Why Quantile Regression Works Well in Economics: A Partial Explanation

Olga Kosheleva[1], Vassilis G. Kaburlasos[2],
Vladik Kreinovich[1], and Roengchai Tansuchat[3]
[1]University of Texas at El Paso,
500 W. University, El Paso, Texas 79968, USA,
olgak@utep.edu, vladik@utep.edu
[2]Department of Computer Science,
International Hellenic University (IHU),
Kavala 65404, Greece, vgkabs@cs.ihu.gr
[3]Faculty of Economics, Chiang Mai University,
Chiang Mai, Thailand, roengchaitan@gmail.com

1. Predictions are important

- One of the main objectives of science is to predict the future state of the world.
- In general, to describe the state of the world, we need to describe the values of the quantities that characterize this state.
- Because of this, usually, prediction means predicting values of different quantities.
- For example, in economics:
  - we want to predict the Gross Domestic Product (GDP),
  - we also want to predict the future values of the stock market indices,
  - we want to predict the agriculture yield.

2. We need to predict the future probability distribution

- In many practical situations, the state of a system cannot be adequately described by a single variable.
- To fully characterize this state, we need to describe a probability distribution.
- For example, to understand the general state of the country's economy:
  - it is not enough to know the average income,
  - we also need to know how income is distributed: what percentage of people lives below poverty level,
  - what percentage is super-rich and what is their share.
- All these factors are important to decide how stable is the economic situation.

3. We need to predict the future probability distribution (cont-d)

- Similarly, in agriculture, it is not enough to know the overall yield of certain crops (e.g., grapes).
- From the economic viewpoint, we need to know how many grapes will be of certain size.

4. Quantiles are natural characteristics of a probability distribution

- We are interested in the proportion of people whose income X is below the poverty level x.
- From the probability viewpoint, it is the value of the cumulative distribution function $\text{Prob}(X \le x)$.
- From this viewpoint, what we want to predict are the values of the cumulative distribution function.
- Describing this function is equivalent to describe the inverse function, i.e., a function that assigns:
  - to each probability $\alpha \in [0, 1]$,
  - the value $x(\alpha)$ for which $\text{Prob}(X \le x(\alpha)) = \alpha$.
- This value $x(\alpha)$ is known as α-*quantile*.
- For $\alpha = 0.5$, we get the *median*, for $\alpha = 0.25$ and $\alpha = 0.75$, we get *quartiles*, etc.

5. Quantile regression: an unexpected success

- For each future quantity of interest y, we want to predict its α-quantiles $y(\alpha)$:
  - based on the available information about the current quantities x_i,
  - i.e., based on the values $x_i(\alpha_i)$ corresponding to different i and different α_i.
- In principle, all this information may be important for the prediction.
- For example:
  - if we use quantiles corresponding to $\alpha = 0, 0.1, 0.2, \ldots, 1.0)$,
  - then to describe each value $y(\alpha)$, we should know all the quantiles of all n current variables:

$$y(\alpha) = f_\alpha(x_1(0), x_1(0.1), \ldots, x_1(1), x_2(0), x_2(0.1), \ldots, x_2(1), \ldots,$$

$$x_n(0), x_n(0,1), \ldots, x_n(1)).$$

6. Quantile regression: an unexpected success (cont-d)

- Interestingly, it turns out that:
  - in many practical situations in economics (and beyond economics),
  - we can get a good prediction of the α-quantile for y by using only quantiles for x_i corresponding to the exact same value α:

$$y(\alpha) = f_\alpha(x_1(\alpha), \ldots, x_n(\alpha)).$$

- Prediction techniques that use this expression are known as *quantile regression.*
- There is no good explanation of why the simplified formula often leads to good predictions.

7. What we do in this talk

- In this talk:
  - we use our experience – of predicting agriculture yields
  - to come up with a partial explanation for the empirical success of quantile regression.
- Specifically, we explain the efficiency of quartile regression in situations when the variables used for prediction are highly correlated.

8. Why this challenge is, in general, difficult

- Many scientists, including many economists, have what is called "physics envy".
- In fundamental physics, we can represent each object as a combination of simple objects.
- E.g., of small body parts or even of molecules.
- For simple objects, researchers have experimentally studied their interactions.
- They came up with simple laws that describe these interactions – starting with well-known Newton's laws.
- Based on these laws, we can predict how complex combinations of simple objects will interact.
- Often, the resulting predictions are very accurate.
- In contrast, in economics, we largely deal with the economy as a whole as a black box.

9. Why this challenge is, in general, difficult (cont-d)

- Yes, the overall economy does consist of individual people.
- However, we do not have the ability to trace every single person's economics-related decisions.
- We cannot perform easy experiments.
- We cannot separate people so that only two of them will interact – as we can do with masses or electric charges.
- From this viewpoint, all we can do is:
  - come up with empirical general laws,
  - often without a clear understanding of why these laws are valid.

10. But there are cases when a detailed study is possible

- It is true that:
  - in most economic phenomena – phenomena that deals with people,
  - a physics-level detailed study of the phenomenon is not possible.
- However, there are situations when such a study *is* possible.
- This was exactly our case study.

11. Description of the case study

- In our case study, we analyzed how the distribution of grapes by size changes with time.
- It turns out to be one of the cases when quantile regression leads to a very good prediction.
- Good news is that:
  - in contrast to other economic situation,
  - we have a detailed record of values corresponding to several selected plants at different moments of time.
- The observation of these plants enable us to explain the effectiveness of quantile regression.

12. How our observations explain the effectiveness of quantile regression: case of a single input

- In our case study, at each moment of time, we observe the values $v_1, \ldots, v_n$ corresponding to different objects (in our case, plants).
- To describe the corresponding quantiles, we need to sort these values in an increasing order: $v_{\pi(1)} < v_{\pi(2)} < \ldots < v_{\pi(n)}$.
- In this order, the median is the value $v_{\pi(n/2)}$.
- In general, the α-quantile $v(\alpha)$ is the value $v_{\pi(n \cdot \alpha)}$.
- How will these values change from the current moment of time to the next?
- There are many factors that affect each object.
- For example, for agriculture predictions, weather is the most important factor.
- There may be some interaction between individual plants as well.

13. Case of a single input (cont-d)

- E.g., one plant can provide shade on another one thus somewhat slowing down the other one's growth.
- However, such effects are small.
- So, in the first approximation, we can safely assume that different objects do not affect each other.
- In other words, in this approximation, the value v_i' of the variable x_i at the next moment of time:
  - does not depend on the values v_j for $j \neq i$,
  - it only depends on the value v_i, and, of course, on the external factors e: $v_i' = f(v_i, e)$ for some function $f(v, e)$.
- This dependence is usually monotonic.
- So in the most cases when we had $v_i < v_j$, in the next moment of time, we will still have $v_i' < v_j'$.
- Thus, the order between these values will remain largely the same.

14. Case of a single input (cont-d)

- In other words, at the next moment of time, we will still largely have the same order: $v'_{\pi(1)} < v'_{\pi(2)} < \ldots < v'_{\pi(n)}$;
  - the smallest value will remain the smallest,
  - the second smallest will remain the second smallest, etc., and
  - the largest value will remain the largest.
- Thus, for each α:
  - the new value $v'(\alpha) = v'_{\pi(\alpha \cdot n)}$ of the α-quantile corresponds to the same index $i = \pi(n \cdot \alpha)$
  - as the value $v(\alpha) = v_{\pi(\alpha \cdot n)}$ of this α-quantile at the previous moment of time.
- So, for this i, the above formula implies that $v'(\alpha) = f(v(\alpha), e)$.
- This is exactly the type of relation that corresponds to quantile regression.

15. The same arguments apply to people, not only to plants

- In general economic situations, we have people, families, or companies instead of plants.
- People do interact with each other.
- However, in general, a person's economic behavior is most affected by general factors – advertisements, general feeling.
- So an effect of immediate neighbors can be, in the first approximation, safely ignored.
- Thus, we can apply the same arguments and conclude that:
  - in a general economic situation in which predictions are based on a single input,
  - we should expect quantile regression to be efficient.

16. What if we have several highly correlated inputs

- In many practical situations, the inputs x_i – that are used to predict y – are highly correlated.
- High correlation implies that:
  - once we sort the objects in the increasing order of one of the inputs – e.g., of the input x_1,
  - then we will have almost the same order for all other inputs x_i.
- Namely, we will have:
  - increasing order if the correlation between x_1 and x_i is positive, and
  - decreasing order if the correlation between x_1 and x_i is negative.
- Thus, similar to the case when we have a single input:
  - the α-quantiles of all the variables at the next moment of time
  - correspond largely to the same objects as in the previous moment of time.

17. What if we have several highly correlated inputs (cont-d)

- So, similarly to the single-input case:
  - if we start with a formula that describes how the values $x_{1i}, \ldots, x_{ni}$ describing object i change with time:

$$x'_{ji} = f_j(x_{1i}, \ldots, x_{ni}, e),$$

  - then we get a similar quantile-regression-type formula describing how α-quantiles change with time:

$$x'_j(\alpha) = f_j(x_1(\alpha), \ldots, x_n(\alpha), e).$$

- So, in the case of highly correlated inputs, we indeed have an explanation for the empirical success of quantile regression.

18. Acknowledgments

- This work was supported in part by the National Science Foundation grants:
  - 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and
  - HRD-1834620 and HRD-2034030 (CAHSI Includes).
- It was also supported by the AT&T Fellowship in Information Technology.
- It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.