Why Is Linear Quantile Regression Empirically Successful: A Possible Explanation

Hung T. Nguyen[1,2], Vladik Kreinovich[3],
Olga Kosheleva[4], and Songsak Sriboonchitta[2]

[1] Department of Mathematical Sciences, New Mexico State University
Las Cruces, New Mexico 88003, USA, hunguyen@nmsu.edu
[2] Department of Economics, Chiang Mai University
Chiang Mai, Thailand, songsakecon@gmail.com
[3] Department of Computer Science, University of Texas at El Paso
500 W. University, El Paso, TX 79968, USA, vladik@utep.edu
[4] University of Texas at El Paso, 500 W. University,
El Paso, TX 79968, USA, olgak@utep.edu

Abstract. Many quantities describing the physical world are related to each other. As a result, often, when we know the values of certain quantities $x_1, \ldots, x_n$, we can reasonably well predict the value of some other quantity y. In many application, in addition to the resulting estimate for y, it is also desirable to predict how accurate is this approximate estimate, i.e., what is the probability distribution of different possible values y. It turns out that in many cases, the quantiles of this distribution linearly depend on the values $x_1, \ldots, x_n$. In this paper, we provide a possible theoretical explanation for this somewhat surprising empirical success of such linear quantile regression.

1 Formulation of the Problem

What is regression: a brief reminder. Many things in the real world are related to each other. As a result, if we know the values of some quantities $x_1, \ldots, x_n$, then we can often reasonable well estimate the value of some other quantity y.

In some cases, the dependence of y on $x_1, \ldots, x_n$ is known. In many other situations, we do not know this dependence, so we need to find this dependence based on the empirical data. The desired dependence of y on $x_1, \ldots, x_n$ is known as a *regression function* $y \approx f(x_1, \ldots, x_n)$, and the methodology of determining the regression function from the empirical data is known as *regression analysis.*

In many practical situations, the dependence of y on $x_1, \ldots, x_n$ is well described by a linear function

$$y \approx \beta_0 + \sum_{i=1}^{n} \beta_i \cdot x_i. \tag{1}$$

Such linear dependence is known as *linear regression.*

What is quantile regression. Traditionally, the emphasis of regression analysis has been on finding the actual dependence $y \approx f(x_1, \ldots, x_n)$. However, finding this dependence is not enough. As we have mentioned earlier, the value $f(x_1, \ldots, x_n)$ is only an *approximation* to y. It is good to know this approximation, but it is also important to know *how accurate* is this approximation. In other words, we want to know not only the estimate of y for given x_i, we also want to know how the conditional probability distribution of y depends on the inputs $x_1, \ldots, x_n$.

One of the empirically efficient ways for finding this dependence is the method of *quantile regression.* One of the possible ways to describe the conditional probability distribution $P(y \,|\, x_1, \ldots, x_n)$ is to describe, for each probability p, the p-th quantile y_p of this distribution, i.e., that value for which the conditional probability $\mathrm{Prob}(y \le y_p \,|\, x_1, \ldots, x_n)$ is equal to p. In particular:

- for $p = 0.5$, we get the median,
- for $q = 0.25$ and $q = 0.75$, we get the quartiles, etc.

One of the most empirically successful methods of describing the dependence of the conditional probability distribution on x_i is a method of *quantile regression*, when, for each p, we find a regression function $y_p = f_p(x_1, \ldots, x_n)$ that describes the dependence of the corresponding quantile y_p on the inputs x_i.

In particular, somewhat surprisingly, in many practical situations, this dependence turns out to be linear:

$$y_p \approx \beta_{0,p} + \sum_{i=1}^{n} \beta_{i,p} \cdot x_i \tag{2}$$

for appropriate coefficients $\beta_{i,p}$; see, e.g., [2–6].

Why is linear quantile regression empirically successful? Why is this linear quantile regression empirically successful in many practical applications? In this paper, we provide a possible explanation for this empirical success.

The structure of this paper is as follows. First, in Section 2, we provide fundamental reasons why linear regression is often empirically successful. In Section 3, we expand this result to the case of *interval uncertainty*, when instead of predicting the exact value of the quantity y, we predict the interval of its possible values. Finally, in Section 4, we show how this result can be expanded from interval to probabilistic uncertainty – thus explaining the empirical success of linear quantile regression.

2 Why Linear Regression Is Often Empirically Successful: A General Explanation

What we do in this section. Our goal is to explain why linear *quantile* regression is empirically successful. To explain this empirical phenomenon, let us first provide a possible explanation of why linear regression *in general* is empirically successful.

Empirical fact. Linear regression is empirically successful in many real-life situations, often situations when the known empirical dependence is non-linear.

In this section, we will provide a possible explanation for this empirical fact.

Basis for our explanation: possibility of different starting points for measuring the corresponding quantities. We are interested in the dependence between the *quantities* x_i and y. To describe this dependence between *quantities*, we describe the dependence between the *numerical values* of these quantities.

The difference between the quantity itself and its numerical value may be perceived as subtle but, as will show, this difference is important – and it provides the basis for our explanation. The reason why there is a difference in the first place is that the numerical value of a quantity depends on the starting point for measuring this quantity. If we change this starting point to the one which is a units earlier, then all the numerical values of this quantity change from the previous value x to the new value $x + a$.

For example, we can start measuring temperature with the absolute zero (as in the Kelvin scale) or with the temperature at which the ice melts (as in the Celsius scale), the corresponding numerical values differ by $a \approx 273$ degrees. Similarly, we can start measuring time with the birth year of Jesus Christ or, as the French Revolution decreed, with the year of the French Revolution.

It may be not so clear, but when we gauge many economic and financial quantities, there is also some arbitrariness in the selection of the starting point. For example, at first glance, unemployment is a well-defined quantity, with a clear starting point of 0%. However, economists who seriously study unemployment argue that starting it from 0 is somewhat misleading, since this may lead to an unrealistic expectation of having 0 unemployment. There is a natural minimal unemployment level of approximately 3%, and a more natural way of measuring unemployment is:

- not by its absolute value,
- but by the amount by which the current unemployment level exceeds its natural minimum.

Similarly, a person's (or a family's) income seems, at first glance, like a well-defined quantity with a natural starting point of 0. However, this does not take into account that 0 is not a possible number, a person needs to eat, to get clothed. Thus, a more reasonable way to gauge the income is:

- not by the absolute amount,
- but by how much the actual income exceeds the bare minimum needed for the person's survival.

The changes in the starting points should not affect the actual form of the dependence. In general, we can have different starting points for measuring each of the input quantities x_i. As a result, for each i, instead of the original numerical values x_i, we can have new values $x'_i = x_i + a_i$, for some constants a_i.

The change in the starting point:

- *changes* the numerical value, but
- *does not change* the actual quantity.

Thus, it is reasonable to require that the exact form of the dependence between x_i and y should not change if we simply change the starting points for all the inputs.

Of course, even for the simplest dependence $y = x_1$, if we change the starting point for x_1, then the numerical value of y will change as well, by the same shift – and thus, while the numerical value of y changes, the quantity y does not change – because the change in the starting point for x_1 simply implies that we correspondingly change the starting point for y.

In general, it is therefore reasonable to require that for each combination of shifts $a_1, \ldots, a_n$:

- once we shift the inputs to $x_i' = x_i + a_i$ and apply the function f to these shifted values,
- the resulting value $y' = f(x_1', \ldots, x_n')$ should simply be obtained from the original pre-shifted value $y = f(x_1, \ldots, x_n)$ by an appropriate shift:

$$y' = y + s(a_1, \ldots, a_n).$$

Thus, we arrive at the following definition.

Definition. *We say that a function $f(x_1, \ldots, x_n)$ is* shift-invariant *if for every tuple $(a_1, \ldots, a_n)$ there exists a value $s(a_1, \ldots, a_n)$ such that for all tuples $(x_1, \ldots, x_n)$, we have*

$$f(x_1 + a_1, \ldots, x_n + a_n) = f(x_1, \ldots, x_n) + s(a_1, \ldots, a_n). \tag{3}$$

The desired dependence should be continuous. The values x_i are usually only approximately known – they usually come from measurements, and measurement are always approximate. The actual values x_i^{act} of these quantities are, in general, slightly different from the measurement results x_i that we use to predict y. It is therefore reasonable to require that when we apply the regression function $f(x_1, \ldots, x_n)$ to the (approximate) measurement results, then the predicted value $f(x_1, \ldots, x_n)$ should be close to the prediction $f(x_1^{\text{act}}, \ldots, x_n^{\text{act}})$ based on the actual values x_i^{act}.

In other words, if the inputs to the function $f(x_1, \ldots, x_n)$ change slightly, the output should also change slightly. In precise terms, this means that the function $f(x_1, \ldots, x_n)$ should be *continuous.*

Now that we have argued that the regression function be shift-invariant and continuous, we can explain why linear regression is empirically successful.

Proposition. *Every shift-invariant continuous function $f(x_1, \ldots, x_n)$ is linear, i.e., has the form*

$$f(x_1, \ldots, x_n) = \beta_0 + \sum_{i=1}^{n} \beta_i \cdot x_i \tag{4}$$

for appropriate coefficients β_i.

Proof. Substituting the values $x_i = 0$ into the equality (3), we conclude that

$$f(a_1, \ldots, a_n) = f(0, \ldots, 0) + s(a_1, \ldots, a_n) \tag{5}$$

for all possible tuples $(a_1, \ldots, a_n)$. In particular, this is true for the tuples $(x_1, \ldots, x_n)$ and $(x_1 + a_1, \ldots, x_n + a_n)$, i.e., we have:

$$f(x_1, \ldots, x_n) = f(0, \ldots, 0) + s(x_1, \ldots, x_n) \tag{6}$$

and

$$f(x_1 + a_1, \ldots, x_n + a_n) = f(0, \ldots, 0) + s(x_1 + a_1, \ldots, x_n + a_n). \tag{7}$$

Substituting the expressions (6) and (7) into the equality (3) and cancelling the common term $f(0, \ldots, 0)$ in both sides of the resulting equality, we conclude that

$$s(x_1 + a_1, \ldots, x_n + a_n) = s(x_1, \ldots, x_n) + s(a_1, \ldots, a_n) \tag{8}$$

Such functions are known as *additive.*

From the equality (5), we conclude that

$$s(a_1, \ldots, a_n) = f(a_1, \ldots, a_n) - f(0, \ldots, 0). \tag{10}$$

Since the function $f(a_1, \ldots, a_n)$ is continuous, we can conclude that the function $s(a_1, \ldots, a_n)$ is continuous as well. So, the function $s(x_1, \ldots, x_n)$ is continuous and additive.

It is known (see, e.g., [1]) that every continuous additive function is a homogeneous linear function, i.e., has the form

$$s(x_1, \ldots, x_n) = \sum_{i=1}^{n} \beta_i \cdot x_i \tag{11}$$

for some real numbers β_i. Thus, from the formula (5), we can conclude that

$$f(x_1, \ldots, x_n) = \beta_0 + s(x_1, \ldots, x_n) = \beta_0 + \sum_{i=1}^{n} \beta_i \cdot x_i, \tag{12}$$

where we denoted $\beta_0 \stackrel{\text{def}}{=} f(0, \ldots, 0)$.

The proposition is proven.

Comment. It is easy to see that, vice versa, every linear function (4) is continuous and shift-invariant: namely, for each such function, we have:

$$f(x_1 + a_1, \ldots, x_n + a_n) = \beta_0 + \sum_{i=1}^{n} \beta_i \cdot (x_i + a_i) =$$

$$\beta_0 + \sum_{i=1}^{n} \beta_i \cdot x_i + \sum_{i=1}^{n} \beta_i \cdot a_i = f(x_1, \ldots, x_n) + s(a_1, \ldots, a_n),$$

where we denoted $s(a_1, \ldots, a_n) \stackrel{\text{def}}{=} \sum\limits_{i=1} \beta_i \cdot a_i$.

3 Case of Interval Uncertainty

Description of the case. In the previous section, we have shown that when we try to predict a numerical value y, then it is often beneficial to use linear regression. As we have mentioned, predicting a single value y is often not enough:

- in addition to the approximate value y,
- it is also necessary to know how accurate is this approximate value, i.e., which values y are possible.

Because of this necessity, in this section, we consider a situation, in which, for each inputs $x_1, \ldots, x_n$:

- instead of predicting a *single* value y,
- we would like to predict the *interval* $\left[\underline{y}(x_1, \ldots, x_n), \overline{y}(x_1, \ldots, x_n)\right]$ of all the values of y which are possible for given inputs $x_1, \ldots, x_n$.

Why linear regression. In the case of interval uncertainty, instead of a *single* regression function $y = f(x_1, \ldots, x_n)$, we have *two* regression functions:

- a regression function $\underline{y} = \underline{f}(x_1, \ldots, x_n)$ that describes the lower endpoint of the desired interval, and
- a regression function $\overline{y} = \overline{f}(x_1, \ldots, x_n)$ that describes the upper endpoint of the desired interval.

It is reasonable to require that each of these two functions is

- continuous, and
- does not change if we change the starting points for measuring the inputs – i.e., is *shift-invariant* (in the sense of the above Definition).

Thus, due to our proposition, each of these functions is linear, i.e., we have

$$\underline{f}(x_1, \ldots, x_n) = \underline{\beta}_0 + \sum_{i=1}^{n} \underline{\beta}_i \cdot x_i \tag{13}$$

and

$$\overline{f}(x_1, \ldots, x_n) = \overline{\beta}_0 + \sum_{i=1}^{n} \overline{\beta}_i \cdot x_i. \tag{14}$$

for appropriate values $\underline{\beta}_i$ and $\overline{\beta}_i$.

4 Case of Probabilistic Uncertainty

Description of the case. We consider the situation in which for each combination of inputs $x_1, \ldots, x_n$, in addition to the set of possible values of y, we also know the probability of different possible values of y. In other words, for each

tuple of inputs $x_1, \ldots, x_n$, we know the corresponding (conditional) probability distribution on the set of all possible values y.

What is the relation between a probability distribution and the set of possible values? From the previous section, we know how to describe regression in the case of interval uncertainty. We would like to extend this description to the case of probabilistic uncertainty. To be able to do that, let us recall the usual relation between the probability distribution and the set of possible values.

This relation can be best illustrated on the example of the most frequently used probability distribution – the normal (Gaussian) distribution, with the probability density

$$\rho(y) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right). \tag{15}$$

The ubiquity of this distribution comes from the Central Limit Theorem, according to which the probability distribution caused by the joint effect of many small independent random factors is close to Gaussian; see, e.g., [7].

From the purely mathematical viewpoint, a normally distributed random variable can attain any real value. Indeed, the corresponding probability density is always positive, and thus, there is always a non-zero probability that we will have a value far away from the mean μ.

However, values which are too far from the mean have such a low probability that from the practical viewpoint, they are usually considered to be impossible. It is well know that:

- with probability 95%, the normally distributed random variable y is inside the two-sigma interval $[\mu - 2\sigma, \mu + 2\sigma]$;
- with probability 99.9%, y is inside the three-sigma interval $[\mu - 3\sigma, \mu + 3\sigma]$, and
- with probability $1 - 10^{-8}$, y is inside the six-sigma interval $[\mu - 6\sigma, \mu + 6\sigma]$.

In general, a usual way to transform a probability distribution into an interval of practically possible values is to use a *confidence interval*, i.e., an interval for which the probability to be outside this interval is equal to some pre-defined small value p_0. A usual way to select such an interval is to select the bounds $\underline{y}$ and $\overline{y}$ for which:

- the probability to have y smaller than $\underline{y}$ is equal to $\frac{p_0}{2}$, and
- the probability to have y larger than $\overline{y}$ is equal to $\frac{p_0}{2}$.

One can easily see that:

- the lower endpoint $\underline{y}$ of this confidence interval is the quantile $y_{p_0/2}$, and
- the upper endpoint $\underline{y}$ of this confidence interval is the quantile $y_{1-p_0/2}$.

Depending on the problem, we can have different probabilities p_0, so we can have all possible quantiles.

Conclusion: why linear quantile regression is empirically successful. For each combination of inputs $x_1, \ldots, x_n$, based on the related (conditional) probability distribution of y, we can form the interval of practically possible values, in which both endpoints are quantiles y_p corresponding to some values p.

In the previous section, we have shown that reasonable requirements imply that each of these endpoints is a linear function of the inputs. Thus, we conclude that for each p, we have

$$y_p \approx \beta_{0,p} + \sum_{i=1}^{n} \beta_{i,p} \cdot x_i, \tag{2}$$

for appropriate values $\beta_{i,p}$.

This is exactly the formula for linear quantile regression. Thus, we have provided the desired first-principles for linear quantile regression formulas. The existence of such a justification can explain why linear quantile regression is empirically successful.

Acknowledgments

We acknowledge the partial support of the Center of Excellence in Econometrics, Faculty of Economics, Chiang Mai University, Thailand. This work was also supported in part by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721.

References

1. J. Aczél and J. Dhombres, *Functional Equations in Several Variables*, Cambridge University Press, 2008.
2. J. D. Angrist and J.-S. Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, Princeton, New Jersey, 2009.
3. C. Davino, M. Furno, and D. Vistocco, *Quantile Regression: Theory and Applications*, Wiley, New York, 2013.
4. B. Fitzenberger, R. Koenker, and J. A. F. Machado (editors), *Economic Applications of Quantile Regression*, Physika-Verlag, Heidelberg, 2002.
5. L. Hao and D. Q. Naiman, *Quantile Regression*, SAGE Publications, Thousands Oaks, California, 2007.
6. R. Koenker, *Quantile Regression*, Cambridge University Press, New York, 2005.
7. D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC, Boca Raton, Florida, 2011.