Departmental Technical Reports (CS)   Computer Science

6-1-2022

Why Quantile Regression Works Well in Economics: A Partial Explanation

Olga Kosheleva
The University of Texas at El Paso, olgak@utep.edu

Vassilis G. Kaburlasos
International Hellenic University (IHU), vgkabs@cs.ihu.gr

Vladik Kreinovich
The University of Texas at El Paso, vladik@utep.edu

Vladik Kreinovich
The University of Texas at El Paso, vladik@utep.edu

Roengchai Tansuchat
Chiang Mai University, roengchaitan@gmail.com

Follow this and additional works at: https://scholarworks.utep.edu/cs_techrep

Part of the Computer Sciences Commons, and the Mathematics Commons

Comments:

Technical Report: UTEP-CS-22-80

Why Quantile Regression Works Well in Economics: A Partial Explanation

Olga Kosheleva, Vassilis G. Kaburlasos, Vladik Kreinovich, and Roengchai Tansuchat

Abstract To get a better picture of the future behavior of different economics-related quantities, we need to be able to predict not only their mean values, but also their distribution. For example, it is desirable not only to predict future average income, but also to predict the future distribution of income. One of the convenient ways to describe a probability distribution is by using α-quantiles such as medians (corresponding to $\alpha = 0.5$), quartiles (corresponding to $\alpha = 0.25$ and $\alpha = 0.75$), etc. In principle, an α-quantile of the desired future quantity can depend on β-quantiles of current distributions corresponding to all possible values β. However, in many practical situations, we can get very good predictions based only on current quantiles corresponding to $\beta = \alpha$; this is known as *quartile regression*. There is no convincing explanation of why quantile regression often works. In this paper, we use an agriculture-related case study to provide a partial explanation for this empirical success – namely, we explain it in situations when the inputs used for prediction are highly correlated.

1 Formulation of the Problem

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

Olga Kosheleva, and Vladik Kreinovich
University of Texas at El Paso, 500 W. University, El Paso, Texas 79968, USA
e-mail: olgak@utep.edu, vladik@utep.edu

Vassilis G. Kaburlasos
HUman-MAchines INteraction (HUMAIN) Lab, Department of Computer Science
International Hellenic University (IHU), Kavala 65404, Greece, e-mail: vgkabs@cs.ihu.gr

Roengchai Tansuchat
Faculty of Economics, Chiang Mai University, Chiang Mai, Thailand
e-mail: roengchaitan@gmail.com

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; see, e.g., [9, 10, 11, 15, 17, 18, 19, 21, 23].

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.

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.

Quantiles are natural characteristics of a probability distribution. The proportion of people whose income X is below the poverty level x is, from the probability viewpoint, a 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.

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:

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

$$x_n(0), x_n(0, 1), \ldots, x_n(1)), \tag{1}$$

for some function f_α. 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:

$$y(\alpha) = f_\alpha(x_1(\alpha), \ldots, x_n(\alpha)). \tag{2}$$

Prediction techniques that use the expression (2) are known as *quantile regression*; see, e.g., [1, 3, 4, 6, 8, 13, 14, 16, 22].

Remaining challenge. There is no good explanation of why the simplified formula (2) often leads to good predictions.

What we do in this paper. In this paper, 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.

2 Case Study and the Resulting Partial Explanation

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; see, e.g., [5, 20]. For simple objects, researchers have experimentally studied their interactions and 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, and often, the resulting predictions are very accurate.

In contrast, in economics, we largely deal with the economy as a whole as a black box. Yes, the overall economy does consist of individual people, but 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.

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, and this was exactly our case study.

Description of the case study. In our case study, we analyzed how the distribution of grapes by size changes with time; see, e.g., [2, 12]. 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 enables us to explain the effectiveness of quantile regression.

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:

$$v_{\pi(1)} < v_{\pi(2)} < \ldots < v_{\pi(n)}. \tag{3}$$

In this order, the median is the value $v_{\pi(n/2)}$, and, 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 – e.g., one plant can provide shade on another one thus somewhat slowing down the other one's growth – but 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:

$$v_i' = f(v_i, e) \tag{4}$$

for some function $f(v, e)$.

The dependence (4) 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. In other words, at the next moment of time, we will still largely have the same order:

$$v_{\pi(1)}' < v_{\pi(2)}' < \ldots < v_{\pi(n)}'; \tag{5}$$

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 formula (4) implies that

$$v'(\alpha) = f(v(\alpha), e), \tag{6}$$

which is exactly the type of relation that corresponds to quantile regression.

The same arguments apply to people, not only to plants. In general economic situations, we have people, families, or companies instead of plants. While people do interact with each other, 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.

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:

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

$$x'_{ji} = f_j(x_{1i}, \ldots, x_{ni}, e), \tag{7}$$

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

$$x'_j(\alpha) = f_j(x_1(\alpha), \ldots, x_n(\alpha), e). \tag{8}$$

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

Acknowledgments

References

1. I. G. N. Agung, *Quantile Regression: Applications on Experimental and Cross Section Data using EViews*, Wiley, Hoboken, New Jersey, 2021.
2. C. Bazinas, E. Vrochidou, C. Lytridis, and V. G. Kaburlasos, "Yield estimation in vineyards using intervals' numbers techniques", In: *Proceedings of the 25th Panhellenic Conference on Informatics PCI'2021*, Volos, Greece, November 26–28, 2021, pp. 454–459.
3. T. J. Cleophas and A. H. Zwinderman, *Quantile Regression in Clinical Research: Complete analysis for data at a loss of homogeneity*, Springer, Cham, Switzerland, 2021.
4. C. Davino, M. Furno, and D. Vistocco, *Quantile Regression: Theory and Applications*, Wiley, Chichester, UK, 2014.
5. R. Feynman, R. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Addison Wesley, Boston, Massachusetts, 2005.
6. B. Fitzenberger, R. Koenker, and J. A. F. Machado (eds.), *Economic Applications of Quantile Regression*, Physica-Verlag, Heidelberg, 2001.
7. M. Furno and D. Vistocco, *Quantile Regression: Estimation and Simulation*, Wiley, Hoboken, New Jersey, 2018.

8. L. Hao and D. Q. Naiman, *Quantile Regression*, Sage Publications, Thousand Oaks, California, 2007.
9. T. Iizumi, Y. Shin, W. Kim, M. Kim, and J. Choi, "Global crop yield forecasting using seasonal climate information from a multi-model ensemble", *Climate Services*, 2018, Vol. 11, pp. 13–23.
10. M. Jayakumar, M. Rajavel, and U. Surendran, "Climate-based statistical regression models for crop yield forecasting of coffee in humid tropical Kerala, India", *International Journal of Biometeorology*, 2016, Vol. 60, pp. 1943–1952.
11. M. A. Jayaram and N. Marad, "Fuzzy inference systems for crop yield prediction", *Journal of Intelligent Systems*, 2012, Vol. 21, pp. 363–372.
12. V. G. Kaburlasos, C. Bazinas, E. Vrochidou, and E. Karapatzak, "Agricultural yield prediction by difference equations on data-induced cumulative possibility distributions", *Proceedings of the Annual Conference of the North American Fuzzy Information Processing Society NAFIPS'2022*, Halifax, Nova Scotia, Canada, May 31 – June 3, 2022.
13. R. Koenker, *Quantile Regression*, Cambridge University Press, New York, 2005.
14. R. Koenker, V. Chernozhukov, X. He, and L. Peng, *Handbook of Quantile Regression*, Chapman and Hall/CRC, Boca Raton, Florida, 2017.
15. J. A. Mathieu and F. Aires, "Assessment of the agro-climatic indices to improve crop yield forecasting", *Agricultural and Forest Meteorology*, 2018, Vol. 253–254, pp. 15–30.
16. D. P. McMillen, *Quantile Regression for Spatial Data*, Springer, Heidelberg, New York, Dordrecht, London, 2013.
17. T. Monga, "Estimating vineyard grape yield from images", In: E. Bagheri and J. C. K. Cheung (eds.), *Advances in Artificial Intelligence. Proceedings of the 31st Canadian Conference on Artificial Intelligence Canadian AI'2018, Toronto, Ontario, Canada, May 8–11, 2018*, pp. 339–343.
18. A. Shastry, H. A. Sanjay, and M. Hegde, "A parameter based ANFIS model for crop yield prediction", In: *Proceedings of the 2015 IEEE International Advance Computing Conference IACC'2015*, Bangalore, India, June 12–13, 2015, pp. 253–257.
19. P. Taherei Ghazvinei, H. Hassanpour Darvishi, A. Mosavi, K. bin Ean Yusof, M. Alizamir, S. Shamshirband, and K.-W. Chau, "Sugarcane growth prediction based on meteorological parameters using extreme learning machine and artificial neural network", *Engineering Applications of Computational Fluid Mechanics*, 2018, Vol. 12, pp. 738–749.
20. K. S. Thorne and R. D. Blandford, *Modern Classical Physics: Optics, Fluids, Plasmas, Elasticity, Relativity, and Statistical Physics*, Princeton University Press, Princeton, New Jersey, 2017.
21. S. M. Upadhya and S. Mathew, "Implementation of fuzzy logic in estimating yield of a vegetable crop", In: *Proceedings of the Third National Conference on Computational Intelligence NCCI'2019*, Bangalore, India, December 5–6, 2019, Bangalore, India, *Journal of Physics: Conference Series*, 2020, Vol. 1427, Paper 0120132020.
22. J. M. Uribe and M. Guillen, *Quantile Regression for Cross-Sectional and Time Series Data: Applications in Energy Markets Using R*, Springer, Cham, Switzerland, 2020.
23. T. van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review", *Computers and Electronics in Agriculture*, 2020, Vol. 177, Paper 105709.