Wake Forest University
Department of Statistical Sciences

Finite sample valid probabilistic inference on quantile regression

Leonardo Cella lcella@wfu.edu

1 - Introduction

Data $Z^n = \{Z_i = (X_i, Y_i) : i = 1, \ldots, n\}$ of n covariates-response pairs are iid with distribution P. Nothing is assumed about P.

- $Q_x(\tau) = x^\top \theta$ is the τ-th quantile of Y given $X = x$.
- **Goal:** make inferences on θ that are *distribution-free* and *valid*.
- **Common solution:** Inferences through *confidence regions*
- Notion of validity is familiar $\rightarrow$ coverage guarantees:

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

2 - Probabilistic Inference

Beyond confidence regions, where we can assign degrees of belief Π to relevant assertions about θ, e.g., $\theta \in A$, $A \subseteq \Theta$.

- Validity: control the assignment of high degrees of belief to false assertions:

$$\sup_{P:\theta(P)\notin A} P^n\big\{\Pi_{Z^n}(A) > 1-\alpha\big\} \leq \alpha, \quad \alpha \in [0,1]. \tag{2}$$

- Bayesian approach? *False Confidence Theorem* says we need imprecision!

3 - Inferential Models

Consider the parametric case where $Z^n = (Z_1, \ldots Z_n)$ are iid with distribution P_ω. IM approach [2] offers valid probabilistic inference for ω.

Two-step IM construction

1. Choose an appropriate $h : (\mathbb{Z}^n \times \Omega) \rightarrow \mathbb{R}$ that determines a partial ordering of candidate values for ω given z^n, e.g.,likelihood ratio:

$$h(z^n, \omega) = L_{z^n}(\omega)/L_{z^n}(\hat{\omega}_{z^n})$$

2. Compute the possibility contour

$$\pi_{z^n}(\omega) = P^n_\omega\{h(Z^n, \omega) \leq h(z^n, \omega)\}$$

$\pi \rightarrow$ valid probabilistic inference and confidence regions for ω

4 - Nonparametric IM for quantile regression

The idea here is to mimic the construction above:

1. Choose an h that orders candidate values for θ given z^n
2. Compute the contour

$$\pi_{z^n}(\theta) = P^n\{h(Z^n, \theta) \leq h(z^n, \theta)\}, \quad \theta \in \Theta. \tag{3}$$

Theorem:

- The degrees of belief obtained from (3) are valid in the sense of (2).
- $\{\theta \in \Theta : \pi_{z^n}(\theta) > \alpha\}$ is a valid confidence region in the sense of (1).

Challenges:

1. What h? No model $\rightarrow$ no likelihood ratio.
2. How to compute (3)? Recall that P is unknown.

A possible solution

- In [1], a bootstrap-based IM construction was proposed
- validity is just achieved asymptotically

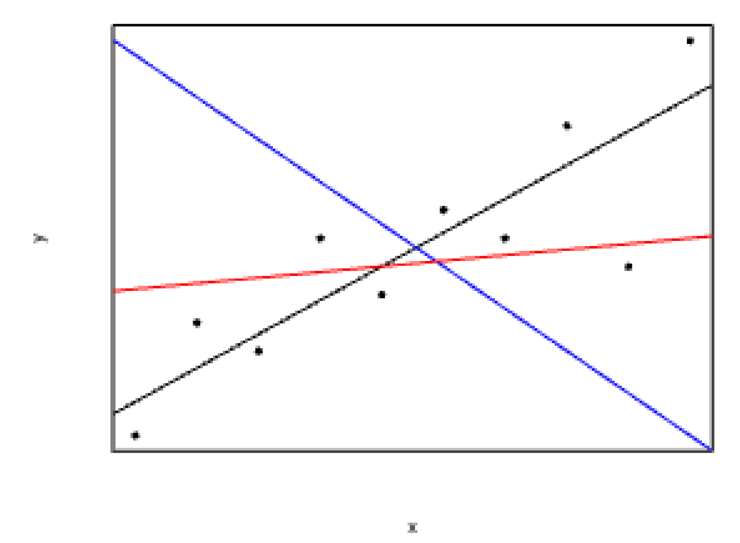
But we want more! $\rightarrow$ IM that achieves validity for any sample size

Strategy: Choose an h whose distribution is known and independent of unknown quantities, so (3) can be computed! More specifically:

- Find $\gamma(\theta, z^n)$ that is a pivot
- h $\rightarrow$ γ's probability mass

4.1 - An intuitive (but bad) solution

- $\gamma = \sum_{i=1}^n I_{(0,\infty)}(Y_i - x_i^\top \theta) \rightarrow \gamma \sim \mathsf{Bin}(n, 1-\tau) \rightarrow h = \binom{n}{\gamma}(1-\tau)^\gamma \tau^{n-\gamma}$
- Very inefficient! For example, for $\tau = 0.5$, any line that splits the data in half, e.g., the **black**, **red** and **blue** below, is equally maximally plausible.



4.2 - A better solution

Let X be discrete with k levels. The idea is to consider the binomials for each level of X separately, and have h as the product of their probability masses.

$$h = \prod_{i=1}^k \binom{n_i}{\gamma_i}(1-\tau)^{\gamma_i}\tau^{n_i-\gamma_i}, \text{ where } \quad \gamma_i = \sum_{j=1}^{n_i} I_{(0,\infty)}(Y_j - x_i\theta). \tag{4}$$

Example: k=3, $n_1 = n_2 = n_3 = 10$, $\tau = 0.5$:

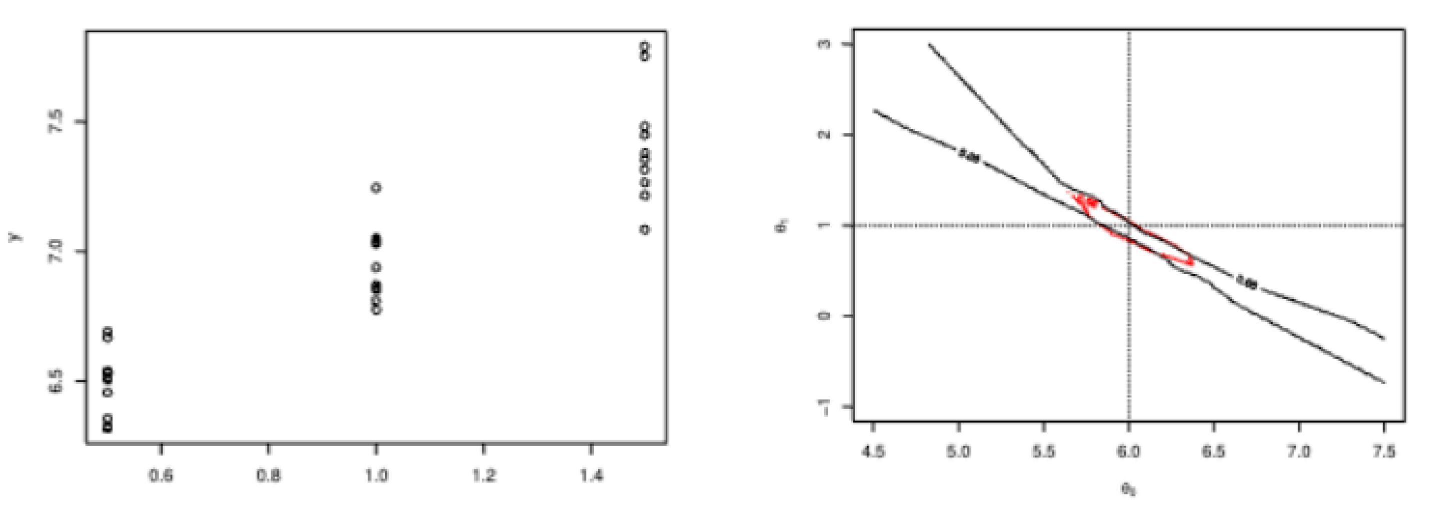


Figure 1: Data set on the left. 95% confidence region for θ on the right.

If X is continuous, there is no replication of Y for any given $X = x$. But we do have replications of Y in neighborhoods of X, so (4) can still be used!

- Form k neighborhoods of X
- Consider each one of the k independent binomials separately
- Use the product of their probability masses as the plausibility order h

Example: Simulation study with $n = 30$, $\tau = 0.3$ and $k = 2$ to compare the coverage probabilities and mean length of 95% interval estimates for the quantile regression coefficients based on the IM and two other methods:

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)

5 - Open questions

Other pivot options? How to best select the neighborhoods of a continuous X? Specifically, does the number of neighborhoods and/or the number of replications per neighborhood impact the efficiency of the IM?

References

[1] 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, 2022.

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