Estimation of scale function for linear quantile regression

Ruizhe Huang - 1006444331

2022-08-12

Abstract

The conventional least-square linear regression model gives only the mean effects of independent variables on the response variable. However, the quantile regression model can give the effects of independent variables on the response variable at any desired percentile level. Moreover, the linear regression model strongly relies on the assumptions of homoscedasticity and normality. Nevertheless, the quantile regression model can deal with heteroscedasticity. The heteroscedasticity form of the quantile regression model was discussed in this study. The heteroscedasticity components were modelled by the product of a scale function and random error. This study simulated the data with the given true parameters and scale function. In this study, the simulation replicates 5000 times to estimate the parameters. After that, the average estimates of the quantile model parameters and scale function were computed and compared to the given true parameters. As a result, the average estimates of the parameters for quantile regression and scale function in the 5000 simulations are fairly accurate, but the estimates of the scale function in every individual simulation may not be as precise, especially the estimated slope of the scale function.

Keywords: Quantile regression, heteroscedasticity, scale function, simulation study

Introduction

The ordinary least square regression model was widely used among investigators in the research field. However, it only gives the conditional mean estimates of the response variable. Conversely, the quantile regression model allows the analysis of the comprehensive conditional distribution effects of the response variable. See Koenker (2005) for a comprehensive review. In addition, quantile regression tends to resist the influence of outlier observations. It can also deal with heteroscedasticity. Hao and Naiman (2007) proposed that the median may be more appropriate than the mean to represent central tendency when the distribution is skewed.

In this study, we are going to focus on the estimates of 0.5 quantile, that is, the median. The linear quantile regression model with heteroscedasticity form was considered in this study. We apply the adaptive approach, carried out by Selvaratnam et al., to the linear quantile regression scale function estimation process. See Selvaratnam et al. (2021) for the recent study of the robust designs for nonlinear quantile regression.

The data were simulated with true parameters of quantile regression and scale function. The simulation process was repeated 5000 times for a more accurate result. In addition, we would compare different scenarios of true parameters of quantile regression and scale function. The data section briefly describes how data was simulated and states the general model with true parameters. Also, we have the data summary to quantify the data. The method section will give the linear quantile regression model formula with heteroscedasticity and how we estimate the parameters and scale function. After that, we will get the average estimate of the quantile model parameters in the result section. In addition, we also estimate the parameters of the scale function by using the adaptive design mentioned above. In the end, we will compare these estimates with the true values to check the accuracy. As a result, the average estimates of the parameters for quantile regression and scale function in the 5000 simulations are fairly accurate, but the estimates of the scale function in every individual simulation may not be as precise, especially the estimated slope of the scale function.

Data

Simulation process

The datasets were simulated. The parameters of quantile regression were preset to three different scenarios. We generate the response variable y based on fixed original x, where x takes 70 distinct values between 40 and 600. Also, we replicate each x for 6 times. Therefore, the total sample size is $70 \times 6 = 420$. Note, for each x, we have a distinct response y. We also set up three different scale functions ($\sigma(x)$). After that, we create a new set of x values to evaluate the estimated scale function. The new x takes 60 distinct values between 40 and 600. Ultimately, we replicate the entire simulation process 5000 times to get the average estimates of the quantile regression model and scale function parameters.

Model description

General model:

$$y_i = \beta_0 + x_i\beta_1 + \sigma(x_i)u_i$$

where β_0 and β_1 represents the true parameters of the quantile regression; $\sigma(x_i)$ represents the true scale function; random error u_i are independent and identically distributed (iid).

Data summary

In our simulated data, we have the original x and the new x for scale function, both ranging between 40 and 600. Also, both of them have the same mean and median at 320. The standard deviation of original x is 164.1798, slightly smaller than the standard deviation of new x is 165.7624. The IQR of the original x is $462 - 178 = 284$, which is slightly larger than the IQR of the new x is $460 - 180 = 280$.

Method

Since any real-valued random variable X may be characterized by its continuous distribution function.

$$F(x) = P(X \leq x)$$

For any $0 < \tau < 1$, we have we have the τ^{th} quantile of X given by:

$$F^{-1}(\tau) = \inf\{x : F(x) \geq \tau\}$$

Now, we will specify the heteroscedasticity form of quantile regression model:

$$y_i = \beta_0 + x_i\beta_1 + \sigma\left(x_i\right)u_i$$

where $\sigma(x)$ is the unknown parameter and u_i are iid.
So the conditional quantile functions of y:

$$Q_y(\tau \mid x) = \beta_0 + x\beta_1 + \sigma(x)F^{-1}(\tau)$$

Where $\tau \sim Uniform(0, 1)$, F_u is the common distribution function of the errors.
Then $\beta(\tau)$ can be estimated by:

$$\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau \left(y_i - x_i^\top \beta\right)$$

Where $\rho_\tau(u)$ is the check loss function which defined as:

$$\rho_\tau(u) = u(\tau - I(u < 0))$$

Equivalently:

$$\rho_\tau(u) = \tau \max(u, 0) + (1 - \tau) \max(-u, 0)$$

(Koenker, 2005)
To estimate the scale function $\sigma(x)$, we will use Gaussian kernel estimates:

$$s_n^2(x) = \sum_{i=1}^{n} w(x - x_{(i)})(Y_i - \hat{Y}_i|\hat{\beta}_\tau)^2$$

where $w\left(t_i\right) = \frac{\exp\left\{-\frac{1}{2}\left(\frac{t_i}{h}\right)^2\right\}}{\sum_{i=1}^{n}\exp\left\{-\frac{1}{2}\left(\frac{t_i}{h}\right)^2\right\}}$, with bandwidth $h = 10$.
Then fit the least square regression $\hat{\theta_0} + \hat{\theta_1}x$ to the data $\left\{s_n\left(x_{(i)}\right)\right\}_{i=1}^{n}$, and find $\hat{\sigma}(x_i) = \hat{\theta_0} + \hat{\theta_1}x_i$ (Selvaratnam et al., 2021).

Result

Quantile regression estimates

Table 1: Estimates of three quantile regression models

scenario	β_0	β_1	$\hat{\beta}_0$	$\hat{\beta}_1$	$\sigma(x)$	u_i	% of CI captures β_0	% of CI captures β_1
s1	2	4	1.787247	3.99904	2x	$\mathcal{N}(0,1)$	83.06%	79.26%
s2	5	0.3	4.937843	0.2998001	0.5x	$\mathcal{N}(0,1)$	83.06%	79.20%
s3	0.8	10	0.7607623	9.999879	0.3x+2	$\mathcal{N}(0,1)$	83.58%	79.76%

From table 1, we can compare the true parameters of quantile regression to the estimates. The estimates are the average estimates of 5000 simulations. All the estimated parameters of quantile regression are close compared to the true values. Nevertheless, around 83% of confidence intervals with the significant level $\alpha = 0.05$ capture the true intercept β_0. Around 79% of confidence intervals with significant level $\alpha = 0.05$ capture the true slope β_1.

Scale function estimates

Table 2: Estimates of three scale functions

scenario	$\sigma(x)$	θ_0	θ_1	$\hat{\theta}_0$	$\hat{\theta}_1$	% of CI captures θ_0	% of CI captures θ_1
s1	2x	0	2	3.552333	1.968469	82.98%	61.34%
s2	0.5x	0	0.5	0.888159	0.4921173	83.00%	61.36%
s3	0.3x+2	2	0.3	2.498877	0.2952948	82.92%	61.44%

From table 2, we can compare the true parameters of the scale function to the estimates. The estimates are the average estimates of 5000 simulations. The estimates of scale function are mostly close compared to the true values. The slight difference between the estimated intercepts and the true intercepts acts on little impact towards the regression.
Nevertheless, around 83% of confidence intervals with the significant level $\alpha = 0.05$ capture the true intercept θ_0. Around 61% of confidence intervals with significant level $\alpha = 0.05$ capture the true slope θ_1.
Overall, the average estimates of parameters based on 5000 simulations approach the true values. However, for every individual simulation, the estimates may not be as precise, especially the estimated slope of scale function.

Discussion

We know that quantile regression allows the analysis of the comprehensive conditional distribution effects of the response variable. Throughout the study, we discussed the heteroscedasticity form of the quantile regression model. Also, we used the simulation study to find out the estimates of parameters of quantile regression and scale function. We performed 5000 simulations to estimate the parameters and got the average estimates are fairly accurate, but the estimates of the scale function in every individual simulation may not be as precise, especially the estimated slope of the scale function. Moreover, distinct true parameters were used to ensure accuracy, and we got similar results. In this study, we have focused exclusively on the 0.5 quantile level, which is the median effect, but the proposed simulations and analysis can be readily extended to any desired quantile level with heteroscedasticity.
Developing such a study in real data could be the next step. It could also be possible to discuss the nonlinear form of scale function in future studies. See the study of Selvaratnam et al. (2021) for a detailed example of nonlinear quantile regression.

References

1. Koenker, R. (2005). *Quantile regression*. Cambridge University Press.

2. Selvaratnam, S., Kong, L., & Wiens, D. P. (2021). *Model-robust designs for non-linear quantile regression*. Statistical Methods in Medical Research, 30(1), 221–232. https://doi.org/10.1177/0962280220948159

3. Jung, Y., Lee, Y., & MacEachern, S. N. (2015). Efficient quantile regression for heteroscedastic models. Journal of Statistical Computation and Simulation, 85(13), 2548–2568. https://doi.org/10.1080/00949655.2014.967244

4. Andriyana, Y., Gijbels, I., & Verhasselt, A. (2018). Quantile regression in varying-coefficient models: Non-crossing quantile curves and heteroscedasticity. Statistical Papers, 59(4), 1589–1621. https://doi.org/10.1007/s00362-016-0847-7

5. Hao, L., & Naiman, D. (2007). Quantile Regression. 2455 Teller Road, Thousand Oaks California 91320 United States of America. SAGE Publications, Inc. https://doi.org/10.4135/9781412985550

Appendix

A1: Ethics Statement

I respect and acknowledge the contributions and intellectual property of others. I have appropriately cited them in the reference and used inline APA format citations.