Anima Indonesian Psychological Journal
2016, Vol. 32, No. 1, 41-54

Get Acquainted with *Quantile Regression*

Agung Santoso
Faculty of Psychology
Universitas Sanata Dharma

Tri Hayuning Tyas
Faculty of Psychology
Universitas Gadjah Mada

This article was written to introduce quantile regression (QR) analysis technique for research in Psychology. The authors present the advantages possessed by QR compared with ordinary least square (OLS) for the regression analysis approach. The QR's main advantage than OLS is the information concerning the effects of the independent variables on the dependent variable at a location other than the mean. QR can also provide information regarding the effect of independent variables on the distribution and skewness of the dependent variable. Another QR's advantage is associated with the robustness against violations of assumptions about the normal distribution of data and homogeneity of variance. These two advantages make the authors feel the need to introduce QR in studies in Psychology. The authors are then applying the QR on real data as an illustration. The results of the analysis in the illustration show the advantages of QR over OLS, especially in providing information on the phenomenon under study.

Keywords: quantile regression, ordinary least squares, assumption violation, robustness, variance heterogeneity

Artikel ini ditulis untuk memperkenalkan teknik analisis regresi kuantil (QR) dalam penelitian di Psikologi. Penulis memaparkan kelebihan-kelebihan yang dimiliki QR dibandingkan analisis regresi dengan pendekatan *ordinary least square* (OLS).Kelebihan utama QR dibandingkan OLS adalah informasi mengenai efek variabel independen terhadap variabel dependen pada lokasi selain rerata.QR juga dapat memberikan informasi mengenai efek variabel independen terhadap sebaran dan kejulingan variabel dependen. Kelebihan QR yang lain terkait dengan ketangguhan (*robustness*) terhadap pelanggaran asumsi mengenai distribusi data yaitu normalitas dan homogenitas varian. Dua kelebihan ini lah yang membuat penulis merasa perlu untuk memperkenalkan QR dalam penelitian-penelitian di Psikologi. Penulis kemudian mengaplikasikan QR pada data riil sebagai ilustrasi. Hasil analisis dalam ilustrasi menunjukkan kelebihan QR dibandingkan OLS khususnya dalam memberikan informasi mengenai fenomena yang diteliti.

Kata kunci: regresi kuantil, ordinary least squares, pelanggaran asumsi, tangguh (*robust*), heterogenitas varian

Regression analysis is a technique that is often used in studies of psychology. Flexibility, compactness and the existence of a variety of software options that further facilitate the application of regression analysis on substantive research is the main reason behind the high frequency of use of this analysis. Regression allows researchers to involve the independent variable (IV) with a continuum and/or discrete of data. The regression model also allows researchers to estimate the non-linear relationship between the IV and the dependent variable (DV), so far as the non-linear relationship does not involve regression coefficients as, for example, rank numbers. For instance, model like $y_i = x_{1i}^{\beta_1} + {}^{\beta_2}\log(x_{2i}) + e_i$ cannot be estimated using regression model. The regression model can easily develop into the more complex analysis, like factor analysis and structural equation modeling.

There are several ways to do parameter estimation in regression, like ordinary least squares (OLS), weighted least square (WLS), maximum likelihood (ML), and least absolute deviation (LAD). Among the various ways, the OLS is an estimation technique that most often used by researchers in psycho-

Correspondence concerning this article should be addressed to Agung Santoso, Universitas Sanata Dharma Paingan, Maguwoharjo, Depok, Sleman, Yogyakarta 55281. E-mail: agungsan_psy@yahoo.com logy. The OLS popularity is due to most of the available software, using the OLS as initial settings (default) in performing a regression analysis. Also, the results of estimation using OLS has several characteristics desired by researchers and not shared by other techniques. The OLS estimation is unbiased and the most efficient estimation known by the acronym BLUE (Best, Linear, Unbiased Estimate) that no other estimation techniques have (Berry & Feldman, 1985; Hao & Naiman, 2007; Pedhazur, 1997). Unbias means that if researchers took samples repeatedly until infinity, the average estimation of all samples would be the same as the parameter in the population. Efficient means the estimation results have the smallest variance estimation variance compared with using other analytical techniques (Casella & Berger, 2001).

However, the characteristics of the results of these estimates can only be obtained if the underlying assumptions of OLS are met; such as the absence of the regression model specification errors, the homogeneity of the variance residuals, normality of the residuals or DV normality, absence of multicollinearity among IVs, the lack of autocorrelation between residuals, etc (Berry, 1993; Pedhazur, 1997). Violation of these assumptions may result in regression estimation results to be biased and/or inefficient anymore. This means less reflect the population and/or has a greater variation. Some violations assumptions can be resolved without replacing the OLS estimation techniques with other techniques. Researchers can avoid misspecification of the model and multicollinearity by the literature review so as to involve IV that had average intercorrelation (Berry, 1993; Pedhazur, 1997).

Several other assumption violations can only be resolved by modifying or replacing the OLS estimation technique. For example, the presence of residual variance heterogeneity and autocorrelation can only be overcome by replacing OLS with WLS while a violation of normality and the presence of outliers can only be resolved by estimating regression parameters to be robust, for example, by weighting Huber-White (Holland & Welsch, 1977). The issue of violations of normality and homogeneity assumptions is necessary because the data in psychological studies often violate this second assumption (Micceri, 1989; Ruscio & Roche, 2012).

Some of these estimation techniques can address violations of normality and homogeneity of variance residue assumptions properly. However, all these techniques were the conditional mean estimate only (conditional mean = mean value of the DV for each value IV) of DV based on one or several IV. Mathematically, this model is expressed as follows (Efron & Tibshirani, 1993):

$$E(y|x) = \alpha + \beta_j x_{ij} \quad (1)$$

α is the intercept of the regression coefficients, β_j is the regression *slope* from IVj, and x_{ij} is participant's score-i for IVj. Estimates that only limited to conditional mean; make the information obtained from data on the effects of IV on DV to be limited. In other words, the researcher should assume that there is only one model that is accepted in the population or that β_j is homogeneous for all individuals.

The assumption that β_j homogeneous to all persons in a population is not necessarily true. As mentioned earlier, the heterogeneity of the residual variance, which is a manifestation of heterogeneity β_j mostly appear in psychological research (Grissom, 2000; Ruscio & Roche, 2012). Grissom (2000) mentioned that psychological intervention becomes more satisfactory if the intervention does cause not only change or average difference but also variance variation. In this case, the within subjects of the intervention group were smaller than the control group variance because there was a greater intervention effect on participants with low-score in the DV compared to the participants with high-score in DV.

Incorporating the moderator variable into the model is a workable solution when researchers suspect the emergence of the heterogeneity of regression slope coefficients. The involvement of the moderator variable in the model allows the researcher to identify the heterogeneity mechanism of the regression slope coefficient which can be explained by the moderator variable (Pedhazur, 1997; Yuan, Cheng, & Maxwell, 2014). However, researchers need first to identify variables that moderate the regression coefficients. The difficulty that arises in this process is that researchers are not always able to determine the moderator variable completely. Although the researcher can determine the moderator variables adequately, the researcher cannot retrieve or use all data from the relevant moderator variable. For example, the client's psychopathology history variable is a variable that has the potential to moderate the effects of therapy but cannot be used because it is qualitative with has many categories and occurs in the past. On the other hand, often the heterogeneity of the regression slope coefficients is not directly related to the moderator variable, for example, related to the developmental span. The variations in cognitive and psychological wellbeing variables will be greater as one gets older (Bornstein & Smircina, 1982; Nelson & Dannefer, 1992).

Analysis using multilevel modeling is one other alternative that can be done to overcome the heterogeneity of regression coefficient. This analysis treats regression slope coefficients as coefficients that are random and vary between individual groups. For example, researchers apply an intervention to many schools and see if the effect of such interventions varies across schools. The multilevel model can be expressed as follows:

$$y_{ik} = \alpha_k + \beta_k x_{ik} + e_{ik} \quad (2)$$

$$\alpha_k = \gamma_0 + \nu_{0k} \quad (3)$$

$$\beta_k = \gamma_1 + \nu_{1k} \quad (4)$$

α_k is the coefficient of regression intersection for each school; β_k is the interaction effect in each school; k; γ_0 is the mean of all school intercept; γ_1 is the average of intervention effect from all schools; $\mathrm{e}_{ik}, \nu_{0k}, \nu_{1k}$ are residuals. Intervention effect variations and regression intersection coefficients are illustrated by the variance of ν_{0k} and ν_{1k}. In that three formula, due to only one IV then the j subscript can be deleted. This approach still requires researchers to have information on the relevant groups in the population. For example, the researcher needs to know whether the intervention effect has different scores between schools or other variables that influence β_k. Therefore, the issues raised earlier also potentially appear in this analysis.

Analysis based on latent class or mixture model can accommodate the heterogeneity of regression slope coefficients, assuming that within the population there are sub-populations. For example, Ding (2006) showed that the relationship between mathematics achievement of grade 5 students with four other variables has the value and statistical significance varied between the identified sub-populations. In one sub-population, only one predictor has a significant relationship, that is the teacher's assessment of students' mathematics competency. While in the second sub-population, the teacher's assessment of the students' mathematical and social competencies has a significant relationship. In this analysis, the author (or researcher?) must assume about the distribution of variables in each sub-population. Although it is theoretically possible to have different forms of distribution between sub-populations, the analysis usually runs with the assumption that the distribution of all sub-populations has a uniform shape. This restriction is carried out for example by assuming the distribution of all sub-populations following a normal distribution, albeit having different mean and variance values. Furthermore, regression analysis based on this mixed model also still assumes the homogeneity of residual variance (Faria & Soromenho, 2010). The analysis based on this model is also exploratory: the results obtained depend on how the researcher assumes the complexity of the data structure in each sub-population. The more complex the data structure, fewer sub-populations are assumed (Ding, 2006).

Apart from several weaknesses mentioned, the main drawbacks of OLS and the alternative approaches discussed earlier are the inability to estimate the regression slope coefficients in other distribution areas other than conditional mean; such as the magnitude of the regression slope coefficients for participants with the DV scores that were in the 10th percentile, which may be the main interest of the research. For example, using the previous example about the school intervention. When the researcher interested to answer the research question about the effect of the intervention in those schools at percentile 75 and percentile 25, the researcher cannot easily get the needed information using either OLS or the alternative approaches mentioned earlier. For example, the research that is asking similar research questions such as research conducted by Buchinsky (1994, 1998).

Some other examples of QR applications can be observed in the following paragraphs. Ramdani and Witteloostujin (2010) conducted a study to see the effect of the independence of the board of directors and the duality of the chief executive officer (CEO) on the firm's performance. The researcher proposes two hypotheses that can only be tested using QR: (1) the independence of the board of directors can effectively improve firms' performance for firms that have high performance but are less effective for firms with low performance; (2) the duality CEOs can effectively improve the performance of low-performing firms but are less effective for high-performing firms.

Other studies that are using QR are performed by Binder and Coad (2011) which revealed that income has a different effect on life satisfaction for participants with different levels of life satisfaction. The income has the strongest effect on satisfaction levels for participants who have low satisfaction rates while for participants with high satisfaction, the income has weaker income effects. An interesting finding in the study was that education has a positive effect on life satisfaction for participants with low levels of life satisfaction, but has a negative effect on participants with high levels of life satisfaction.

A study to uncover the perceptions effects of the social quality towards subjective well-being by Yuan and Golpelwar (2013) also using QR as an analytical technique. They found that one of the perception aspects about social quality that is the attribution of success has a significant relationship only to participants with low subjective well-being. Other findings indicate that in participants with high welfare, those who have had higher education have higher subjective well-being than those who have never had higher education.

This article introduces another approach that may address issues related to violation of assumptions underlying OLS regression and heterogeneity β_j. This method is known as Quantile Regression (QR). The authors felt the need to introduce this method because QR offers better estimation accuracy when assumption violations occur and more information is obtained from the data. In this article, the authors present estimation and inference techniques in QR as well as additional information that can be obtained using the analysis. The author will also conduct QR analysis on real data to demonstrate QR applications in substantive research. The command to run QR in R Program by using the quantreg package is also included as an example.

Quantile Regression

Before presenting QR, it is necessary for the authors to explain first about quantile. Quantile is one of the parameters that are very close to the understanding with the percentile, only the percentile has a value between 0 to 100, while the quantile has a value between 0 to 1. Quantile is defined as a value below which there are some observations with a certain proportion (Davino, Furno, & Vistocco, 2013; Gilchrist, 2000; Koenker, 2005). For example, if the A's intelligence score is on quantile with $\tau = .25$ in a group, then 25% of the people in the group have an intelligence score equal to or below A. In the standard normal distribution, as many as 97.5% of the scores will be below the value of 1.96, therefore the value 1.96 is quantile with $\tau = 0.975$.

Quantiles are usually obtained by first sorting a group's score followed by finding value with the proportion of observations at a value smaller or equal to that value, equal to the desired proportion. This technique cannot be applied immediately to estimate the effect of a variable on another variable. Koenker and Hallock (2001) developed an estimation method that allows the application to estimate the effect of a variable on other variables.

Quantile Estimates and their Applications on QR

How to estimate the quantile in QR described by Koenker and Hallock (2001) is in line with the way OLS estimates the mean. OLS estimates the mean by finding the values that can minimize the following functions:

$$M(\mu) = \sum_{i=1}^{n}(x_i - \mu)^2 \qquad (5)$$

in the equation, μ is the estimated target value and x_i is the score of the variables studied. The sample mean, or $\sum_{i=1}^{n} x_i / n$, is a value that can minimize function (5) (proof is given in supplement 1). If the absolute value used in the function, it can be written as:

$$M(\mu) = \sum_{i=1}^{n} |x_i - \mu| \qquad (6)$$

then the value that can minimize the function (6) is the median (proof is given in supplement 1).

QR uses function (6) developed to estimate not only the median but also the quantile. The function to estimate this quantile is formulated as follows:

$$Q_\tau(q) = \tau \sum_{x_i \geq q} |x_i - q| + (1 - \tau) \sum_{x_i < q} |x_i - q| \qquad (7)$$

In that function, τ is the proportion of observations that are below the quantile to be searched with range $\tau = (0,1)$, q is the target quantile value to be sought. Quantile with the proportion of underlying observations equal to the value τ; will minimize the function (proof is given in supplement 1).

Function (7) can be applied immediately in estimating the regression coefficients by replacing q with $\boldsymbol{x}_i\boldsymbol{\beta}$, same as μ replaced the same vector in the

OLS regression. Function (7) then as follows:

$$Q_\tau(\beta) = \tau \sum\nolimits_{y_i \geq \boldsymbol{x}_i \beta^{(\tau)}} \left| y_i - \boldsymbol{x}_i \boldsymbol{\beta}^{(\tau)} \right| + (1 - \tau) \sum\nolimits_{y_i < \boldsymbol{x}_i \beta^{(\tau)}} \left| y_i - \boldsymbol{x}_i \boldsymbol{\beta}^{(\tau)} \right| \qquad (8)$$

In that function, y_i is dependent variable, $\boldsymbol{\beta}^{(\tau)}$ is the vector of regression coefficients on quantile- τ, $\boldsymbol{x}_i$ is score vector independent variable for i individuals with i = 1, 2, 3, … , n. With minimalize the function (8), the $\boldsymbol{\beta}^{(\tau)}$ coefficient can be acquired. The $\boldsymbol{\beta}^{(\tau)}$ vector is the unique vector for every τ. This means the model in QR allows the vector $\boldsymbol{\beta}^{(\tau)}$ to variate on different τ, while OLS provides only one vector $\boldsymbol{\beta}$ to represent all τ.

There are at least two other ways to estimate the recently adopted QR coefficients, the maximum likelihood by using an asymmetric Laplace distribution (Geraci & Bottai, 2013) and estimates using Bayesian techniques (Yu, van Kerm, & Zhang, 2005). Both estimation techniques will not be discussed in this article because the core idea of estimation is similar to that developed by Koenker and Hallock (2001).

The Inference of QR Coefficient

Koenker (2005) comparing the performance of the eight inference methods of QR estimation with the Monte Carlo method. The comparison results show that the bootstrap based inference method has the most confidence interval coverage closer to the expected coverage and has better calculation efficiency. These results also supported by Hahn (1995) that showed the bootstrap method has a higher degree of accuracy than other methods. Therefore, the author uses this bootstrap method in conducting inference. In this case, the authors chose to use confidence intervals created by the bootstrap corrected and accelerated (BCa) method which has high accuracy than other methods (Efron & Tibshirani, 1993).

Estimation of Distribution Shifting and Skewness

Estimation of shifting distribution and skewness can be done by using a representation of these parameters based on the quantile. Data distribution can be represented by interquartile range (IQR), which is formulated as follows (Gilchrist, 2000; Hao & Naiman, 2007):

$$\mathrm{IQR} = \mathrm{Q}_{.75}(\mathrm{q}) - \mathrm{Q}_{.25}(\mathrm{q}) \qquad (9)$$

$\mathrm{Q}_{.75}(\mathrm{q})$ represents the third quartile (or percentile 75) and $\mathrm{Q}_{.25}(\mathrm{q})$ represents the first quartile (or percentile 25). While the skewness can be represented by using Galton's measures of skewness as formulated as follows (Gilchrist, 2000):

$$SK = \frac{Q_{.75}(q) + Q_{.25}(q) - 2 * Q_{.5}(q)}{IQR} \qquad (10)$$

with $Q_{.5}(q)$ is median (or percentile 50).

IQR application in QR can provide information on how broad the distribution of the dependent variable widened or shrank when the score of IV increased one point or called IQR Shift (P-IQR). If $y_1^{.75}$ and $y_1^{.25}$is the magnitude of y in the third and the first quartile when $x_1 = X$; $y_0^{.75}$ and $y_0^{.25}$ the magnitude of y in the third and the first quartile $x_0 = (X - 1)$; α is the coefficient of regression intersection, and β is the slope of regression, so P-IQR can be formulated as follows:

$$P - IQR = \left[y_1^{.75} - y_0^{.75}\right] - \left[y_1^{.25} - y_0^{.25}\right]$$

$$= \left[[\alpha^{.75} + \beta^{.75}x_1] - [\alpha^{.75} + \beta^{.75}x_0]\right] - \left[[\alpha^{.25} + \beta^{.25}x_1] - [\alpha^{.25} + \beta^{.25}x_0]\right] = \left[[\alpha^{.75} + \beta^{.75}x_1] - [\alpha^{.75} + \beta^{.75}(X - 1)]\right] - \left[[\alpha^{.25} + \beta^{.25}x_1] - [\alpha^{.25} + \beta^{.25}(X - 1)]\right]$$

$$= [\beta^{.75}X - \beta^{.75}X + \beta^{.75}] - [\beta^{.25}X - \beta^{.25}X + \beta^{.25}]$$

$$= \beta^{.75} - \beta^{.25} \qquad (11)$$

P-IQR can be interpreted as the magnitude of the difference between quartiles of each difference of one point from IV. The positive and significant value of P-IQR means the greater the IV's value, the greater the distance between the first and third quartiles depicting the greater the distribution of data. The negative and significant value of P-IQR means the greater the IV's value, the smaller the distance between the first and third quartiles depicting the smaller the distribution of data. The interpretation of P-IQR illustration as follows. The study was conducted to investigate the effect of new teaching methods on improving student's learning achievement, in which the control group was coded 0, and the treatment group was coded 1. If the results showed significant effect with a negative and meaningful value of P-IQR, this means the treatment not only improve student achievement but also reduce the gap between those with low achievement and high achievement. The conclusions about the decrease in inequalities were obtained from information on negative P-IQR, which means the interquartile range (IQR) of students achievement scores were smaller in the treatment group (with code 1) than the control group (code 0).

SK application in QR provides information on how much the skewness of the dependent variable changes when the score of IV increases one point or is called SK (P-SK). P-SK can be formulated as follows:

$$P - SK = \frac{[y_1^{.75} + y_1^{.25}] - 2y_1^{.50} - [(y_0^{.75} + y_0^{.25}) - 2y_0^{.50}]}{P - IQR}$$

$$= \frac{[\alpha^{.75} + \beta^{.75}X + \alpha^{.25} + \beta^{.25}X - 2*(\alpha^{.50} + \beta^{50}X)]}{P - IQR} -$$

$$\frac{\left[\alpha^{.75} + \beta^{.75}(X-1) + \alpha^{.25} + \beta^{.25}(X-1) - 2*\left(\alpha^{.50} + \beta^{50}(X-1)\right)\right]}{P - IQR}$$

$$= \frac{\beta^{.75} + \beta^{.25} - 2*\beta^{.50}}{P - IQR} \tag{12}$$

P-SK can be interpreted as the difference in the extent of the skewness of the distribution of DV every one point difference from the two values of IV. The positive and significant value P-SK means the greater the value of IV, the more positive the DV data skewed. In other words, the greater the value of IV the more participants with a DV score higher than average.

Conversely, the negative and significant P-SK value means the greater the value of IV, the more negative the skewness of the DV. In other words, the larger the IV, the more participants with smaller DV scores than the mean. The illustration of P-SK interpretation is given below. For example, in the previous example, the new teaching methods treatment, the analysis results show that treatment effects have insignificant values while P-SK has a positive and significant value. This means that, although the treatment did not lead to an inter-group average difference, the skewness of distribution in the experimental group became more positive than the control group. Because the more positive skewness means larger scores than average being more in the treatment group, then the researcher can still conclude that the treatment has a positive effect, even it is not towards the mean. If the researchers only analyze data using OLS regression, then the researchers will not get evidence of the treatment effects.

Distribution Shifting and Skewness Interpretation

The study of inference from the two previous parameter estimations has not been done. The authors have just found a reference about the inference from these two statistics in Hao and Naiman (2007). They recommend the use of the bootstrap method to estimate the standard error of the estimates of both parameters. The authors chose to use confidence intervals constructed with the bootstrap correction and acceleration (BCa) method (Efron & Tibshirani, 1993) because it tends to be more accurate than other bootstrap methods, especially if the form of the parameter estimate distribution is unknown. (Authors' Note: The one that described in this paragraph is the inference of P-IQR and P-SK, while the explanation in the previous section is related to the inference from the estimation of the slope coefficient and the regression intersection).

The Advantages and Limitations of QR Compared to OLS

There are two advantages of QR over OLS, namely: (1) accuracy and precision of QR estimation results when assumption violations occur; and (2) information can be obtained from the data. The first advantage that owned by QR as a consequence of the use of absolute values, not quadratic values, in minimized functions. The use of this absolute value also makes the effect of observation with the extreme value to be minimal to the estimation of the regression coefficient. Also, the blend between QR and inference using a bootstrap approach can improve the efficiency of the estimated QR (Hahn, 1995; Hao & Naiman, 2007; Koenker, 2005). Violation of assumptions is an important issue in statistical analysis, especially regression, because it can cause bias and reduced efficiency of estimation results (Berry, 1993; Pedhazur, 1997; Zu & Yuan, 2010), while violation of assumptions often occurs in research data in psychology (Bryk & Raudenbush, 1988; Grissom, 2000; Hampel, Ronchetti, Rousseeuw, & Stahel, 1986; Micceri, 1989; Ruscio & Roche, 2012). Therefore, QR becomes one of the alternatives analysis that should be considered due to its firm against violation of assumptions.

QR has the potential to provide more information than OLS because for each sample; the researcher can estimate more than one location from the distribution of the dependent variable. For example, in

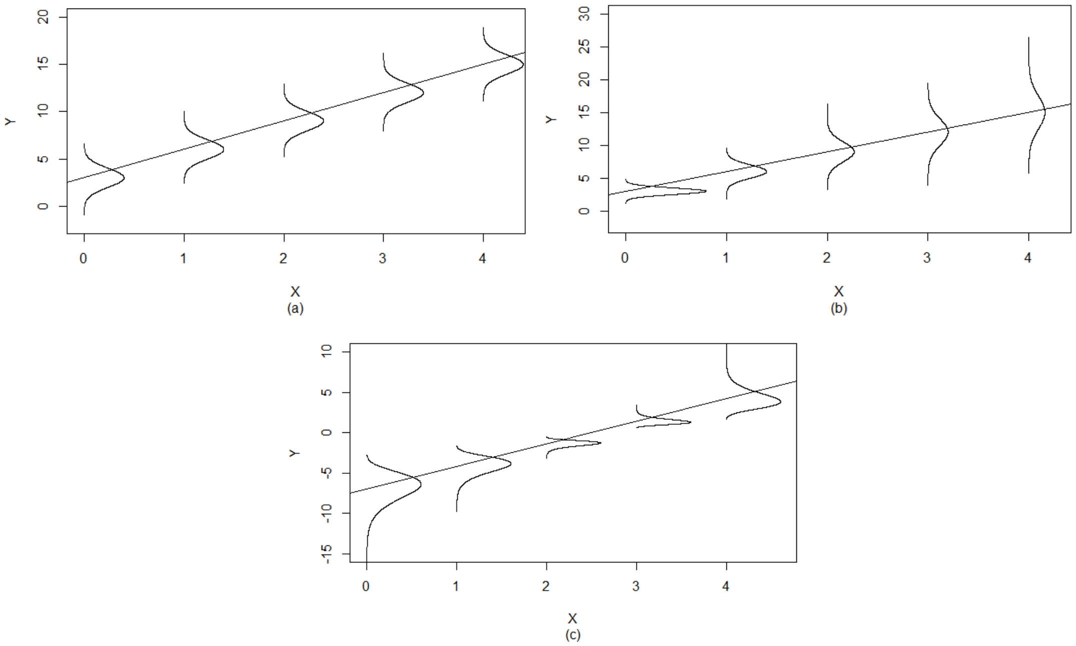

Figure 1. Illustration of a shift (a) location, (b) the location and distribution, and (c) the location and skewness.

previous examples of interventions in schools, researchers were not only able to obtain information on the average effects of interventions, but also the effects of interventions at other locations such as the 10th, 25th, 75th, 90th percentiles by the research question posed. The estimations at some of these sites will provide information on the effects of IV towards DV on different groups of individuals. Also, using the size of the distribution, such as the interquartile range (IQR) and the skewness based on quantile (Gilchrist, 2000; Hao & Naiman, 2007), the researcher can see the effect of IV towards DV on the two additional information. Therefore QR not only can provide information about location shifts but also shifting distribution, and the skewness. The third possible illustration of the shift can be seen in Figure 1.

Figure 1.a. is an illustration when data only shifts location from Y along X. Researchers often assume this location shift in analyses based on OLS, such as regression, variance analysis, and so on. Figure 1.b. and 1.c. provide a data illustration when experiencing location shifts as well as shifting distribution, and the skewness. For example, in Fig. 1.b., the greater the X, not only the mean of Y which becomes larger but the distribution of Y is also wider. In Figure 1.c., the larger the value of X, the skewness of Y changed from positive, when $X = 0$, to negative, when $X = 5$. The shift of these two parameters can not be easily estimated using OLS so that QR is a promising alternative analysis to supplement the information obtained from OLS. The estimation of shifting distribution and skewness as illustrated in Figure 1.b. And 1.c are discussed in the next subsection.

QR has at least four limitations when compared to the OLS regression approach. When the assumptions underlying the OLS regression can be met by the data, QR estimates tend to have a larger standard error than the OLS regression. This means QR tends to report fewer regression coefficient significance than OLS regression when the coefficient is not equal to zero in the population. QR also can not be immediately applied to the analysis involving the dependent variable with the dichotomy score. The QR analysis results tend to be complex because it reports more than one coefficient for one IV. This complexity increases exponentially when QR is applied to examine the effects of mediation (Imai, Keele, & Tingley, 2010; Santoso & Fairchild, 2016; Shen, Chou, Pentz, & Berhane, 2014). For example, if the researcher is interested to see the mediation effect on the three quantile mediation variables and the five quantile dependent variables, mediation analysis using QR will result in 5 × 3 = 15 coefficients of mediation effect. The last limitation that

Table 1

Regression Coefficient of Seven Quantile with 95% Level of Confidence

		Regression Coefficient	
		Intercept (β_0)	Slope (β_1)
QR	τ = .01	4,190* (3.,867,4,834)	0,328* (0,299,0,338)
	τ = .05	3,665* (2,393,4,429)	0,417* (0,393,0,460)
	τ = .25	1,939* (0,863,2,563)	0,588* (0,569,0,617)
	τ = .5	0,928* (0,236,1,270)	0,707* (0,690,0,744)
	τ = .75	-0,737 (-1,963,0,264)	0,863* (0,858,0,869)
	τ = .95	-1,945* (-3,787,-0,719)	1,105* (1,067,1,165)
	τ = .99	-5,919* (-11,423,-1,061)	1,418* (1,264,1,596)
OLS		5,033* ($SE = 0{,}212$)	0,614* ($SE = 0{,}006$)

Note. The values in parentheses under the successive QR coefficients are the lower and upper bounds of the 95% level of confidence. The value within the brackets below the OLS coefficient is a standard error.

the authors can find is the less popular software to run QR. As far as the author's knowledge, there are only three softwares that provide this QR analysis: SAS, Stata, and R, all of which are not widely used or known by researchers in the field of psychology in Indonesia.

Illustration of Quantile Regression Application on Real Data

The authors use data from the Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (N = 11.933, National Center for Education Statistics, 2009) as an illustration in applying QR. The authors chose this data because this data has an error distribution with heterogeneous variance, so it can provide a clearer illustration of the benefits of using QR. The author analyzes two variables of the data; Namely: (1) Mathematics Test (MAT) as DV and (2) Reading Test result (BACA) as IV.

The authors apply QR to estimate the regression coefficients in the seven quantile locations that are quantile with $\tau = \{ .05, .25, .5, .75, .95\}$ which illustrates the Five-Number Summary (*Five Number Summary,* Howell, 2009) also quantile with τ = { .01, .99} to check the regression coefficients in the outlier area. A Five-Number Summary is a five-digit that considered could summarize the overview of distribution, which includes the median, the first and third quartiles, then the 5th and 95th percentiles. The author takes 200 bootstrap samples as recommended by Efron and Tibshirani (1993).

The authors used one of the program packages in R that is quantreg (Koenker, 2015) in estimating the QR coefficients. The authors devise their program to make estimates of P-IQR and P-SK, and the inference of estimated results using the BCa technique since the quantreg program package does not provide those functions. The program can be downloaded from http://www3.nd.edu/~asantos1/Quantile%20 Regression%20BCa/ QR_ BCa.R. The authors also included the program code in this article supplement 2. The description of the command line for running the program is given in Appendix A.

Analysis Results

QR Coefficient Estimation and Inferential on Seven Quantiles and OLS

The first results of the analysis can be seen in Table 1 and Figure 2. Table 1 shows the results of estimation and inference of regression coefficients using QR and OLS. Figure 2 shows the regression lines based on the estimation of regression coefficients using QR and OLS.

Table 1 and Figure 2 showed that the magnitude of β_0 and β_1 vary between quantiles. Smaller quantile regression coefficients tend to have smaller values than coefficients in larger quantile. For example, in quantiles with τ = *.01*, $\beta_1 = 0.328$ ($p <$ .05), while in quantiles with τ = *.75*, $\beta_1 = 0.863$ ($p <$ .05). It can be interpreted as one point increasing in BACA score followed by 0.328 point of MAT score for student whose quantile with τ = *.01*. If β_1 can be interpreted as an effect, then the BACA effect on MAT is greater for students with a higher MAT, while the BACA effect on MAT is smaller for students with lower MAT.

Table 1 shows that the OLS regression provides limited information about the READ effect on the MAT, that is the effect on the conditional mean only. The magnitude of the BACA effect against the MAT estimated using OLS ($\beta_1 = 0.614, p < .05$) is quite different from the effect on the other locations. This shows the different BACA effects on MAT for various levels of MAT capability. If the authors

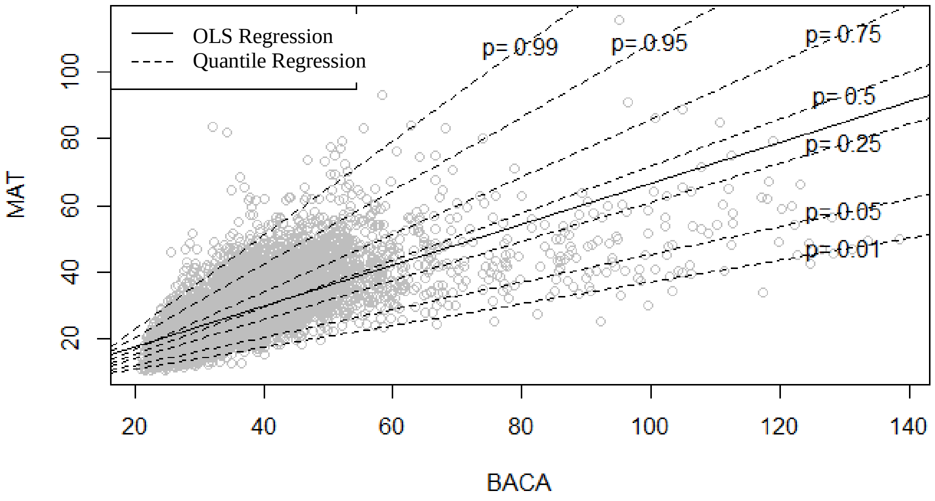

Figure 2. Scatter plot of data BACA score and MAT score with eight regression lines.

used only OLS regression, then the conclusions drawn about the BACA effect on MAT would be mistaken for most of the study subjects. In this research data, OLS regression results apply only to subjects that are in the locality of the conditional mean only.

Differences between Quartiles

The authors continue the analysis by examining the differences of β_1 from two different quantiles, comparing β_1 from quantile with τ = *,25* (or the first quartile) and β_1 from quantile with τ = *.75* (or the third quartile). This difference also illustrates the magnitude of the increase in P-IQR on the MAT for every single point of difference BACA. The results of the analysis showed significant P-IQR values ($P - IQR = 0.275, p < .05$) with $95\%\ CI = (0.273, 0.277)$. This P-IQR value can be interpreted as follows: every one-point BACA score increase followed by an increase of 0.275 IQR from MAT. The participant with the smallest BACA score (21.01), had IQR of MAT of 3.102, while participant with the largest BACA score (138.51) had an IQR of MAT of 35.414. The increase in the BACA score will be followed by the widening of the MAT distribution, or in other words the greater the BACA score, the greater the student's MAT score gap.

Skewness Shift

The next analysis result showed there was significant skewness shift ($P - SK = 0.135, p < .05$) with $95\%\ CI = (0.067, 0.200)$. The *P-SK* value can be interpreted as follow: every one point increased in the BACA score will be followed by increased skewness on MAT data of 0.135 points. The increase in the P-SK score means the greater the BACA score, the MAT score will be more positively skewed, or in other words, the proportion of students with low MAT scores will tend to be larger.

The OLS regression does not have the ability to extract information about the shifting of distribution and skewness. This inability is due to the limited estimation generated by this OLS regression, which presents only one estimate β_1 for each variable in the conditional mean.

In this study, the coefficient of regression intersection has no practical meaning, so it is not too important to interpret. However, in experimental research design, the intersection coefficient of regression can have practical significance. This practical importance is due to the value 0 being often used to represent the control group, so the intersection coefficient provides information about the mean score of DV in the control group (if using OLS regression) and the distribution of participant scores in the control group if the researcher estimates the intersection coefficient on some quantiles using QR.

Discussion

This paper has shown both theoretically and empirically the advantages of QR compared to the frequently used regression techniques in the field of psychology (OLS). QR has these advantages because of the peculiarity of loss function which is minimized in estimating its coefficients. For example, because using absolute values (not quadratic values used in regression analysis in general), the QR coefficients become more immune to the presence of extreme values in the data. Weighted loss function allows QR to estimate regression coefficients at locations other than the distribution center's location.

The results of the analysis using real data show the advantages of QR in providing information from research data. QR may supply information on the effects of IV on DV at locations other than the distribution center's location. For example, in the authors' analysis, QR contains information on the effect of IV on DV on seven locations moving from the lowest quintiles ($\tau = .01$) until the highest ($\tau = .99$), while the OLS regression OLS only gives information about the effects on the conditional mean only.

The second information that can be obtained from the QR is the magnitude of the IQR shift which illustrates the extent of the increase in the inter-quartile of DV scores per one point of increase in IV score. For example, in the analysis results found that P-IQR of 0.275 and statistically significant. This means that every single point of difference in BACA score will be followed by 0.275 IQR increase from MAT. Thus, the researchers can conclude that the greater the BACA, the greater the MAT score gap between students with high MAT scores and students with low MAT scores.

The third information obtained from the QR is the magnitude of the displacement of skewness shifting that describes the difference in DV skewness in IQR units per one point of increase in IV score. In the analysis results found that P-SK of 0.135 and statistically significant. This means that every single point of difference in BACA score will be followed by an increase of 0.135 P-SK from MAT. The authors can conclude from these findings that the greater the BACA, the more positive the MAT skewness, which means more participants have lower MAT scores than average. Meanwhile, the result of regression analysis using OLS approach can only present information about the effect of IV towards DV on conditional mean only. Information on the shifting of distribution and skewness can not be obtained from this approach. Furthermore, if the researcher only uses information derived from the OLS approach, then (s)he can draw the conclusion less precisely. For example, the researcher will conclude that the magnitude of IV variable effect on DV is homogeneous for all MAT distribution. Hao and Naiman (2007) also mentioned that means are not a significant measure of central tendency when data have skewed distributions. Therefore, in a skewed data condition, the OLS will provide inaccurate information about the BACA effect on the MAT at the central tendency location. QR estimates in the median provide a more meaningful descriptive than OLS. This is because in a skewed distribution, the mean value, as a result of OLS estimation, is not really in the middle location. The average location will tend to follow the direction of the data skewness. While the median tends to remain in the middle of the distribution because it is not affected by the extreme value (see Figure 3).

Conclusion

Based on these theoretical and empirical illustrations, the author can take some conclusions. When data satisfy assumptions about the distribution of data, such as normality and homogeneity of distribution, the OLS approach regression provides accurate and best-precision results. However, if there is a violation of these assumptions, the OLS approach will no longer produce satisfactory results. QR is an analytics alternative that has several advantages over OLS, including the firm nature of violating assumptions about data distribution. QR also provides more information than the OLS approach; as the information about the effects of IV on DV on areas other than areas around the central tendency of distribution, IV effects on data distribution of DV, and IV effects on DV data skewness. This three information can enrich the researchers' understanding of the phenomenon being studied.

The authors feel it is necessary to state that the use of the word 'effect' in this article is not intended to demonstrate the ability of QR to establish causal relationships. The authors use the word 'effect' only to show the theoretically derived effect. The researcher remains in a position to support the view that a causal relationship can only be established through the use of an adequate research design.

Recommendation

The authors encourage the use of QR in studies in psychology alongside the use of OLS to obtain richer information about the phenomenon under study. The use of QR becomes crucial especially when the

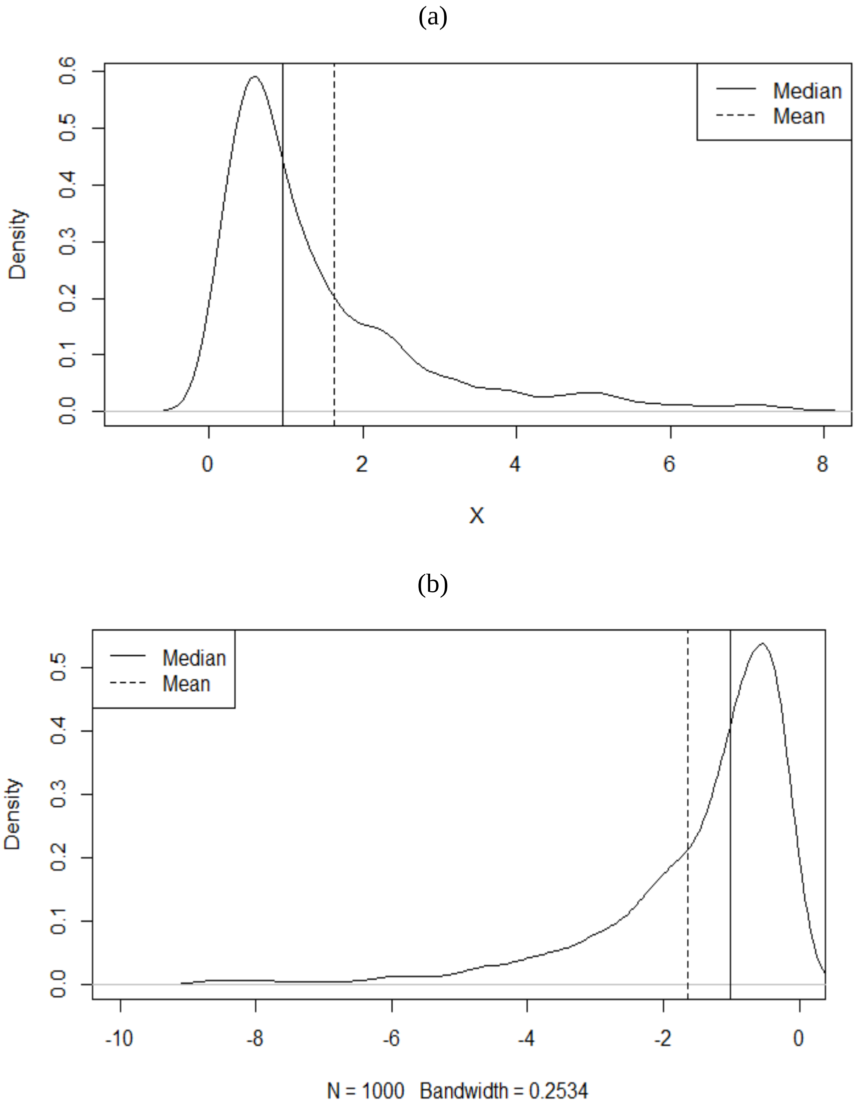

Figure 3. The illustration of the comparison between mean and median of the right skewed (2a) distribution, and left skewed (2b) distribution.

research data violates the assumption of data distribution or when the researcher has questions about the effect of IV on DV on a location other than the mean.

The authors do not advise the researchers to abandon the OLS approach at all, especially when assumptions about the distribution of data are met. The authors recommend the use of QR side by side with OLS to increase the wealth of information that can be obtained from the data.

The researchers interested in using QR in their research need to pay attention to several things:

(1) The sample size taken needs to adjust to the number of QR coefficients to be estimated. The more QR coefficients to be estimated, the larger the sample size the researcher needs to take, especially if the QR coefficient in the tail area of the distribution is also estimated. As far as the author's knowledge, no research investigates the link between sample size and power of analysis within the QR framework. Therefore, the authors have not been able to provide advice on the size of the sample should be taken. The authors suspect the rules regarding the sample size applicable to OLS regression can also be applied to this QR.

(2) Some of the assumptions underlying OLS still apply to this QR approach, so researchers need to pay attention to the fulfillment of these assumptions. For example, QR still assumes that the measurement of independent variables contains small

measurement errors, and the linear equations of independent variables are sufficient to explain the variation of the dependent variable. Although QR can overcome the assumptions of normality and homogeneity of distribution, researchers still need to check whether the data follows both assumptions. The use of QR in data that satisfies both assumptions is less useful unless the researcher has a typical research question related to the between variables relationship in a particular distribution area.

(3) Reporting of QR analysis results may follow the reporting guidelines of regression analysis results in general. Researchers can pair the results of QR analysis with OLS in one table to facilitate comparison of the coefficients obtained from both methods.

(4) The selection of the number of quantiles to be estimated highly dependent on the distribution of the dependent variable data and research questions to be answered.

(a) If the researcher anticipates the effect of independent variables on the dependent variable being heterogeneous in different distribution areas, the estimation on the quantile of five numbers summary can be done to cover all distributions of the dependent variable, so that the effect heterogeneity can be observed completely.

(b) If researchers anticipate the occurrence of floor effect or ceiling effect on the data, then the estimated QR coefficient on the median will be more accurate than the use of OLS regression. If the researcher is interested in examining the area on which the floor/ceiling effect occurs, then the estimated QR coefficient on the quantity of the same proportion can be performed.

(c) If the researcher identifies the observation with the extreme value then the QR coefficient estimate on the quantile in the distribution tail area, such as the quantile with the proportion of 0.01 or 0.99, can be done. It should be noted that the quantile estimates on the tail distribution area have the lowest efficiency (Santoso & Fairchild, 2016), so a larger sample size is needed to improve the power of analysis.

References

Berry, W. D. (1993). *Understanding regression assumptions*. Newbury Park: SAGE Publications.

Berry, W. D., & Feldman, S. (1985). *Multiple regression in practice*. Beverly Hills: SAGE Publications.

Binder, M., & Coad, A. (2011). From Average Joe's happiness to Miserable Jane and Cheerful John: Using quantile regressions to analyze the full subjective well-being distribution. *Journal of Economic Behavior & Organization*, *79*(3), 275–290. http://dx.doi.org/10.1016/j.jebo.2011.02.005

Bornstein, R., & Smircina, M. T. (1982). The status of the empirical support for the hypothesis of increased variability in aging populations. *The Gerontologist*, *22*(3), 258–260. http://dx.doi.org/10.1093/geront/22.3.258

Bryk, A. S., & Raudenbush, S. W. (1988). Heterogeneity of variance in experimental studies: A challenge to conventional interpretations. *Psychological Bulletin*, *104*(3), 396–404. http://dx.doi.org/10.1037/0033-2909.104.3.396

Buchinsky, M. (1994). Changes in the U.S. wage structure 1963-1987: Application of Quantile Regression. *Econometrica*, *62*(2), 405–458. http://dx.doi.org/10.2307/2951618

Buchinsky, M. (1998). The dynamics of changes in the female wage distribution in the USA: A quantile regression approach. *Journal of Applied Econometrics*, *13*(1), 1–30.

Casella, G., & Berger, R. L. (2001). *Statistical inference* (2nd edition). Australia ; Pacific Grove, CA: Duxbury Press.

Davino, C., Furno, M., & Vistocco, D. (2013). *Quantile regression: Theory and applications*. Chichester: Wiley.

Ding, C. X. (2006). Using regression mixture analysis in educational research. *Practical Assessment, Research and Evaluation*, *11*(11), 1 – 11.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Faria, S., & Soromenho, G. (2010). Fitting mixtures of linear regressions. *Journal of Statistical Computation and Simulation*, *80*(2), 201–225. http://dx.doi.org/10.1080/00949650802590261

Geraci, M., & Bottai, M. (2013). Linear quantile mixed models. *Statistics and Computing*, *24*(3), 461–479. http://dx.doi.org/10.1007/s11222-013-9381-9

Gilchrist, W. (2000). *Statistical modelling with quantile functions*. Boca Raton: Chapman and Hall/CRC.

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology*, *68*(1), 155–165. http://dx.doi.org/10.1037/0022-006X.68.1.155

Hahn, J. (1995). Bootstrapping quantile regression estimators. *Econometric Theory*, *11*(1), 105–121.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions*. New York: John Wiley & Sons.

Hao, L., & Naiman, D. Q. (2007). *Quantile regression*. Thousand Oaks, Calif: SAGE Publications, Inc.

Holland, P. W., & Welsch, R. E. (1977). Robust regression using iteratively reweighted least-squares. *Communications in Statistics - Theory and Methods*, *6*(9), 813–827. http://dx.doi.org/10.1080/03610927708827533

Howell, D. C. (2009). *Statistical methods for psychology* (7 edition). Australia : Belmont, CA: Cengage Learning.

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, *15*(4), 309–334. http://dx.doi.org/10.1037/a0020761

Koenker, R. (2005). *Quantile regression*. Cambridge ; New York: Cambridge University Press.

Koenker, R. (2015). quantreg: Quantile Regression. R package. (Version 5.11) [Computer Software]. Retrieved from http://CRAN.R-project.org/package=quantreg

Koenker, R., & Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives*, *15* (4), 143–156. http://dx.doi.org/10.1257/jep.15.4.143

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156–166. http://dx.doi.org/10.1037/0033-2909.105.1.156

National Center for Education Statistics. (2009). Early childhood longitudinal study, Kindergarten Class of 1998–99 (ECLS-K)[Data file]. Retrieved from http://nces.ed.gov/ecls/kindergarten.asp

Nelson, E. A., & Dannefer, D. (1992). Aged heterogeneity: Fact or fiction? The fate of diversity in gerontological research. *The Gerontologist*, *32* (1), 17–23. http://dx.doi.org/10.1093/geront/32.1.17

Pedhazur, E. J. (1997). *Multiple regression in behavioral research* (3rd ed.). Fort Worth: Wadsworth Publishing.

Ramdani, D., & van Witteloostuijn, A. (2010). The impact of board independence and CEO duality on firm performance: A quantile regression analysis for Indonesia, Malaysia, South Korea and Thailand: Board independence, CEO duality and firm performance. *British Journal of Management*, *21*(3), 607–627. http://dx.doi.org/10.1111/j.1467-8551.2010.00708.x

Ruscio, J., & Roche, B. (2012). Variance heterogeneity in published psychological research: A review and a new index. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *8*(1), 1–11. http://dx.doi.org/10.1027/1614-2241/a000034

Santoso, A., & Fairchild, A. J. (2016). *Mediation analysis using quantile regression.* (Manuscript in peparation.)

Shen, E., Chou, C.-P., Pentz, M. A., & Berhane, K. (2014). Quantile mediation models: A comparison of methods for assessing mediation across the outcome distribution. *Multivariate Behavioral Research*, *49*(5), 471–485. http://dx.doi.org/10.1080/00273171.2014.904221

Yuan, H., & Golpelwar, M. (2013). Testing subjective well-being from the perspective of social quality: Quantile regression evidence from Shanghai, China. *Social Indicators Research*, *113*(1), 257–276. http://dx.doi.org/10.1007/s11205-012-0091-z

Yuan, K.-H., Cheng, Y., & Maxwell, S. (2014). Moderation analysis using a two-level regression model. *Psychometrika*, *79*(4), 701–732. http://dx.doi.org/10.1007/s11336-013-9357-x

Yu, K., van Kerm, P., & Zhang, J. (2005). Bayesian quantile regression: An application to the wage distribution in 1990s Britain. *Sankhyā: The Indian Journal of Statistics (2003-2007)*, *67*(2), 359–377.

Zu, J., & Yuan, K.-H. (2010). Local influence and robust procedures for mediation analysis. *Multivariate Behavioral Research*, *45*(1), 1–44. http://dx.doi.org/10.1080/00273170903504695

(Appendix follows)

Appendix A

Application of the R Program to Perform Quantile Regression Analysis With Inference Using BCa

The content of Appendix A is kept by the first author. Interested readers may contact the first author to obtain it.

Supplement 1

Proofs

Supplement 2

Program Code

The content of Supplement 1 and Supplement 2 are kept by the first author. Interested readers may contact the authors to obtain it.