Article

LANGUAGE
TESTING

Language Testing
2014, Vol 31(1) 63–87
© The Author(s) 2013
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0265532213493623
ltj.sagepub.com
SAGE

Principles of quantile regression and an application

Fang Chen
East China Normal University, China

Micheline Chalhoub-Deville
University of North Carolina at Greensboro, USA

Abstract

Newer statistical procedures are typically introduced to help address the limitations of those already in practice or to deal with emerging research needs. Quantile regression (QR) is introduced in this paper as a relatively new methodology, which is intended to overcome some of the limitations of least squares mean regression (LMR). QR is more appropriate when assumptions of normality and homoscedasticity are violated. Also QR has been recommended as a good alternative when the research literature suggests that explorations of the relationship between variables need to move from a focus on average performance, that is, the central tendency, to exploring various locations along the entire distribution. Although QR has long been used in other fields, it has only recently gained popularity in educational statistics. For example, in the ongoing push for accountability and the need to document student improvement, the calculation of student growth percentiles (SGP) utilizes QR to document the amount of growth a student has made. Despite its proven advantages and its utility, QR has not been utilized in areas such as language testing research. This paper seeks to introduce the field to basic QR concepts, procedures, and interpretations. Researchers familiar with LMR will find the comparisons made between the two methodologies helpful to anchor the new information. Finally, an application with real data is employed to demonstrate the various analyses (the code is also appended) and to explicate the interpretations of results.

Keywords

Growth modeling, linear regression, math and reading, quantile regression, relationships, scores

Traditionally, when the research interest is to examine the relationship between and among variables or when one wants to estimate how independent variables influence changes in a dependent variable, least squares mean regression (LMR) is the standard

Corresponding author:
Fang Chen, English Department, East China Normal University, 210, School of Foreign Languages, 500 Dongchuan Road, Minhang District, Shanghai, 200241, China.
Email: fennycf@yahoo.com

tool. However, it is sometimes difficult in the social sciences to meet two of the required regression assumptions, that is, normality and homoscedasticity (where the standard deviations of the error terms are assumed to be constant, or expressed differently, residuals are considered to be approximately equal for all predicted dependent variable scores). QR relaxes the need for these assumptions (Hao & Naiman, 2007). Also, regular regression focuses on the mean. However, with changes in higher order moments such as skewness or kurtosis of the distribution, the median is likely to be a more appropriate measure of central tendency than the mean (Edgeworth, 1888; Fox, 1997; Hao & Naiman, 2007; Koenker, 2005). In reality, it is also commonly observed that the relationship between variables can change at different points in the distribution. In that case, a single, average pattern cannot adequately represent a complex relationship that shifts rather than stays constant along the distribution.

So, if assumptions of normality and homoscedasticity are violated or previous research suggests the need to explore the relationship of variables across the distribution, quantile regression (QR) is a better alternative. This article aims to introduce this statistical tool to the language testing community. It is hoped that this introduction will encourage researchers to examine the usefulness of this tool to further their explorations and to expand the knowledge base in the field.

To demonstrate the application of QR, the paper employs data from the National Center for Education Statistics, that is, The Early Childhood Longitudinal Study Kindergarten Class (ECLS-K) Program (http://nces.ed.gov/ecls/). The data set used is the released full sample data posted on the website http://nces.ed.gov/ecls/kinderdatainformation.asp, which includes English language learners (ELLs) as well as non-English language learners (N-ELLs). The present application addresses the relationship between language proficiency and math achievement.

A quick survey of the published literature (Abedi & Gandara, 2006; Abedi & Lord, 2001; Bailey, 2005; Kato, Albus, Liu, Guven, & Thurlow, 2004; Kieffer, Lesaux, Rivera, & Francis, 2009; Solano-Flores, 2011; Stevens, Butler, & Castellon-Wellington, 2000; Wright & Li, 2008) shows that the relationship between language proficiency and math achievement has recently attracted a great deal of attention. Language proficiency has been investigated as a factor that contributes to the math achievement gap between ELLs and N-ELLs (Abedi & Lord, 2001; Cottrell, 1968). This differential relationship has been observed within and at various grade levels (Freeman & Crawford, 2008; Kopriva, Bauman, Cameron, & Triscari, 2009) and with different content areas (Abedi & Gandara, 2006; Abedi & Leon, 1999; Bailey, 2005). The published literature underscores the fact that the relationship between language ability and math achievement is not static, which suggests that the relationship should not be summarized by one average pattern based on the mean, but modeled along the full ability distribution, an undertaking for which QR is best suited. The article introduces QR and models its procedures by investigating the role that language ability plays in order to achieve in math.

A word is in order about the application presented in the paper. The application is part of a larger study in which several independent variables are investigated. Given the amount of space typically allotted to an article, it is not feasible to present the larger study and to introduce the methodology. Upon consideration of the newness, deemed importance, and complexity of the literature discussing the QR methodology, we decided to focus the first publication on introducing QR and to use a limited number of ECLS-K variables to explicate the methodology and its utility. The larger investigation will be prepared for publication in the near future. In the present application, the dependent variable is math achievement scores. Reading, which is used as a proxy of language proficiency, and gender are included as the independent variables. The gender variable is chosen primarily to move the demonstration beyond a simple analysis, to multiple quantile regression, and to make accessible to the field relevant computer program syntax and graphs.

Historical overview of quantile regression research

Roger Koenker (2005), the author of the first book devoted to QR, traced the procedure back to the mid-1700s by a Jesuit priest, Boscovich. This means that QR actually predates the introduction of least squares regression. In that first attempt to "ever *do* regression" (Koenker, 2005, p. 2), Boscovich estimated the slope coefficient through a process, which Laplace (1818) later noted as the "method of situation" because the model was an interesting mixture of central tendency measures. Although the slope was estimated based on the median, the intercept was still estimated as a mean. In 1888, Edgeworth improved Boscovich's and Laplace's ideas by proposing a process to minimize the sum of absolute residuals in both intercept and slope parameters. Thus, QR formally started.

Linear programming and technology advances have made efficient computation a manageable task and facilitated the use of QR with large-scale applications. QR has become a common statistical tool in many fields, such as medicine (Austin et al., 2005), biology (Wei et al., 2006), environmental studies (Pandey & Nguyen, 1999), survival analysis (Koenker & Geling, 2001), finance (Chevapatrakul, Kim, & Mizen, 2009), and economics (Koenker & Bilias, 2001). It is regarded as "the standard tool in wage and income studies in labor economics" (Yu, Lu, & Stander, 2003, p. 339) because of the less stringent assumptions and the advantages mentioned above.

The use of QR in the educational field is relatively new. Most of the earliest QR investigations focused on equality issues and appeared in journals of education economics (Haile & Nguyen, 2008; Wößmann, 2005). For example, Haile and Nguyen (2008) studied the achievement gap among different ethnic groups and the impact of gender. Results from traditional LMR analyses were consistent with established findings that Asian students scored on average better than White students in mathematics, regardless of gender. The QR results, however, offer a more nuanced depiction of this relationship. The QR analyses revealed that out of the five quantiles investigated (0.1, 0.25, 0.5, 0.75, and 0.9), a significant score difference was found between Asian and White males only at the lowest ability level, that is, the 0.1 quantile. On the other hand, Asian female students outperformed their White counterparts at all the other ability or quantile levels investigated (0.25, 0.5, 0.75, and 0.9).

In recent years, Damian Betebenner and colleagues (Betebenner, 2009a; Linn, Baker, & Betebenner, 2002) have used QR to formulate an innovative growth model of student achievement utilizing student growth percentiles (SGP), which has been embraced for accountability purposes in the United States. In 2005, former U.S. Secretary of Education, Margaret Spellings (Spellings, 2005), endorsed the Growth Model Pilot Program as an alternative for states to comply with NCLB achievement mandates. (For a detailed discussion of NCLB, see the special issue of *Language Testing* (Deville & Chalhoub-Deville, 2011).) Basically, and as opposed to simply reporting the percentage of proficient students every year, SGPs have been adopted by many states because, with the help of QR, it allows the documentation and investigation of the relative amount of growth that students make across a distribution. Growth models are now being promoted as a central component in the next reauthorization of NCLB (see Betebenner & Linn, 2010).

Language testers have also expressed interest in empirical research to document the abilities underlying performance as well as those contributing to growth, for example, diagnostic assessment programs such as DIALANG (Alderson, 2005), hierarchical/developmental language levels and descriptors such as the CEFR (North, 2000), and the development of proficiency scales for ELLs as part of NCLB, Title III tests such as WIDA-*ACCESS for ELL*s and ELDA (Bunch, 2011). The present article seeks to introduce the QR methodology to help spark interest among language researchers to explore its potential to address research questions (e.g., What is the magnitude of relationship between reading and content achievement at different points of a distribution? How can growth patterns be investigated in language achievement or performance?) that have interested researchers in the field.

Technical overview of quantile regression and software

Quantile is an equivalent term to percentile, where the median is the 50th quantile. Similarly, the 25th and 75th quantiles correspond to the first and third quartiles. QR modeling is a term for a series of QR alternatives. Quantiles are order-statistics and are relatively resistant to outliers. If errors follow a normal distribution, results of LMR and QR at the median coincide. If errors are not normally distributed or homoscedasticity does not hold, QR provides a more efficient and accurate estimate of parameters. In comparison to LMR, QR can uncover differences in the nature of a relationship at different points in the distribution. QR can better handle the unequal variation observed with one or more independent variables at various points of a dependent variable.

A conventional approach to explore the differential relationship between reading and math is to divide the population into subgroups based on students' math scores and conduct a series of classical regressions. Heckman (1979) argues strongly that such an approach could create biased parameter estimates. Figure 1 provides a visual presentation of the issue.

Figure 1 includes four LMR regression plots, including the total group and three subgroups, which document the relationship between reading and math for a sample of grade 5 students in US schools. The LMR total data plot shows a stronger relationship between the reading and math scores. The relationship looks different, however, when the students are divided into three math ability groups with equal numbers. Compared to the overall regression line based on the complete data, the flatter regression lines in the subgroup analyses indicate a weaker relationship between reading and math scores. Such results suggest that the truncated LMR fails to discover the strong relationship between reading and math scores at different ability levels. QR is a more appropriate analytic tool to study a changing relationship, such as the one observed in the present example.

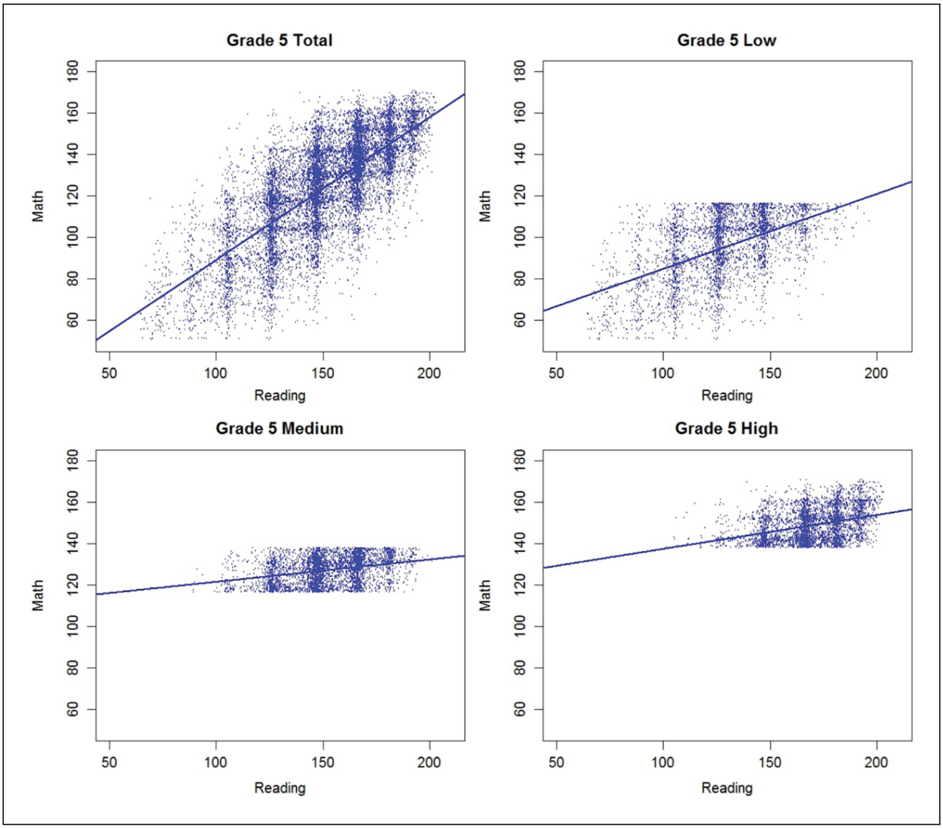


Figure 1. Total mean regression versus subgroup mean regression plots.

(The graphical results are expected based on the restricted range in the three subgroups. The main point, however, is that QR can model the relationship among the variables more efficiently—namely with one analysis—and can reveal where the relationship of the variables differs within the distribution).

Several software packages are available to perform QR analyses. These include commercial programs such as SAS and Stata. Free programs are also available, for example, the Quantreg package in R (cran.r-project.org/package=quantreg) and Blossom (www.fort.usgs.gov/products/software/blossom/). R has the most complete and easy-to-carry-out functions. Quantile process plots can be obtained with all these programs. The ggplot2 package in R (cran.r-project.org/package=ggplot2) is especially useful for tailored plots. Appendix A provides R or Stata code to run the analyses reported on in the present paper. Code for the Figure 3 example is too long to be included but is available upon request. In summary, "With today's fast computers and wide availability of statistical software packages … fitting a quantile regression model to data has become easy. However, we have so far had no introduction … to the method to explain what quantile-regression is all about" (Hao & Naiman, 2007, p. vii). This quotation represents quite adequately the state of affairs with regard to QR in fields such as language testing. The present paper seeks to remedy this situation.

Application data and data layout

The present QR illustration employs the grade 5 ECLS-K data, which is a partial set of the Kindergarten–Eighth Grade Full Sample Public-Use Data File especially prepared for longitudinal studies. More information about this data set can be located in the Combined User's Manual for the ECLS-K Eighth-Grade and K–8 Full Sample Data Files and the Electronic Codebooks (Tourangeau, Nord, Lê, Sorongon, & Najarian, 2009) and the ECLS-K Psychometric Report for the Eighth Grade (NCES 2009-002) (Najarian, Pollack, & Sorongon, 2009). The ECLS-K was developed under the sponsorship of the U.S. Department of Education, Institute of Education Sciences and National Center for Education Statistics.

The selected data set includes 11,265 students in grade 5. The students represent those currently classified as ELLs, formerly classified as ELLs, and N-ELLs. Their respective percentages are approximately 3%, 14%, and 83%. An examination of the math distribution shows that whereas former ELL and N-ELL students are distributed evenly, more ELLs fall into the lower end of the math distribution.

The central research question in the present application is how language ability, operationalized as a reading score, contributes to performance in a content area, specifically math. The reading and math scores are based on assessment instruments designed for the ECLS-K program. Scores are derived using the three-parameter IRT model and are vertically scaled from kindergarten to the 8th grade. The score range for reading is 64–203 and for math is 51–171. In terms of the gender variable, the data are coded 0 for males and 1 for females.

The data for LMR as well as for QR are organized in a similar fashion for analyses to be carried out. It is the computation algorithm, not the data set-up that makes the difference in the types of analyses conducted. Figure 2 provides a snapshot of the data layout used in the current paper. Finally, to help the reader better anchor the QR concepts introduced, LMR modeling is presented first and a comparison is made between the two statistical tools.

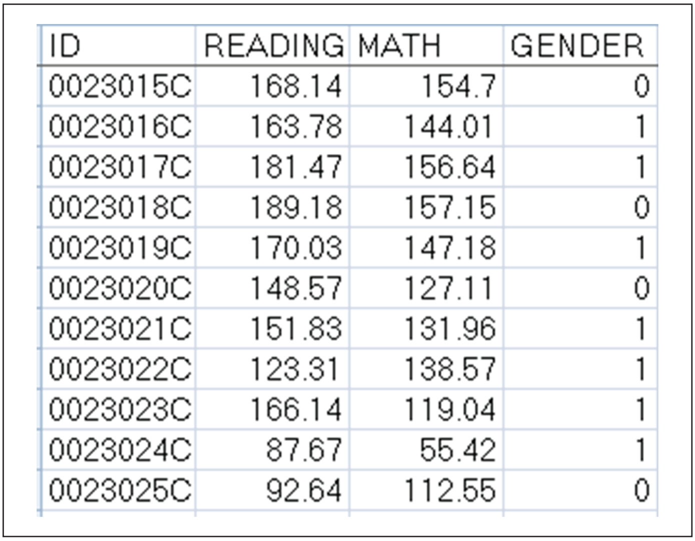
Finally, and with regard to the quantiles chosen in the present application, seven quantile points are selected. These quantiles, commonly seen in the QR literature (e.g., Buchinsky, 1994; Haile & Nguyen, 2008; Koenker & Hallock, 2001; Konstantopoulos, 2009; Wößmann, 2005), include .05, .10, .25, .50, .75, .90, and .95.

Equations

Using one independent variable as an example, a simple LMR model can be written as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{1}$$

β_1 is the slope (i.e., the steepness of the regression line), which represents the strength of the relationship between variables x and y and β_0 is its intercept on the y-axis. For the ECLS-K example, the formula is written as



ID	READING	MATH	GENDER
0023015C	168.14	154.7	0
0023016C	163.78	144.01	1
0023017C	181.47	156.64	1
0023018C	189.18	157.15	0
0023019C	170.03	147.18	1
0023020C	148.57	127.11	0
0023021C	151.83	131.96	1
0023022C	123.31	138.57	1
0023023C	166.14	119.04	1
0023024C	87.67	55.42	1
0023025C	92.64	112.55	0

Figure 2. A snapshot of data layout for QR regression.

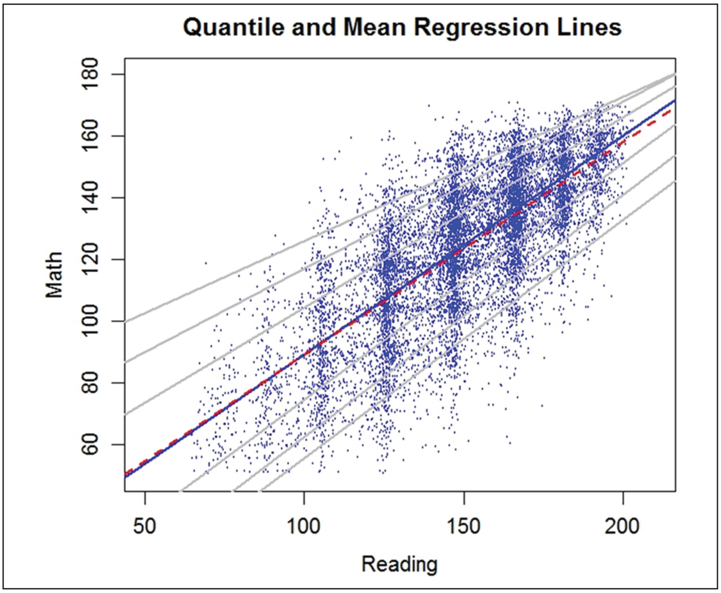


Figure 3. Quantile and least squares mean regression.

$$MATH_i = \beta_0 + \beta_1 READING_i + \varepsilon_i \tag{2}$$

Conceptually, we wish to investigate the relationship between reading proficiency and math test scores. The data are used to find a single regression line that minimizes the error term (thus also the least squared function). Algebraically, the goal is to find the point where the first derivative of the mean squared deviation is zero with respect to the mean. Graphically, the resulting regression line minimizes the sum of squared vertical distances of all response observations from the regression line. The best fitting line is the one that passes through the expected means of the response distributions conditioned at every value of the independent variable.

In comparison, a QR model can be written as

$$y_i = \beta_0^{(p)} + \beta_1^{(p)} x_i + \varepsilon_i^{(p)} \tag{3}$$

Or using the ECLS-K example,

$$MATH_i = \beta_0^{(p)} + \beta_1^{(p)} READING_i + \varepsilon_i^{(p)} \tag{4}$$

The only notational difference between Equations as in equations plural 1 and 3 is the extra superscript "*p*", which specifies the *p*th QR model.[2] Usually a predetermined set of QR models are compared to detect the different effect of the independent variable on the dependent variable at various quantiles of the response distribution. It is important to note that all the data points are used for every QR modeling.[3] Taking the ECLS-K as an example, where the reading score is the independent variable and math the dependent variable, the best fitting line for $p = .5$ passes the conditional 50th percentile (the median) of the math score distribution. In other words, half of the math scores lie above the median regression line and half below the line. For the regression line at $p = .75$, 75% of the cases are below the best fitting line and 25% are above. Similar interpretations apply to other QR *p*s.

Figure 3 shows the plots of the LMR as well as seven QR lines for the ECLS-K example. The seven QR lines correspond to, from the bottom up, the regression modeling with conditional math percentiles at .05, .10, .25, .50, .75, .90, and .95. The LMR line (the dotted line) is very close to the median QR line (the solid line in the middle). However, the other QR lines (solid gray) all have different intercepts and slope coefficients. The slopes indicate that there is a differential relationship between reading and math scores at different parts of the distribution. For instance, the relationship seems stronger for students with low math scores (see the bottom line) than for those with high math scores (see the top line). The LMR subgroup models in Figure 1, on the other hand, show that the low, medium, and high slopes are relatively flat, thus indicating a weak relationship between language and math achievement,

Parameter estimation

In LMR modeling, estimates of the intercept and slope coefficients of the best-fitting line are the ones that minimize the sum of squared errors and is written as

$$\sum_i^n \varepsilon_i^2 = \sum_i^n (y_i - (\beta_0 + \beta_1 x_i))^2 \tag{5}$$

The estimates can be shown to be $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_0 = \bar{x}$ and $\widehat{\beta}_1 = \dfrac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^n (x_i - \bar{x})^2}$. When the assumptions of linearity, constant variance, and independence of *x* values are met, ordinary least square estimation provides the best, unbiased estimators of the population parameters.

In QR modeling, estimates of the intercept and slope coefficients that correspond to the best-fitting line are the ones that minimize the *weighted* sum of *absolute* errors

$$\sum_{i}^{n} w_p \left|\varepsilon_i\right| = \sum_{i}^{n} w_p \left|y_i - (\beta_0^{(p)} + \beta_1^{(p)} x_i)\right| \tag{6}$$[4]

$$\text{where } w_p = \begin{cases} p & when \quad y_i \geq (\beta_0^{(p)} + \beta_1^{(p)} x_i) \\ 1-p & when \quad y_i < (\beta_0^{(p)} + \beta_1^{(p)} x_i) \end{cases}$$

$$\text{or } p \sum_{y_i \geq p} \left|y_i - (\beta_0^{(p)} + \beta_1^{(p)} x_i)\right| + (1-p) \sum_{y_i < p} \left|y_i - (\beta_0^{(p)} + \beta_1^{(p)} x_i)\right| \tag{7}$$

When p = .5, both simplify to $\sum_{i}^{n} \left|y_i - (\beta_0^{(0.5)} + \beta_1^{(0.5)} x_i)\right|$

The solution that minimizes the weighted sum of absolute distance is when $\hat{y}_i = \beta_0^{(p)} + \beta_1^{(p)} x_i$ equals the *p*th percentile. For more detailed information on QR parameter estimation, see Koenker (2005) and Hao and Naiman (2007).

Several algorithms are available to estimate the QR parameters, for example, simplex (Koenker & d'Orey, 1987, 1994), interior point (Portnoy & Koenker, 1997), and the smoothing method (Chen, 2007). The default algorithm in both the Quantreg package in R and SAS is simplex. However, this method is computationally demanding and thus not recommended for large sample sizes. For sample sizes larger than 5000 observations and 50 variables, interior point is considered more efficient (SAS, 2008, p. 5400).

The estimates for the QR coefficients using the interior point algorithm are summarized in Table 1. The slope for the reading score in LMR is .69. The QR slopes of the reading scores for different math ability students vary from .77 (at the .05 quantile) to .47 (at the .95 quantile). The values, as noted, are larger at lower quantiles than at higher quantiles. This implies that the relationship between reading and math is stronger for low-scoring math students and weaker for high-scoring math students. In comparison, the LMR average-centered slope of .69 underestimates the relationship for low math ability students but overestimates it for high math ability students.

The QR parameter estimates quantify the differential relationship between reading and math performance observed in Figures 1 and 3. The QR estimates indicate that reading plays a more integral role with low-achieving math students and much less so with high-achieving math students. Given the QR estimates, we can clearly state that the LMR average-centered slope diminishes the importance of reading ability and the role it plays in predicting math achievement for low-scoring math students. On the other hand, for high-scoring math students, the LMR slope depicts an exaggerated role for reading in terms of students' math performance. The differences in LMR and QR findings are not statistical nuances but critical flag posts that advise differentiated and more focused language attention for students to be able to improve their math achievement. Given the

Table 1. Slope coefficients.

	LMR	QR						
		.05	.10	.25	.50	.75	.90	.95
Intercept	20.44	−21.47	−15.88	−1.93	18.38	42.80	63.19	79.29
Slope	.69	.77	.78	.77	.71	.62	.54	.47
SE	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)
p	.00	.00	.00	.00	.00	.00	.00	.00

present QR findings, students with low math scores do not simply need more math; they need more language instruction.

Standard errors and confidence intervals

Once the coefficients are estimated, standard errors are calculated to help test the statistical significance of the strength of the relationship, that is, the slope coefficient estimate $\hat{\beta}_1^p$. The null hypothesis specifies that the slope coefficient is equal to 0, which means there is no linear relationship between the independent and dependent variables. In this section, the discussion focuses on the standard errors and the confidence interval estimation. The section that follows presents this information in a graphical form. Subsequent sections address hypothesis testing and goodness-of-fit.

In LMR, the standard error for the coefficient β_1 is calculated by assuming a normal distribution of the error term. That is, the ε_i in equation 1 is regarded as independently and identically distributed across all covariate values with a mean of 0 and a constant variance of σ_ε^2 (In fact, the subscript "*i*" can be dropped.) The σ_ε^2 is not known but the variance of the residuals, s_ε^2, provides an unbiased estimator of σ_ε^2 (Fox, 1997) .

$$s_\varepsilon^2 = \frac{\sum \varepsilon_i^2}{n-2} = \frac{\sum\left[y_i - (\beta_0 + \beta_1 x_i)\right]^2}{n-2} \qquad (8)$$

The sampling variance of $\beta_1$can then be estimated as

$$V\hat{a}r_{\beta_1} = \frac{s_\varepsilon^2}{\sum(x_i - \bar{x})^2} \qquad (9)$$

and the estimated standard error of the slope coefficient is just the square root of the sampling variance.

$$SE_{\hat{\beta}_1} = \frac{s_\varepsilon}{\sqrt{\sum\left(x_i - \bar{x}\right)^2}} \qquad (10)$$

$(\beta_1 - \beta_1^{null})/SE_{\beta_1}$ is assumed to follow a student's *t* distribution with *n* − *2* degrees of freedom and thus the 100(1 − α)% confidence interval for β_1 is given by

$\hat{\beta}_1 \pm t_{\alpha/2} SE_{\hat{\beta}_1}$. The confidence interval helps assess the precision of the estimated $\hat{\beta}_1$, that is, the extent to which our estimated slope coefficient represents the population value.

As reported in Table 1, the LMR slope $\hat{\beta}$ is .69 and the standard error is .01. For the 95% confidence interval, $t_{.05/2}$ is almost the same as $z_{.05/2}$, which is 1.96. Thus, the final 95% confidence interval falls between .69 − 1.96 × .01 and .69 + 1.96 × .01, which correspond to .67 and .71. This narrow band of confidence interval can be clearly seen in Figure 4 by the solid horizontal line and the dotted lines closely above and below it. This is also consistent with the significant *p* values ($p = .00$) in Table 1, which implies high precision of estimation.

One reason for using QR modeling over LMR is that the conditional response distribution is skewed rather than normal. In such cases, the traditional approach for calculating the standard error is not appropriate for QR modeling. Instead, it is recommended to use bootstrap methodology (Efron, 1979), such as the xy-pair technique, which does not require a specific distributional form.[5] The observed data set is regarded as the population and the algorithm bootstraps pairs of observations (e.g., a reading score with a corresponding math score) from the data repeatedly and generates multiple samples. Every sample gives a parameter estimate, which yields a distribution of the $\hat{\beta}_1 s$. The standard deviation of these $\hat{\beta}_1$ is taken as the standard error of the parameter β_1. As the number of bootstrapped samples increases, the sampling distribution of the $\hat{\beta}_1$ is approaching normal distribution,[6] and the confidence interval follows the form of $\hat{\beta}_1 \pm z_{\alpha/2} SE_{\hat{\beta}_1}$ (Koenker & Bassett, 1982).

Another approach to determining the confidence interval that does not require that estimates be normally distributed and capitalizes on the set of bootstrap samples obtained entails taking the empirical values from the distribution of the estimated $\hat{\beta}_1$and locating the corresponding empirical percentiles. For example, the 95% confidence interval of the parameter β_1 starts from the 2.5th percentile of all the $\hat{\beta}_1 s$ from the empirical samples and ends at the 97.5th percentile of the estimated $\hat{\beta}_1 s$ from these samples.

The standard errors of estimation for the ECLS-K example are provided in Table 1 in parentheses. These estimates are produced, by default, in Stata by randomly sampling the data (StataCorp, *Base Reference Manual*, 2009, p. 1457). As the table shows, all the standard error values are .01, which indicates that when these confidence bands are plotted, they will be very close to their regression lines. This is clearly the case, as depicted in Figure 4, which is discussed in the next section.

The standard errors indicate a high degree of precision in each of the LMR and the QR plots. This precision, while desirable in terms of the quality of the estimates obtained, is not reassuring because the models do not yield comparable results. In the absence of the QR information, researchers are likely to move forward with a less than accurate depiction of that relationship. It is only when we analyze the type of information provided by each of the two models that we can discern that QR facilitates a more nuanced understanding of the role played by reading ability for students with different math achievement scores. These findings impact our understanding of the nature of the relationship and the consequent instructional practices needed.

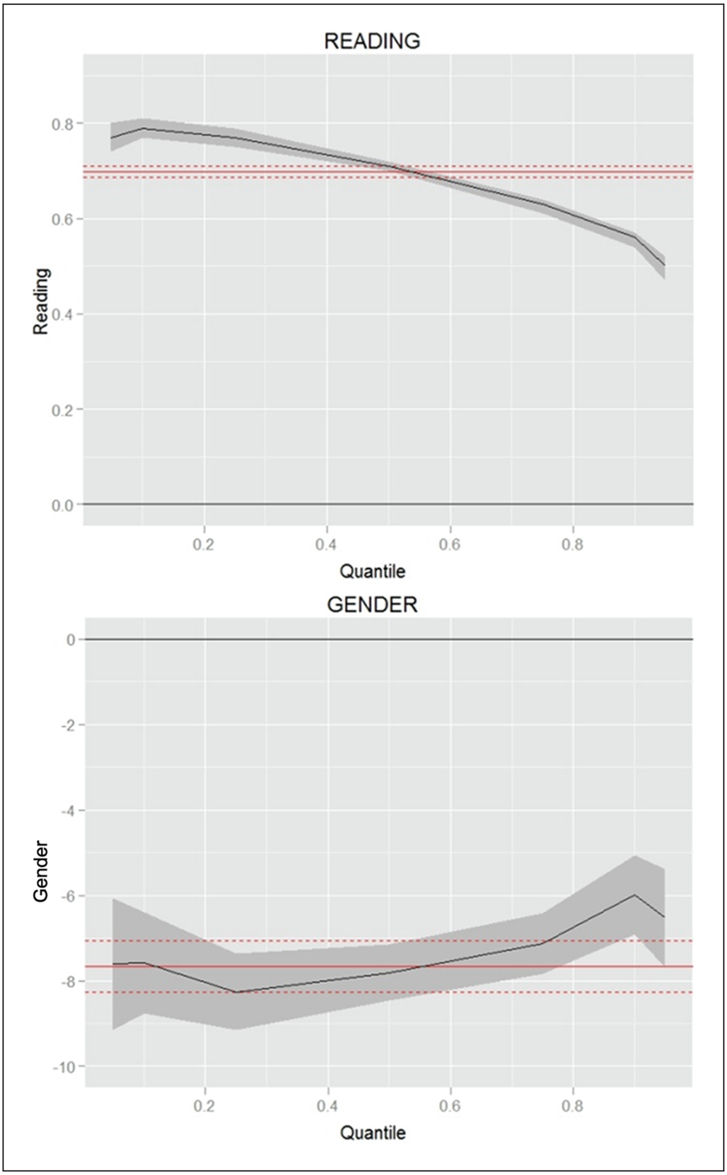


Figure 4. Quantile process plots.

Notes: The black solid line with gray areas represents the quantile regression slope estimates at each quantile and the confidence interval for the estimates respectively.

The horizontal solid red line with dotted red lines above and below represents the mean LSR regression estimates and confidence band.

The horizontal line at zero is the reference line for hypothesis testing against a slope value of 0.

Graphs

For LMR and QR, when there is only one independent variable, as shown in equation 4, the regression lines can be plotted out directly. The difference is that in LMR there is only one regression line representing the mean pattern of relationship between the dependent and independent variables, whereas in QR there are usually many quantile regression lines corresponding to the relationship at several quantiles of interest (see Figure 3).

For quantile models that involve at least two independent variables or covariates, a unique form of graph called a *quantile process plot* is used to present the complex set of regression lines that depict the changes in the slope coefficients at each quantile. This quantile process plot can show more clearly how the coefficients differ across quantiles. Both the *quantreg* package in R and SAS can produce process plots.

Figure 4 is an example of a quantile process plot based on the QR model defined by equation 11. This model includes an additional variable, Gender, as a covariate.

$$MATH_i = \beta_0^{(p)} + \beta_1^{(p)} READING_i + \beta_2^{(p)} GENDER_i + \varepsilon_i^{(p)} \tag{11}$$

In Figure 4, the x-axis includes the seven quantiles from .05 to .95 and the y-axis presents the corresponding slope coefficients from these QRs. Figure 4 shows that the reading slopes are all positive (all the values are above the 0 reference line). The slope coefficients for reading decrease from .77 to .50 as students' conditional math ability moves up in quantile values. The narrow confidence band (gray area), similar to the standard errors addressed above, indicates that the estimations are quite precise. The plot also shows little overlap with the LMR confidence band (area between the two dotted horizontal lines), which illustrates how LMR and QR differ in modeling the relationship between math and reading. As already stated, in comparison to QR, the LMR model underestimates the relationship for low-scoring math students and overestimates that relationship for high-scoring math students. These results, however, need statistical significance confirmation, which is presented under the 'Hypothesis testing' section.

The slope coefficients for the Gender variable are all negative. Since males are coded as 0 and females as 1, the negative slope coefficient means females tend to score lower than males in math. This pattern is true for both LMR and for the QR models at all seven quantiles. However, the quantile process plot shows that the difference in math scores between males and females is smaller for high-scoring math students, such as at the conditional quantile of .90 as compared to the quantile of .25. These results, similar to those with reading, provide a more nuanced depiction of the nature of the relationship between these two variables.

Comparing the reading and the Gender plots in Figure 4, it is evident that the confidence band for the Gender slope coefficients (i.e., the width of the gray area) reveals less precision of estimation compared to the estimation of reading. The Gender coefficients are still statistically significant since the confidence band does not cross the 0 reference line (hypothesis of the slope coefficient being 0). Finally, the interaction effect between Reading and Gender is not modeled here, but it can be explored and graphed.

Hypothesis testing

With LMR, hypothesis testing for the significance of a single independent variable draws on the central limit theorem and follows regular regression procedures. The *t*-statistic is calculated as follows:

$$t = \frac{\hat{\beta}_1 - \beta_1^{null}}{SE_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - 0}{SE_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{SE_{\hat{\beta}_1}} \qquad (12)$$

The result is compared against the critical *t* with *n* −2 degrees of freedom under the null distribution. The hypothesis testing for a single independent variable in QR follows the same pattern with the only difference being that $SE_{\hat{\beta}_1}$ is estimated based on bootstrapped samples as described in a previous section. Using the model in equation 4 again, the results of the hypothesis of $\beta_1^{null} = 0$ is summarized in Table 1. All the Reading slopes are statistically significant at the level of $p = .00$, whether it be LMR or QR.

However, because in QR several quantiles are modeled, additional hypotheses are of interest such as the equivalence of the various slope coefficients across quantiles. In QR, the xy-pair bootstrapping method described previously is used to produce a covariance matrix of the cross-quantile estimates, which can then be employed to perform this hypothesis testing, known as the test of equivalence. For a given variable, the covariance matrix allows the examination of whether any difference between the slope coefficients of any pair of quantiles is statistically different. For example, we can investigate whether there is a statistical difference in how reading performance predicts math scores when the students are at the 75th versus the 90th percentile of the math score distribution. The Wald statistic, shown in equation 13, is used for the test of equivalence.

$$\text{Wald statistic} = \frac{(\hat{\beta}_1^{(p)} - \hat{\beta}_1^{(q)})^2}{\hat{\sigma}^2_{\hat{\beta}_1^{(p)} - \hat{\beta}_1^{(q)}}} \qquad (13)$$

In this equation, $\hat{\beta}_1^{(p)}$ is the parameter estimate from the *p*th QR model and $\hat{\beta}_1^{(q)}$ is the parameter estimate from the *q*th quantile regression model (i.e., any given pair of quantiles). The denominator is the variance of the difference between the two coefficients for the *p*th and *q*th quantile regressions. Obviously, the Wald statistics can be extended to a joint test of equality of slopes at the same time. In that test, the null hypothesis becomes $H_0 : \beta_1^{.05} = \beta_1^{.10} \cdots = \beta_1^{.90} = \beta_1^{.95}$ which is an omnibus test.

In a regression model with one independent variable, the Wald statistic follows a χ^2 distribution with one degree of freedom. In a model with *p* independent variables, the Wald statistic follows a χ^2 distribution with *p* degrees of freedom (Koenker & Machado, 1999). Thus, the Wald statistic can be readily extended for more complicated models for the test of equivalence of coefficients between quantiles. The Wald test is readily available in computer programs such as Stata and the Quantreg package in R (Koenker, 2009). Stata uses the *sqreg* command and Quantreg uses the command *anova.rq* to test the equivalence of coefficients between quantiles.

Table 2. Test of equivalence.

	Overall	.05 = .10	.10 = .25	.25 = .50	.50 = .75	.75 = .90	.90 = .95
READING	86.86	1.75	4.14	82.67	187.63	78.40	33.22
	.000	.186	.042	.000	.000	.000	.000
	***			***	***	***	***

The first row presents the test statistics; the second row presents the *p*-values.
***indicates significance level at or below .01.

Using the ECLS-K data, Stata produces the results based on the model in equation 4 in Table 2 on the equivalence of coefficients. The overall statistics is the omnibus test, which shows that there is statistically significant difference between some or all of the slope coefficients. Furthermore, pair-wise Wald tests reveal that the reading slope is statistically different between quantile .25 and .50 ($w = 82.67$, $p = .000$). This means the relationship between reading and math scores is different between students whose conditional math ability is at the 25th percentile versus those at the 50th percentile. In addition, there is a significant difference between all upper pairs.

This significance for all the pair-wise comparisons of.25 and above provides statistical confirmation with regard to the differential relationship between math and reading at different points of the math score distribution. This statistical significance is strong evidence that analyses of the relationship between reading and math should be explored beyond the mean of the distribution. In conclusion, QR is a more appropriate methodology when a differential rather than an average relationship is thought to exist between and among variables.

Substantively, the lack of statistical significance for the two lowest quantile pairs could perhaps be interpreted as the incapacity to differentiate among students' math achievement for those who possess low reading ability. The statistically significant pair-wise comparisons observed with the other quantile pairs indicate that the relationship is not uniform at the conditional math ability percentiles investigated. These results suggest that the use of reading performance to predict math achievement needs to be considered at the specified points of the math distribution. Given the parameter and other statistical estimates reported earlier, we can conclude that for students struggling with math, one seeming course of action is to attend to developing these students' reading skills to help promote their capability to engage and achieve in the content. Furthermore, for students with high math scores, evidence shows that prediction of their math performance using reading ability is less strong. These students seem to have attained a requisite language level and thus are less impacted in terms of their math performance.

QR goodness-of-fit index

For LMR, R² is the usual measure of goodness-of-fit. It is defined as follows:

$$R^2 = \frac{SSR}{SST} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{SSE}{SST} \qquad (14)$$

R^2 is the ratio of the sum of squares due to regression (SSR) and the sum of squares of the total model (SST). As is commonly known, R^2 represents the proportion of variance in the response variable being explained by the independent variables in the regression model. Inversely, R^2 can be seen as the proportion of error variance (SSE/SST) subtracted from 1 (this notion comes into play in the following paragraph). R^2 ranges between 0 and 1. Higher values indicate a stronger relationship.

In QR, a similar index is suggested by Koenker and Machado (1999), which is the likelihood ratio of the sum of weighted absolute distances for the full *p*th QR model $V^1(p)$ and the sum of the weighted absolute distances for a model with only the intercept $V^0(p)$. Stata labels this ratio *pseudo*-R^2 to distinguish it from the LMR R^2. The default Stata output includes R^2 and *pseudo*-R^2. The equation for *pseudo*-R^2 is[7]

$$Pseudo - R^2 = 1 - \frac{V^1(p)}{V^0(p)} = 1 - \frac{p \sum_{y_i \geq \hat{y}_i} \left| y_i - (\beta_0^p + \beta_1^p) \right| + (1-p) \sum_{y_i < \hat{y}_i} \left| y_i - (\beta_0^p + \beta_1^p) \right|}{p \sum_{y_i \geq \hat{Q}^{(p)}} \left| y_i - \beta_0^p \right| + (1-p) \sum_{y_i < \hat{Q}^{(p)}} \left| y_i - \beta_0^p \right|} \qquad (15)$$

For the model $V^0(p)$, the intercept is the sample *p*th quantile $\hat{Q}^{(p)}$ of the dependent variable. In the ECLS-K example here, the intercept for the *p*th quantile regression is the reading score at the *p*th percentile. Both $V^0(p)$ and $V^1(p)$ are nonnegative since they are the sum of absolute values. $V^1(p)$ is always equal to or smaller than $V^0(p)$ since a covariate is supposed to explain some variance of the dependent variable. Similar to the R^2, the *pseudo*-R^2 range is 0–1, with a larger value indicating better model fit.

The goodness-of-fit results of the ECLS-K data for the two models defined by equation 4 and 11 are summarized in Table 3. The results in Table 3 show that with regard to the LMR R^2, reading helps to explain 54% of the variance in math scores, while gender only explains an additional 2% (the difference between .56 and .54 in column 1) of the total variance. These LMR results show that while both variables are statistically significant, their meaningfulness is quite different. The explanatory magnitude of reading (i.e., the amount of variance in the math dependent variable that can be predicted from students' performance on the independent variable of reading) is substantial and deserves serious consideration. The small 2% increment renders the contribution of the other independent variable, gender, to be effectively meaningless. Having said that, one could still argue that this increment, while exceedingly small, points to a differential reading and math relationship for males and females, which may be important in some contexts.

The QR *pseudo*-R^2 is interpreted as a measure of the relative effectiveness or goodness-of-fit of the model in explaining the data at the *p*th quantile. The results in Table 3 show that R^2 is always higher than *psuedo*-R^2s. Such a pattern is typically observed in quantile regression studies (see Drescher & Goddard, 2011). These values, which range between .20 and .36, are still respectable and underscore the practical value of having reading performance as a predictor of students' math scores. Finally, similar to the LMR R^2, the QR *pseudo*-R^2 at each quantile indicates that the gender variable does not contribute much to the explanation of the total variance in the data once reading is controlled for.

Table 3. R^2 and *pseudo-*R^2.

	R^2	Pseudo-R^2 at each p						
Model		.05	.10	.25	.50	.75	.90	.95
$MATH_i = \beta_0^{(p)} + \beta_1^{(p)} READING_i + \varepsilon_i^{(p)}$	.54	.34	.36	.36	.33	.29	.24	.20
$MATH_i = \beta_0^{(p)} + \beta_1^{(p)} READING_i + \beta_2^{(p)} GENDER_i + \varepsilon_i^{(p)}$	.56	.36	.37	.38	.35	.31	.26	.22

Summary and remarks

In a traditional LMR, a single, mean-based slope is estimated to describe the relationship between a dependent, response variable and an independent, predictor variable(s). With this approach to modeling, the statistics are restricted in terms of their portrayal of a relationship and cannot address whether the variables show significantly different patterns of association at points other than the mean. The traditional LMR analysis at the conditional mean is expanded with QR to provide a richer picture of variable relationships.

QR is considered a methodological improvement because it can depict a more detailed picture of the relationship between variables by estimating multiple slopes along the entire response distribution. The purpose of this article is to introduce the QR methodology and to show researchers in the field how it can be utilized when multiple quantiles in a conditional distribution of the response variable are of interest. The paper explicates how to compute QR estimates and related statistics through a modification or extension of the familiar LMR methodology.

Special attention was given in the article to formulating interpretations of the quantitative findings, a critical issue with a less familiar methodology. However, given the illustrative nature of the present application, the reader is cautioned when using these preliminary findings to draw conclusions about the relationship between reading and math. Whereas the results are accurate, they are restricted in scope given the delimited independent variables employed. The interpretations should be viewed more as modeling potential results of QR principles and procedures.

Language testing research/applications

The application employed in the present paper highlights how a relationship can be explored at various ability levels of a distribution. This relationship is modeled given a group's one-time performance: one set of reading and math scores. In an alternative, hypothetical implementation of QR, this statistical tool can be used to model growth across time for students/learners at different points of an initial distribution of performance. In other words, the one-time performance could be modeled longitudinally with scores at several points in time to delineate a growth pattern. We can modify the graph in Figure 4 (the reading plot) to look at the performance of a given subgroup of students over time, for example, those at the .25 quantile. The horizontal x-axis of the figure is no longer quantiles, but instead specifies longitudinal assessment administration points. In this hypothetical revised plot we can observe the performance of this group of .25 quantile students longitudinally and track their growth pattern. Moreover, this growth pattern or trajectory can be used to help make predictions about expected future performance.

The SGP model, mentioned at the beginning of the article, links performance scores over time utilizing QR (SGPs have been popular with numerous states such as Colorado (www.schoolview.org/GMFAQ.asp) and New Jersey (www.nj.gov/education/njsmart/performance/) to generate individual as well as group—buildings and districts—accountability reports). The SGP model is said to provide a norm- and a criterion-referenced depiction of student growth (Betebenner, 2009a). Normatively, the model provides "the relative location of a student's [or a group's] current score compared to the current scores of students with similar score histories." The location in this reference group of "academic peers" is expressed as a percentile rank. For example, a student earning an SGP of 80 performed as well as or better than 80 percent of her academic peers" (Castellano & Ho, 2012, p. 87). Additionally, SGPs are portrayed against criterion-referenced information such as proficiency descriptors/categories. This norm- and criterion-referenced depiction of growth is illustrated in Figure 5, which was adapted from Betebenner (2009b). Betebenner generated his figure based on real student achievement data but we make use of it here to illustrate hypothetical applications in language testing.

Figure 5 provides an illustrative language testing example of SGP growth trajectories that we could produce using QR with longitudinal assessment data. The white SGP lines depict longitudinal patterns of growth for students at different quantiles (percentile levels) along the y-axis. The x-axis comprises performances collected at different points in time, labeled as 'longitudinal administration points' in Figure 5. These longitudinal administration points could be progressive levels of language assessments related to a suite of certificate-based exams, a sequenced textbook series of tests provided by a publisher, successive course assessment offered at a language school, or annual exams for ELLs as required by NCLB.

Students' achievement SGPs are portrayed over four (gray-scale) levels that represent proficiency categories, which are typically derived from a framework (e.g., the ACTFL Guidelines, the CEFR, or descriptors of language expected in a content domain) and quantified using some standard setting procedure (see Cizek & Bunch, 2007). By connecting the achievement-based SGPs to proficiency standards, we can then chart the potential for growth to the next proficiency level. The seven black lines in Figure 5 illustrate a variety of growth trajectories that one might expect (given the actual rates of progress observed with the SGP achievement data) for a student at the 30th percentile/Below Proficient–Proficient threshold. The black growth trajectories across the longitudinal administration points show that a typical growth trajectory, 50th percentile, is needed for the student to reach the Proficiency category and that a remarkable growth potential, 90th percentile, can move the student by longitudinal administration point 8 to an Advanced Proficient category. To sum up, these growth percentile trajectories account for students' current status, offer multiple scenarios of growth, quantify the likelihood of achieving an amount of growth, and depict the level of proficiency—and related descriptor information—attained.

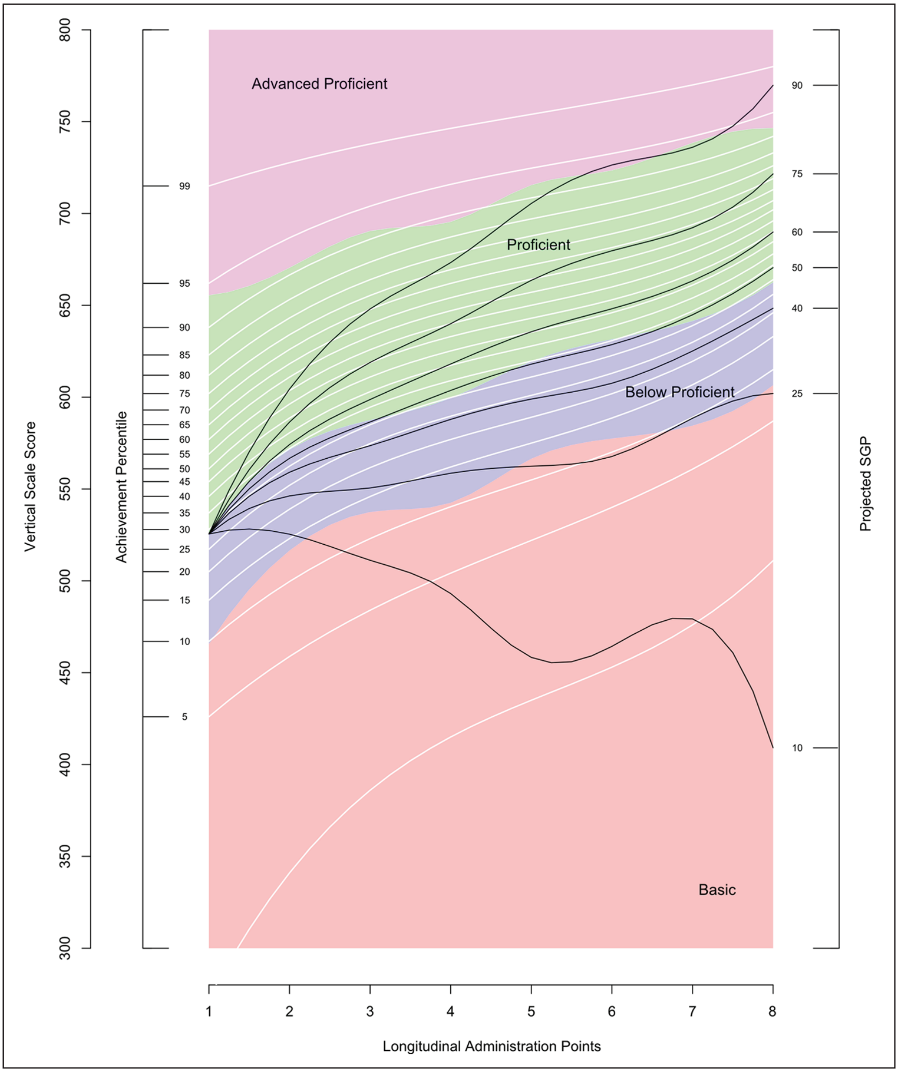


Figure 5. Language growth modeling.

QR is a statistical methodology that can facilitate the analysis of a variety of research questions and lead to worthwhile applications. Language testers are encouraged to consider the usefulness of QR as a tool that can strengthen and expand their research investigations, enrich the field's understanding of phenomena of interest, and enhance score interpretation and use.

Acknowledgement

We would like to acknowledge and thank Craig Deville, Arianto Wibowo, and the anonymous reviewers for their comments and suggestions on earlier drafts of the manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Notes

1. The methodology sections benefit from and correspond to the structure found in Hao and Naiman (2007).
2. Koenker and other authors use the letter τ rather than p. For ease of communication, p is used here to reference a given percentile value.
3. It is a misconception that only a subset of the observations is used for every quantile regression. It is standard procedure to use all observations in a given data set to locate a quantile given that it is the the pth value in the ordered observations. Also, the quantile regression analysis is a minimization of the weighted sum of absolute residuals for all the observations.
4. Koenker's notation for this concept is $\sum_{i=1}^{n} \rho_\tau (y_i - \xi)$. The notation used in this paper is more consistent with notations commonly seen in equations for least squares regressions in the social science literature.
5. Other techniques are available such as the Parzen, Wei, and Ying's (1994) version of the xy-pair bootstrap and the Markov chain marginal bootstrap by He and Hu (2002) and by Kocherginsky, He, and Mu (2005). Non-bootstrap methods have also been developed. For more details about these methods, the reader is referred to Koenker (2005).
6. The normal distribution here refers to the distribution of the $\hat{\beta}_1 s$ from all the bootstrapped samples. This is related to the features of *sampling distribution*, which as the central limit theorem describes, will lead to a normal distribution of the $\hat{\beta}_1 s$ if we repeat the sampling procedure enough times. This is different from the mean regression normality assumption of where the *error terms* are required to be normally distributed.
7. In Hao and Naiman (2007), this is denoted as $R(p)$.

References

Abedi, J., & Gandara, P. (2006). Performance of English language learners as a subgroup in large-scale assessment: Interaction of research and policy. *Educational Measurement: Issues and Practice*, *25*, 36–46.

Abedi, J., & Leon, S. (1999). *Impact of students' language background on content-based performance: Analyses of extant data*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, *14*, 219–234.

Alderson, J. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. New York: Continuum.

Austin, P., Tu, J., Daly, P., & Alter, D. (2005). The use of quantile regression in health care research: A case study examining gender differences in the timeliness of thrombolytic therapy. *Statistics in Medicine*, *24*, 791–816.

Bailey, A. (2005). Language analysis of standardized achievement tests: Considerations in the assessment of English language learners. In F. Butler, J. Abedi & A. Bailey (Eds.), *The validity of administering large-scale content assessments to English language learners: An investigation from three perspectives* (CSE Technical Report 663; pp. 79–100). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Betebenner, D. (2009a). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, *28*, 42–51.

Betebenner, D. (2009b). *Growth, standards, and accountability*. Retrieved from www.nciea.org/publications/growthandStandard_DB09.pdf

Betebenner, D., & Linn, R. (2010). *Growth in student achievement: Issues of measurement, longitudinal data analysis and accountability*. Retrieved from www.k12center.org/

Buchinsky, M. (1994). Changes in the U.S. wage structure 1963–1987: Application of quantile regression. *Econometrica*, *62*, 405–458.

Bunch, M. (2011). Testing English language learners under No Child Left Behind. *Language Testing*, *28*, 323–342.

Castellano, K., & Ho, A. (2012). *A practitioner's guide to growth models*. Retrieved from http://scholar.harvard.edu/andrewho/publications/practitioners-guide-growth-models

Chen, C. (2007). A finite smoothing algorithm for quantile regression. *Journal of Computational and Graphical Statistics*, *16*, 136–164.

Chevapatrakul, T., Kim, T., & Mizen, P. (2009). The Taylor principle and monetary policy approaching a zero bound on nominal rates: Quantile regression results for the United States and Japan. *Journal of Money, Credit and Banking*, *41*(8), 1705–1723.

Cizek, G., & Bunch, M. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: SAGE Publications.

Cottrell, R. S. (1968). *A study of selected language factors associated with arithmetic achievement of third grade students*. (Unpublished doctoral dissertation). Syracuse University, Ann Arbor, MI.

Deville, C., & Chalhoub-Deville, M. (2011). Standards-based assessment in the United States. *Language Testing*, 28.

Drescher, L., & Goddard, E. (2011). Heterogeneous demand for food diversity: A quantile regression analysis. *Research in Agricultural and Applied Economics*. Retrieved from http://purl.umn.edu/114484

Edgeworth, F. (1888). On a new method of reducing observations relating to several quantities. *Philosophical Magazine*, *25*, 184–191.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*, 1–26.

Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Thousand Oaks, CA: Sage Publications.

Freeman, B., & Crawford, L. (2008). Creating a middle school mathematics curriculum for English-language learners. *Remedial and Special Education*, *29*(1), 9–19.

Haile, G.A., & Nguyen, A. N. (2008). Determinants of academic attainment in the United States: A quantile regression analysis of test scores. *Educational Economics*, *16*, 29–57.

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

He, X., & Hu, F. (2002). Markov chain marginal bootstrap. *Journal of the American Statistical Association*, *97*, 783–795.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*(1), 153–161.

Kato, K., Albus, D., Liu, K., Guven, K., & Thurlow, M. (2004). *Relationships between a statewide language proficiency test and academic achievement assessments* (LEP Projects Report 4). Minneapolis, MN: University of Minnesota, National Center for Educational Outcomes.

Kieffer, M.J., Lesaux, N. K., Rivera, M., & Francis, D. J. (2009). Effectiveness of accommodations for English language learners taking large-scale assessments. *Review of Education Research*, *79*(3), 1168–1201.

Kocherginsky, M., He, X., & Mu, Y. (2005). Practical confidence intervals for regression quantiles. *Journal of Computational and Graphical Statistics*, *14*, 41–55.

Koenker, R. (2005). Quantile regression. *Econometric Society Monographs, 48*. Cambridge: Cambridge University Press.

Koenker, R. (2009). Quantreg: Quantile regression. *R package version 4*.27. Retrieved from http://CRAN.R-project.org/package=quantreg

Koenker, R., & Bassett, G. (1982). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica*, *50*, 43–61.

Koenker, R., & Bilias, Y. (2001). Quantile regression for duration data: A reappraisal of the Pennsylvania Reemployment Bonus Experiments. *Empirical Economics*, *26*, 199–220.

Koenker, R., & d'Orey, V. (1987). Algorithm AS 229: Computing regression quantiles. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *36*, 383–393.

Koenker, R., & d'Orey, V. (1994). Remark AS R92: A Remark on Algorithm AS 229: Computing dual regression quantiles and regression rank scores. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *43*, 410–414.

Koenker, R., & Geling, R. (2001). Reappraising medfly longevity: A quantile regression survival analysis. *Journal of American Statistical Association*, *96*, 458–468.

Koenker, R., & Hallock, K. F. (2001). Quantile regression. *The Journal of Economic Perspectives*, *15*(4), 143–156.

Koenker, R., & Machado, J.A. F. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, *94*(448), 1296–1310.

Konstantopoulos, S. (2009). The mean is not enough: Using quantile regerssion to examine trends in Asian–White differences across the entire achievement distribution. *Teachers College Record*, *111*, 1274–1295.

Kopriva, R. J., Bauman, J., Cameron, C., & Triscari, R. (2009). *Final research report: Obtaining necessary parity through academic rigor in science*. PR/Award # 368A06007. Center for Applied Linguistics, Washington, DC.

Laplace, P. S. de (1818). *Deuxième Supplément a la Thèorie Analytique des Probabilités*.Paris: Courcier. Reprinted (1847) in *Oeuvres de Laplace 7*, pp. 569–623. Paris: Imprimerie Royale; (1886) in Oeuvres Complètes de Laplace 7, pp. 531–580. Paris: Gauthier-Villars.

Linn, R., Baker, E., & Betebenner, D. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher*, *31*, 3–16.

Najarian, M., Pollack, J. M., & Sorongon, A. G. (2009). *Early childhood longitudinal study, kindergarten class of 1998–99 (ECLS-K), psychometric report for the eighth grade (NCES 2009–002)*. National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

North, B. (2000). *The development of a common framework scale of language proficiency*. New York: Peter Lang.

Pandey, G., & Nguyen, V. (1999). A comparative study of regression based methods in regional flood frequency analysis. *Journal of Hydrology*, *225*, 92–101.

Parzen, M. I., Wei, L. J., & Ying, Z. (1994). A resampling method based on pivotal estimating functions. *Biometrika*, *81*, 341–350.

Portnoy, S., & Koenker, R. (1997). The Gaussian hare and the Laplacian tortoise: Computation of squared-error vs. absolute-error estimators. *Statistical Science*, *12*, 279–300.

SAS Institute Inc. 2008. *SAS/STAT® 9.2 User's guide*. Cary, NC: SAS Institute Inc.

Solano-Flores, G. (2011). Language issues in mathematics and the assessment of English language learners. In K. Tellez, J. Moschkovich & M. Civil (Eds.), *Latino/as and mathematics education* (pp. 283–314). New York: Information Age Publishing.

Spellings, M. (2005, Nov). *Secretary Spellings announces growth model pilot* [Press Release]. U.S. Department of Education. Retrieved from www.ed.gov/news/pressreleases/2005/11/1182005.html.

StataCorp. 2009. *Base reference manual*. Statistical software. College Station, TX: StataCorp LP.

Stevens, R. A., Butler, F. A., & Castellon-Wellington, M. (2000). *Academic language and content assessment: Measuring the progress of English language learners*. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Tourangeau, K., Nord, C., Lê, T., Sorongon, A. G., & Najarian, M. (2009). *Early childhood longitudinal study, kindergarten class of 1998–99 (ECLS-K), combined user's manual for the ECLS-K eighth-grade and K–8 full sample data files and electronic codebooks* (NCES 2009–004). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Wei, Y., Pere, A., Koenker, R., & He, X. (2006). Quantile regression methods for reference growth charts. *Statistics in Medicine*, *25*, 1369–1382.

Wößmann, L. (2005). The effect heterogeneity of central examinations: Evidence from TIMSS, TIMSS-Repeat and PISA. *Education Economics*, *13*(2), 143–169.

Wright, W. E., & Li, X. (2008). High-stake math tests: How No Child Left Behind leaves newcomer English language learners behind. *Language Policy*, *7*, 237–266.

Yu, K., Lu, Z., & Stander, J. (2003). Quantile regression: Applications and current research areas. *Journal of the Royal Statistical Society, Series D (The Statistician)*, *52*, 331–350.

Appendix A: Codes for most of the analyses for the examples

Code using R program:

```
# Code to generate the LSR line for the total group in Figure 1 #
G5<-read.csv(“G5all.csv”)
attach(G5)
G5Slm<-lm(MATH~READING,data=G5)
plot(READING,MATH,cex=0.25,type=“n”,xlab=“Reading”,ylab=“Math”,main=“Grade 5 Total”,xlim=c(50,210),ylim=c(50,180))
points(READING,MATH,cex=0.25,col=“blue”)
abline(lm(MATH~READING),lwd=2,col=“blue”)
# Code for the LSR line for Grade 5 Low math ability #
G5low<-read.csv(“G5low.csv”)
attach(G5low)
G5Slmlow<-lm(MATH~READING,data=G5low)
plot(READING,MATH,cex=0.25,type=“n”,xlab=“Reading”,ylab=“Math”,main=“Grade 5 Low”,xlim=c(50,210),ylim=c(50,180))
points(READING,MATH,cex=0.25,col=“blue”)
abline(lm(MATH~READING),lwd=2,col=“blue”)
# Code for the LSR line for Grade 5 Medium math ability #
G5medium<-read.csv(“G5medium.csv”)
attach(G5medium)
G5Slmmedium<-lm(MATH~READING,data=G5medium)
plot(READING,MATH,cex=0.25,type=“n”,xlab=“Reading”,ylab=“Math”,main=“Grade 5 Medium”,xlim=c(50,210),ylim=c(50,180))
points(READING,MATH,cex=0.25,col=“blue”)
abline(lm(MATH~READING),lwd=2,col=“blue”)
```

```
# Code for the LSR line for Grade 5 High math ability #
G5high<-read.csv(“G5high.csv”)
attach(G5high)
G5Slmhigh<-lm(MATH~READING,data=G5high)
plot(READING,MATH,cex=0.25,type=“n”,xlab=“Reading”,ylab=“Math”,main=“Grade 5 High”,xlim=c(50,210),ylim=c(50,180))
points(READING,MATH,cex=0.25,col=“blue”)
abline(lm(MATH~READING),lwd=2,col=“blue”)
# Code to generate Figure 3 #
plot(READING,MATH,cex=0.25,type=“n”,xlab=“Reading”,ylab=“Math”,main=“Quantile and Mean Regression Lines”,xlim=c(50,210),ylim=c(50,180))
points(READING,MATH,cex=0.25,col=“blue”)
abline(rq(MATH~READING,tau=0.5),lwd=2,col=“blue”)
abline(lm(MATH~READING),lty=2,lwd=2,col=“red”)
taus<-c(0.05,0.1,0.25,0.75,0.9,0.95)
for (i in 1:length(taus)){
abline(rq(MATH~READING,tau=taus[i]),lwd=2, col=“gray”)
}
# Code to generate Figure 4#
G5Flm<-lm(MATH~READING+GENDER,data=G5)
G5Full<-summary(rq(MATH~READING+GENDER,tau = 1:19/20,data=G5, method=“fn”),se=“boot”,R=500,bsmethod=“xy”)
# “method=‘fn’” stands for the Frisch-Newton interior point estimation, “se=‘boot’” means the SE is estimated using the bootstrapping method, “R=500” means 500 samples are generated, “bsmethod=‘xy’” means the specific bootstrapping method used here is xy-pair. #
plot(G5Full)
abline(lm(MATH~READING+GENDER),lty=2,col=“red”)
savePlot(“G5Full”,type=“jpeg”)
```

Code in Stata:

```
# Code to run the model with one covariate and produce statistics in Table 1 #
insheet using “ G5all.csv”;
set seed 12345;
regress math reading;
sqreg math reading, q(.05 .10 .25 .5 .75 .90 .95) reps(500);
# Code to run the model with two covariates and produce statistics in Table 2 #
set seed 12345;
regress math reading gender;
sqreg math reading gender, q(.05 .10 .25 .5 .75 .90 .95) reps(500);
test [q5]reading=[q10]reading;
test [q10]reading=[q25]reading,accum;
test [q25]reading=[q50]reading,accum;
test [q50]reading=[q75]reading,accum;
test [q75]reading=[q90]reading,accum;
test [q90]reading=[q95]reading,accum;
test [q5]reading=[q10]reading;
test [q10]reading=[q25]reading;
test [q25]reading=[q50]reading;
test [q50]reading=[q75]reading;
test [q75]reading=[q90]reading;
test [q90]reading=[q95]reading;
log close;
clear;
```