Quantile Regression Model

Abstract

This methodological framework was developed to expand the range of models used to improve the accuracy of estimates of the fundamental value of housing prices in assessing conditions in Uzbekistan's real estate market. In this context, the framework describes the quantile regression model, which enables more robust estimation of the effects of explanatory variables when the distribution of relative errors deviates from normality and the observed values exhibit substantial variation over the sample period. In particular, the parameters of the quantile regression model are estimated by minimizing the weighted sum of absolute deviations between the observed values and the conditional quantile function.

Keywords: absolute, quantile, relative error, regression, simplex table, variable, conditional mean.

Prepared by the Financial Stability Department of the Central Bank of Uzbekistan.

If you have any comments, please send them to udjumanazarov@cbu.uz.

Quantile Regression Model

The purpose of regression analysis is to demonstrate the relationship between dependent and independent variables. The classic linear regression model estimates the conditional mean[1]. The process of modeling the conditional mean involves applying the method of OLS, which simplifies the computation. When conducting regression analysis using the method of OLS, nonnormality of residuals and the presence of periods with sharp variations during the observation period may lead to a decrease in the accuracy of the results, resulting in higher residuals. Therefore, applying the quantile regression model rather than OLS model in regression analysis increases the accuracy of the results[2].

The parameters of the quantile regression model are estimated by minimizing the weighted sum of the absolute values of residuals when estimating the unknown parameters of the conditional quantile function. The quantile regression model evaluates variations at conditional quantiles. Conditional median regression is a special case of quantile regression, in which the conditional median corresponds to the 50th quantile function. In general, it is possible to use any other quantiles to characterize non-central positions of the distribution and conduct analysis of the relationship between dependent and independent variables at different quantiles.

The equation of the quantile regression model does not differ from the simple linear regression equation. The regression equation of the quantile regression model takes the following form:

$$y_i = x_i\beta_q + \varepsilon_i$$

Where,
y_i – $nx1$ vector of dependent variables;
x_i – nxk matrix of independent variables;
β_i – $kx1$ vector of unknown parameters;
ε_i – $nx1$ vector of residuals;
q – quantile.

[1] Hastie, T., Tibshirani, R., & Friedman, J. (2017). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics.

[2] Cameron, A., & Trivedi, P. (2005). Microeconometrics: Methods and Applications. Cambridge University Press.

When estimating the parameters of the quantile regression model, different weights are assigned to positive and negative errors. The parameters of the quantile regression are obtained by minimizing the weighted sum of the absolute values of residuals $\left(Q(\beta_q)\right)$, and take the following form:

$$Q(\beta_q) = \sum_{i:y_i \geq x_i\beta_q} q\left|y_i - x_i\beta_q\right| + \sum_{i:y_i < x_i\beta_q} (1-q)\left|y_i - x_i\beta_q\right|.$$

Determining the unknown parameters in the equation requires using the linear programming type of an optimization model in mathematical programming.

Solving linear programming problems requires first defining the objective function and the boundary conditions of the model. In the conditional quantile function of quantile regression, positive and negative errors are assigned asymmetric weights[3]:

$$Q(\beta_q) = \sum_{i:y_i \geq x_i\beta_q} q\left|y_i - x_i\beta_q\right| + \sum_{i:y_i < x_i\beta_q} (1-q)\left|y_i - x_i\beta_q\right|.$$

The optimal values of the unknown parameters are determined by minimizing this sum. The objective function of linear programming $\left(Q(\beta_q)\right)$ is obtained as follows[4]:

$$Q(\beta_q) = \sum_{i:y_i \geq x_i\beta_q} q s_i^+ + \sum_{i:y_i < x_i\beta_q} (1-q) s_i^-.$$

In the equation, s_i^+ and s_i^- are the positive and negative errors, respectively, and are obtained by the following formulas:

$$s_i^+ = y_i - x_i\hat{\beta}_q$$

[3] Asymmetric weight – different types of weights that are not equal to each other.

[4] Abdullahi, I. (2015). Analysis of quantile regression as alternative to ordinary least squares regression. Department of Mathematics Ahmadu Bello University.

$$s_i^- = -(y_i - x_i\hat{\beta}_q)$$

$s_i^+ \geq 0, s_i^- \geq 0.$

Where, $\hat{\beta}_q$ is the parameter estimated by the model.

Based on the boundary conditions of the linear programming model, the following equation is obtained:

$$y_i = x_i\hat{\beta}_q + s_i^+ - s_i^-$$

Under this boundary condition, the positive and negative errors cannot be non-zero at the same time. In this case, one or both parameters may take a value of zero.

The boundary conditions remain constant across all quantiles and only the objective function changes. Minimizing the function $Q(\beta_q)$ is considered equivalent to the problem of minimizing the objective function in linear programming. Therefore, it is sufficient to find the optimal value of the parameters β_q in the linear programming problem.

The simplex method for determining the optimal values of unknown parameters uses iteratively replacing processes. The expansion of the data series increases the accuracy of the regression parameters. In this case, sum $\left(Q(\beta_q)\right)$ in the objective function of the linear programming problem is minimized the under the boundary condition $Ax = B$.

Simplex method

The simplex method is one of the main methods for solving linear programming problems. By moving step-by-step toward optimality, it finds the optimal value of the variables $x_1, x_2, \ldots, x_n$, which result in the objective function reaching its maximum or minimum value[5].

A linear programming problem that yields a maximum value is expressed as follows:

$$\sum_{j=1}^{n} a_{ij}x_j \leq y_i, \ (i = 1,2, \ldots, m)$$

[5] Winston, W. (2004). Operations Research: Applications and Algorithms. Indiana University.

$$x_j \geq 0, (j = 1,2, \dots, n)$$
$$Z = \sum_{j=1}^{n} c_{ij} x_j \ \rightarrow max.$$

A simplex tableau is constructed for solving the problem. In constructing the simplex tableau, the given problem is first converted into canonical form. In this case, all the boundary conditions are transformed into equalities in order to obtain the canonical form. Inequalities are converted into equalities by adding slack variables to the inequalities. As a result of adding slack variables, the values of $n - m$ variables become zero, and the remaining m variables' unique values are found by determining the optimal values of the equations. After adding the slack variables, the inequalities are converted into the following equalities:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \dots + a_{1s}x_s + \dots + a_{1n}x_n + x_{n+1} = y_1 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2s}x_s + \dots + a_{2n}x_n + x_{n+2} = y_2 \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{ms}x_s + \dots + a_{mn}x_n + x_{n+m} = y_m \end{cases}$$

$x_j \ \geq 0 \quad (i = 1, 2, \dots, m) \quad (j = 1, 2, \dots, n)$

$$Z = c_1x_1 + \ c_2x_2 + \dots + \ c_nx_n \ \longrightarrow max.$$

Where,
x_{n+i} – slack variables.

In order to distinguish the slack variables from the other variables, they are denoted as e_1, e_2, … , e_m, respectively. The expression on the right side of the objective function equation is transferred to the left side of the equation:

$$Z - c_1x_1 - \ c_2x_2 - \dots - \ c_nx_n = 0.$$

After the changes, the simplex tableau can be constructed in the following form:

	x_1	x_2	…	x_s	…	x_n	b_i
e_1	a_{11}	a_{12}	…	a_{1s}	…	a_{1n}	b_1
e_2	a_{21}	a_{22}	…	a_{2s}	…	a_{2n}	b_2
…	…	…	…	. . .	…	…	…
e_r	a_{r1}	a_{r2}	…	a_{rs}	…	a_{rn}	b_r
…	…	…	…	. . .	…	…	…
e_m	a_{m1}	a_{m2}	…	a_{ms}	…	a_{mn}	b_m
Z	$-c_1$	$-c_2$	…	$-c_s$	…	$-c_n$	0

All b_i constants are required to be positive. If a constant is not positive, both sides of the equation are multiplied by -1 to convert it into a positive form.

As the given problem has a maximization problem, the optimization process continues until all parameters in the Z row take non-negative values.

Initially, a basic feasible solution is determined. In this case, the first column of slack variables is considered the basic solution column, and all variables not located in this column are assigned a value of zero. In the simplex tableau, the basic feasible solutions e_i are determined:

$$e_i = b_i \quad \text{and} \quad x_j = 0$$

This basic feasible solution is considered optimal if all the coefficients in the Z row are nonnegative. If any of these coefficients is negative, the next simplex tableau is constructed. Initially, the pivot element is identified at the intersection of the pivot row and pivot column. The pivot column is determined by choosing the negative parameter in the objective function row that has the greatest absolute value.

For instance, if the element with the largest absolute value among the negative elements is c_s , this column s is considered as the pivot column. Once the pivot column is identified, the pivot row is determined. To find the pivot row, elements of the constants column is divided by corresponding pivot column elements, and the smallest positive value is selected.

$$\min\left\{\frac{b_1}{a_{1s}};\frac{b_2}{a_{2s}};\dots;\frac{b_r}{a_{rs}};\dots;\frac{b_m}{a_{ms}}\right\}.$$

If the smallest positive of these ratios is $\frac{b_r}{a_{rs}}$, then the row containing the element a_{rs} is called the pivot row, and the element a_{rs} itself is the pivot element. After determining the pivot element, the new simplex tableau is filled gradually.

The construction of the second simplex tableau is carried out in several steps. Initially, the positions of the pivot row and pivot column variables (x_s and e_r) are interchanged. Then, the elements of the pivot column are divided by the pivot element and written to the new tableau with negative values as follows:

$$a'_{is} = -\frac{a_{is}}{a_{rs}} \qquad (i \neq r)$$

The elements of the pivot row are also divided by the pivot element and written to the new tableau:

$$a'_{rs} = \frac{a_{rj}}{a_{rs}} \qquad (j \neq s)$$

The pivot element is replaced by its reciprocal.

$$a'_{is} = \frac{1}{a_{rs}}$$

The remaining elements of the new simplex tableau are calculated using the following formula:

$$a'_{ij} = \frac{a_{ij}a_{rs} - a_{is}a_{rj}}{a_{rs}} = a_{ij} - \frac{a_{is}a_{rj}}{a_{rs}}$$
$$(j \neq s), \quad (i \neq r).$$

Where,
a_{ij} – the corresponding element of a'_{ij} in the previous tableau;

a_{is} – the element at the intersection of the column of the pivot element a_{rs} and the row containing a_{ij};
a_{rj} – the element at the intersection of the row of the pivot element a_{rs} and the column containing a_{ij};
a_{rs} – pivot element.

If all elements of the basic solution in the *Z* row of simplex tableau, except for the constant, are nonnegative, then this basic solution is unique and optimal. If at least one element in the *Z* row is negative, the optimization process is continued according to the procedure described above until the optimal value is derived:

$$Z = \sum_{j=1}^{n} c_j \, x_j \longrightarrow max.$$

Additionally, a maximization problem can be converted into a minimization problem by multiplying it by -1:

$$-Z = \sum_{j=1}^{n} - c_j \, x_j \longrightarrow min.$$

The optimal parameters obtained using the simplex method of linear programming are considered the optimal estimates of the unknown parameters of quantile regression. Using these parameters, the estimated values of the quantile regression are determined.

The 50th quantile of the following quantile regression equation was used to determine the fundamental price per square meter of houses:

$$y_t = \beta_{1,q} y_{t-1} + \beta_{2,q} x_{1,t} + \beta_{3,q} x_{2,t} + \varepsilon_t \,.$$

Where,
y_t – market price of houses in period t;
y_{t-1} – market price of houses in period $t-1$;
$x_{1,t}$ – annual growth rate of household loans;
$x_{2,t}$ – annual growth rate of household loans;

$\beta_{1,q}, \beta_{2,q}, \beta_{1,q}$ – unknown coefficients;
q – quantile;
ε_t – residual.