PREDICTORS OF EDUCATIONAL ATTAINMENT IN INDONESIA: COMPARING OLS REGRESSION AND QUANTILE REGRESSION APPROACH

by

AGUNG SANTOSO

(Under the Direction of Jonathan Templin)

ABSTRACT

The current study applied quantile regression analysis to estimate the relationship between educational attainment and its predictors, and compared the results to parameter estimates using OLS regression. OLS regression is a regression technique that uses conditional mean as a solution to minimize the error variance. Predictors of educational attainment used were socioeconomic status represented by parent's education, parent's occupation, and family income, hours of study at school, intelligence, and students' employment. Results from the quantile regression analysis showed that for several variables parameter estimates were significant only for certain quantiles. Parameters for hours of study at school and family income were significant only for lower quantiles, while intelligence and managerial/professional were significant for higher quantiles. There were variables that had significant parameters on OLS but not on all quantile from quantile regression. Significance tests of difference between quantiles showed non-significant values. Therefore, an analysis to estimate scale and skewness shift were not reasonable to be conducted.

INDEX WORDS: quantile regression, Educational attainment, Indonesia, Socioeconomic status, Student employment, Learning hours, Parent's education, Parent's occupation, Family income, Intelligence, EBTANAS.

PREDICTORS OF EDUCATIONAL ATTAINMENT IN INDONESIA: COMPARING OLS REGRESSION AND QUANTILE REGRESSION APPROACH

by

AGUNG SANTOSO

BA., Gadjah Mada University, Indonesia 2001

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirement for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2008

PREDICTORS OF EDUCATIONAL ATTAINMENT IN INDONESIA: COMPARING OLS REGRESSION AND QUANTILE REGRESSION APPROACH

by

AGUNG SANTOSO

Major Professor :	Jonathan Templin
Committee:	Deborah Bandalos
	Seock-Ho Kim

Electronic Version Approved

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2008

ACKNOWLEDGEMENTS

Scientific efforts have and will always be communal efforts. So did many individuals have contributed to current study.

I would like to thank my advisor, Dr. Jonathan Templin, for his support, encouragement, trust, and for introducing me to new thoughts. And also to Dr. Deborah Bandalos, Dr. Seock-Ho Kim and all other professors for their examples, teachings, and support.

I could not forget to thank my family for their support, trust, and love that have given me courage to face difficult times. And to dearest Fang, for her patience, love and encouragement that strengthen my motivation to finish my study.

Thanks to all of my friends for their help and kindness as I walked with you along the way, especially Hye-Jeong, Adeya, Young-soon, Flora Liu, Mushtaq, Ai jun, Fei ming, In Heok, and Maurice. Thank you for helping and teaching me many things that you may not even realize.

To Jesus, my Lord and my friend. My life has been a testimony of Your faithfulness, kindness and grace.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS iv

LIST OF TABLES vii

LIST OF FIGURES viii

CHAPTER

1. INTRODUCTION 1

Research Questions 7

2. LITERATURE REVIEW 8

Factors Predicting Academic Achievement 8

Quantile Regression 10

Comparing Quantile Regression and OLS regression 14

3. PROCEDURE 21

Variables 21

Data 22

Computer Program 23

Descriptive Statistics 24

4. RESULTS AND DISCUSSION 26

Assumption Check and Diagnostic for OLS regression 26

Comparison of OLS regression and Median Regression 28

Comparison to Other Quantiles 29

Difference between Parameters on Certain Quantiles 32

5. CONCLUSIONS AND SUGGESTIONS....33
Conclusions....33
Suggestions....34
REFERENCES....36
TABLES....39
FIGURES....46
APPENDICES....57
A. SUMMARY TABLES FOR OLS REGRESSION AND QUANTILE REGRESSION ON EACH QUANTILE....57
B. R SCRIPT FOR DESCRIPTIVE STATISTICS....61
C. R SCRIPT FOR CORRELATION MATRIX....63
D. R SCRIPT FOR OLS REGRESSION AND QUANTILE REGRESSION ANALYSIS....65

LIST OF TABLES

Table 1. Descriptive Statistics....39

Table 2. Correlation Matrix 40

Table 3. Parameter Estimates of OLS regression with Potential Outliers Included and Excluded....41

Table 4. *R*-Square and *F* Values of OLS regression with Potential Outliers Included and Excluded....42

Table 5. Parameter Estimates of OLS regression, OLS regression Excluding Outliers, and Median Regression....43

Table 6. Parameters of OLS regression and Quantile Regression....44

Table 7. Significance Test for Parameter Difference Between Quantile .25, .5, and .75....45

LIST OF FIGURES

Figure 1. Relationship between cumulative distribution function and quantile function 46

Figure 2. Plots of *y* against mean of squared deviation 47

Figure 3. Plots of *y* against mean of absolute deviation 48

Figure 4. Plots of *y* against weighted mean of absolute deviation for quantile .1 49

Figure 5. Illustration of skewness shift 50

Figure 6. Distribution of IFLS3 data.....51

Figure 7. Plots of residuals against predicted values 52

Figure 8. Q-Q plot for normality of error term 53

Figure 9. Plots of standardized residuals against Leverage 54

Figure 10. Plots of standardized residuals and Cook's *D* 55

Figure 11. Parameters of OLS regression and quantile regression with 95% confidentia Intervals.....56

CHAPTER 1

INTRODUCTION

Education is one of many problems for the Indonesian people since the falling of Soeharto's regime in 1998. Many people, including practitioners, academicians, and politicians, have shown their concern regarding the educational process and educational attainment in Indonesia. These problems were reflected in reports about the educational attainment of Indonesian students as compared to students from other countries. The Programme for International Student Assessment (PISA) report showed that Indonesia ranks 50th out of 57 countries on students' achievement in mathematics, 49th of 57 in reading, and 50th of 57 in science (2006). The Trends in International Mathematics and Science Study (TIMSS) report showed that Indonesia ranks 34th of 45 countries on eight-grade student mathematics scores and 36th of 45 countries on science (International Association for Evaluation of Educational Achievement, 2003).

Problems related to education quality in Indonesia are attributed to many factors. One factor that is usually suspected as the biggest problem is the small education expenditure of government. The Indonesian government expenditure for education is only 9% of the national budget compared to 20% in Malaysia and 16% in Bangladesh (United Nations Statistics Division, 2008). Other factors thought to be the sources of the problem are the disparity between provinces regarding teachers, facilities and fund distribution, and economic hardships caused by the 1998 economic crisis in Asia.

Unfortunately, there are few research studies investigating the relationship of these factors with education quality represented by educational attainment in Indonesia. Discussions and government policies related to education quality were based mostly on discourse and theory or even "common sense." It is important, therefore, to conduct research related to this topic in order to give more empirically based evidence to factors predicting educational attainment. This master's thesis is dedicated to that purpose and is aimed at finding factors that predict educational attainment for students in Indonesia.

Research studies investigating predicting factors of educational achievement have been conducted in both developing and underdeveloped countries. The results of these research studies were not in agreement (Hanushek, 1995; Sirin, 2005; Velez, Schiefelbein, and Valenzuela, 1993; White, 1982). There were studies that gave evidence of the relationships between educational achievement with some factors, and then others that did not show enough evidence for those relationships. Some of these factors were even well known predictors of educational achievement, such as socioeconomic status, school facilities, and teachers' level education.

There are some explanations for the large variations in the results. First, it was suspected that there was actually no relationship between some of these factors and academic achievement (Hanushek, 1979, 1995). The variation of the degree of relationships between research studies was attributed to a problem of sampling error. Second, there were flaws in the studies that have been conducted related to the methods, including measures of response and explanatory variables (Velez et al., 1993). These variable results were also attributed to moderation by other variables included in the analysis such as minority status, grade level, and school location (Sirin, 2005).

All of these studies used moment-based statistical techniques such as the Pearson product moment correlations or regression analysis based on ordinary least square (OLS) estimation. Moment-based techniques use the conditional mean as their optimal value, a solution for the minimization problem posed by examining the negative log-likelihood function in each. In situations in which the model assumptions were not met or there were outliers, the conditional mean cannot accurately reflect the conditional distribution of the data. If the effect of predictors was different across varying percentiles of conditional distributions, then the effect of the predictors on the upper tail of the distribution may be cancelled out by the effect or predictors on the lower tail of the distribution, which in turn make the effects seem to be zero.

The misspecification errors, as it was pointed out by Sirin (2005), related to exclusion of relevant variables, also tend to make the analysis biased. Sirin showed that the effect of socio-economic status (SES) can be different in different levels of other variables, e.g. grade level. It was also shown by Cade and Noon (2003) that the more variables that are excluded from the model, the more heterogeneous the error variances would be. This means that the results from the moment-based analyses can give incomplete and inaccurate information about the relationship between the response and outcome variables. For the results of the study to be accurate, what is needed is a statistical technique that can provide more information about the relationships between variables at varying locations of the distributions of the data. Another limitation in parameter estimation using OLS that was not investigated in the previous research is that of parameters estimated by OLS procedure being influenced by outliers. The existence of outliers violates one key model assumption: only one regression line is needed to represent the relationships for the whole distribution (Hao and Naiman, 2007). The outliers can alter the correlation coefficient or the regression parameters to be smaller or larger than the parameters

estimated when the outliers were not included in the data (Moore, 2007; Pedhazur, 1997). It was suggested that the outliers can be excluded from the analysis if they were, from thorough investigations, proven to be non-valid observations. But when the outliers are valid, it can give new insights about the nature of the data (Pedhazur, 1997). It means that a statistical technique is needed that will capture the outliers in the analysis and yet is less influenced by their presence.

An alternative technique that has more capability to solve some issues mentioned earlier is called quantile regression. Quantiles are values that give us information about location of a case in a distribution related to proportion of cases having smaller values (Koenker and F.Hallock, 2000). It was developed from a conditional median regression introduced by Boscovich in the 18th century, even before the idea of least squares regression estimators emerged (Koenker, 2005; Koenker and Bassett, 1978). Quantile regression was developed by applying estimation and minimization methods for the conditional median, which is quantile .5, and to other quantiles, rather than the conditional mean, as is done in OLS regression.

Quantile regression has some advantages over OLS regression. It provides information of location shift not only in terms of central tendency location but also other quantile locations (Hao and Naiman, 2007; Koenker, 2005; Koenker and Bassett, 1978). This means that we may have more than one regression line can be modeled, covering the whole conditional distribution including the outliers. For this reason, quantile regression may give us more information about relationship between variables, not only the relationship in term of location shift but also distributional shift including scale shift and skewness shift (Hao and Naiman, 2007). Furthermore, we do not need to assume certain characteristics of the data we used, especially homogeneity of error variances along explanatory variables and normality of error distribution (Hao and Naiman, 2007; Koenker and Bassett, 1978). Another advantage of using quantile

regression is its monotone equivariance property. Hao and Naiman (2007) explain that if we apply a monotone transformation to the outcome variable and then conduct a quantile regression analysis, the predicted values from this procedure will be approximately the same with predicted values from a procedure in which we conduct quantile regression first and then apply monotone equivariance to its prediction.

Applications of Quantile Regression are still limited to economics or environmental studies, but currently there are more and more studies using Quantile Regression as a data analysis tool. To date, few research studies in education have been conducted using Quantile Regression. One of them was conducted by Haile and Nguyen (2008) investigating the effect of family background and race on educational attainment in USA. They also tried to compare results from OLS regression and quantile regression and found that although results using quantile regression resembled OLS regression, the parameters varied across quantiles. For example, the parameter for race (black compared to white students) was -.849 (p <.01) using OLS regression and -1.128, -1.172, and -1.146 using quantile regression for $Q_{.5}$, $Q_{.75}$, and $Q_{.9}$ respectively. It can be seen that the OLS regression provided smaller values than those from quantile regression for median and higher quantiles. It means that the relationship between race and educational attainment were different across different quantiles of the conditional distribution.

The current research project will apply quantile regression on data from the Indonesia Family Life Survey 3 (Strauss et al., 2004) to estimate the relationship between educational attainment and some of its predictors. The term relationship is meant to be used in a broader sense: not only relationships in term of conditional locations but also conditional distributions.

The study will also compare information provided by this method to those provided by OLS regression to get the sense of how both methods provide different information about the data.

Predictors included in current research were those that have consistently shown a significant relationship with educational attainment and those still with contrasting results in previous research. Predictors that have been shown a significant relationship were intelligence (Chen, Lee, and Stevenson, 1996; Diseth, 2002; Laidra, Pullmann, and Allik, 2007; Rohde and Thompson, 2007) and hours of study at school (Gettinger, 1985; Gettinger and White, 1979; Wiley and Harnischfeger, 1974). Predictors considered to still have mixed results were SES (Hanushek, 1979; Sirin, 2005; White, 1982) and student employment (Cooper, Valentine, Nye, and Lindsay, 1999; D'Amico, 1984).

Educational attainment in the current study will be represented by student's score on EBTANAS (National Learning Evaluation) taken in the 6th grade. SES was measured by three indicators which were family income, parents' highest education, and parents' occupation. Student employment will be represented by how many hours of work students perform in a week. Hours of study at school was measured by hours of schooling while intelligence was measured using Raven's Standard Progressive Matrices.

Research Questions

There are several questions to be answered by the current study. They are:

1. What is the relationship between educational attainment and its predictors either using OLS regression or quantile regression methods?
2. Are there any differences in information given by OLS regression and quantile regression?

CHAPTER 2

LITERATURE REVIEW

Factors Predicting Academic Achievement

The second chapter presents a discussion of the factors predicting educational attainment. There are some predictors that have been acknowledged to have a close relationship to educational attainment in many studies that turn out to be controversial in other studies such as teacher experience (Hanushek, 1995), socioeconomic status (Sirin, 2005; White, 1982), etc. Other predictors were consistently reported to have a strong relationship with educational attainment such as intelligence (Velez, Schiefelbein, and Valenzuela, 1993).

Socioeconomic Status (SES)

The relationship between SES and educational attainment was reported to have widely varying levels, from a very high to a non significant correlation (Haile and Nguyen, 2008; Halle, Kurtz-Costes, and Mahoney, 1997; Sirin, 2005; Velez et al., 1993; White, 1982). There was a tendency for the relationship between SES and educational attainment to be weaker when the units of analysis were students, and higher when the unit of analysis were the aggregate of students (e.g. school, district, etc.).

There were also various indicators used to represent socioeconomic status (SES). Indicators that were used consistently to represent SES were parent's education, parent's occupation and parent's or family income. These indicators were also reported to have stronger relationships with educational attainment among other indicators (Haile and Nguyen, 2008; Sirin, 2005; White, 1982).

Intelligence

Intelligence is consistently reported as having a significant relationship with educational attainment. Velez et al. (1993) conducted a meta-analysis on 5 studies that include intelligence in the regression model and found that in all of them intelligence has a significant relationship with educational attainment. The size of the effect, however, was not mentioned. Other studies reported correlations of .3 to .7 between intelligence and educational attainment (Chen, Lee, and Stevenson, 1996; Diseth, 2002; Rohde and Thompson, 2007). Therefore, it is expected that the relationship between intelligence and educational attainment in the current study will be significant.

Hours of study at School

Several studies have shown evidence of a relationship between hours of study at school and educational attainment. Wiley and Harnischfeger (1974) reported that there was a positive relationship between educational attainment and the amount of days in of schooling a year. Gettinger (1985) and Gettinger and White (1979) both found a significant relationship between hours of schooling and educational attainment. The relationship was reported to be stronger than the relationship between intelligence and educational attainment (Gettinger and White, 1979).

Student Employment

School-age children's employment has become a more serious problem in Indonesia since the economic crisis of 1998. The amount of elementary school-aged children employed outside the home has risen from 764,386 in 2006 to 2,854,123 in 2007 (Suara Pembaruan, 2008). Many of these children go to school in the morning then go to work after school.

Few research studies have investigated the relationship between employment and educational attainment, especially in elementary school. Studies that have been conducted have

been in high school or college and have shown a non-linear relationship between hours of work and employment. High school students who worked less than 20 hours a week would have more benefit from their work as the hours increased, while those who worked more than 20 hours would suffer from insufficient time for study (Brown and Steinberg, 1991; D'Amico, 1984; Mortimer and Finch, 1996). One study conducted on 6th grade and 12th grade students found that there were significant negative relationships between time spent working and several measures of educational attainment. Yet, the relationships were considered to be weak. The relationships between time spent working with achievement test scores, teacher assigned grade, and grade after controlling achievement score were -.12, -.17, and -.14 respectively (Cooper, Valentine, Nye, and Lindsay, 1999).

Quantile Regression

Quantile and Quantile Function

A quantile is a value that gives us information about the location of a case or a score in a group "… corresponds to a specified proportion of the sample or population" (Gilchrist, 2000, p. 1). A person's score on a test is said to be in the *p*-th quantile in his/her group if his/her score in the test is bigger than a proportion of *p* of his/her group and smaller than a proportion *(1-p)* of his/her group (Koenker and F. Hallock, 2000). The median is at the .5 quantile, because there is half of the group that have values bigger than the median, and half of the group that have values smaller than the median. The lower quartile is at .25 quantile and the higher quartile is at .75 quantile.

A function that gives us the value of a certain quantile is called a quantile function (QF) denoted as $Q^{(p)}$. For example, if a median of a group has a value of 50, it can also be said that

$Q^{(.5)}$ is 50. The quantile function is an inverse of cumulative distribution function (CDF) denoted as $F(x)$. The CDF can show us a proportion of a group that has a value equal to or smaller than a certain value of *x*. It can be formulated as:

$$F(x) = P(\mathrm{X} \leq \mathrm{x}). \quad (1)$$

The relationship between quantile function and CDF can be depicted in Figure 1 and denoted as

$$Q^{(p)} = F^{-1}(p). \quad (2)$$

For example, if $F(90) = .25$ then $Q^{(.25)} = 90$.

Quantiles as Solutions of Minimization Problems

It has been shown that the mean is a solution for a minimization problem. The arithmetic mean minimizes the mean of squared deviations (MSD) in a single distribution. To illustrate this point, data from 10 cases have been generated from a normal distribution with a mean of 10 and SD of 5. The mean of the sample was 10.052 and the standard deviation was 4.898. The MSD was counted for each value as it was used as the solution for minimization. The results were presented graphically in Figure 2.

The *x*-axis was the value of $\bar{y}$, while the y-axis represented the value of MSD produced if one used a certain value of *y* as the solution for minimization. From the curve, we could see that the mean of *y* (*y*=10.052) produced the smallest mean squared deviation (*MSD*=23.99) compared to other points in the distribution. The median (*y=10.782*) produced slightly higher mean square deviation (*MSD*=24.58).

The median is also a solution for a different minimization problem. It minimizes the mean of absolute deviations (MAD). For illustration, the means of different absolute deviations were counted from the same data mentioned above. The results were presented in Figure 3 with the *x*-axis as the median value, and *y*-axis as the value of absolute deviations using every point in

the data. It can be seen that the median (y=10.782) has the smallest value of mean absolute deviation (MAD=3.61) when compared with all other values. The mean of y (y=10.052) had a slightly higher value of MAD (MAD=3.68) compared to median.

The idea of the median as the solution for a minimization problem has been generalized to other quantiles. For other quantiles to be solutions of minimization problems, different weights should be applied for values less and bigger than the quantiles. For example, for the quantile of .1 to be the solution, all values less than quantile .1 are weighted by 1-.1=.9, and values bigger than the quantile .1 are weighted by .1. The general notation of the solution is expressed as follows:

$$(Weighted)\ MAD = \frac{1-p}{n}\sum_{y_i<q}|y_i - q| + \frac{p}{n}\sum_{y_i>q}|y_i - q|. \quad (3)$$

The weighted MAD for quantile .1 was counted from previous data to illustrate this point. The results were graphically depicted in Figure 4. From this figure, we can see that quantile .1 (y=6.273) had the smallest value of MAD (MAD = 1.19) compared to other points. Hao and Naiman (2007) provide a proof of quantile as the solution of minimization problem using derivative of MAD.

Quantile Regression

The idea of quantiles as solutions to certain minimization problems was then applied on conditional distributions to investigate the relationship between an outcome variable and a set of response variables (Cade and Noon, 2003). In other words, this idea was extended to conditional quantile functions expressing "quantiles of the conditional distribution of the response variables as functions of observed covariates" (Koenker and F. Hallock, 2000, p. 2).

Quantile regression (QR) is similar to ordinary least squares regression (OLS-R) in a sense that both of them investigate relationships between variables and the rate of the outcome variable following response variables represented by regression coefficients denoted as β s. The main difference is that OLS-R chooses parameter values that have the least squared deviation from the regression line as the parameter estimates, expressed by

$$\min \sum_{i=1}^{n} (y_i - \hat{y})^2 \tag{4}$$

while QR chooses parameter values that have the least absolute deviation/error

$$\min \left[(1-p) \cdot \sum_{\hat{y}_i < \hat{y}_i^{(p)}} \left| y - \hat{y}^{(p)} \right| + (p) \cdot \sum_{\hat{y}_i \geq \hat{y}_i^{(p)}} \left| y - \hat{y}^{(p)} \right| \right] \tag{5}$$

(Hao and Naiman, 2007; Koenker, 2005). In Equation 4, $\hat{y}$ is the predicted value of y using regression line, while $\hat{y}^{(p)}$ in Equation 5 is the predicted value of y on quantile p.

OLS-R uses conditional mean $E(y_i \mid x_i)$ as the solution for minimization problem while QR uses conditional quantiles $Q^{(p)}(y_i \mid x_i)$. As a result, OLS-R will produce only one regression line, which is the regression line on the conditional mean,

$$E(y_i \mid x_i) = \beta_0 + \sum \beta_{ij} x_i \,, \tag{6}$$

while QR can produce more than one regression line, one for any quantile of interest:

$$Q^{(p)}(y_i \mid x_i) = \beta_0^{(p)} + \sum \beta_{ij}^{(p)} x_i \,. \tag{7}$$

Statistical Inference in Quantile Regression

There are several methods used to make inferences about parameter estimates in QR, including asymptotic distributions, Wald tests, ranks, and bootstrap methods. The bootstrap has been found to provide desirable results (Hahn, 1995; Koenker, 2005) especially when dealing with skewed distributions (Hao and Naiman, 2007). The bootstrap also facilitates the opportunity

to test additional hypotheses, like hypotheses related to difference between parameter estimates, scale shift or skewness shift (Hao and Naiman, 2007). Tests for parameter differences between multiple quantiles are conducted using Wald tests (Koenker, 2008).

The bootstrap is a method of estimating the sampling distribution of parameter estimates calculated from a sample drawn with replacement from and with size equal to the original data set (Hao and Naiman, 2007). This method provides reliable estimates of standard deviations for each parameter, especially when the distribution of the population cannot be identified as identically independently distributed (IID). The pair form (x_i, y_i) of this method provides simple and effective ways of drawing samples with replacement of pair (x_i, y_i) from the joint distribution of the original samples, with size *n* as large as *n* of the original sample. Each pair was drawn with the same probability of 1/*n* (Koenker, 2005).

Using the bootstrap may give two alternatives to make inferences about parameters. The first alternative is counting standard deviation of parameters and using it to obtain a t-value and its p-value of related parameters. Confidence intervals (CI) can be approximated using this method. The second alternative is by constructing 95% CI (or other CIs) using 97.5th percentile and 2.5th percentile of the samples of bootstrap estimates. If the CI captured the parameter, we can make the inference that the parameter is significant on α=.05 (Hao and Naiman, 2007).

Comparing Quantile Regression and OLS regression

Limitation of OLS regression

OLS-R is claimed to produce parameters with desirable characteristics - best, linear, unbiased estimators (BLUE). This means that parameters estimated using OLS-R have the smallest variance, model a linear relationship between response and outcome variables, and