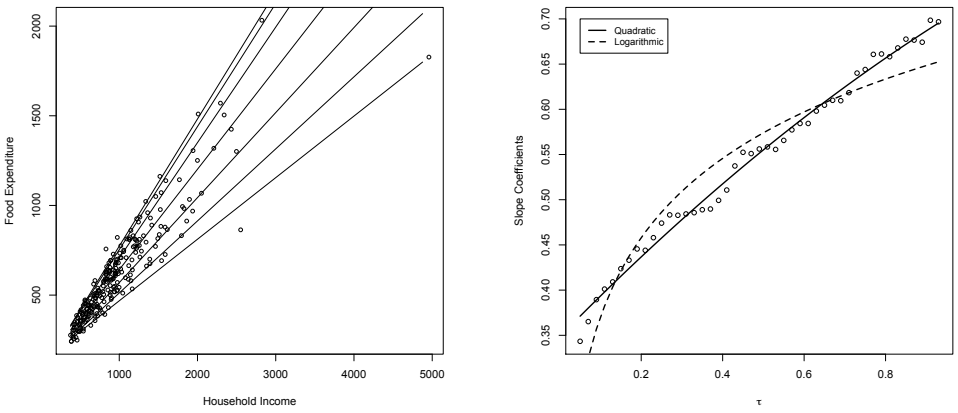



Figure 1: Left box plots QR fitted lines for different quantiles. Right box plots the QR slope coefficient as function of $\tau$, as well as quadratic and logarithmic models.

of $Y$ are natural in many empirical examples. For instance, in survival function analysis, the restrictions across quantiles come from the functional form of the survival function. The literature on estimation of tail risk in finance usually imposes restrictions on the extreme quantiles of the distribution of $Y$. Now we provide two brief examples to illustrate the parametric restrictions across quantiles in more details.

**Example 2.1.** First, we use the well-known Engel curve example in Koenker (2005). This is a classical data set in economics based on 235 budget surveys of 19th century working-class households. The left panel in Figure 1 plots the data and several fitted lines for a linear QR model. Household expenditure on food is measured on the vertical axis, and household income is measured on the horizontal axis. When examining the slopes coefficients, it is possible to see that, as the quantiles increase, the slopes increase but at decreasing rates. The points in right panel in Figure 1 are the slope coefficients as a function of quantiles $\tau$, and we superpose lines with a quadratic and a logarithmic model. Thus, the pictures suggest a concave relationship between the slope coefficients and the quantiles $\tau$. Our method will allow to estimate all the quantile effects simultaneously, imposing the concave curvature across quantiles.

**Example 2.2.** Consider the simple case of a linear location-scale-shift QR model with one regressor. The model can the written as

$$Y = b_0 + b_1 X + \sigma(X)U,$$

where $\sigma(X) = (\gamma_0 + \gamma_1 X)$. Thus, the conditional quantile function can be written as following

$$\begin{aligned} Q_Y(\tau|X) &= b_0 + b_1 X + \sigma(X) F_U^{-1}(\tau) \\ &= (b_0 + \gamma_0 F_U^{-1}(\tau)) + (b_1 + \gamma_1 F_U^{-1}(\tau)) X \\ &= \beta_0(\tau) + \beta_1(\tau) X, \end{aligned}$$

where $\beta_0(\tau) = (b_0 + \gamma_0 F_U^{-1}(\tau))$ and $\beta_1(\tau) = (b_1 + \gamma_1 F_U^{-1}(\tau))$. For simplicity, let $F_U$ be the Logistic distribution with location and scale parameters zero and one, respectively. Then, $F_U^{-1}(\tau) = \ln \tau/(1-\tau)$, and we have both the constant and slope coefficients varying with the quantile $\tau$. The function $g(\theta, \tau)$ is a simple linear function as

$$\beta_0(\tau) = g_0(\theta, \tau) = \theta_1 + \theta_2 \ln \frac{\tau}{1-\tau}, \quad \beta_1(\tau) = g_1(\theta, \tau) = \theta_3 + \theta_4 \ln \frac{\tau}{1-\tau}.$$

In Section 5 below, we provide an empirical illustration to the proposed methods with an application to the estimation of the effects of various covariates on birthweight at the extreme of the distribution, which is motivated by this location-scale example.

# 3 GMM Quantile Regression Estimation

This section presents a generalized method of moments quantile regression (GMM-QR) estimator when the number of moments restrictions is fixed. First, we discuss the population moment conditions of interest. Second, we present a new partitioning argument of the quantile space that we use for estimation. This partition facilitates practical implementation. Third, we present the GMM-QR estimator and establish its asymptotic properties. Finally, we briefly discuss the optimal GMM, and show that for a fixed number of partitions the optimal GMM is more efficient than the GMM-QR.

## 3.1 Moment Conditions

We consider the model in equation (2.3). In order to describe the model in a formal way, we impose the following assumptions.

**A1.** There is a random sample of size $n$ such that the data $\{(Y_i, X_i)\}_{i=1}^n$ are independent and identically distributed (i.i.d.).

**A2.** The conditional density of $Y_i$ given $X_i$ ($f_{Y_i|X_i}(y|x)$) exists almost surely for every $i = 1, \ldots, n$. Additionally, it is bounded above, bounded away from zero, and continuous differentiable in $Y$, uniformly over the support of $X_i$.

Assumptions **A1** and **A2** are standard in the literature. We impose **A1** only to make presentation of the results simpler. We also impose the following restriction on the slope function of the QR model.

**A3.** $\beta_0(\tau) = g(\theta_0, \tau)$ for $\theta_0 \in \mathbb{R}^{d_\theta}$ for all $\tau \in \mathcal{T} \subset (0,1)$, and $g(\cdot,\cdot)$ is continuous and twice differentiable in both arguments, with $g_\theta(\cdot,\cdot)$ being the partial derivative with respect to the first argument, and $g_u(\cdot,\cdot)$ being the partial derivative with respect to the second argument.

**A4.** There is a unique $\theta_0$ such that $Q_Y(\tau|X) = X^\top g(\theta_0, \tau)$, which is an increasing function on $\tau$.

Assumption **A3** describes the restriction on the coefficients of interest as a function of the quantiles. These restrictions are of interest for modeling purposes. Note that $g(\cdot)$ is a $K$-vector of functions that are differentiable in both arguments. We assume that the functional form of the functions into vector $g(\cdot)$ are known up to the vector of parameters $\theta$, which has dimension size $\mathbb{R}^{d_\theta}$. For each component of the vector $\beta(\tau)$ ($\beta_j(\tau)$, for $j = 1, ..., K$), we can have a different format for the function $g_j(\cdot)$. When we impose assumption **A3**, the vector of unknown parameter is $\theta$, and not $\beta(\tau)$. Condition **A4** states a correct model specification.

The QR linear model in (2.3) can be represented by a conditional moment restriction as:

$$\mathrm{E}\left[\tau - 1\left\{Y - X^\top \beta_0(\tau) \leq 0\right\} \middle| X\right] = 0, \tag{3.1}$$

for all $\tau \in (0,1)$. This conditional moment restriction implies many unconditional moment conditions that are going to be used to estimate $\beta(\tau)$. For a given $\tau$, let

$$m\left(W; \beta, \tau\right) = q_M(X, \tau)\left(\tau - 1\left\{Y - X^\top \beta(\tau) \leq 0\right\}\right), \tag{3.2}$$

where $q_M(X,\tau)$ is a $M_\tau$-vector of subset of the conditioning (or instrumental) variables of $X$, and $W = [Y, X]$. For a fixed quantile $\tau$, $M_\tau \geq K$. In addition, notice that, in the standard linear QR case, $q_M(X,\tau) = X$ for all $\tau$.

In this paper, for simplicity of exposition, we work with exogenous regressors, $X \in \mathbb{R}^K$, for constructing the conditioning variables set. Nevertheless, we note that one is able to interpret the conditioning set in (3.1) as a vector of instrumental variables. The results presented in the paper extend to the case of instrumental variables as long as the moment conditions are still satisfied.[3]

Using equation (3.2), we can then define the following unconditional moment condition for the QR problem,

$$\mathrm{E}\left[m\left(W;\beta_0,\tau\right)\right]=0. \tag{3.3}$$

Several other papers in the literature, see e.g., Kaplan and Sun (2017), Chen and Liao (2015), Chen and Pouzo (2012, 2009), use an unconditional moment condition similar to (3.3) for estimation and inference in quantile models.

Finally, when we impose the structure $g(\cdot)$ across quantiles, we obtain the set of moments conditions in (3.3) with the following $m(\cdot)$ function where, for a given $\tau$, $M_\tau \geq d_\theta$

$$m\left(W;g\left(\theta,\tau\right),\tau\right):=q_M(X,\tau)\psi(W,\theta,\tau), \tag{3.4}$$

where $\psi(W,\theta,\tau):=\left(\tau-1\left\{Y-X^\top g\left(\theta,\tau\right)\leq 0\right\}\right)$.

To make estimation for multiple quantile practical in applications, we use a partition of the space of quantiles, which we discuss in detail in the next section. From equation (3.4), let

$$\begin{aligned}\mathsf{m}(W,\theta,\tau_1,...,\tau_{L-1}) &:= \left[m\left(W;g\left(\theta,\tau_1\right),\tau_1\right)^\top,...,m\left(W;g\left(\theta,\tau_{L-1}\right),\tau_{L-1}\right)^\top\right]^\top, \\ &= \left[q_M(X,\tau_1)^\top\psi(W,\theta,\tau_1),...,q_M(X,\tau_{L-1})^\top\psi(W,\theta,\tau_{L-1})\right]^\top,\end{aligned} \tag{3.5}$$

be the $\sum_{l=1}^{L-1} M_{\tau_l}$-vector of moments restrictions for $L-1$ values of $\{\tau_1,\ldots,\tau_{L-1}\}\subset(0,1)$. The index $L$ defines the number of partitions of the quantile space, hence $L-1$ is the number of quantiles. For instance, when $L=4$ we have the three quartiles and four partitions. For simplicity, we consider $M_{\tau_1}=M_{\tau_2}=...M_{\tau_{L-1}}=M$ for now, so the dimension of vector of moments is $M\cdot(L-1)$.

We note that the weights $q_M(X,\tau)$ may be a function of the quantile $\tau$ in the general statement of the problem in (3.5). Nevertheless, it is common in the QR literature for linear

[3]de Castro et al. (2018) develop a GMM estimator for instrumental variables QR models where they smooth the indicator function using the idea of Horowitz (1998).

models to impose $q_M(X,\tau) = X$ and use the moment restrictions (3.4) for a unique fixed $\tau$. For example, to investigate the effect at the median, we fix $\tau = 0.5$ and use the moment restrictions at this specific value of $\tau$. In this case, in order to obtain identification, the dimension of $X$ needs to be at least equal to the dimension of $\theta$. If we assume a very flexible function form for $g(\theta,\tau)$, we may not be able to achieve identification if we use the moment for a specific $\tau$ as in the standard case. For instance, suppose that $X$ is a scalar, and the function $g$ is linear in $\tau$, $g(\theta,\tau) = \theta_0 + \theta_1 \cdot \tau$. In this case for a fixed $\tau$, the dimension of $X$ is smaller than the dimension of $\theta$, and we cannot obtain identification. In this case, we need to consider the moments associated with different values of $\tau$ at the same time. In this case, identification comes from the restrictions imposed across different values of $\tau$.

In this section, we follow the QR literature, and with abuse of notation, set $q_M(X,\tau) = q_M(X)$ for all $\tau$. Nevertheless, we will see below that for the optimal GMM, the weights are a function of the quantiles. When the function $q_M$ does not depend on $\tau$, the vector $\mathsf{m}(\cdot)$ in (3.5) can be written as

$$\mathsf{m}(W,\theta,\tau_1,...,\tau_{L-1}) = \Big[\psi(W,\theta,\tau_1),...,\psi(W,\theta,\tau_{L-1})\Big]^\top \otimes q_M(X). \tag{3.6}$$

In this paper, we move away from the standard QR literature, and our parameter of interest is the vector of parameters $\theta$ that solves the following minimization problem considering a set of values of $\tau$, $\tau_1,...,\tau_L$ (and not only one specific $\tau$)

$$\theta_0 = \underset{\theta}{\operatorname{argmin}}\, \mathrm{E}\left[\mathsf{m}(W,\theta,\tau_1,...,\tau_{L-1})\right]^\top \Omega(\theta,\tau_1,...,\tau_{L-1})^{-1} \mathrm{E}\left[\mathsf{m}(W,\theta,\tau_1,...,\tau_{L-1})\right], \tag{3.7}$$

where $\mathsf{m}(W,\theta,\tau_1,...,\tau_{L-1})$ is given in (3.6), and $\Omega(\theta,\tau_1,...,\tau_{L-1})$ is the weighting matrix that is equal to the variance-covariance matrix of the moments conditions. Except for Koenker (1984, 2004) and Yang and He (2012), the studies in the QR literature estimate separate regression models for each value of $\tau$. These simple approaches correspond to solving a GMM set-up for a single given value of $\tau$ separately.

Note that in the system of unconditional moments restrictions in (3.7), the dimension of unknown parameters is equal to the dimension of $\theta$, $d_\theta$. In this case, the dimension of the vector of instruments $M$ needs to be at least as large as $d_\theta$. In our case, $\theta$ is the same for all values of $\tau$, and it is more efficient to estimate a model that considers the moments conditions for different values of $\tau$ all together. When we consider the moments conditions

for all $\tau$ together, even in the case that $M = d_\theta$, we have a overidentified model since the number of equations is larger than the number of parameters.[4]

## 3.2 The Partitioning Argument

In order to make estimation over multiple quantiles $\tau$ feasible and simple, we use a partition of the quantile space. We partition off the interval $(0, 1)$ into $L$ equally spaced intervals. In this case, there are $L-1$ quantiles, and the $K$ vector $\beta(j/L) = g(\theta, j/L)$, $j = 1, ..., L-1$. Let

$$\begin{aligned}\mathsf{H}_L(W;\theta) &:= \left[\frac{1}{L} \ -1\left\{Y - X^\top g\left(\theta, \frac{1}{L}\right) \leq 0\right\}, ..., \frac{L-1}{L} - 1\left\{Y - X^\top g\left(\theta, \frac{L-1}{L}\right) \leq 0\right\}\right]^\top \\ &= \left[\psi\left(W, \theta, \frac{1}{L}\right), ..., \psi\left(W, \theta, \frac{L-1}{L}\right)\right]^\top.\end{aligned}$$

Based on this partition of the quantile space, and the moment conditions in (3.4), we can construct a $M \cdot (L-1) \times 1$ vector of moments denoted by $\mathsf{m}_L(W; \mathsf{g}_L(\theta))$ and defined as follows:

$$\mathsf{m}_L(W;\mathsf{g}_L(\theta)) := \left[m\left(W; g\left(\theta, \frac{1}{L}\right), \frac{1}{L}\right)^\top, m\left(W; g\left(\theta, \frac{2}{L}\right), \frac{2}{L}\right)^\top, ..., m\left(W; g\left(\theta, \frac{L-1}{L}\right), \frac{L-1}{L}\right)^\top\right]^\top. \tag{3.8}$$

We also define the following:

$$\begin{aligned}\mathsf{g}_L(\theta) &:= \left[g(\theta, 1/L)^\top, g(\theta, 2/L)^\top, ..., g(\theta, (L-1)/L)^\top\right]^\top \\ \Omega_L(\theta) &:= \mathrm{E}\left[\mathsf{m}_L(W;\mathsf{g}_L(\theta))\mathsf{m}_L(W;\mathsf{g}_L(\theta))^\top\right],\end{aligned}$$

where $\mathsf{g}_L(\theta)$ is a $K \cdot (L-1)$ vector and $\Omega_L(\theta)$ is a $M \cdot (L-1) \times M \cdot (L-1)$ covariance matrix.

Our goal is to rewrite the population version of the GMM-QR minimization problem in (3.7) using the partition argument. As mentioned in the previous section, the GMM-QR considers the simple case of the moment condition in equation (3.6) where $q_M(X, \tau) =$

[4] In this case, we can still have a just identified case if $M$ times the number of quantiles is equal to the dimension of the vector of unknown parameters, $\theta$.

$q_M(X)$, that is, weights are independent of the quantiles. To do so, we impose one more assumption:

**A5.** For fixed $M$, $q_M(X,\tau) = q_M(X)$, and the matrix $\Sigma_M := \mathrm{E}[q_M(X)q_M(X)^\top]$ is positive definite.

Assumption **A5** guarantees that the weighting matrix in this overideintified GMM is well-defined. Assumption **A5** together with **A4** impose identification for a fixed $L$. For a general discussion on identification we refer the reader to Chen et al. (2014). This weighting matrix will be the Kronecker product of the matrix in assumption **A5** and a matrix of dimension $(L-1)\times(L-1)$ that will represent the part of the weight matrix related to number of partitions of unit interval. Thus, our parameter of interest can be written as the solution of the following problem

$$\theta_0 = \underset{\theta}{\operatorname{argmin}}\, Q_L(\theta), \tag{3.9}$$
$$Q_L(\theta) := \mathrm{E}[\mathsf{m}_L(W;\mathsf{g}_L(\theta))]^\top \Omega_L^{-1}(\theta_0)\,\mathrm{E}[\mathsf{m}_L(W;\mathsf{g}_L(\theta))].$$

Note that equation (3.9) is equal to equation (3.7) using the partition of the quantile space.

Next we present a result where we compute the population function $Q_L(\theta)$ as a function of the number of partitions of the quantile space. Define

$$\Sigma_L^{-1} := L \cdot \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ 0 & -1 & \ddots & \ddots & 0 \\ \vdots & \vdots & \ddots & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{bmatrix}. \tag{3.10}$$

**Lemma 3.1.** *Under conditions **A1**-**A5**, the objective function of the partition GMM problem $Q_L(\theta)$ in (3.9) can be expressed as:*

$$Q_L(\theta) = \left(\mathrm{E}\left[\mathsf{H}_L(W;\theta)\otimes q_M(X)\right]\right)^\top \cdot \left(\Sigma_L^{-1}\otimes\Sigma_M^{-1}\right)\cdot\left(\mathrm{E}\left[\mathsf{H}_L(W;\theta)\otimes q_M(X)\right]\right).$$

The result in Lemma 3.1 shows an important intermediate result in this paper. By using the partition in (3.8), the objective function $Q_L(\theta)$ in (3.9) simplifies substantially and becomes very tractable for statistical analysis. The main simplification is on the inverse of weighting matrix, $\Omega_L(\theta_0)$, which can be written as $\Sigma_L^{-1}\otimes\Sigma_M^{-1}$. Therefore, the weighting matrix in the population function does not depend on the parameter $\theta$.

## 3.3 GMM-QR Estimator

Now we suggest a GMM-QR estimator based on a set of the moments that are the sample analog of the ones in equation (3.9). The GMM-QR estimator is defined as,

$$\widehat{\theta}^{M,L}_{GMM} = \underset{\theta}{\operatorname{argmin}} \left( \frac{1}{n} \sum_{i=1}^{n} \mathsf{m}_L(W_i; \mathsf{g}_L(\theta)) \right)^{\top} (\Sigma_L^{-1} \otimes \widehat{\Sigma}_M^{-1}) \left( \frac{1}{n} \sum_{i=1}^{n} \mathsf{m}_L(W_i; \mathsf{g}_L(\theta)) \right), \quad (3.11)$$

with

$$\widehat{\Sigma}_M = \frac{1}{n} \sum_{i=1}^{n} q_M(X_i) q_M(X_i)^{\top},$$

$\mathsf{m}_L(\cdot)$ given in (3.8), and $\Sigma_L^{-1}$ given in (3.10).

In addition, as we showed in Lemma 3.1, the weighting matrix in this GMM problem does not depend on unknown parameters and part of the weighting matrix does not need to be estimated.

By imposing two additional assumptions in our model, we can establish consistency and asymptotic normality of GMM-QR estimator. Consider the following conditions

**A6.** $\theta_0 \in \text{int}(\Theta)$ where $\Theta$ is a compact subset of $\mathbb{R}^{d_\theta}$.

**A7.** Let $h(X, \theta, U) = X^{\top} g(\theta, U)$. For all $\theta, \theta' \in \Theta$, $|h^{-1}(X, \theta, Y) - h^{-1}(X, \theta', Y)| \leq \kappa(X, Y) \|\theta - \theta'\|$, where $\mathrm{E}[\kappa(X, Y)^2] < \infty$.

Assumptions **A6**-**A7** are standard in the GMM literature. **A6** requires compactness of the parameter space. Condition **A7** imposes restriction on $h^{-1}(X, \theta, Y)$ requiring it to be Lipschitz.

Before we present the results on the limiting properties of the estimator, we define the several population quantities. Let

$$G_{lL}(X, \theta_0) := f_{Y|X} \left( X^{\top} g \left( \theta_0, \frac{l}{L} \right) \right) X^{\top} g_\theta \left( \theta_0, \frac{l}{L} \right)$$
$$G_L(X, \theta_0) := (G_{1L}(X, \theta_0)^{\top}, \ldots, G_{L-1,L}(X, \theta_0)^{\top}),$$

where $g_\theta \left( \theta_0, \frac{l}{L} \right)$ is a $K \times d_\theta$ matrix, $G_{lL}(X, \theta_0)$ is a $1 \times d_\theta$ vector and $G_L(X, \theta_0)$ is a $d_\theta \times (L-1)$ matrix.

Now we state the result for asymptotic normality of the GMM-QR estimator.

**Theorem 3.1.** *Let Assumptions* ***A1****–****A7*** *hold and let* $\frac{\partial}{\partial \theta^\top}\mathrm{E}[\mathsf{m}_L(W_i;\mathsf{g}_L(\theta)]$ *have full row rank when* $\theta = \theta_0$*. Then, as* $n \to \infty$

$$\widehat{\theta}^{M,L}_{GMM} \xrightarrow{p} \theta_0,$$

*and*

$$\sqrt{n}(\widehat{\theta}^{M,L}_{GMM} - \theta_0) \xrightarrow{d} N\left(0, V_L^{-1}\right).$$

*where* $V_L := (\mathrm{E}[G_L(X,\theta_0) \otimes q_M(X)^\top])(\Sigma_M \otimes \Sigma_L)^{-1}\mathrm{E}[G_L(X,\theta_0) \otimes q_M(X)^\top]^\top$.

The proof of Theorem 3.1 is established by verifying that the required conditions of Theorem 2.6 in Newey and McFadden (1994) for consistency of a M-estimator are satisfied and that the conditions for asymptotic normality of this class of estimators stated in Theorem 7.2 of Newey and McFadden (1994) are also satisfied.

An interesting implication of Theorem 3.1 is the convergence of the estimator for the parameters $\beta$ in restriction (2.4).

**Corollary 3.1.** *Let* $\widehat{\beta}(\tau) = g(\widehat{\theta}, \tau)$*. Under conditions of Theorem 3.1, as* $n \to \infty$

$$\sqrt{n}(\widehat{\beta}(\tau) - \beta(\tau)) \xrightarrow{d} N\left(0, g_\theta(\theta_0,\tau)^\top V_L^{-1} g_\theta(\theta_0,\tau)\right).$$

The standard approach used to estimate linear QR model can be viewed as estimating a moment restriction for one quantile at the time. In this case, as we mentioned before, we need the dimension of $q_M(X)$ to be at least equal to the dimension of the vector of unknown parameters, $\theta$. Note that estimating equation by equation (or quantile by quantile) is less efficient than estimate many moments at the same (multiple quantiles simultaneously). In the second case, we are always in the overidentified case and could use a weight matrix that minimizes the variance of the estimator. Estimating the model quantile by quantile is a particular case of using multiple quantile moments restrictions in the estimation. In order to get this particular case, in the GMM overidentified case we should use a specific weight matrix that put weight equals one in the moment restriction associated with the specific quantile and zero in the other restrictions. From the GMM theory, we know that this particular weight matrix does not give the GMM estimator with lowest variance among all the GMM that considers the set of full moments.

## 3.4 Optimal GMM-QR Estimator

In this section, we use the standard GMM theory (see, e.g., Hall (2005)) to derive the optimal GMM estimator based on the condition moment restrictions (3.1) for all the $L-1$ quantiles. We show that this estimator will depend on the estimation of the conditional density function of $Y$ given $X$ and it attains the efficient bound.

We can define the vector of conditional moment restrictions at the true value $\theta_0$ as,

$$\mathrm{E}\left[\mathsf{H}_L(W;\theta_0)|X\right]=0.$$

This vector of conditional moments conditions implies a vector of unconditional moments conditions:

$$\mathrm{E}\left[q_M^*(X,L)\mathsf{H}_L(W,\theta_0)\right]=0,$$

where $q_M^*(X,L)$ is a $d_\theta\times(L-1)$ matrix of functions of $X$ that minimizes the asymptotic variance of the GMM estimator. We write $q_M^*(X,L)$ instead of $q_M^*(X,\tau)$ to emphasize the use of the partition of the quantile space.

Using the optimal GMM theory, $q_M^*(X,L)$ is given by:

$$\begin{aligned} q_M^*(X,L) &= \frac{\partial}{\partial\theta^\top}\mathrm{E}\left[\mathsf{H}_L(W;\theta)|X\right]|_{\theta=\theta_0}\cdot\mathrm{E}\left[\mathsf{H}_L(X,Y,\theta_0)\cdot\mathsf{H}_L(X,Y,\theta_0)^\top\Big|X\right]^{-1}, \\ &= G_L(X,\theta_0)\cdot\Sigma_L^{-1}, \end{aligned} \tag{3.12}$$

since $\mathrm{E}\left[\mathsf{H}_L(X,Y,\theta_0)\cdot\mathsf{H}_L(X,Y,\theta_0)^\top\Big|X\right]=\Sigma_L$, and $\frac{\partial}{\partial\theta^\top}\mathrm{E}\left[\mathsf{H}_L(W;\theta)|X\right]|_{\theta=\theta_0}=G_L\left(X,\theta_0\right)$.

The first part of equation (3.12), together with the standard theory on GMM, imply that a feasible optimal GMM would involve estimating the conditional density and evaluating it at the estimated conditional quantile. The estimate of the $\tau$-th conditional quantile would be $X^\top g\left(\widehat{\theta},\tau\right)$, where $\widehat{\theta}$ would be, for instance, the consistent GMM-QR estimator proposed above.

From the optimal GMM theory, an optimal estimator reaches the efficiency bound for the optimal GMM, which is given by the inverse of the following variance-covariance matrix[5]

$$V_L^*=\mathrm{E}[G_L(X,\theta_0)\Sigma_L^{-1}G_L(X,\theta_0)^\top]. \tag{3.13}$$

The GMM-QR estimator in (3.11) does not use the optimal weighting matrix, hence, as

[5]See Supplemental Appendix A.4 for a brief discussion of the optimal GMM estimator.

the result in Theorem 3.1 shows it is not efficient, $V_L$ does not reach the lower bound in (3.13). However, the GMM-QR estimator has the advantage that it does not need estimates of the density function as required by the optimal GMM. Thus, practical implementation of GMM-QR is simpler.

Now we formally compare the variance of the GMM-QR, $V_L$, given in Theorem 3.1 with the variance of the optimal GMM, $V_L^*$, given in equation (3.13). The next result shows that, for a fixed number of partitions $L$, the variance of the optimal GMM is smaller than that of the GMM-QR.

**Lemma 3.2.** *Under Assumptions **A1**–**A7**, we have that for a fixed $L$, $V_L^* \geq V_L$, in the positive semidefinite sense.*

# 4 Many Moments

In this section, we develop methods for a large number of quantiles, that is, partitions $L$, as well as variables in the conditioning set, $M$. First, we investigate the properties of the GMM-QR estimator defined in the previous section when $L$ diverges to infinity. In order to establish the efficient bound for this problem, we briefly write the maximum likelihood estimator (MLE), and show its relationship with the GMM-QR. Then, we propose an alternative smooth GMM estimator for a large number of moments restrictions. The asymptotic variance matrix of the smooth GMM-QR coincides with the asymptotic variance of the MLE.

## 4.1 GMM Objective Function with Large Number of Quantiles

We start by investigating the properties of the GMM population objective function $Q_L(\theta)$ in (3.9) when $L$ diverges to infinity and the partitioning set becomes dense. The dimension of the vector of conditioning variables or instruments ($q_M(X,\cdot)$) is kept constant. In other words, $L$ increases, and $M$ is fixed.

**Lemma 4.1.** *Assume **A1**-**A5** and that*

$$\begin{aligned}
&\lim_{\tau\downarrow 0} \frac{1}{\tau} \cdot \mathrm{E}\left[m(W; g(\theta,\tau),\tau)^\top\right] \Sigma_M^{-1}\ \mathrm{E}\left[m(W; g(\theta,\tau),\tau)\right] = \\
&\lim_{\tau\uparrow 1} \frac{1}{1-\tau} \cdot \mathrm{E}\left[m(W; g(\theta,\tau),\tau)^\top\right] \Sigma_M^{-1}\ \mathrm{E}\left[m(W; g(\theta,\tau),\tau)\right] \\
&= 0.
\end{aligned}$$

*Then, for fixed $\theta$,*

$$\lim_{L\to\infty} Q_L(\theta) = Q(\theta),$$

*where*

$$Q(\theta) = \int_0^1 \left( \mathrm{E}\left[ \left( \frac{f_{Y|X}\left(X^\top g(\theta;\tau);\theta_0\right)}{f_{Y|X}\left(X^\top g(\theta;\tau);\theta\right)} \right) q_M(X)^\top \right] \Sigma_M^{-1}\ \mathrm{E}\left[ q_M(X) \left( \frac{f_{Y|X}\left(X^\top g(\theta;\tau);\theta_0\right)}{f_{Y|X}\left(X^\top g(\theta;\tau);\theta\right)} \right) \right] \right) d\tau$$
$$- \mathrm{E}\left[q_M(X)^\top\right] \Sigma_M^{-1}\ \mathrm{E}\left[q_M(X)\right],$$

*which is uniquely minimized at $\theta_0$ and $Q(\theta_0) = 0$.*

The conditions in Lemma 4.1 requiring the lower and upper limits of the population objective function being negligible as $\tau \to 0$ and $\tau \to 1$, respectively, rule out the boundary cases. These conditions relate with the standard requirement in the QR literature that restricts analysis only on the open interval of $\tau \in (0,1)$. We can interpret these assumptions in Lemma 4.1 as identification conditions, since they exclude possible lack of identification at the boundaries.

Lemma 4.1 shows that when the number of partitions diverge to infinity, $Q_L(\theta)$ converges to population function that is minimized at the true parameter value, $\theta_0$. This is important to guarantee that the objective function has a unique minimum at the true parameter when the number of partitions become dense as long as the contribution of the objective function evaluated at quantile values at the boundary is close to zero. This result guarantees identification of the parameters of interest when $L$ diverges to infinity and the partitioning set becomes dense. In addition, Lemma 4.1 provides an interesting form of the population objective function when the number of partitions is large. To fix the ideas consider the case where $q_M(X) = X = 1$, i.e., the case with the marginal distribution of $Y$ and not with the conditional distribution of $Y$ on $X$. In this particular case, called unconditional case, from Lemma 4.1, we can write the objective function as:

$$\begin{aligned} Q(\theta) &= \int_0^1 \left( \frac{f_Y\left(g(\theta;\tau);\theta_0\right)}{f_Y\left(g(\theta;\tau);\theta\right)} \right)^2 d\tau - 1 \\ &= \mathrm{E}\left[ \frac{f_Y(Y;\theta_0)}{f_Y(Y;\theta)} \right] - 1, \end{aligned} \tag{4.1}$$

which is a Kullback–Leibler divergence criterion.[6]

The result of Lemma 4.1 can also be derived by an application of results in (Parzen, 1970, p. 30), (Carrasco and Florens, 2000, p. 816) and (Sacks and Ylvisaker, 1968, p. 86) on Reproducing Kernel Hilbert Spaces.[7]

## 4.2 Efficiency Bound

Now we derive the efficiency bound from the calculation of the asymptotic variance-covariance matrix of the MLE. By definition, the MLE maximizes a recentered version of the average log-likelihood function,

$$\begin{aligned}\widehat{\theta}_{MLE} &= \arg\max_{\theta} \frac{1}{n}\sum_{i=1}^{n} \ln\left(f_{Y|X}\left(Y_i, X_i; \theta\right)\right) \\ &= \arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n} \ln\left(\frac{f_{Y|X}\left(Y_i, X_i; \theta_0\right)}{f_{Y|X}\left(Y_i, X_i; \theta\right)}\right). \end{aligned} \tag{4.2}$$

In order to derive the properties of the MLE, we use again the representation $Y = X^{\top} g(\theta, U)$ where $U \sim \text{Unif}[0,1]$ independently of $X$. We start by deriving the log-likelihood function. From $h\left(X, \theta, U\right) = X^{\top} g(\theta, U)$ in assumption **A7**, we have that

$$\begin{aligned} P(Y \leq y | X = x; \theta) &= P(x^{\top} g(\theta, U) \leq y | X = x; \theta) \\ &= P(h\left(x, \theta, U\right) \leq y | X = x; \theta). \end{aligned}$$

Assuming that $h$ has a well-defined inverse function,

$$P(Y \leq y | X = x; \theta) = P(U \leq h^{-1}\left(x, \theta, y\right)),$$

[6]Recall that the classical Kullback–Leibler criterion between the densities $f_{Y_i}\left(\cdot; \theta_0\right)$ and $f_{Y_i}\left(\cdot; \theta\right)$ is

$$\int \ln\left(\frac{f_{Y_i}\left(Y_i; \theta_0\right)}{f_{Y_i}\left(Y_i; \theta\right)}\right) f_{Y_i}\left(Y_i; \theta_0\right) dy \approx \int \left(\frac{f_{Y_i}\left(Y_i; \theta_0\right)}{f_{Y_i}\left(Y_i; \theta\right)} - 1\right) f_{Y_i}\left(Y_i; \theta_0\right) dy = Q\left(\theta\right)$$

and the approximation is the one from the log of the ratio to the rate of change.

[7]See Supplemental Appendix A.7 for a proof of this result.

and the conditional density of $Y$ given $X$ at $\theta$ is

$$\begin{aligned} f_{Y|X}(y|x;\theta) &= \frac{\partial}{\partial y} h^{-1}(x,\theta,y) \\ &= \frac{1}{x^\top g_u(\theta, h^{-1}(x,\theta,y))}, \end{aligned} \tag{4.3}$$

where $g_u(x,u) = \frac{\partial}{\partial u} g(x,u) \in \mathbb{R}^K$. Using this derivation, the log-likelihood at a given $\theta$ is

$$l_{y|X}(y|x;\theta) = \sum_{i=1}^{n} -\ln x^\top g_u(\theta, h^{-1}(x,\theta,y)).$$

We can also derive the score function $S_{Y|X}(y|x;\theta) \in \mathbb{R}^{d_\theta}$ as[8]

$$S_{Y|X}(y|x;\theta)^\top = -\left( \frac{x^\top g_{u\theta}(\theta, h^{-1}(x,\theta,y))}{x^\top g_u(\theta, h^{-1}(x,\theta,y))} - \frac{x^\top g_{uu}(\theta, h^{-1}(x,\theta,y))}{x^\top g_u(\theta, h^{-1}(x,\theta,y))} \frac{x^\top g_\theta(\theta, h^{-1}(x,\theta,y))}{x^\top g_u(\theta, h^{-1}(x,\theta,y))} \right).$$

Define $S(\cdot)$ as the score evaluated at the true parameter $\theta_0$ as

$$S(X_i, Y_i, \theta_0) := S_{Y|X}(Y_i|X_i;\theta_0).$$

Using the fact that $h(X_i,\theta,U_i) = X_i^\top g(\theta,U_i)$ in Assumption **A7**, we obtain the following expression for $S(X_i,U_i,\theta_0)$,

$$S(X_i,U_i,\theta_0) = \left( \frac{X_i^\top g_{uu}(\theta_0,U_i)}{X_i^\top g_u(\theta_0,U_i)} \frac{X_i^\top g_\theta(\theta_0,U_i)}{X_i^\top g_u(\theta_0,U_i)} - \frac{X_i^\top g_{u\theta}(\theta_0,U_i)}{X_i^\top g_u(\theta_0,U_i)} \right)^\top.$$

Based on the score evaluated at the true parameter, we are able to find the efficiency bound for this problem based on the following information matrix,

$$\mathcal{I}(\theta_0) = \mathrm{E}[S(X,U,\theta_0) S(X,U,\theta_0)^\top], \tag{4.4}$$

where $\mathcal{I}(\theta_0)$ is the information matrix. The asymptotic variance of the MLE is the inverse of the information matrix, and the inverse of the information matrix represents the efficiency bound for this quantile problem.

[8] See Supplemental Appendix A.8 for details of the derivation.

## 4.3 Efficiency of GMM-QR with Large Number of Quantiles

Now we study the properties of the GMM estimators discussed in the previous section when $L$ is large. First, we show that, as $L$ diverges to infinity, the GMM-QR estimator does not reach the efficiency bound. Second, we show that the variance of the optimal GMM estimator reaches the lower bound.

The following result relates the asymptotic variance of the GMM-QR estimator described in Section 3.3 to the lower bound derived in Section 4.2

**Lemma 4.2.** *Under Assumptions **A1**–**A7**, and*

$$\lim_{\tau\downarrow 0}\frac{1}{\tau}\cdot \mathrm{E}\left[f_{Y|X}\left(X^{\top}g\left(\theta_0,\tau\right)\right)g_{\theta}^{\top}\left(\theta_0,\tau\right)X\otimes q_M(X)^{\top}\Sigma_M^{-1/2}\right]\mathrm{E}\left[f_{Y|X}\left(X^{\top}g\left(\theta_0,\tau\right)\right)X^{\top}g_{\theta}\left(\theta_0,\tau\right)\otimes\Sigma_M^{-1/2}q_M(X)\right]=$$
$$\lim_{\tau\uparrow 1}\frac{1}{1-\tau}\cdot \mathrm{E}\left[f_{Y|X}\left(X^{\top}g\left(\theta_0,\tau\right)\right)g_{\theta}^{\top}\left(\theta_0,\tau\right)X\otimes q_M(X)^{\top}\Sigma_M^{-1/2}\right]\cdot\mathrm{E}\left[f_{Y|X}\left(X^{\top}g\left(\theta_0,\tau\right)\right)X^{\top}g_{\theta}\left(\theta_0,\tau\right)\otimes\Sigma_M^{-1/2}q_M(X)\right]$$
$$=0.$$

*As $L\to\infty$, we have that, $\lim_{L\to\infty}V_L\leq\mathcal{I}(\theta_0)$, in the positive semidefinite sense.*

Now we establish a result on the variance of the optimal GMM estimator discussed in Section 3.4. The result in Lemma 3.2 shows that, for a fixed number of partitions $L$, the variance of the GMM-QR estimator is larger than the efficient bound. In the next result we show that, when $L$ diverges to infinity, the variance of the optimal GMM estimator reaches the efficiency bound.

**Lemma 4.3.** *Under assumptions **A1**–**A7**, and*

$$\lim_{\tau\downarrow 0}\frac{1}{\tau}\cdot \mathrm{E}\left[\left[f_{Y|X}\left(X^{\top}g\left(\theta_0,\tau\right)\right)g_{\theta}^{\top}\left(\theta_0,\tau\right)X\right]\cdot\left[X^{\top}f_{Y|X}\left(X^{\top}g\left(\theta_0,\tau\right)\right)g_{\theta}\left(\theta_0,\tau\right)\right]\right]=$$
$$\lim_{\tau\uparrow 1}\frac{1}{1-\tau}\cdot \mathrm{E}\left[\left[f_{Y|X}\left(X^{\top}g\left(\theta_0,\tau\right)\right)g_{\theta}^{\top}\left(\theta_0,\tau\right)X\right]\cdot\left[X^{\top}f_{Y|X}\left(X^{\top}g\left(\theta_0,\tau\right)\right)g_{\theta}\left(\theta_0,\tau\right)\right]\right]$$
$$=0.$$

*As $L\to\infty$, we have that*

$$\lim_{L\to\infty}V_L^{*}=\mathcal{I}(\theta_0),$$

*where $\mathcal{I}(\theta_0)$ is the Fisher-information matrix.*

Lemma 4.3 shows that if we use the optimal GMM estimator considering the conditional moment on $X$, when $L$ goes to infinity, the asymptotic variance of the GMM estimator attains the information matrix. Note that this optimal GMM estimator is based on the

original moments, writing the moment function as a function of $Y_i$, and not $U_i$. In the next subsection, we investigate the asymptotic properties of an alternative smooth GMM-QR estimator, which is a smooth GMM for QR models for a different set of moments restrictions that is defined in the objective function in (4.5) below.

## 4.4 Smooth GMM-QR Estimator

In this section, we generalize the estimators developed in the paper for fixed $L$ and $M$, and suggest an efficient, smooth GMM-QR estimator that considers a large number of moments restrictions. This alternative estimator is based on the approach proposed by Poirier (2017) to obtain a different GMM estimator based on smooth basis functions that cover all the information embedded in the identifying restriction. In this case, we modify the moment restrictions used to compute our estimator.

When $Y = X^\top g(\theta_0, U) \equiv h(X, \theta_0, U)$ is strictly increasing in $U$, we can equivalent write $U = h^{-1}(X, \theta_0, Y)$ such that $U$ is independent of $X$. This independence restriction between the unobservable $U$ and $X$ and the normalization that $U$ follows a uniform distribution on $[0,1]$, $U \sim \text{Unif}[0,1]$, will form the basis of the estimator below. Independence of $U$ and $X$ is equivalent to $\text{E}\left[(r(U) - \text{E}[r(U)])l(X)\right] = 0$ holding for any functions $r(\cdot)$ and $l(\cdot)$. Let $q_{mM}(x)$ be basis functions for $x$ and let $p_{lL}(u)$ be mean-zero basis functions of $u$. For example, consider polynomials, splines, or other sets of functions which can continuously approximate continuous functions. Given that $U$ is distributed Unif$[0,1]$, any basis function $\tilde{p}_{lL}(u)$ can be demeaned by taking $p_{lL}(u) = \tilde{p}_{lL}(u) - \int_0^1 \tilde{p}_{lL}(v)dv$. Also let $p_L(u) = [p_{1L}(u), ..., p_{L-1,L}(u)]$ and $q_M(x) = [q_{1M}(x), ..., q_{MM}(x)]$. Note that in this new estimator, $L$ and $M$ indicates the smooth degree of the basis functions of $U$ and $X$, so when $L$ and $M$ go to infinity, all the information embedded in the independence restriction is used to form the estimator.

Using these smooth basis functions for $u$ and $x$, we use the vector of moments $\text{E}\left[\mathsf{m}_{L,M}(W_i;\theta)\right]$ in order to define a new smooth GMM-QR (SGMM-QR) estimator,

$$\widehat{\theta}^{M,L}_{SGMM} = \underset{\theta\in\Theta}{\text{argmin}} \left(\frac{1}{n}\sum_{i=1}^{n} \mathsf{m}_{L,M}(W_i;\theta)\right)^\top \left(\widehat{\Sigma}_L^{-1} \otimes \widehat{\Sigma}_M^{-1}\right) \left(\frac{1}{n}\sum_{i=1}^{n} \mathsf{m}_{L,M}(W_i;\theta)\right), \qquad (4.5)$$

where $\mathsf{m}_L(W;\theta) = p_L(h^{-1}(X,\theta,Y)) \otimes q_M(X)$, $\widehat{\Sigma}_M^{-1}$ is as defined in model (3.11), and $\widehat{\Sigma}_L = \frac{1}{n}\sum_{i=1}^n p_L(h^{-1}(X_i,\widetilde{\theta},Y_i))p_L(h^{-1}(X_i,\widetilde{\theta},Y_i))^\top$, with $\widetilde{\theta}$ in the last equation being a first stage

root-$n$ consistent estimator for $\theta$. One example of $\widetilde{\theta}$ is the GMM-estimator for fixed $M$ and $L$ proposed in the previous sections.

This is the estimator in Poirier (2017) with one difference, the demeaned basis functions $p_{lL}(u)$. This implies that the estimating equations correspond to moment equations instead of covariance equations as in Poirier (2017). Under his assumptions, we can show the same result as in Poirier (2017), except to a smaller asymptotic variance matrix. This smaller asymptotic variance is due to the fact that we know the baseline distribution of $h^{-1}(X, \theta, Y)$ under $\theta = \theta_0$, which is uniform, instead of being unrestricted. The next example provides an illustration of this model.

**Example 4.1.** Let $Y = \beta_0(U) + X\beta_1(U)$ where $\beta_0(\tau) = \theta_1 + \theta_2 \log \frac{\tau}{1-\tau}$, $\beta_1(\tau) = \theta_3 + \theta_4 \log \frac{\tau}{1-\tau}$ for $\tau \in (0, 1)$. For a given parameter value $\tilde{\theta} = (\tilde{\theta}_1, \tilde{\theta}_2, \tilde{\theta}_3, \tilde{\theta}_4)$, we can write unobservable $U$ as

$$\begin{aligned} U &= h^{-1}(X, \tilde{\theta}, Y) \\ &= \exp\left(\frac{Y - \tilde{\theta}_1 - \tilde{\theta}_3 X}{\tilde{\theta}_2 + \tilde{\theta}_4 X}\right) / \left(1 + \exp\left(\frac{Y - \tilde{\theta}_1 - \tilde{\theta}_3 X}{\tilde{\theta}_2 + \tilde{\theta}_4 X}\right)\right). \end{aligned}$$

Given that $\theta$ is point identified in this model under mild additional assumptions, we obtain that $h^{-1}(X, \tilde{\theta}, Y)|X \sim \text{Unif}[0, 1]$ is equivalent to $\tilde{\theta} = \theta$ and, moreover, all the model restrictions are embedded in this independence restriction.

The estimator we propose below considers the approximate independence restriction

$$\mathrm{E}\left[\left\{p_{lL}\left[\exp\left(\frac{Y - \tilde{\theta}_1 - \tilde{\theta}_3 X}{\tilde{\theta}_2 + \tilde{\theta}_4 X}\right) / \left(1 + \exp\left(\frac{Y - \tilde{\theta}_1 - \tilde{\theta}_3 X}{\tilde{\theta}_2 + \tilde{\theta}_4 X}\right)\right)\right] - \int_0^1 p_{lL}(v)dv\right\} q_{mM}(X)\right] = 0$$

for $l = 1, \ldots, L$, and $m = 1, \ldots, M$. Note the form of this approximate independence restriction is a moment restriction. Letting $L, M \to \infty$, this approximate independence restriction becomes exact, assuming some regularity conditions about basis functions we make formal below.

Here are some additional assumptions to allow to use of Theorem 3.4 in Poirier (2017):

**B1.** (Basis function) $p_{lL}(\cdot)$ and $q_{mM}(\cdot)$ are bounded and continuously differentiable everywhere;

**B2.** (P-Donsker Classes) $\{q_{mM}(X_i) : m \in \mathbb{N}\}$, $\{p_{lL}(h^{-1}(X,\theta,Y)) : l \in \mathbb{N}, \theta \in \Theta\}$ and $\{p'_{lL}(h^{-1}(X,\theta,Y))\frac{\partial}{\partial\theta}h^{-1}(X,\theta,Y) : l \in \mathbb{N}, \theta \in \Theta\}$ are uniformly bounded P-Donsker classes;

**B3.** (Approximation) For any bounded continuous function $f(z,u,\theta)$ there exists $\beta^{ML}(\theta)$, a $ML \times 1$ vector such that $\sup_{x\in\mathcal{X},u\in\mathcal{U},\theta\in\Theta} |f(x,u,\theta) - (p_L(u) \otimes q_M(x))\beta^{ML}(\theta)| \to 0$ as $L$ and $M$ goes to infinity;

**B4.** (Eigenvalues) The minimum eigenvalues of $\mathrm{E}[p_L(h^{-1}(X,\theta,Y))p_L(h^{-1}(X,\theta,Y))^\top]$ and $\mathrm{E}[q_M(X)q_M(X)^\top]$ are bounded above and bounded below by the function $C/\zeta(L)$ and $C/\zeta(M)$ uniformly in $\theta$, where $\zeta(L)$ is a known function with $\zeta(L) \to \infty$ as $L \to \infty$, and $C > 0$ is a constant;

**B5.** (Differentiability) $h^{-1}(X,\theta,Y)$ is twice continuously differentiable in $\theta$ a.s. - $X$.

Examples of basis functions that satisfy these assumptions include polynomials, or cubic splines: see Propositions 3.2 and 3.3 in Poirier (2017). Under some conditions, for cubic splines the corresponding function is $\zeta(m) = m$, while for power series it is $\zeta(m) = m^2$. See Donald et al. (2003). We require these functions to be smooth approximating basis functions and therefore indicators are ruled out by these assumptions.

Under these assumptions and an additional condition regarding the rate of growth of the number of basis functions, we can show that the GMM estimator in (4.5) is asymptotically efficient attaining the efficient bound in equation (4.4).

**Theorem 4.1** (Consistency and Asymptotic Normality)**.** *Let Assumptions **A1**-**A7** and **B1**-**B5** hold, and let $\widetilde{\theta}$ be a preliminary first-step estimator of $\theta_0$ that satisfies $\|\widetilde{\theta} - \theta_0\| = O_p(c_n)$. Then, if $M^2\zeta(M)^2L^2\zeta(L)^2\left(c_n + \frac{1}{\sqrt{n}}\right) \to 0$ and $M, L \to \infty$ as $n \to \infty$, then*

$$\widehat{\theta}^{M,L}_{SGMM} \xrightarrow{p} \theta_0,$$

*and*

$$\sqrt{n}(\widehat{\theta}^{M,L}_{SGMM} - \theta_0) \xrightarrow{d} N\left(0, V^*\right),$$

*where $V^* = \mathcal{I}(\theta_0)^{-1}$ is the efficiency bound for this problem.*

The first-step, consistent estimator can be for example $\widehat{\theta}^{M,L}_{GMM}$ which converges to $\theta_0$ at the rate $\tau_n = n^{-1/2}$.

In an Online Supplemental Appendix, we compare the finite sample performances of the proposed GMM-QR, the optimal GMM-QR, the SGMM-QR, and the standard QR estimator which is estimated quantile by quantile using a Monte Carlo exercise.

## 4.5 Selecting the Number of Moments

For all the estimators proposed in this paper, one must select the number of moments, which is either $L-1$ or $(L-1)\times M$. In some cases this number must converge to infinity subject to rate constraints, but the finite sample choice of these numbers can be done using a number of different approaches. Nagar (1959) proposes a method for obtaining the finite sample MSE of estimators using higher-order expansions of the objective function. Donald and Newey (2001) and Donald et al. (2009) use this method to propose a selection criterion for the number of moments in models with instruments, also see Donald et al. (2008). An interesting alternative would be the use of cross-validation, which has the advantage of being fully data-driven. Recent work by Komiyama and Shimao (2018) explores this possibility. In the Monte Carlo exercises, in the Supplemental Appendix, we explore the leave-many cross-validation similar to the one suggested by Komiyama and Shimao (2018) to choose $L$ for a fixed value of $M$. For a given sample, we randomly split it into $J$ partitions. We denote these partitions by $\{S_j\}_{j=1}^{J}$. Let $\widehat{\theta}(L)$ denote an estimator computed using $L-1$ vector of moments. Denote $\widehat{\theta}_{-S_j}(L)$ the estimator computing using all the sample except the partition $S_j$, $j\in\{1,2\dots J\}$. We choose the optimal $L$ that minimizes the mean squared error (MSE):

$$\sum_{j=1}^{J}\sum_{i\in S_j}\int_0^1\left(y_i-x_i^\top g\left(\tau;\widehat{\theta}_{-S_j}(L)\right)\right)^2 d\tau. \tag{4.6}$$

Alternatively, we can also choose the $L$ by minimizing the $\ell_1$ loss function,

$$\sum_{j=1}^{J}\sum_{i\in S_j}\int_0^1\rho_\tau\left(y_i-x_i^\top g\left(\tau;\widehat{\theta}_{-S_j}(L,M)\right)\right)d\tau, \tag{4.7}$$

where $\rho_\tau(u)=(\tau-1\{u<0\})\cdot u$ is the check function.

For sample size equals to $1,000$ and $1,000$ simulations, the results of the Monte Carlo in the Supplemental Appendix show that using both criteria choose $L$ that delivers estimators with low values for the Root Mean Square Error (RSME) and good coverage. So, the leave-

many cross-validation seems a reasonable criterion to choose $L$ for a fixed value of $M$.

# 5 Application

In this section we illustrate the practical use of the proposed GMM quantile regression (GMM-QR) methods by estimating the effects of various covariates on birthweight of live infants at the extreme bottom of the conditional distribution.[9] For comparison, we also report results for the extremal quantile of Chernozhukov and Fernández-Val (2011). The failure of the Gaussian laws for extremal quantile has been extensively documented, and more accurate approximations for extreme quantiles have been developed (see, e.g., Chernozhukov and Fernández-Val (2011)). The methods we propose in this paper constitute an alternative to extremal quantile, since standard asymptotic inference is valid for the proposed GMM-QR estimators, even at the tails.

Recently birthweight has been shown to be the foremost telltale of infant health. Unhealthy births have large economic costs in both immediate medical costs and longer care costs. Infants are classified as low birthweight (LBW) when weighing less than 2.5 kilograms at birth. There exists empirical evidence showing that the direct medical costs of LBW are very high. Almond et al. (2005) document that the hospital costs for newborns are elevated: the expected cost of delivery and initial care of a baby weighing one kilogram at birth can exceed $100,000 (in year 2000 dollars). The costs remain elevated even among babies weighing 2–2.1 kilograms; an additional pound (454 grams) of weight is still associated with a $10,000 difference in hospital charges for inpatient services.[10] The infant mortality rate also increases at lower birthweights.

We replicate Chernozhukov and Fernández-Val (2011) empirical application using the June 1997 Detailed Natality Data published by the National Center for Health Statistics. We select a sample of 31,912 children born in the United States to black mothers, aged 18 to 45 years old. We focus on extremely low birthweight quantiles, considering percentiles within the subset (0, 0.025). The dependent variable is the birthweight (in kilograms). The set of covariates includes a dummy variable that takes the value one if the mother smoked during

[9]Previous quantile estimation approaches to estimating birthweight outcome regressions include, among others, Abrevaya (2001), Koenker and Hallock (2001), and Chernozhukov and Fernández-Val (2011).

[10]Expenditures, such as radiological, pharmaceutical, respiratory, and laboratory fees, greatly extend the costs of intensive care for LBW infants (see, e.g., Behrman et al. (2007)).

pregnancy (Smoker), and another variable to account for the average cigarettes smoked per day (Cigarette's / Day). In addition, two dummy covariates that capture whether the mother was married (Married) and whether the child is male (Boy), and a dummy that takes the value one for mothers with no prenatal visits are included (No Prenatal).

Since we are interested on inference at the extreme tail, we compute the GMM-QR estimator over the subset $(0, \tau^*)$ (GMM-Standard-$\tau^*$), where we set $\tau^*$ equal to 0.03. The estimator uses moment conditions for $\tau \in (0, \tau^*)$.[11] We choose $L$ (number of partitions in the GMM-QR) using the leave-many-out cross-validation, where we randomly split our dataset in *two* halves. We use both MSE and check-function-based criteria and consider values for $L$ between 20 and 100. We report the results from the MSE criterion. The estimated model can be motivated by a simple location-scale model as $y = g(\theta;\tau)'X = X'\beta + (X'\gamma)e$ where $X'\gamma$ is almost surely nonnegative. We assume that $e$ is independent from $X$ and follows a Weibull distribution.[12]

The empirical results are displayed in Figure 2, where we plot the GMM-QR estimates of the conditional quantile partial effects (CQPE) and their corresponding 90% confidence intervals for the selected covariates, over the interval (0,0.025). We also report estimates from standard QR point estimates together with its corresponding extremal inference procedure of Chernozhukov and Fernández-Val (2011).

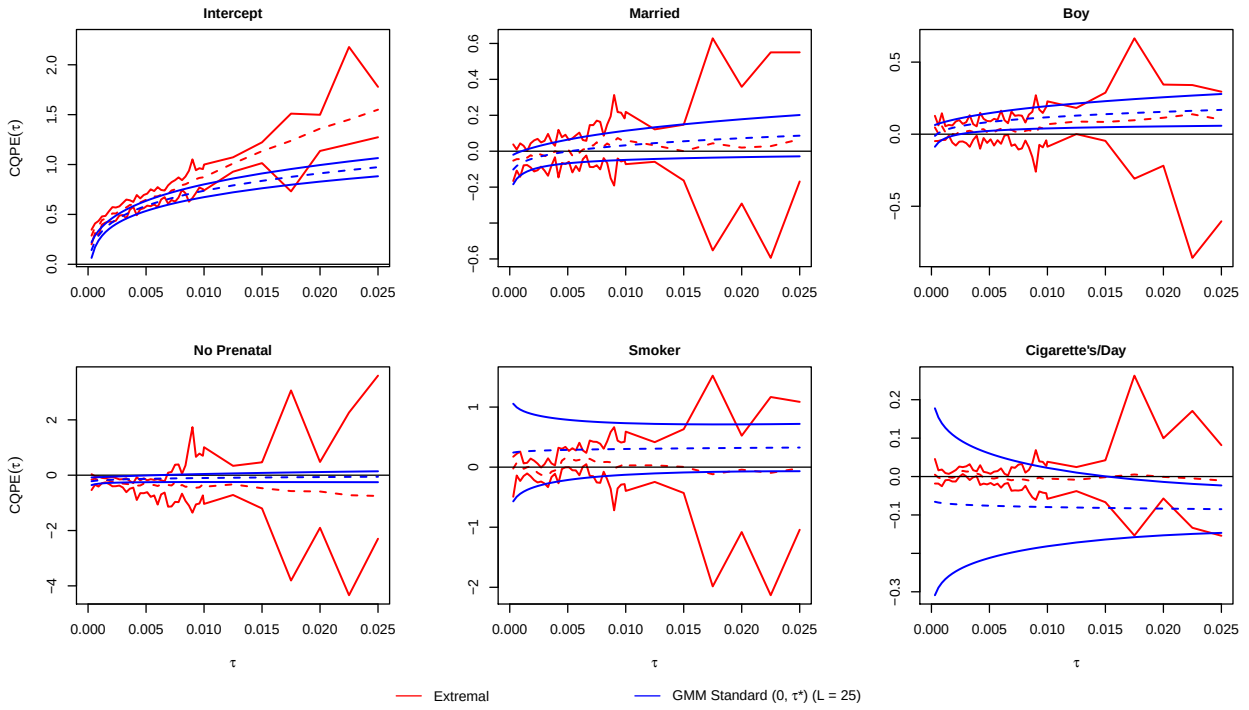
Figure 2 shows that the estimates of CQPE are relatively flat along the bottom tail of the distribution for all considered covariates. Also, confidence intervals from the GMM-QR are narrower than those of Chernozhukov and Fernández-Val (2011) except for variables "Smoker" and "Cigarette's/Day" at the very bottom extreme ($\tau < 0.015$). Neither the extremal nor the GMM-QR procedures captures significant effects of "Smoker", but the latter finds a statistically significant and negative effect of "Cigarette's/Day" for $\tau \in (0.015, 0.025)$ while the former does not. According to both estimators the effect of "Married" and "No prenatal visits" are not significant at the 90% level, while the dummy "Boy" is significant and positive for the GMM-Standard but it is not for the extremal inference.

Another advantage of the GMM-QR that can be observed in Figure 2 is that it is smooth

[11] The weighting matrix $\Sigma_L^{-1}$ is not exactly the same as when estimating the entire (0,1) interval for $\tau$, but it is still deterministic.

[12] The Weibull specification is given by $g(\theta;\tau)^\top X = \left[\theta_1 + \theta_2\left(-\ln(1-\tau)^{\frac{1}{k}}\right)\right]^\top X$, where $\theta$ and $k$ are unknown parameters to be estimated. We also considered other specifications like Normal, Logit or linear.

Figure 2: GMM-QR estimates of CQPE at the bottom tail of live infants' birthweights




(by construction), while confidence intervals from the extremal quantile approach are not, and thus they must be smoothed after estimation.

Summarizing, given the difficulty to perform inference at the extremes of the distribution, the GMM-QR estimates capture important statistically *significant* effects of some of the considered covariates at the bottom tail and the signs of the effects are in line with intuition: positive effects of "Boy" and negative effects of "Cigarette's /Day". These results allow us to conclude that there are efficiency gains in our approaches relative to other procedures that estimate the CQPE separately for each $\tau$, as we formally demonstrate in this paper.

# 6 Conclusion

In this paper, we develop generalized method of moments (GMM) estimation and inference procedures for QR models, allowing for parametric restrictions across the quantile. First, we propose an estimator that calculates simultaneously all the quantile effects for a fixed number of quantiles. We show that this estimator is $\sqrt{n}$-consistent, but it does not attain the efficient

bound. Using a large number of moments and a large basis functions for the explanatory variables ($X$), we propose a smooth GMM estimator that attains the efficient bound. Our methods can be applied to many examples, including survival analysis and structural models. In addition, this method is very useful in the estimation and inference of extreme quantiles. Using our method, the researcher can estimate a model, imposing restrictions across the extreme quantiles. The advantage is that the estimator is $\sqrt{n}$-consistent if only part of the quantile process is estimated (like only the extreme quantiles). Also, our method allows to estimate all the quantiles together, imposing different restrictions in different parts of the quantile process.

Monte Carlo simulations, in the Online Supplemental Appendix, show numerical evidence of the finite sample properties of the methods. Finally, we apply the proposed methods to estimate the effects of various covariates on birthweight of live infants at the extreme bottom of the conditional distribution.

# References


Abrevaya, J. (2001). The effects of demographics and maternal behavior on the distribution of birth outcomes. *Empirical Economics 26*, 247–257.

Almond, D., K. Y. Chay, and D. S. Lee (2005). The costs of low birth weight. *Quarterly Journal of Economics 120*, 1031–1083.

Behrman, R. E., A. S. Butler, and Committee on Understanding Premature Birth and Assuring Healthy Outcomes (2007). *Preterm Birth: Causes, Consequences, and Prevention*. Washington, D.C.: The National Academies Press.

Bradic, J., J. Fan, and W. Wang (2011). Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *Journal of the Royal Statistical Society, Series B 73*, 325–349.

Buchinsky, M. (1998). Recent advances in quantile regression models: A practical guideline for empirical research. *Journal of Human Resources 33*, 88–126.

Carrasco, M. and J.-P. Florens (2000). Generalization of gmm to a continuum of moment conditions. *Econometric Theory 16*(6), 797–834.

Carrasco, M. and J. P. Florens (2014). On the asymptotic efficiency of gmm. *Econometric Theory 30*, 372–406.

Chen, L.-Y. and S. Lee (2017). Exact computation of gmm estimators for instrumental variable quantile regression models. mimeo.

Chen, X., V. Chernozhukov, S. Lee, and W. K. Newey (2014). Local identification of nonparametric and semiparametric models. *Econometrica 82*, 785–809.

Chen, X. and Z. Liao (2015). Sieve semiparametric two-step gmm under weak dependence. *Journal of Econometrics 189*, 163–186.

Chen, X. and D. Pouzo (2009). Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *Journal of Econometrics 152*, 46–60.

Chen, X. and D. Pouzo (2012). Estimation of nonparametric conditional moment models with possibly nonsmooth moments. *Econometrica 80*, 277–322.

Chen, X., A. T. K. Wan, and Y. Zhou (2015). Efficient quantile regression analysis with missing observations. *Journal of the American Statistical Association 110*, 723–741.

Chernozhukov, V. (2005). Extremal quantile regression. *The Annals of Statistics 33*(2), 806–839.

Chernozhukov, V. and I. Fernández-Val (2011). Inference for extremal conditional quantile models, with an application to market and birthweight risks. *Review of Economic Studies 78*, 559–589.

Chernozhukov, V., I. Fernandez-Val, and T. Kaji (2017). Extremal quantile regression: An overview. In R. Koenker, V. Chernozhukov, X. He, and L. Peng (Eds.), *Handbook of Quantile Regression*. Chapman and Hall.

Chernozhukov, V. and H. Hong (2003). An mcmc approach to classical estimation. *Journal of Econometrics 115*, 293–346.

Chiang, H. D. and Y. Sasaki (2017). Causal inference by quantile regression kink designs. Mimeo.

de Castro, L., A. F. Galvao, D. Kaplan, and X. Li (2018). Smoothed gmm for quantile models. *Journal of Econometrics*, forthcoming.

Donald, S. G., G. W. Imbens, and W. K. Newey (2003). Empirical likelihood estimation and consistent tests with conditional moment restrictions. *Journal of Econometrics 117*(1), 55–93.

Donald, S. G., G. W. Imbens, and W. K. Newey (2008). Choosing the number of moments in conditional moment restriction models. Working Paper.

Donald, S. G., G. W. Imbens, and W. K. Newey (2009). Choosing instrumental variables in conditional moment restriction models. *Journal of Econometrics 152*(1), 28–36.

Donald, S. G. and W. K. Newey (2001). Choosing the number of instruments. *Econometrica 69*(5), 1161–1191.

Donald, S. G. and H. J. Paarsch (1993). Piecewise pseudo-maximum likelihood estimation in empirical models of auctions. *International Economic Review 34*(1), 121–148.

Firpo, S. and C. Pinto (2015). Identification and estimation of distributional impacts of interventions using changes in inequality measures. *Journal of Applied Econometrics 31*, 457–486.

Galvao, A. F., G. Montes-Rojas, and J. Olmo (2011). Threshold quantile autoregressive models. *Journal of Time Series Analysis 32*, 253–267.

Hall, A. R. (2005). *Generalized Method of Moments*. Advanced Texts in Econometrics Series. Oxford University Press.

Horowitz, J. L. (1998). Bootstrap methods for median regression models. *Econometrica 66*(6), 1327–1351.

Kaplan, D. M. and Y. Sun (2017). Smoothed estimating equations for instrumental variables quantile regression. *Econometric Theory 33*, 105–157–325.

Koenker, R. (1984). A note on l-estimates for linear models. *Statistics and Probability Letters 2*, 323–325.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis 91*, 74–89.

Koenker, R. (2005). *Quantile Regression*. New York, New York: Cambridge University Press.

Koenker, R. and G. Bassett (1978). Regression quantiles. *Econometrica 46*(1), 33–50.

Koenker, R. and O. Geling (2001). Reappraising medfly longevity: A quantile regression survival analysis. *Journal of the American Statistical Association 96*(454), 458–468.

Koenker, R. and K. Hallock (2001). Quantile regression. *Journal of Economic Perspectives 15*, 143–156.

Komiyama, J. and H. Shimao (2018). Cross validation based model selection via generalized method of moments. arXiv preprint arXiv:1807.06993.

Nagar, A. L. (1959). The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica 27*, 575–595.

Newey, W. K. (1988). Adaptive estimation of regression models via moment restrictions. *Journal of Econometrics 38*(3), 301–339.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of applied econometrics 5*(2), 99–135.

Newey, W. K. (1993). Efficient estimation of models with conditional moment restrictions. In C. R. R. G. S. Maddala and H. D. Vinod (Eds.), *Handbook of Statistics, Vol. 11*. Elsevier.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. In R. F. Engle and D. L. McFadden (Eds.), *Handbook of Econometrics vol. 4, chap. 36*. Elsevier.

Parzen, E. (1970). *Statistical Inference on Time Series by Rkhs Methods*. Defense Technical Information Center.

Poirier, A. (2017). Efficient estimation in models with independence restrictions. *Journal of Econometrics 196*, 1–22.

Portnoy, S. and R. Koenker (1989). Adaptive l-estimation of linear models. *Annals of Statistics 17*, 362–381.

Qu, Z. and J. Yoon (2017). Uniform inference on quantile effects under sharp regression discontinuity designs. *Journal of Business and Economic Statistics forthcoming*.

Sacks, J. and D. Ylvisaker (1968). Designs for regression problems with correlated errors: Many parameters. *Ann. Math. Statist. 39*(1), 49–69.

Xu, G., T. Sit, L. Wang, and C.-Y. Huang (2017). Estimation and inference of quantile regression for survival data under biased sampling. *Journal of the American Statistical Association 112*, 1571–1586.

Yang, Y. and X. He (2012). Bayesian empirical likelihood for quantile regression. *The Annals of Statistics 40*(2), 1102–1131.

Zhang, Y. (2016). Extremal quantile treatment effect. mimeo.

Zhao, Z. and Z. Xiao (2014). Efficient regressions via optimally combining quantile information. *Econometric Theory 30*, 1272–1314.

Zou, H. and M. Yuan (2008). Composite quantile regression and the oracle model selection theory. *Annals of Statistics 36*, 1108–1126.