Finite Sample Valid Probabilistic Inference on Quantile Regression

Leonardo Cella CELLAL@WFU.EDU
Department of Statistical Sciences, Wake Forest University, USA

Abstract

In most applications, uncertainty quantification in quantile regression take the form of set estimates for the regression coefficients. However, often a more informative type of uncertainty quantification is desired where other inference-related tasks can be performed, such as the assignment of (imprecise) probabilities to assertions of interest about (any feature of) the regression coefficients. Validity of these probabilities, in the sense that their values are well-calibrated in a frequentist sense, is fundamental to the trustworthiness of the drawn conclusions. This paper presents a nonparametric Inferential Model (IM) construction that offers provably valid probabilistic uncertainty quantification in quantile regression, even in finite sample settings. It is also shown that this IM can be used to derive finite sample confidence regions for (any feature of) the regression coefficients. As a result, regardless of the type of uncertainty quantification desired, the proposed IM offers an appealing solution to quantile regression problems.

Keywords: quantile regression, inferential models, possibility measure, nonparametric, finite sample, validity, model-free, confidence region

1. Introduction

Regression analysis is a crucial statistical method in data science that uses the values of explanatory variables $X \in \mathbb{R}^p$, where $p \geq 1$, to explain or predict the mean values of a quantitative response variable Y. While the primary focus of regression is usually on estimating the conditional mean of Y given X, there may be instances where it is more appropriate to estimate a conditional quantile of Y given X. For example, if Y has outliers, modeling the median as a measure of central tendency, rather than the mean, may be a more effective approach. More generally, in situations where the conditions of linear regression are not met, quantile regression can be a useful alternative, as it does not require any assumptions about the distribution of the target variable.

Fix a probability $\tau \in (0, 1)$ and let $Q_X(\tau)$ denote the τ^{th} conditional quantile of Y, given $X = x$. Then the quantile regression model says

$$Q_x(\tau) = x^\top \theta, \tag{1}$$

where $\theta = \theta_\tau \in \mathbb{R}^p$ is the vector of regression coefficients of interest. Quantification of uncertainty about the value of θ is desired. Moreover, because quantiles are quantities whose existence does not depend on assumed parametric models for the distribution of Y given $X = x$, nonparametric/distribution-free solutions are often preferred in order to avoid model misspecification bias.

The quantification of uncertainty about θ is often carried out by producing a suitable family of set estimates that are at least approximate confidence regions, i.e., set estimates that have coverage probability guarantees at least in an asymptotic sense. There are multiple techniques for generating these regions, including bootstrapping, covariance matrix estimation of quantile estimates, and rank-test inversion methods. A comprehensive overview of these and other methods can be found in Koenker (2005), Chapter 3. In Chernozhukov et al. (2009), a similar approach to the rank-test method is proposed, but with improved coverage guarantees in finite samples.

In some application, there may be interest in a more complete type of uncertainty quantification, one that allows for assignments of (imprecise) probabilities to all relevant assertions of interest about θ. The Bayesian approach is, of course, probabilistic, but its application to quantile regression faces the challenge of having to circumvent the specification of a parametric likelihood. Several authors have attempted different working likelihoods for Bayesian or "Bayesian-like" quantile regression, e.g., asymmetric Laplace distributions (Yu and Moyeed, 2001; Yang et al., 2016), empirical likelihood (Yang and He, 2012), Dirichlet process mixture models (Kottas and Krnjajić, 2009), an infinite mixture of normals (Reich et al., 2009), exponentiated empirical risk (Martin and Syring, 2022), among others. But before committing to one these approaches for probabilistic inference on quantile regression, where the posterior distributions obtained from the working likelihoods above can be used to derive posterior probabilities for assertions about θ, it is crucial to determine the validity of these posterior probabilities. This means that it is important to verify whether they are accurately calibrated in a frequentist sense.

The same way validity of set estimates is always desirable, so that erroneous conclusions are controllably rare, so is, or should be, validity of probability assignments. When a

posterior distribution is used to evaluate the probability of a certain statement about θ, the size of this probability will be used to draw inference about whether the statement is true or false. To reduce the likelihood of making incorrect conclusions, it is desirable to control the rate at which the posterior distribution assigns small probabilities to true statements and large probabilities to false statements. This topic is further discussed in Section 2.

Unfortunately, the aforementioned Bayesian or "Bayesian-like" solutions to quantile regression do not examine the validity of probability assignments. Instead, they only verify if the posterior regions, derived from the posterior distribution, are approximate confidence regions. For example, Empirical Likelihood quantile regression has been shown to have asymptotically calibrated posterior regions for the components of θ (Yang and He, 2012), while Yang et al. (2016) acknowledge the lack of calibration of posterior regions based on other working likelihoods and recommend adjustments to achieve approximate confidence regions when using Laplace distributions. Similarly, when the exponentiated empirical risk function is adopted, Martin and Syring (2022) recommend adjustments for the same purpose.

However, the calibration of posterior regions for θ does not guarantee the validity of posterior probabilities for statements about θ. The *false confidence theorem* in Balch et al. (2019) states that probabilistic inferences based on additive probabilities are at risk of being invalid, indicating the need for new considerations if the goal is to make valid probabilistic inferences in quantile regression.

Inferential Models (IMs) is a relatively new probabilistic approach to statistical inference; see Martin and Liu (2013, 2015) for the first considerations, and Martin (2019) for a modern perspective. Contrary to the Bayesian approach, the IM's probability assignments are non-additive. This and the specific way that they are constructed make IMs dodge the false confidence bullet, being, to my knowledge, the only probabilistic approach with validity guarantees, even in finite sample settings. However, the original IMs construction requires specifying a parametric likelihood for the data, which can limit its application to quantile regression.

This issue was addressed in Cella and Martin (2022a) when a general IMs construction for model-free problems was proposed, which only requires a suitable loss function that defines the unknown quantity of interest. Koenker and Bassett (1978) show that the quantile regression coefficient θ is a risk-minimizer with respect to the loss function

$$\ell_\theta(x, y) = |y - x^\top\theta| - (2\tau - 1)x^\top\theta.$$

A nonparametric IM for quantile regression can be then obtained from it, as Cella and Martin do in Section 4.3 of their paper. The main limitation of this approach is that it only achieves validity in an asymptotic sense, due to its dependence on bootstrap samples. As mentioned above, finite sample validity is a unique feature of parametric IMs, so, here, I attempt to maintain this property in a new nonparametric IM construction for quantile regression. This is just a single application of a general framework that aims to deliver valid probabilistic inference for a broad range of model-free problems. The full details of the framework will be presented in Cella and Martin (2023).

More specifically, my goal in the present paper is to develop a nonparametric IM for finite sample valid probabilistic inference on quantile regression. The specific construction, presented in Section 3, is surprisingly simple, relying primarily on calculations involving the binomial distribution. Before that, a background on IMs is given in Section 2. I conclude in Section 4 with a brief summary and discussion of some open problems.

2. Background on IMs

To set the scene, let data $Z^n = (Z_1, \ldots, Z_n)$ take values in a general enough space $\mathbb{Z}^n$. That is, besides the case we are mainly interested here, where $Z_i = (X_i, Y_i)$ is a predictor and response variable pair, $\mathbb{Z}^n$ can take other forms, e.g., the real line, a matrix space, etc. Following the modus operandi in most applications, consider that a statistical model P^n_ω, indexed by a parameter $\omega \in \Omega$, is assumed for Z^n. Note that the unknown parameter ω completely specifies the distribution of Z^n, so ω is the inferential target. To put it another way, any quantity of interest $\theta \in \Theta$ related to the distribution of Z^n, such as quantiles or moments, are invariably a function of ω, i.e., $\theta = \theta(\omega)$. Inferences on θ are, therefore, obtained indirectly from inferences on ω.

As stated in Section 1, my focus here is on *probabilistic* inference, where (imprecise) probability assignments to all sorts of assertions about ω can be obtained. Moreover, I will consider the cases where no prior information about ω is assumed or available. The question then arises: can a probabilistic approach provide validity guarantees in this prior-free setting?

Several attempts to answer this question, that Bradley Efron calls the "most important unresolved problem in statistics" (Efron, 2013), have been made. Perhaps more popular are the ones whose output probability assignments are *additive*, e.g., Fisher's fiducial approach (Fisher, 1935), generalized fiducial (Hannig et al., 2016), confidence distributions (Xie and Singh, 2013; Schweder and Hjort, 2016), and Bayes with default priors (Berger, 2006). However, these approaches may be problematic in light of the *False confidence* theorem presented in Balch et al. (2019). This theorem is a significant result that highlights the potential invalidity of probabilistic inference based on additive probabilities, suggesting that the key to a definitive solution to Efron's most important problem may be *non-additivity*.

Despite the existence of various frameworks that output non-additive probabilities, e.g., Dempster–Shafer theory (Dempster, 2014, 1967, 1968, 2008; Shafer, 1976) and other belief function frameworks (Denœux and Li, 2018; Denœux, 2014), *Inferential Models* (Martin and Liu, 2015) is, to the best of my knowledge, the only framework that provides assurance of validity. In what follows, the notion of validity is made precise, and a brief overview of the IM's specific construction to achieve it is exposed.

Let $A \subseteq \Omega$ be an assertion of interest about ω. For the observed data $Z^n = z^n$, denote the IM's non-additive output for the claim "$\omega \in A$" by $\overline{\Pi}_{z^n}(A)$. Consider a scenario where a small $\overline{\Pi}_{z^n}(A)$ is encountered. This indicates that the assertion A is implausible, which may lead a data scientist to conclude that A is false. However, if $\overline{\Pi}_{Z^n}(A)$, as a function of data $Z^n \sim P^n_\omega$, tends to be small even when A is actually true, the data analyst risks making "systematic misleading conclusions" (Reid and Cox, 2015). The goal of the validity criteria is to mitigate this risk by ensuring that erroneous conclusions are controllably rare. More formally, a valid IM with output $\overline{\Pi}_{z^n}$ satisfies

$$\sup_{\omega \in A} P^n_\omega\{\overline{\Pi}_{Z^n}(A) \leq \alpha\} \leq \alpha, \quad \begin{cases} \text{for all } \alpha \in [0,1], \\ \text{for all } A \subseteq \Omega, \\ \text{for all } n. \end{cases} \tag{2}$$

The IM's upper probability $\overline{\Pi}_{z^n}$ takes the mathematical form of a possibility measure (Dubois and Prade, 1988). Interestingly, Martin (2021) argued that IMs of this form are the most efficient, so there is no need to consider non-additive measures with more general structures.

The possibilistic nature of the IM's output has two important implications. First, it implies the existence of a possibility contour π_{z^n}, which fully determines the possibility measure for any assertion of interest through

$$\overline{\Pi}_{z^n}(A) = \sup_{\omega \in A} \pi_{z^n}(\omega).$$

Set estimates for ω can be readily derived from the IM's plausibility contour. Its validity, i.e.,

$$P^n_\omega\{\pi_{Z^n}(\omega) \leq \alpha\} \leq \alpha, \quad \text{for all } \alpha \in [0,1],$$

guarantees the frequentist error rates control of these set estimates. In other words, set estimates obtained from the IM's plausibility contour are confidence regions. More details about this will be presented in Section 3. Second, possibility measures have a dual $\underline{\Pi}_{z^n}(A) = 1 - \overline{\Pi}_{z^n}(A^c)$, known as necessity measures. As a result, IMs can also be characterized in terms of them. More importantly, the "for all $A \subseteq \Omega$" clause in (2) makes possible an equivalent validity result for the IM's necessity measures:

$$\sup_{\omega \in A} P^n_\omega\{\underline{\Pi}_{Z^n}(A) \geq 1 - \alpha\} \leq \alpha, \quad \begin{cases} \text{for all } \alpha \in [0,1], \\ \text{for all } A \subseteq \Omega, \\ \text{for all } n. \end{cases}$$

In words, false assertions tend to be assigned relatively low necessity with respect to the posited statistical model. This prevents systematically misleading conclusions.

The duality between possibility and necessity measures, as well as their equivalent validity does not diminish the importance of either measure. In fact, considering both measures can help prevent the misuse of statistics in practical applications, especially those related to p-values. For further details, see Cella and Martin (2022b).

When the ultimate quantity of interest is $\theta = \theta(\omega)$, possibility measures for assertions about θ can be obtained via marginalization through the mapping $\omega \to \theta$:

$$\overline{\Pi}_{z^n}(\omega : \theta(\omega) \in A), \quad A \subseteq \Theta. \tag{3}$$

Once again, since (2) includes the "for all $A \subseteq \Omega$" clause, the IM's validity property carries over immediately to marginal inferences on θ. Of course, this marginal validity is contingent on correctly specifying the distribution P^n_ω. However, this assumption may not be reasonable depending on the nature of θ. For instance, if θ doesn't determine the data's distribution, it might be better to use nonparametric/distribution-free solutions to avoid biases caused by model misspecification.; see Section 3.

But how are IMs constructed? The first approach, presented in Martin and Liu (2013, 2015), introduces an auxiliary variable U^n with a known distribution and expresses the statistical model in terms of it:

$$Z^n = a(\omega, U^n), \quad U^n \sim \mathsf{P}^n_{known}.$$

While the value of U^n cannot be observed, knowing its distribution allows for educated predictions about its value. Martin and Liu suggest using appropriate random sets (e.g., Nguyen, 2006; Molchanov, 2005) to predict the value of U^n, and the properties of these sets are crucial to ensuring the validity of the possibility measures in (2). More recently, it has been recognized that IMs can also be constructed directly from possibility measures that quantify uncertainty about U^n (Liu and Martin, 2021). A more recent alternative construction skips the specification of auxiliary variables and is motivated by the *probability-to-possibility transform* presented in Hose and Hanss (2020, 2021); see Martin (2022a,b). Here, I will focus on this direct approach.

Let $h : (\mathbb{Z}^n \times \Omega) \to \mathbb{R}$ be a measurable function and define the possibility contour

$$\pi_{z^n}(\omega) = P^n_\omega\{h(Z^n, \omega) \leq h(z^n, \omega)\}, \quad \omega \in \Omega. \tag{4}$$

According to Martin (2022b), this simple construction yields the plausibility contour of a valid IM. However, the choice of h plays a crucial role, as it determines a partial ordering of candidate values for ω given z^n. Martin (2022b) refers to this ordering as the *plausibility order*.

To determine the best h for a given situation, where "best" is related to the efficiency of the resulting plausibility contour, Martin (2022b) appeal to what Hose calls the *Principle of Plausibility*, which suggests choosing the h that represents the plausibility order inherent in the assumed statistical model P^n_ω. Therefore, it is straightforward to take h to be the probability mass or density function of P^n_ω.

3. Nonparametric IMs for Quantile Regression

My focus on the present paper is on quantile regression, so data $Z^n = (Z_1, \ldots, Z_n)$ consist of n covariate/response pairs, i.e., $Z_i = (X_i, Y_i)$. The goal is to make inferences on the quantiles of the conditional distribution of Y given X. More formally, let $Q_x(\tau)$ in (1), $\tau \in (0,1)$, denote the τ^{th} conditional quantile of Y, given $X = x$. The vector of regression coefficients $\theta = \theta_\tau \in \mathbb{R}^p$ is, therefore, the inferential target.

The IM construction presented in Section 2 is powerful, but it has limitations that make it unsatisfactory for quantile regression applications. The primary limitation is that this approach requires the specification of a parametric statistical model for the data, namely, a distribution P^n_ω that is indexed by a parameter $\omega \in \Omega$ for Z^n. However, since the quantile regression parameters θ do not uniquely determine the distribution of the data, ω is not equivalent to θ. As a result, inferences on θ must be obtained indirectly through (3). This indirect approach to inferring θ can be problematic if P^n_ω is misspecified, as the IM validity property in (2) relies on the assumed model being accurate.

To avoid potential biases resulting from model misspecification, it is preferable to proceed without explicitly specifying a model. My assumption is that Z^n consists of independent and identically distributed (iid) components, with $Z^n \sim P^n$. Note that i) P is free to be any distribution, no constraints due to dependence on a parameter ω; and ii) the regression coefficients of the quantile regression $\theta = \theta(P)$ are a functional of the underlying distribution.

Is it possible to develop a nonparametric IM that can generate valid probabilistic inferences for θ even in finite sample settings? First, it is essential to understand how different the desired validity in a model-free context is from (2). Let $\overline{\Pi}_{z^n}(A)$ be the potential nonparametric IM possibilistic quantification of uncertainty about an assertion A of interest. While it may seem reasonable to infer A^c when $\overline{\Pi}_{z^n}(A)$ is small, this approach may not be trustworthy if $\overline{\Pi}_{Z^n}(A)$ tends to be small for $Z^n \sim P^n$ when $\theta \in A$. As discussed in Section 2, the purpose of validity is to make these erroneous conclusions controllably rare. Hence, the desired validity property can be defined as follows:

$$\sup_{P:\theta(P)\in A} P^n\{\overline{\Pi}_{Z^n}(A) \leq \alpha\} \leq \alpha, \quad \begin{cases} \text{for all } \alpha \in [0,1], \\ \text{for all } A \subseteq \Theta, \\ \text{for all } n. \end{cases} \tag{5}$$

It is worth noting that a nonparametric IM construction for quantile regression has been proposed in Cella and Martin (2022a). However, this approach relies on bootstrap samples, and as a result, it can only achieve the validity property in (5) as n approaches infinity. The objective here is to develop a method that can achieve validity for any sample size.

As discussed in Section 2, the IM construction for parametric problems relies on a crucial step, namely the specification of a real-valued function h that establishes a plausibility order for potential values of the model parameters based on the observed data. By applying the probability calculation in (4), a valid IM plausibility contour can be obtained. In essence, the proposed nonparametric IM construction for quantile regression will follow the same logic. Notably, this approach is just one specific instance of a general framework for finite sample valid probabilistic inference in distribution-free problems that will be outlined in Cella and Martin (2023).

Let $h : (\mathbb{Z}^n \times \Theta) \to \mathbb{R}$ be a measurable function that provides a plausibility order for candidate values of the regression coefficients θ given the observed data z^n. As we'll see below, it will be the case that $h(Z^n, \theta)$ is discrete. Moreover, consider that its distribution, as a function of Z^n, is known and independent of unknown quantities. Then

$$\pi_{z^n}(\theta) = P^n\{h(Z^n, \theta) \leq h(z^n, \theta)\}, \quad \theta \in \Theta, \tag{6}$$

is the plausibility contour of a finite sample valid and nonparametric IM for θ.

Theorem 1 *The nonparametric IM for θ defined above, with contour given by (6), is valid in the sense of (5).*

Proof Fix P and let $\theta = \theta(P)$. Also, let G denote the (known) distribution of $h(Z^n, \theta)$. Note that $\pi_{z^n}(\theta)$ in (6) can be written as $\pi_{z^n}(\theta) = G(h(z^n, \theta))$. Therefore, it follows immediately that $\pi_{Z^n}(\theta)$, as a function of $Z^n \sim P^n$, is stochastically no smaller than $\mathsf{Unif}(0,1)$, i.e.,

$$P^n\{\pi_{Z^n}(\theta) \leq \alpha\} = P^n\{G(h(Z^n, \theta)) \leq \alpha\} \leq \alpha. \tag{7}$$

Now, for any assertion A that contains θ, the monotonicity of plausibility measures imply that $\overline{\Pi}_{z^n}(A) \geq \pi_{z^n}(\theta)$. Therefore,

$$P^n\{\overline{\Pi}_{Z^n}(A) \leq \alpha\} \leq \alpha.$$

Taking a supremum over all P such that $\theta(P) \in A$ on the left-hand side above completes the proof. ■

Besides being the key to the IM possibilistic quantification of uncertainty about general assertions of interest about θ, the plausibility contour in (6) can also be used to generate set estimates for θ that have coverage probability guarantees in finite sample settings. As the following corollary shows, if $C_\alpha(z^n)$ denote the α level sets of the possibility contour, i.e.,

$$C_\alpha(z^n) = \{\theta \in \Theta : \pi_{z^n}(\theta) > \alpha\}, \quad \alpha \in [0,1], \tag{8}$$

then the IM validity imply that $C_\alpha(z^n)$ is a finite sample $100(1-\alpha)\%$ confidence region for θ.

Corollary 2 *The α level sets of the possibility contour in (8) are finite sample $100(1-\alpha)\%$ confidence regions in the sense that*

$$\sup_P P^n\{C_\alpha(Z^n) \not\ni \theta(P)\} \leq \alpha, \quad \alpha \in [0,1].$$

Proof Fix P and let $\theta = \theta(P)$. Observe that $C_\alpha(Z^n) \not\ni \theta$ if and only if $\pi_{Z^n}(\theta) \leq \alpha$. Then the claim follows immediately from (7). ■

In certain applications, there may be particular features of the regression coefficients θ that are of primary interest. Denote them by $\phi = \phi(\theta)$. For example, in a single covariate case, such features might include:

- $\phi = \theta_1$, which represents the slope of the quantile regression line;
- $\phi = \theta_0 + \theta_1 \mathrm{x}$, which represents the quantile of Y given $X = x$;
- $\phi = \frac{q-\theta_0}{\theta_1}$, which represents the value of X that yields a quantile equal to q.

A marginal IM for any feature ϕ can be obtained from that for θ. That is, define the possibility contour

$$\pi^\phi_{z^n}(\varphi) = \sup_{\vartheta:\phi(\vartheta)=\varphi} \pi_{z^n}(\vartheta), \quad \varphi \in \phi(\Theta).$$

Note that, as a direct consequence of (7), $\pi^\phi_{Z^n}(\varphi)$, as a function of $Z^n \sim P^n$, is also stochastically no smaller than $\mathsf{Unif}(0,1)$. Therefore, finite sample valid possibility measures can be assigned to assertions about ϕ as it was done before for θ:

$$\overline{\Pi}^\phi_{z^n}(A) = \sup_{\varphi \in A} \pi^\phi_{z^n}(\varphi), \quad A \subseteq \phi(\Theta).$$

Moreover, the set

$$\{\varphi \in \phi(\Theta) : \pi^\phi_{z^n}(\varphi) > \alpha\} \tag{9}$$

constitutes a finite sample $100(1-\alpha)\%$ confidence region for ϕ.

The above results suggest that the proposed nonparametric IM is a powerful alternative to quantile regression, regardless of the type of uncertainty quantification that is desired. More specifically, the proposed IM can be utilized for both probabilistic inference on (features of) the regression coefficients θ and for the provision of set estimates for (features of) θ. Both types of uncertainty quantification are accurately calibrated in a frequentist sense, irrespective of the sample size.

But the question of how to obtain a suitable h, a fundamental element in the IM construction, remains to be addressed. In Cella and Martin (2023), a general strategy suitable for various model-free problems will be provided. For quantile regression, this strategy involves identifying a function $\gamma = \gamma(Z^n, \theta)$ of the data Z^n and the regression coefficients θ that is a pivot, i.e., that has a distribution that is independent of θ or any other unknowns. Once a suitable pivot has been identified, the plausibility order h can be selected based on its probability mass.

Perhaps the most intuitive pivot in quantile regression is

$$\gamma = \sum_{i=1}^n I_{(0,\infty)}(Y_i - x_i^\top \theta), \tag{10}$$

where $I_B(A)$ is the indicator that even A belongs to B. In words, (10) is the sum of the indicators of the sign of $Y_i - x_i^\top \theta$. Given the independence of the Y's given $X = x$ and the assumption that $x^\top \theta$ is the true τ-th quantile, (10) follows a $\mathsf{Bin}(n, 1-\tau)$ distribution. One can then use the probability mass of this binomial as the plausibility order h, i.e.,

$$h = \binom{n}{\gamma}(1-\tau)^\gamma \tau^{n-\gamma}, \tag{11}$$

and derive an IM plausibility contour from (6). Even though Theorem 1 guarantees the validity of this solution, I'll argue next that it can be very inefficient, so new considerations are needed. To build intuition, it will be helpful to consider, separately, the case where the levels of the covariates are fixed, in the sense that there are replications of Y given $X = x$, and the case where at least one of the covariates is continuous, so there is no replication of Y for any $X = x$.

3.1. Fixed Covariates

Let's assume that the levels of the covariates are fixed, in the sense that there are replications of Y in the levels of X. For example, there may be interest in studying the effect of k specific doses $x_1, x_2, \ldots, x_k$ of a certain medication X in some response variable Y. For a given x_i, $i = 1, \ldots, k$, n_i instances of Y are observed, i.e., $Y_{i1}, \ldots, Y_{in_i}$. In this setting, (10) can be rewritten as

$$\sum_{i=1}^k \sum_{j=1}^{n_i} I_{(0,\infty)}(Y_{ij} - x_i^\top \theta).$$

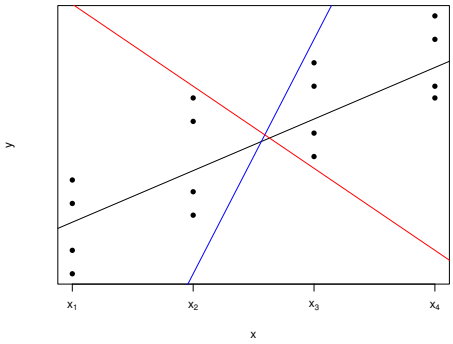


Figure 1: Simple illustration for the median regression when the levels of X are fixed.

Note that this is the sum of k independent binomial distributions with size n_i and success probability $1-\tau$. Therefore, it still follows a $\mathsf{Bin}(n, 1-\tau)$ distribution, where $\sum_{i=1}^{k} n_i = n$.

Figure 1 illustrates this scenario with $k = 4$ and $n_1 = n_2 = n_3 = n_4 = 4$. For simplicity but without loss of generality, suppose that interest is in the median regression, i.e., $\tau = 0.5$. The lines in the graph aid in understanding why using h as in (11) is not optimal. This choice results in a plausibility order where any line that splits the data in half — with half of the points above and half below the line — is equally maximally plausible. As a result, such a h cannot distinguish between the three lines in Figure 1, even though it is clear that they are not equally desirable. This inefficiency motivates the consideration of an alternative approach.

Towards this, I suggest considering each one of the k independent binomials separately, and using the product of their probability masses as the plausibility order h. In other words, my suggestion consists of taking h to be the likelihood function of the independent binomially distributed pivots that arise at each level of X:

$$h = \prod_{i=1}^{k} \binom{n_i}{\gamma_i} (1-\tau)^{\gamma_i} \tau^{n_i - \gamma_i}, \tag{12}$$

where

$$\gamma_i = \sum_{j=1}^{n_i} I_{(0,\infty)}(Y_j - x_i\theta), \quad i = 1, \ldots, k.$$

It is worth noting that, unlike (11), (12) yields different plausibilities for the lines in Figure 1. Specifically, the black line, which is the most desirable option, maximizes (12), while the blue and red lines minimize it.

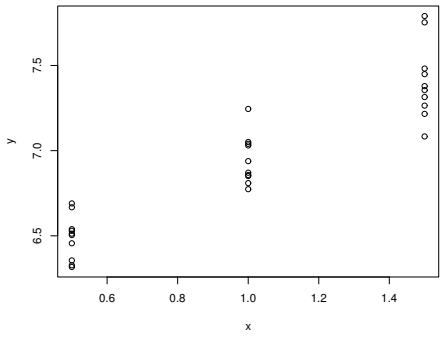
As an illustration, let $X = (0.5, 1, 1.5)$ and $n_1 = n_2 = n_3 = 10$. Let $Y_i = \mu(X_i) + \epsilon(X_i)$, where $\mu(x) = 6 + x$, and $\epsilon(x) \sim \mathsf{N}\big(0, (0.1 + 0.1x)^2\big)$. Figure 2(a) displays one simulated data set. Suppose the interest is in the median regression line, so $\tau = 0.5$ and $\theta = (6, 1)$. Figure 2(b) shows the empirical distribution function of $\pi_{z^n}(\theta)$, with h as in (12), in a simulation study where the above scenario is repeated 1000 times and $\pi_{z^n}(\theta)$ is calculated as in (6), in each replication. Note that (7) and, therefore, validity is verified. The same simulation is repeated for $\tau = 0.25$ and $\tau = 0.75$, showing that the proposed IM's validity is not specific to the median regression. Figure 2(c) shows, in red, the 95% confidence region for θ obtained from (8) with h as in (12), for the data in Figure 2(a). The 95% confidence region in black is obtained from using h as in (11), confirming the lack of efficiency that arises from such choice.

3.2. Continuous Covariates

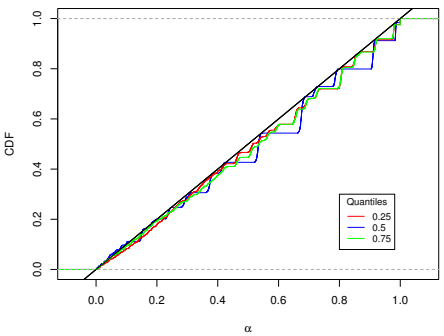
Let us now consider the scenario where at least one of the covariates is continuous, so that there is no replication of Y for any given $X = x$. The problem with choosing h as in (11), discussed in Section 3.1 above, still persists in this case. Once again, for simplicity, but without loss of generality, suppose that interest is in the median regression. Here, any line that divides the data in half — with half of the points above and half below the line — such as the three lines in Figure 3, would be equally good. The alternative plausibility order in (12), shown to be effective when the levels of X are fixed, does not solve the problem here; it actually exacerbates it. Specifically, for any candidate θ, (12) is equal to 0.5^n, as $n_i = 1$ for all i. As a result, not only are the three lines in Figure 3 equally good, but any line, including those that do not divide the data in half.

It appears that an alternative solution is necessary when the levels of X are not fixed. Luckily, this alternative solution need not be entirely distinct from the one in Section 3.1. This is because, while we do not have independent replications of Y for each $X = x$, we do have independent replications of Y in neighborhoods of X. If k neighborhoods of X are formed, the idea is to consider each one of the independent binomials that arise in each of the k neighborhoods separately, and to use the product of their probability masses as the plausibility order h. For example, in the case of one continuous covariate, we can consider a set of increasing real numbers $l_1, l_2, \ldots, l_{k-1}$ and consider the n_i replications of Y given that $l_{i-1} < X < l_i$, $i = 1, \ldots, k$, where l_0 and l_k are $-\infty$ and $+\infty$, respectively. Then

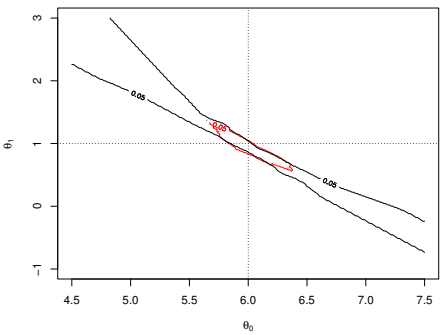
$$\gamma_i = \sum_{j=1}^{n_i} I_{(0,\infty)}(Y_j - x_j\theta), \quad \begin{cases} l_{i-1} < x_j < l_i, \\ i = 1, \ldots, k, \end{cases} \tag{13}$$



(*a*)



(*b*)



(*c*)

Figure 2: Panel (a): Data. Panel (b): Empirical CDF of the plausibility contour evaluated at the true quantile regression line based on 1000 Monte Carlo sample. Panel (c): 95% confidence regions for θ, obtained from (8).

are independent and follow a $\mathsf{Bin}(n_i, 1-\tau)$ distribution. The plausibility order in (12) can then be used. Consider $k = 2$

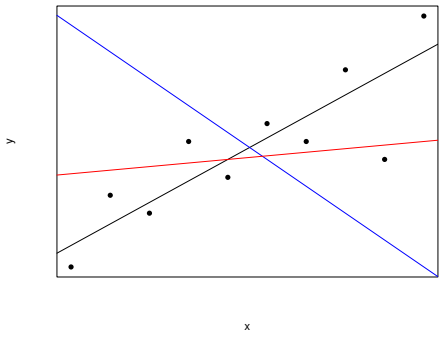


Figure 3: Simple illustration for the median regression when X is continuous.

and l_1 to be the median of the observed X's in Figure 3. In this case, γ_1 and γ_2 are iid $\mathsf{Bin}(5, 0.5)$. Note how, based on (12), the black, red and blue candidate median regression lines in Figure 3 have, as desired, a descending plausibility order.

When dealing with multiple explanatory variables, i.e., $X \in \mathbb{R}^p$ with $p \geq 2$, there are various possible ways to construct neighborhoods around X. Perhaps an effective one is to use unsupervised clustering algorithms like *K-means* (Celebi and Aydin, 2016). These techniques can efficiently group data points into clusters, making them a natural choice for creating neighborhoods in higher dimensional spaces. Additional details and examples of these approaches in high-dimensional settings will be presented elsewhere.

As an illustration, let $X_i \overset{\text{iid}}{\sim} \mathsf{Unif}(0, 5)$, $i = 1, \ldots, n$, with $n = 30$, and let $Y_i = \mu(X_i) + \epsilon(X_i)$, where $\mu(x) = 6 + 0.1x$, and $\epsilon(x) \sim \mathsf{N}\big(0, (0.1 + 0.1x)^2\big)$. Figure 4(a) displays one simulated data set. Interest here is in $\theta = \theta_\tau$ for $\tau = 0.3$ which, in this case, is roughly equal to $(5.95, 0.05)$. In (13), I will consider $k = 2$ and l_1 equal to the median of X and then use h as in (12) for the IM construction. To check validity of the resulting plausibility contour, I simulate 1000 data sets according the above scheme and calculated $\pi_{z^n}(\theta)$ as in (6) in each replication. Figure 4(b) shows the empirical distribution of these values. It is clear that it is stochastically no smaller than the uniform distribution, confirming Theorem 1. The same simulation is repeated for $\tau = 0.6$ and $\tau = 0.9$. Validity is verified in all scenarios. For further illustration, I also extracted 95% interval estimates for the components of θ through (9) from these 1000 data sets. This type of uncertainty quantification is popular in quantile regression applications. The goal was to compare the estimated coverage probabilities and mean length of the IM intervals with those obtained from the *quantreg* and

θ	IM	Rank	Bayes
θ_0	0.99 (1.11)	0.88 (0.43)	0.96 (0.44)
θ_1	0.98 (0.48)	0.83 (0.19)	0.88 (0.18)

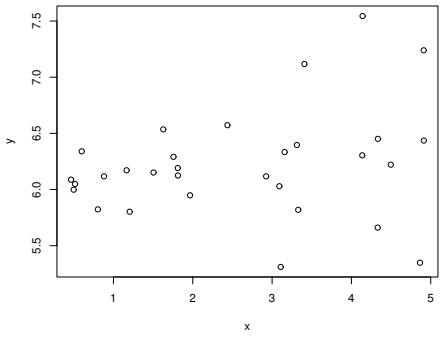
Table 1: Estimated coverage probabilities and mean length of 95% interval estimates for the quantile regression coefficients based on the proposed IM, the method based on the inversion of rank tests and the Bayesian-like method that uses the asymmetric Laplace distribution as the working likelihood.

bayesQR packages in R (Koenker et al., 2022; Benoit and Van den Poel, 2017), which are based on, respectively, the inversion of rank tests (Gutenbrunner et al., 1993; Hušková, 1994; Koenker, 2005) and the use of asymmetric Laplace distribution as the working likelihood for a "Bayesian-like" solution (Yu and Moyeed, 2001; Yang et al., 2016). The results are summarized in Table 1. As expected, the IM intervals are finite sample confidence intervals. However, the intervals obtained from the other two approaches are not always wide enough to achieve the correct coverage probability for $n = 30$. Finally, Figure 4(c) shows, in red, the 95% confidence regions for θ, obtained from (8), for the single data set displayed in Figure 4(a). The 95% confidence region in black is obtained from using h as in (11), confirming, once again, the lack of efficiency that arises from such choice.

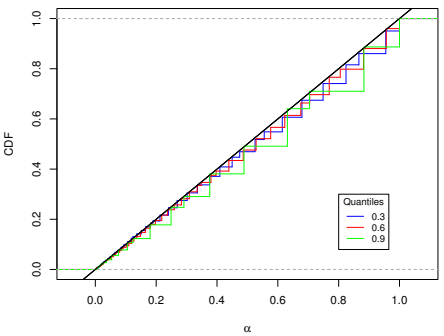
4. Conclusion

This paper introduces a new nonparametric IM construction for probabilistic inference on quantile regression. It is demonstrated that this approach is valid, in the sense of Theorem 1. Specifically, the IM's possibility assignments to all assertions about (features of) the quantile regression coefficients are calibrated in a statistical sense. Importantly, this calibration is not just asymptotic, but holds for any sample size. Additionally, this achieved validity does not exclude, but rather complements, the often-desired calibration of set estimates. This means that the proposed nonparametric IM can also be used to obtain finite sample confidence regions for (features of) the quantile regression coefficients.

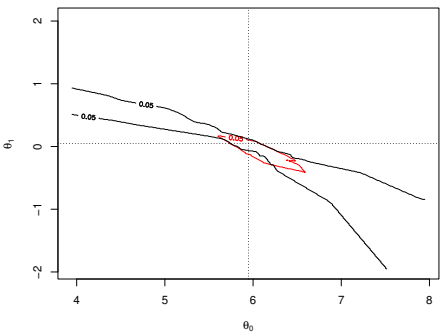
It's worth mentioning that the framework presented in this paper, which focuses on quantile regression, is actually more versatile than just this one context. The approach of identifying a function of data and inferential target that acts as a pivot and using its likelihood to establish the plausibility order in the IM construction can be applied to a wide range of relevant model-free problems, thereby producing finite sample valid probabilistic inferences for these problems. The full details of this general framework will be presented in Cella and Martin (2023).



(*a*)



(*b*)



(*c*)

Figure 4: Panel (a): Data. Panel (b): Empirical CDF of the plausibility contour evaluated at the true quantile regression line based on 1000 Monte Carlo samples. Panel (c): 95% confidence regions for θ, obtained from (8).

This section concludes with a brief discussion of some open questions. Firstly, while I consider my choice of pivot

to be very intuitive, it's worth noting that there may be other possible choices that could potentially be more effective. An interesting follow-up project could investigate these other options, comparing them in terms of efficiency, and even exploring the possibility of an optimal solution. On the topic of efficiency, a second open question is how to best select the neighborhoods of X that replicate Y in (13). Specifically, does the number of neighborhoods and/or the number of replications per neighborhood impact the efficiency of the IM? Thirdly, a more careful investigation is needed into the method of neighborhood formation when dealing with multiple explanatory variables. Lastly, while evaluating the plausibility contour in (6) is simple and only involves calculations related to the binomial distribution, obtaining marginal inferences for components of a potential high-dimensional θ in an efficient manner remains an open question.

Acknowledgments

The author thanks Professor Ryan Martin and the four anonymous reviewers for their valuable feedback on earlier versions of this manuscript.

References

M. S. Balch, R. Martin, and S. Ferson. Satellite conjunction analysis and the false confidence theorem. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 475(2227):20180565, 2019.

D. F. Benoit and D. Van den Poel. bayesQR: A bayesian approach to quantile regression. *Journal of Statistical Software*, 76(7):1–32, 2017.

J. Berger. The case of objective bayesian analysis. *Bayesian Analysis*, 1(3):385–402, 2006.

M.E. Celebi and K. Aydin. *Unsupervised Learning Algorithms*. Springer International Publishing, 2016.

L. Cella and R. Martin. Direct and approximately valid probabilistic inference on a class of statistical functionals. *International Journal of Approximate Reasoning*, 151: 205–224, 2022a.

L. Cella and R. Martin. Valid inferential models offer performance and probativeness assurances. In Sylvie Le Hégarat-Mascle, Isabelle Bloch, and Emanuel Aldea, editors, *Belief Functions: Theory and Applications*, pages 219–228, Cham, 2022b. Springer International Publishing.

L. Cella and R. Martin. Distribution-free inferential models for direct and valid probabilistic inference. In preparation, 2023.

V. Chernozhukov, C. Hansen, and M. Jansson. Finite sample inference for quantile regression models. *Journal of Econometrics*, 152(2):93–103, 2009. Nonparametric and Robust Methods in Econometrics.

A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.

A. P. Dempster. A generalization of Bayesian inference. (With discussion). *Journal of the Royal Statistical Society, Series B*, 30:205–247, 1968.

A. P. Dempster. The Dempster–Shafer calculus for statisticians. *International Journal of Approximate Reasoning*, 48(2):365 – 377, 2008.

A. P. Dempster. Statistical inference from a Dempster–Shafer perspective. In Xihong Lin, Christian Genest, David L. Banks, Geert Molenberghs, David W. Scott, and Jane-Ling Wang, editors, *Past, Present, and Future of Statistical Science*, chapter 24. Chapman & Hall/CRC Press, 2014.

T. Denœux. Likelihood-based belief function: justification and some extensions to low-quality data. *International Journal of Approximate Reasoning*, 55(7):1535–1547, 2014.

T. Denœux and S. Li. Frequency-calibrated belief functions: review and new insights. *Internat. J. Approx. Reason.*, 92:232–254, 2018.

D. Dubois and H. Prade. *Possibility Theory*. Plenum Press, New York, 1988.

B. Efron. Discussion: "confidence distribution, the frequentist distribution estimator of a parameter: A review". *International Statistical Review*, 81(1):41–42, 2013.

R. A. Fisher. The fiducial argument in statistical inference. *Annals of Eugenics*, 6(4):391–398, 1935. doi:10.1111/j.1469-1809.1935.tb02120.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1935.tb02120.x.

C. Gutenbrunner, J. Jurečková, R. Koenker, and S. Portnoy. Tests of linear hypotheses based on regression rank scores. *Journal of Nonparametric Statistics*, 2(4):307–331, 1993.

J. Hannig, H. Iyer, R. C. S. Lai, and T. C. M. Lee. Generalized fiducial inference: A review and new results. *Journal of the American Statistical Association*, 111 (515):1346–1361, 2016.

D. Hose and M. Hanss. On data-based estimation of possibility distributions. *Fuzzy Sets and Systems*, 399: 77–94, 2020. Fuzzy Intervals.

D. Hose and M. Hanss. A universal approach to imprecise probabilities in possibility theory. *International Journal of Approximate Reasoning*, 133:133–158, 2021.

M. Hušková. Some sequential procedures based on regression rank scores. *Journal of Nonparametric Statistics*, 3 (3-4):285–298, 1994.

R. Koenker. *Quantile Regression*. Cambridge University Press, Cambridge, 2005. ISBN 978-0-521-60827-5; 0-521-60827-9.

R. Koenker and G. Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978. ISSN 00129682, 14680262.

R. Koenker, S. Portnoy, P. T. Ng, B. Melly, A. Zeileis, P. Grosjean, C. Moler, Y. Saad, V. Chernozhukov, I. Fernandez-Val, and B. D Ripley. *quantreg: Quantile Regression*, 2022. URL https://cran.r-project.org/package=quantreg. R package version 5.94.

A. Kottas and M. Krnjajić. Bayesian semiparametric modelling in quantile regression. *Scandinavian Journal of Statistics*, 36(2):297–319, 2009.

C. Liu and R. Martin. Inferential models and possibility measures. *Handbook of Bayesian, Fiducial, and Frequentist Inference*, to appear; arXiv:2008.06874, 2021.

R. Martin. False confidence, non-additive beliefs, and valid statistical inference. *International Journal of Approximate Reasoning*, 113:39–73, 2019.

R. Martin. An imprecise-probabilistic characterization of frequentist statistical inference. *Researchers.One*, https://researchers.one/articles/21.01.00002, 2021.

R. Martin. Valid and efficient imprecise-probabilistic inference with partial priors, i. first results. *arXiv*, 2022a.

R. Martin. Valid and efficient imprecise-probabilistic inference with partial priors, ii. general framework. *arXiv*, 2022b.

R. Martin and C. Liu. Inferential models: A framework for prior-free posterior probabilistic inference. *Journal of the American Statistical Association*, 108:301–313, 2013.

R. Martin and C. Liu. *Inferential Models: Reasoning with Uncertainty*. Monographs in Statistics and Applied Probability Series. Chapman & Hall/CRC Press, 2015.

R. Martin and N. Syring. Chapter 1 - direct gibbs posterior inference on risk minimizers: Construction, concentration, and calibration. In Arni S.R. Srinivasa Rao, G. Alastair Young, and C.R. Rao, editors, *Advancements in Bayesian Methods and Implementation*, volume 47 of *Handbook of Statistics*, pages 1–41. Elsevier, 2022.

I. Molchanov. *Theory of Random Sets*. Probability and Its Applications (New York). Springer-Verlag London Ltd., London, 2005.

H. T. Nguyen. *An Introduction to Random Sets*. Chapman & Hall/CRC, Boca Raton, FL, 2006.

B. J. Reich, H. D. Bondell, and H. J. Wang. Flexible Bayesian quantile regression for independent and clustered data. *Biostatistics*, 11(2):337–352, 11 2009.

N. Reid and D. R. Cox. On some principles of statistical inference. *International Statistical Review*, 83(2):293–308, 2015.

T. Schweder and N. Lid Hjort. *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2016.

G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, N.J, 1976. ISBN 9780691100425.

M. Xie and K. Singh. Confidence distribution, the frequentist distribution estimator of a parameter: A review. *International Statistical Review*, 81(1):3–39, 2013.

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

Y. Yang, H. J. Wang, and X. He. Posterior inference in bayesian quantile regression with asymmetric laplace likelihood. *International Statistical Review*, 84(3):327–344, 2016.

K. Yu and R. A. Moyeed. Bayesian quantile regression. *Statistics & Probability Letters*, 54(4):437–447, 2001.