

Linear Regression for Quantity Quantiles ^(*)

La Regressione Lineare per i Quantili di Quantità

Paolo Radaelli and Michele Zenga

Dipartimento di Metodi Quantitativi per le Scienze Economiche ed Aziendali
Università degli Studi di Milano-Bicocca
e-mail: paolo.radaelli@unimib.it

Riassunto: La regressione dei quantili introdotta da Koenker e Basset (1978) (cfr. Koenker, 2005) può essere vista come estensione al contesto della regressione della definizione del quantile campionario come soluzione ad un problema di minimo. In questo lavoro si mostra come, seguendo una procedura del tutto analoga, sia possibile definire anche i quantili di quantità di una variabile non negativa come soluzione ad un problema di minimo e come i risultati ottenuti possano essere estesi nell'ambito della regressione lineare.

Keywords: quantile regression, quantity quantile.

1. Introduction

The crucial feature in quantile regression, as introduced in Koenker and Basset (1978), is the definition of the θ th quantile ζ_θ of the real valued random variable (r.v.) Y with distribution function:

$$F_Y(y) = \int_0^y f_Y(t) dt \quad (1)$$

as the solution of the minimization problem:

$$\min_{c \in \mathbb{R}} \mathbb{E} [\ell_\theta(Y - c)] \quad (2)$$

where $\ell_\theta(u)$ is the *asymmetric absolute loss function*:

$$\ell_\theta(u) = [\theta I\{u > 0\} + (1 - \theta) I\{u \leq 0\}] |u| = [\theta - I\{u \leq 0\}] u \quad (3)$$

with $0 < \theta < 1$ and I denoting the indicator function.

Let $Y_1, \dots, Y_i, \dots, Y_n$ be a random sample from Y , the natural estimator for ζ_θ is the corresponding sample quantile $\hat{\zeta}_\theta$ that can be obtained simply by replacing the distribution function F_Y in (2) with the empirical distribution function:

$$F_n(y) = n^{-1} \sum_{i=1}^n I\{y_i \leq y\} \quad (4)$$

(*) Work partially supported by Fondo d'Ateneo per la Ricerca anno 2005 - Università degli Studi di Milano-Bicocca

obtaining the minimization problem:

$$\begin{aligned} \min_{\zeta_\theta \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n \ell_\theta(y_i - \zeta_\theta) &\equiv \min_{\zeta_\theta \in \mathbb{R}} \frac{1}{n} \left[\sum_{[i: y_i > \zeta_\theta]} \theta |y_i - \zeta_\theta| + \sum_{[i: y_i \leq \zeta_\theta]} |1 - \theta| |y_i - \zeta_\theta| \right] \\ &\equiv \min_{\zeta_\theta \in \mathbb{R}} \frac{1}{n} \left[\sum_{[i: y_i > \zeta_\theta]} \theta (y_i - \zeta_\theta) + \sum_{[i: y_i \leq \zeta_\theta]} (\theta - 1) (y_i - \zeta_\theta) \right] \end{aligned} \quad (5)$$

The point of departure of quantile regression is thus the definition of the θ th sample quantile as the solution of a minimization problem instead of the usual procedure that implies the ordering of the sample observations.

As pointed out in Koenker and Hallock (2001) the solution to problem (5) is an estimate of the unconditional θ th quantile of Y . Suppose now to have p explanatory variables $X_1, \dots, X_p$ and that the conditional quantile function of Y is linear and given by:

$$B_Y(\theta|\mathbf{x}) = \mathbf{x}^T \boldsymbol{\beta}_\theta \quad 0 < \theta < 1. \quad (6)$$

In the quantile regression (linear) model the unknown parameters $\boldsymbol{\beta}_\theta$ in the conditional quantile function are estimated simply by replacing the scalar ζ_θ in (5) by the function $\mathbf{x}^T \boldsymbol{\beta}_\theta$. Thus the θ th regression quantile is defined as any solution to the following minimization that can be solved by linear programming:

$$\begin{aligned} \min_{\boldsymbol{\beta}_\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \ell_\theta(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_\theta) \\ \equiv \min_{\boldsymbol{\beta}_\theta \in \mathbb{R}^p} \frac{1}{n} \left[\sum_{[i: y_i > \mathbf{x}_i^T \boldsymbol{\beta}_\theta]} \theta (y_i - \mathbf{x}_i^T \boldsymbol{\beta}_\theta) + \sum_{[i: y_i \leq \mathbf{x}_i^T \boldsymbol{\beta}_\theta]} (\theta - 1) (y_i - \mathbf{x}_i^T \boldsymbol{\beta}_\theta) \right] \end{aligned} \quad (7)$$

where $\mathbf{y} = [y_1, \dots, y_n]$ is a vector of responses on the random variable Y and $X = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T$ is the known $n \times p$ matrix of the regressors.

2. Quantity quantiles

When one face with a nonnegative random variable, the income for instance, it is useful to consider not only the quantiles as defined in the previous section but the quantity quantiles as well.

From now on Y is supposed to be a nonnegative r.v. with density function f_Y and with finite and strictly positive mean value. The distribution function of Y is given by (1) and the so-called *first incomplete moment* (or first-moment distribution) is:

$$Q_Y(y) = \int_0^y \frac{t}{\mu} f_Y(t) dt. \quad (8)$$

Thus, if Y represent the income, $F_Y(y)$ gives the fraction of the population with income no greater than y while $Q_Y(y)$ gives the share of the total income accruing to the

population with income no greater than y . In this framework it is possible to define the θ th quantile ζ_θ and the θ th *quantity quantile* η_θ :

$$\zeta_\theta = \inf \{y \in \mathfrak{R} \mid F_Y(y) \geq \theta\} \quad 0 < \theta < 1 \quad (9)$$

$$\eta_\theta = \inf \{y \in \mathfrak{R} \mid Q_Y(y) \geq \theta\} \quad 0 < \theta < 1. \quad (10)$$

This approach has been widely used in the study of income and wealth distribution and many concentration measures have been derived from the comparison of (1) and (8), see Zenga (1990) for a short review, and from the comparison of (9) and (10), see Zenga (1984) and Kleiber and Kotz (2003).

As the quantile can be defined as any solution to the minimization problem that makes use of the loss function (3), so the quantity quantile η_θ can be defined as any solution to a minimization problem that attaches a suitable weight to the loss function (3).

In particular let W be a r.v. with the same support as Y and density function $q_W(y) = f_Y(y) \frac{y}{\mu}$. The distribution function of W is given by:

$$F_W(y) = \int_0^y q_W(t) dt = \int_0^y \frac{t}{\mu} f_Y(t) dt = Q_Y(y)$$

which is the first incomplete moment (8) of Y , therefore the θ -th quantile of the r.v. W is the θ -th quantity quantile of the r.v. Y .

It is now straightforward to define the quantiles of the r.v. W in the same way as we showed in the previous section. Thus, according to (2), the θ th quantile of W can be defined as the solution to the minimization problem:

$$\min_{c \in \mathfrak{R}} \mathbb{E} [\ell_\theta(W - c)]. \quad (11)$$

The expected loss in (11) can be rewritten as follows:

$$\mathbb{E} [\ell_\theta(W - c)] = \int_{-\infty}^{\infty} \ell_\theta(y - c) \frac{y}{\mu} f_Y(y) dy = \mathbb{E} \left[\ell_\theta(Y - c) \frac{Y}{\mu} \right]$$

i.e. the θ -th quantity quantile of the r.v. Y can be obtained by minimizing the expect loss in which the distances $|y - c|$ are weighted not only according to the asymmetric absolute loss function (3) but with the additional nonnegative weight y/μ .

Consider now n sorted observations $0 \leq y_{(1)} \leq \dots \leq y_{(i)} \leq \dots \leq y_{(n)} > 0$, obtained from a random sample $Y_1, \dots, Y_i, \dots, Y_n$ from Y and let $T = \sum y_i > 0$ and $\bar{y}$ denote, respectively, the total amount and the arithmetic mean of the observations.

The natural estimator for η_θ is the corresponding sample quantity quantile:

$$\hat{\eta}_\theta = \inf \{y_{(i)} : \hat{Q}(y_{(i)}) \geq \theta\}$$

where:

$$\hat{Q}(b) = \frac{1}{n \bar{y}} \sum_{j: y_j \leq b} y_j = \frac{1}{T} \sum_{j: y_j \leq b} y_j.$$

Thus traditionally, in order to obtain the θ -th sample quantity quantile, one should sum the sorted values $y_{(i)}$ until at least the share θ of the total T is reached. Nevertheless by replacing the empirical distribution function (4) of Y with:

$$G_n(y) = n^{-1} \sum_{i=1}^n I \{y_i \leq y\} \frac{y_i}{\bar{y}} = \hat{Q}(y) \quad (12)$$

the θ -th sample quantity quantile can be obtained as any solution to the minimization problem:

$$\min_{b \in \mathbb{R}} \sum_{i=1}^n \ell_{\theta} (y_i - b) \frac{y_i}{n\bar{y}} \equiv \min_{b \in \mathbb{R}} \left[\sum_{[i: y_i > b]} \theta (y_i - b) \frac{y_i}{n\bar{y}} + \sum_{[i: y_i \leq b]} (\theta - 1) (y_i - b) \frac{y_i}{n\bar{y}} \right] \quad (13)$$

It can be easily shown that the expression within square brackets in (13) reaches its minimum for b that satisfies $\hat{Q}(b) = \theta$, i.e. $b = \hat{\eta}_{\theta}$.

3. Linear regression for quantity quantiles

In section 1 has been shown that the (linear) quantile regression model can be viewed as an extension to the regression context of the definition of the sample quantile as the solution to a problem of minimum. Considering the results obtained in the previous section, it is straightforward to define the coefficients β_{θ} of the hyperplane for the θ th conditional quantity quantile as the solution of the minimization problem obtained replacing (4) with (12) in (7):

$$\begin{aligned} & \min_{\beta_{\theta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \ell_{\theta} (y_i - \mathbf{x}_i^T \beta_{\theta}) \frac{y_i}{\bar{y}} \\ & \equiv \min_{\beta_{\theta} \in \mathbb{R}^p} \frac{1}{n} \left[\sum_{[i: y_i > \mathbf{x}_i^T \beta_{\theta}]} \theta (y_i - \mathbf{x}_i^T \beta_{\theta}) \frac{y_i}{\bar{y}} + \sum_{[i: y_i \leq \mathbf{x}_i^T \beta_{\theta}]} (\theta - 1) (y_i - \mathbf{x}_i^T \beta_{\theta}) \frac{y_i}{\bar{y}} \right] \end{aligned} \quad (14)$$

The problem (14) has a linear programming formulation.

References

- Kleiber C., Kotz S. (2003) *Statistical Size Distributions in Economics and Actuarial Sciences*, Wiley Series in Probability and Statistics. John Wiley & Sons.
- Koenker R. (2005) *Quantile Regression*, Cambridge University Press.
- Koenker R., Basset G. J. (1978) Regression Quantiles, *Econometrica*, 46(1), 33–50.
- Koenker R., Hallock K. (2001) Quantile Regression, *Journal of Economic Perspectives*, 15(4), 143–156.
- Zenga M. (1984) Proposta per un Indice di Concentrazione Basato sui Rapporti tra Quantili di Popolazione e Quantili di Reddito, *Giornale degli Economisti e Annali di Economia*, 43, 301–326.
- Zenga M. (1990) Income and Wealth Distribution, Inequality and Poverty, in: *Proceedings of the Second International Conference on Income Distribution by Size: Generation, Distribution, Measurement and Applications, held at the University of Pavia, Italy, September 28-30, 1989*, Dagum C., Zenga M. (Eds.), Springer-Verlag