Chapter 8
Quantile Regression: Analyzing Changes in Distributions Instead of Means

Stephen R. Porter

Introduction

For the past several decades, ordinary least squares (OLS) has been the workhorse of quantitative postsecondary research. OLS has several features that make it especially appealing to applied researchers, such as its ability to control for multiple independent variables, its ease of estimation and interpretation, and its robustness to violations of underlying assumptions. Open any education journal featuring empirical articles on postsecondary topics, and you will find numerous papers using OLS, or one of its variants, such as logistic regression, instrumental variables, hierarchical linear models, or fixed effects models.

As applied researchers, we rarely think deeply about what a regression coefficient tells us; we tend to assume that it just tells us the effect of x on y, ceteris paribus. From a technical perspective, however, this is not exactly correct. A regression coefficient tells us the effect of x on the *mean* of y controlling for other x's, not just "y". This may seem like a subtle distinction, but it is not, as a simple example demonstrates.

Access and completion are two major areas of focus in postsecondary research, and interventions that aim to prepare students for college success, such as summer bridge programs and developmental education classes, are widely used across the country. Suppose we are studying a program to increase incoming students' math skills, a common deficit area for new students. We are interested in the effect of the program on math performance; in other words, does participation in the program increase math proficiency? One approach to assessing the effect of the program

S.R. Porter (✉)
Department of Leadership, Policy, and Adult and Higher Education, North Carolina State University, Raleigh, NC 27695, USA
e-mail: srporter@ncsu.edu

M.B. Paulsen (ed.), *Higher Education: Handbook of Theory and Research*,
Higher Education: Handbook of Theory and Research 30,
DOI 10.1007/978-3-319-12835-1_8

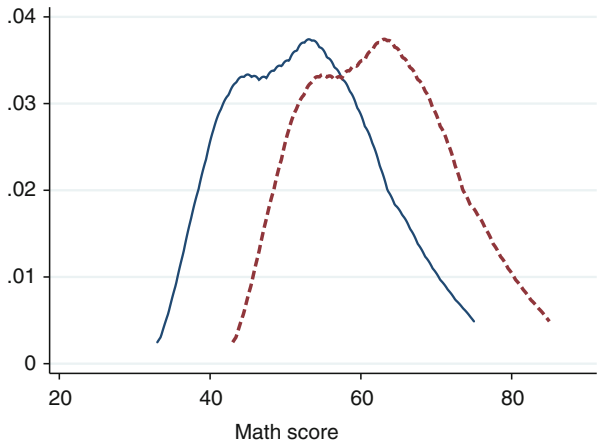

Fig. 8.1 Hypothetical math score distributions with and without remediation

would be estimating a regression model with performance on a math exam as the dependent variable, a dummy variable indicating program participation as the main independent variable of interest, and a set of control variables (assume that these control variables are such that we are not worried about omitted variable bias).

A positive and statistically significant coefficient on the dummy variable would tell us that performance was larger for students participating in the program. Figure 8.1 illustrates this possibility. Two hypothetical distributions are shown, for participants (dashed line) and non-participants (solid line). As can be seen, participation in the program shifts test scores for participating students to the right; that is, remediation appears to increase math proficiency. For the sake of this example, assume the increase is 10 points on a 100-point scale. Based on our regression results, we would conclude that the program was successful in increasing math proficiency. Technically, however, we can only conclude that program participation had an effect on the mean of the test score distribution; we can say nothing about other points of the distribution.

Why is this potentially problematic? We can consider three alternative scenarios, depending upon how the remediation program affects different individuals. First, we can imagine a scenario in which the distribution for the treated shifts such that there is still a 10-point increase at the mean, but the program has the strongest effect for students at the lower end of the distribution, increasing their test scores by 20 points. This is consistent with the idea that math remediation will have the strongest effect for students who have deficits and will likely struggle with college-level math, and little effect for those highly proficient. Second, the opposite could occur, with no increase at the low end, a 10-point increase at the mean, and a 20-point increase at the high end of the distribution. Here, math remediation helps those students already comfortable and successful at math, with little effect at the low end of the distribution. This is consistent with the idea that remediation helps students who are already proficient at math, but does little for non-proficient students. Third, aspects

of both trends could occur: the intervention could help the average student (with a 10-point increase in scores at the mean), but with no changes in proficiency at the high and low ends of the distribution.

From a policy perspective, distinguishing between the four possibilities is important. Does offering additional math instruction raise all boats, help those most in need, help those who need it least, or just help the average student? Most postsecondary researchers and administrators would agree that an intervention that largely helps students already proficient at math is not a wise use of funds. Yet using OLS to study math outcomes would not tell us whether this was happening, only whether students at the mean were experiencing an increase in proficiency. Indeed, in using OLS to analyze the four scenarios, we would reach the exact same conclusion, even though the math intervention is having very different effects on students at other parts of the distribution in each scenario.

Quantile regression is one approach to analyzing changes in distributions that is becoming increasingly popular with applied researchers. As with OLS, quantile regression estimates the effect of an independent variable on an outcome, while allowing for covariates as controls. Unlike OLS, quantile regression provides estimates of these effects at different points of the distribution of y, such as the 5th percentile, 25th percentile, 95th percentile, etc. Quantile regression thus allows the researcher to understand how an independent variable affects the entire distribution of an outcome, rather than just the average. In addition, these models are easily estimated by most statistical packages and can be widely used by postsecondary researchers.

This chapter reviews the two main types of quantile regression models used by researchers, the conditional and unconditional quantile regression models. The latter is more widely used by researchers, because it focuses on changes to the unconditional distribution of the dependent variable. After reviewing estimation, interpretation, and sensitivity analyses for unconditional quantile regression models, I discuss and demonstrate the use of instrumental variables within the quantile regression context.

Why Use Quantile Regression?

Due to the tremendous increase in computing power in recent years, a wide variety of advanced statistical techniques are now available at the touch of a drop-down screen. Researchers can feel a bit overwhelmed at the dizzying array of choices for analyzing their data, and skeptical of new approaches, which often tend to be seen as faddish at best. Quantile regression should not be viewed as a fad, but rather as a more informative approach to analyzing educational data than more familiar techniques such as OLS. Indeed, the technique dates to the late 1970s, and has been used in the field of economics for many years. Given recent advances in estimation and interpretation of quantile regression models, including the ability to deal with endogenous regressors, the technique will soon be commonplace. Quantile regression is generally seen as having two advantages over OLS.

Table 8.1 Sensitivity of OLS estimates to outliers on Y

	Sample unchanged		One male score changed to 1,000	
	OLS	CQR	OLS	CQR
Female	4.870 (1.304)***	5.000 (2.080)**	−5.383 (9.622)	5.000 (2.080)**
Intercept	50.121 (.963)***	52.000 (1.535)***	60.374 (7.103)***	52.000 (1.535)***

Note: Cell entries are coefficients, with standard errors in parentheses
** $p < 0.05$; *** $p < 0.01$

First, an advantage of quantile regression is its insensitivity to outliers on y (Davino, Furno, & Vistocco, 2014; Fröhlich & Melly, 2010). Recalling the formulas for the mean and median, this makes intuitive sense. If we analyzed a sample of incomes and added a single billionaire to the sample, the mean would change quite a bit, because the billionaire's income is used explicitly in the calculation of the mean. Repeating the process but using the median instead, the addition of the billionaire would simply shift the value of the median from the income of the person at the 50th percentile to the next highest income in the distribution, resulting in a small change in the median. Or, if the next highest income were identical to the median, result in no change at all.

To illustrate, Table 8.1 shows OLS and conditional quantile regression estimates for a bivariate model using gender to predict performance on a writing exam, based on a sample dataset (n = 200) from the High School and Beyond survey.[1] The first two columns use the sample dataset with no changes to the observations. Both OLS and conditional quantile regression yield similar results, with females scoring about five points higher than males (note that this similarity is not surprising, given that the mean and median of Y are 53 and 54, respectively).

The highest writing scores in the sample are 67 points, and there are two males in the dataset with these scores. The last two columns of the table demonstrate how the model coefficients change when the writing score for one of these males is changed from 67 points to 1,000 points. As we can see, the OLS estimates change drastically. The predicted writing score for males increases from 50 to 60, and the gender difference switches direction, with females now scoring five points lower than males, as opposed to higher. The conditional quantile regression estimates, however, remain unchanged. The one male whose test score increased was already one of the two highest-scoring males in the dataset, so drastically increasing his score had no effect on the estimates of the effect of gender on the median of test scores. As with estimation of a simple median, increasing scores for observations above the median leaves the median (conditional quantile regression) estimates unchanged.

A second advantage of quantile regression, however, is its ability to allow us to see how the entire distribution of y changes when x changes, rather than just seeing

[1]This example is based on the discussion at http://www.ats.ucla.edu/stat/stata/faq/quantreg.htm

how the mean changes. Some examples from the literature illustrate its advantages over OLS: estimating the effect of class size on student achievement, and estimating the effect of spending on college graduation rates.

Providing additional funding to school districts is one approach to increasing K-12 student achievement, but how these additional funds should be allocated is not at all clear. For example, a district could increase salaries to attract more experienced teachers, or it could maintain current salary levels and use the funds to hire more teachers in order to reduce average class size. Project STAR (the Tennessee Student/Teacher Achievement Ratio experiment) randomly assigned students in public elementary schools to small classrooms (13–17 students), regular classrooms (22–25 students), and regular classrooms with the addition of a full-time aide. Mueller (2013) uses data from the experiment to analyze the effect of class size on math and reading test scores. Conditioning on teacher experience, the OLS estimates indicate that assignment to a small classroom increases math and reading test scores by about .15 standard deviations (see his Table 2, p. 48); these are the effects of small class size at the mean of the test score distributions. The quantile regression estimates, however, tell a different story. Small class size increases test scores about .10 standard deviations at the lowest decile of the math and reading distributions, with the effects almost doubling in size at higher points along the distributions. Small class size, in other words, increases student achievement for all students, but it also increases inequality, with smaller gains at the low ends of the math and reading distributions.

Funding issues also dominate much of the discussion in higher education, especially in terms of recent proposals to develop a national rating system for colleges and universities based on how well they graduate their students. While previous research indicates that expenditures per student are positively associated with higher graduation rates, less is known about the effects of specific categories of expenditures, such as spending on instruction. Webber and Ehrenberg (2010) use IPEDS Finance and Completions data to estimate and compare the effect of instructional, academic support, research, and student services expenditures on institutional 6-year graduation rates. OLS estimates suggest that increasing student services expenditures by $100 per student would increase graduation rates by .2 percentage points (e.g., from 80 to 80.2 %), while the same amount for instructional expenditures would increase graduation rates by only .06 percentage points (their Table 3, p. 953). The quantile regression estimates reveal that the effect of expenditures varies across the distribution of graduation rates (their Table 4, p. 954). The effect of a $100 increase in student services expenditures, for example, is largest at the bottom half of the graduation rate distribution, about .6 percentage points, and declines rapidly to zero from the 50th percentile to the 90th percentile. The effect of instructional expenditures is largest between the 20th and 80th percentiles, with no effect at the top and bottom of the distribution. These results suggest that institutions with low graduation rates would benefit most from increasing expenditures on student services, while increasing expenditures in both areas would achieve little for institutions with very high graduation rates.

As these examples demonstrate, understanding how an independent variable affects an outcome can differ depending on whether the researcher uses OLS or quantile regression. The former only allows us to understand the effect of an independent variable at the mean of an outcome, while the latter allows to observe how the effect varies at different quantiles of the distribution. From an applied researcher's perspective, it is precisely these varying effects in which we are most interested, especially in terms of implementing good policies. Does a treatment affect all students equally, or only some students along the distribution of interest? If the treatment shows positive effects, does it also increase inequality by having the weakest effects for those students at one end of the distribution? Quantile regression can help us begin to answer these important and policy-relevant questions, while OLS cannot.

Conditional Quantile Regression

Conditional quantile regression has been used by researchers for several decades. While the interpretation of the results is somewhat similar to OLS, the estimation approach is not. As Koenker and Hallock (2001, p. 145) note,

> Quantiles seem inseparably linked to the operations of ordering and sorting the sample observations that are usually used to define them. So it comes as a mild surprise to observe that we can define the quantiles through a simple alternative expedient as an optimization problem.

The optimization approach to finding a quantile q (such as the median) can be achieved by using the following equation, and finding the value of β that yields the minimum value for a group of observations y:

$$\sum_{i:y_i \geq \beta}^{N} q|y_i - \beta| + \sum_{i:y_i < \beta}^{N} (1-q)|y_i - \beta|. \tag{8.1}$$

Suppose we have three observations in a sample with the values of 1, 2 and 3, and wish to know the median. The median is obviously 2 by inspection, and we can use Eq. 8.1 instead to estimate the median via optimization. Beginning with the first observation as a possible answer for the median, we use only the first part of Eq. 8.1, as there are no values of y less than 1 in this sample,

$$\sum_{i:y_i \geq 1}^{N} .5|y_i - 1| + \sum_{i:y_i < 1}^{N} (1-.5)|y_i - 1|$$

$$\sum_{i:y_i \geq 1}^{N} .5|y_i - 1|$$

$$.5|1-1| + .5|2-1| + .5|3-1| = 1.5$$

while for the second observation,

$$\sum_{i:y_i \geq 2}^{N} .5|y_i - 2| + \sum_{i:y_i < 2}^{N} (1 - .5)|y_i - 2|$$
$$.5|2 - 2| + .5|3 - 2| + .5|1 - 2| = 1$$

and for the third observation,

$$\sum_{i:y_i \geq 3}^{N} .5|y_i - 3| + \sum_{i:y_i < 3}^{N} (1 - .5)|y_i - 3|$$
$$.5|3 - 3| + .5|1 - 3| + .5|2 - 3| = 1.5.$$

Of the three observations, the value of 2 minimizes Eq. 8.1, and we can conclude that it is the value of the 50th quantile, or median.

While this may seem like an overly complicated solution to the relatively simple problem of finding the median of y, this approach can be used to find the quantile regression estimator (Cameron & Trivedi, 2005), in that minimizing

$$\sum_{i:y_i \geq \boldsymbol{x}_i'\boldsymbol{\beta}}^{N} q|y_i - \boldsymbol{x}_i'\boldsymbol{\beta}| + \sum_{i:y_i < \boldsymbol{x}_i'\boldsymbol{\beta}}^{N} (1 - q)|y_i - \boldsymbol{x}_i'\boldsymbol{\beta}| \tag{8.2}$$

yields the quantile regression coefficient β, where $\boldsymbol{x}_i'$ and $\boldsymbol{\beta}$ indicate a matrix of independent variables and a vector of quantile regression coefficients. Note that the expressions within the absolute value symbols are deviations, so that this approach can also be viewed as a least absolute deviations estimator (as opposed to OLS, which uses squares instead of absolute deviations).

An alternative version that is often cited in articles is

$$\arg\min \sum_{i=1}^{N} \rho_\tau(y_i - x_i\beta) \tag{8.3}$$

where τ is a particular quantile, ρ_τ is an absolute value function $\rho_\tau(u) = u \cdot (\tau - \mathbf{1}(u < 0))$ and $\mathbf{1}(u < 0)$ is an indicator function taking a value of 1 if $u < 0$, 0 otherwise. This simply means that Eq. 8.3 expands into two parts

$$\arg\min \sum \tau \cdot (y_i - x_i\beta) \text{ when } y_i - x_i\beta > 0 \text{ and}$$
$$\arg\min \sum (\tau - 1) \cdot (y_i - x_i\beta) \text{ when } y_i - x_i\beta < 0$$

based on the sign of $y_i - x_i\beta$.

More specifically, the approach outlined above is known as the *conditional quantile regression* approach to studying changes in distributions. In terms of estimation, conditional quantile regression takes a different approach than OLS. If we view the OLS regression model as a mathematical function, we can find the value of β that minimizes the function by using calculus to find the derivative. Unlike OLS, the conditional quantile regression function cannot be differentiated and instead is estimated via linear programming methods (Cameron & Trivedi, 2005). Linear programming is "a subset of mathematical programming facing the efficient allocation of limited resources to known activities with the objective of meeting a desired goal, such as minimizing cost or maximizing profit" (Davino et al., 2014, p. 23). This approach, and related optimization techniques, are widely used for many practical applications, such as determining the optimal driving route between two different locations on a map.

Linear programming typically consists of a series of equations that can be solved to find the solution set. The most common approach is the simplex method, which uses an iterative process to find a solution. Similar to maximum likelihood estimation, multiple solutions are tested until the software fails to find a better solution. This is why the statistical output for conditional quantile regression resembles the output for logistic regression, listing the iterations that have been used to reach a solution. Conditional quantile regression, however, focuses on minimizing the absolute deviations (as seen in Eq. 8.2), not maximizing the likelihood.

Conditional quantile regression models can be estimated with the following statistical packages:

- Stata uses the qreg command, but the estimated standard errors assume homoskedasticity. The vce(robust) option should be used to ensure the correct standard errors.
- SAS uses the quantreg procedure (Chen, 2005).
- R has the package quantreg (http://cran.r-project.org/web/packages/quantreg/index.html); SPSS version 17 allows SPSS users to invoke R packages within SPSS.

Interpretation

One of the most important distinctions to understand when estimating quantile regression models is the difference in interpretation between conditional versus unconditional regression models (described below). For conditional quantile regression, interpretation of the coefficients is in relation to the quantiles of the distributions defined by the covariates (the conditional distribution), rather than the unconditional distribution of y.

Continuing with the developmental math example, suppose we estimated a conditional quantile regression model at the median with math proficiency as the dependent variable, a developmental math dummy variable, and a dummy variable

for gender as regressors. The coefficient for the developmental math dummy variable is not the effect of developmental math at the median of the test score distribution. Instead, it can be thought of as the average of the effect at the median of the distribution for males and at the median of the distribution for females. Why is this problematic? Suppose that females score higher on the test than males, such that the median female score is 85 whereas the median male score is 70. The conditional quantile regression coefficients are effects at these medians, which differ quite a bit. So we would interpret the effect of the program for one group of students scoring at 85 (females), as well as for another scoring at 70 (males). Typically, however, we would like to know the effect at the median of the unconditional distribution; that is, what is the effect for students who perform at the median of the overall score distribution, not for students who score at the median of groups defined by whatever covariates we include in the model (in this case, developmental math and gender).

This conditional definition of effect can be difficult to interpret in many applied settings. The previous example had one treatment variable and only one control variable; with additional control variables, interpretation becomes even more complex. More importantly, this interpretation is typically not what most educational researchers seek. Just as OLS yields the effect of a variable at the mean of y, we also wish to know the effect at other quantiles of y, not quantiles of y defined within subgroups. The main issue here is that inclusion of control variables in a conditional regression model is necessary to deal with selection bias, just as in the case of OLS, yet inclusion of these covariates changes the interpretation of the quantiles. Moreover, as additional covariates are included, the interpretation of the quantiles changes, making comparisons across different model specifications problematic.

The growing consensus in the literature is that many researchers have inadvertently misused conditional quantile regression for many years, by interpreting the results as if they came from an unconditional quantile regression model. In other words, they have interpreted their coefficients as if they were the effect on the quantile of y, rather than quantiles of y defined within groups based on their set of covariates.

Two very recent examples from the literature demonstrate how conditional quantile regression has been misapplied. Maclean, Webber, and Marti (2014) estimate a state-level panel model to understand the effect of state cigarette taxes on cigarette consumption. Cigarette taxes have been an important public health tool used to reduce smoking, but the effects of tax increases in the literature are not clear, especially as previous research has tended to focus on the effect at the mean.

Some previous researchers in this area have used conditional quantile regression to study cigarette taxes, and Maclean et al. (2014) illustrate the drawbacks of this method with a thought experiment. Suppose the researcher estimated a conditional quantile regression model using only a set of dummy variables for each state. This model

> ...effectively yields an average of the treatment effects for observations at the, say, 10th quantile of the 51 state-specific smoking distributions, some of which may deviate substantially from the 10th quantile in the national distribution of smokers. For example, the 10th quantile smoker in Kentucky rises to the 20th quantile in the national distribution, while

> the 10th quantile California smoker falls to the 6th quantile [of the national distribution]. Thus, [conditional quantile regression] at the 10th quantile produces an estimate of cigarette tax increases on smokers who smoke 30 cigarettes per month in California, 150 cigarettes per month in Kentucky, and many values in between for other states.

For most applications, we would not want to know the effect of taxes on smokers at differing absolute levels of smoking (e.g., 30 cigarettes per month, 150 cigarettes per month, etc.), even though these levels of smoking represent the 10th percentile within each state. Instead, we would want to know the effect at the 10th percentile of the national distribution, 60 cigarettes per month.

Budig and Hodges (2010) use conditional quantile regression to analyze wages for females in an effort to determine the "motherhood penalty" – the loss in compensation that women experience if they have children. They find that mothers at the low end of the wage distribution experience larger penalties than higher income females. In a critique, Killewald and Bearak (2014) point out that their interpretation of the penalty from their conditional quantile regression model is actually for the unconditional distribution of wages. Their thought experiment is a simple conditional quantile regression model with motherhood and level of education as covariates. The estimates of the motherhood penalty from this model are not the estimates for workers at different quantiles of wages. Instead, they are the estimates of the motherhood penalty at different quantiles of wages within each education group. The problem lies in the fact that a specific quantile wage for college-educated women will be much larger than the same quantile wage for high-school dropouts. In other words, the 50th quantile wage for college-educated women will be much higher than the 50th quantile wage for high-school dropouts.

At this point, it may seem somewhat confusing that the interpretation changes when covariates are added; doesn't the same thing occur with OLS? When interpreting an OLS coefficient, our overall interpretation may change slightly as we add covariates (would we say, for example, "controlling for independent variables A, B and C" instead of "controlling for independent variables A and B"), but regardless of the number of control variables, the interpretation of an OLS coefficient is always the effect of x on the mean of y. With conditional quantile regression, we lose this simple and clear interpretation of the regression coefficient.

From the perspective of many researchers, conditional quantile regression may not seem very useful, because as control variables are added to the equation to deal with selection, the quantiles and thus the interpretation of the coefficients change. There are, however, other uses of these models besides the estimation of treatment effects via covariate controls. One of the most common in K-12 is the use of conditional quantile regression to track student growth in standardized test scores.

Standardized tests are ubiquitous in K-12, and one approach to accountability is providing parents with their child's test score. Depending on the difficulty of the test, the raw score may not be useful. Suppose a student scores an 80 on a 100-point test. If many other students scored above 80, then this student did not perform very well. Conversely, if many other students scored below 80, then the student

performed well. The interpretation of issue of absolute versus relative performance naturally leads to the use of percentiles in reporting student scores, so that each student is scored relative to other students who took the test.

At the state-level, the educational accountability movement has pushed for measurement and reporting of student test performance and growth over time. One approach uses past and present student test results and conditional quantile regression to estimate student growth scores, referred to as "student growth percentiles"; a dozen states have adopted it for reporting purposes (Castellano & Ho, 2013a). Suppose we estimate a model using conditional quantile regression, in which a student's test score in a grade is regressed on his test score from the previous grade. We can think of the resulting predicted quantile for the student as where they scored on the current grade's test, not in relation to all test-takers across the state, but in relation to all test takers who scored the same as the student in the previous grade. Higher quantiles are then interpreted that a student is scoring higher than his or her academic peers, where academic peers are defined by those other students achieving the same test score as the student. Typically these models use several years of prior testing data, so the comparison group is students with similar score histories (see Castellano and Ho (2013b) for an accessible discussion of this and other methods for calculating student growth).

Note that because of the use of conditional quantiles, scoring higher than a majority of your peers using these models does not mean that student growth has actually occurred. Suppose that for some reason students tended to do poorly this year in relation to last year (e.g., experienced learning loss). If a particular student's loss is much less relative to his peers, then his student growth percentile would be high (implying growth), even though an analysis of absolute test scores would reveal a loss in learning.

As this review makes clear, estimation and correct interpretation of conditional quantiles can be a tricky business. Thus, many researchers have turned to unconditional quantile regression models.

Unconditional Quantile Regression Assuming Exogeneity

Given that the interpretation of conditional quantile regression coefficients depends on the group of covariates used in the model, and that most researchers are instead interested in the effects on the unconditional distribution of y, *unconditional quantile regression* (Firpo, Fortin, & Lemieux, 2009) is becoming the popular choice among applied researchers. Unconditional quantile regression is based on a transformation of the dependent variable into the recentered influence function (RIF)

$$RIF(y; q_t) = q_\tau + \frac{\tau - \mathbf{1}\{y \leq q_\tau\}}{f_Y(q_\tau)}, \tag{8.4}$$

where τ indicates a specific quantile (say the 40th, or .40), q_τ is the value of the dependent variable at that specific quantile, $\mathbf{1}\{y \leq q_\tau\}$ is a function that equals 1 when an observation's value of y is less than or equal to the value of the dependent variable at quantile τ, 0 otherwise, and $f_Y(q_\tau)$ is the density of y at quantile τ. All of these quantities are easily calculated except for the density, which is estimated from the sample using a kernel density estimator.

Table 8.2 demonstrates how the RIF is calculated for three writing test scores from the High School and Beyond dataset. Three scores are shown, one at the 25th percentile (45.5), one at the 50th percentile (54), and one at the 75th percentile (60). We wish to estimate an unconditional quantile regression for the effect of an independent variable on the median of y; τ is set to .50, and we choose the writing score at the median (54) as q_τ. Taking the test score for each student, we check to see whether it is less than or equal to the median score of 54. The first two observations meet this criterion, so $\mathbf{1}\{y \leq q_\tau\}$ is set to 1 for these observations. The third observation scored 60, which is higher than the median of 54, so $\mathbf{1}\{y \leq q_\tau\}$ is set to 0 for this student.

Next, we estimate the density of y when $Y = 54$; the number in the table is estimated using Stata's `kdensity` command, with a Gaussian kernel and an arbitrary bandwidth of 2. The histogram for the writing test score variable is displayed in Fig. 8.2, along with the estimated density. The vertical line is drawn where the writing test score equals 54, and the density (listed on the y-axis) is

Table 8.2 Calculating the recentered influence function

Y	Quantile	τ	q_τ	$\mathbf{1}\{y \leq q_\tau\}$	$f_Y(q_\tau)$	RIF
45.5	.25	.50	54	1	0.03534932	39.8555
54	.50	.50	54	1	0.03534932	39.8555
60	.75	.50	54	0	0.03534932	68.1445

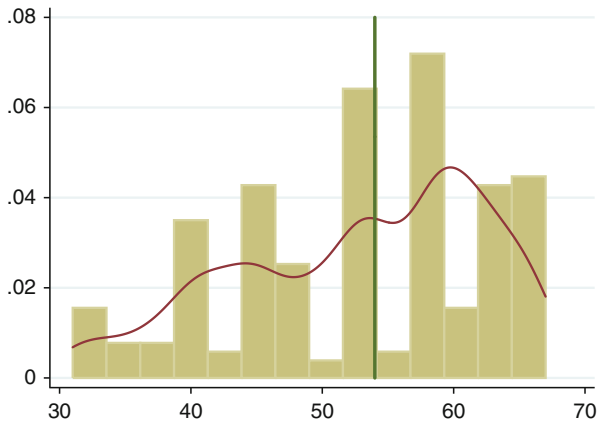

Fig. 8.2 Distribution of writing test scores

equal to .035. Comparing the estimated density to the histogram illustrates one potential disadvantage of the RIF function. We do not know the density of y in the population, so we must rely on estimating it using our sample. But to estimate the density using a kernel function, we must make some distributional and optimal bandwidth assumptions that may or may not be correct. These assumptions, in turn, will determine the quantile regression results.

Using these four quantities, the RIF can be calculated for each student. The formula results in only two values for the dependent variable, depending on whether an observation falls above or below the specified quantile. Once the RIF has been calculated for each observation, it is used as the dependent variable in an OLS model, regressing the RIF on a set of independent variables.

Interpretation

Close examination of Eq. 8.4 provides an intuitive understanding as to why the RIF produces the effect of x on the unconditional distribution of Y, in contrast to conditional quantile regression. Note that in Eq. 8.4, the dependent variable is transformed without reference to any covariates (there are no x's in the equation), so changing the mix of covariates in the model does not change the interpretation of β, other than the fact that the set of control variables has changed. Thus, the value of unconditional quantile regression estimates is that they are interpreted much like OLS estimates; the interpretation is not within groups, as with conditional quantile regression.

Use of unconditional quantile regression can sometimes yield very different conclusions compared to conditional quantile regression. In their seminal paper outlining their unconditional quantile regression estimator, Firpo et al. (2009) analyze the effect of unionization on wages. Misinterpreting the conditional quantile regression results (i.e., ignoring that these are within-group estimates), one would conclude that unionization has a declining linear effect on wages across the distribution, in that unionization greatly raises wages at the low end of the wage distribution, with this effect lessening along the distribution to be lowest at the high end of the wage distribution. Unionization, it would appear, has the biggest impact on those with low wages. Unconditional quantile regression estimates, however, tell a different story, with unionization increasing wages in the middle part of the wage distribution, but actually decreasing wages at the high end of the distribution.

Similarly, the reanalysis of the motherhood penalty using unconditional quantile regression indicates a different effect than the conditional estimates. With the conditional estimates, there is a strong linear effect along the female wage distribution, with motherhood having the strongest negative effects for the lowest quantiles. The unconditional estimates reveal much more similar effects across the distribution, with the strongest negative penalty occurring at the middle of the distribution, rather than the lower end (Killewald & Bearak, 2014).

Interpreting coefficients from unconditional quantile regression models as effects at different points of the distribution of y is a useful feature, but can be easily confused with interpretations of nonlinear OLS models. For example, in an OLS model with an interaction term, the effect of X_1 on Y may increase or decrease, depending on the value of X_2, the variable with which it is interacted. The effect of academic ability on engagement may vary by level of socioeconomic status, if ability and socioeconomic status have been interacted. Similarly, when including a quadratic term, the effect of X_1 on Y increases or decreases, depending on the value of X_1 that is plugged into the quadratic term $X_1 + X_1^2$. The effect of age is commonly specified as a quadratic function in the social sciences, allowing its effect to increase, level off, and decrease as age increases. In both of these examples, the effect of X varies depending on values of other *independent* variables.

With unconditional quantile regression, the effect of X on Y also varies, but it varies *depending on the value of Y*. We interpret the effect of X on a particular quantile of Y, rather than the effect of X conditional on the value of another independent variable. As with nonlinear specifications of X, the effect (as measured by the regression coefficient β) varies, but the effect becomes weaker or stronger depending on the location in the distribution of Y.

Estimation

As the example in Table 8.2 demonstrates, the value of the RIF depends crucially on the estimated density of y. The density refers to the probability distribution of y, such that the area under the density curve equals 1. Figure 8.3 illustrates the difficulties in estimating the density, using histograms for the writing test score variable, and bins with a width of two, four, and six points, respectively. The shape of the distribution varies considerably, depending on the width of the bins. Most obviously, the histograms are not smooth, which is a useful property when trying to estimate the density of y at a particular value of y. The discrete nature of the histogram bins makes it likely that the estimated density of y will be off, compared to a smooth estimate of the density of y.

Kernel density estimators are a non-parametric approach to solving this problem. Non-parametric here refers to the fact that the estimator does not yield a fixed set of parameters. Suppose we had a variable x, and wished to estimate the density of x over the entire distribution of x. Rather than use a histogram, we can estimate a kernel density function such that

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x - x_i}{h}\right) \tag{8.5}$$

where $k(\cdot)$ refers to a kernel function and h is a parameter known as the bandwidth (StataCorp LP, 2013, p. 1009). The bandwidth is the crucial part of this formula, as

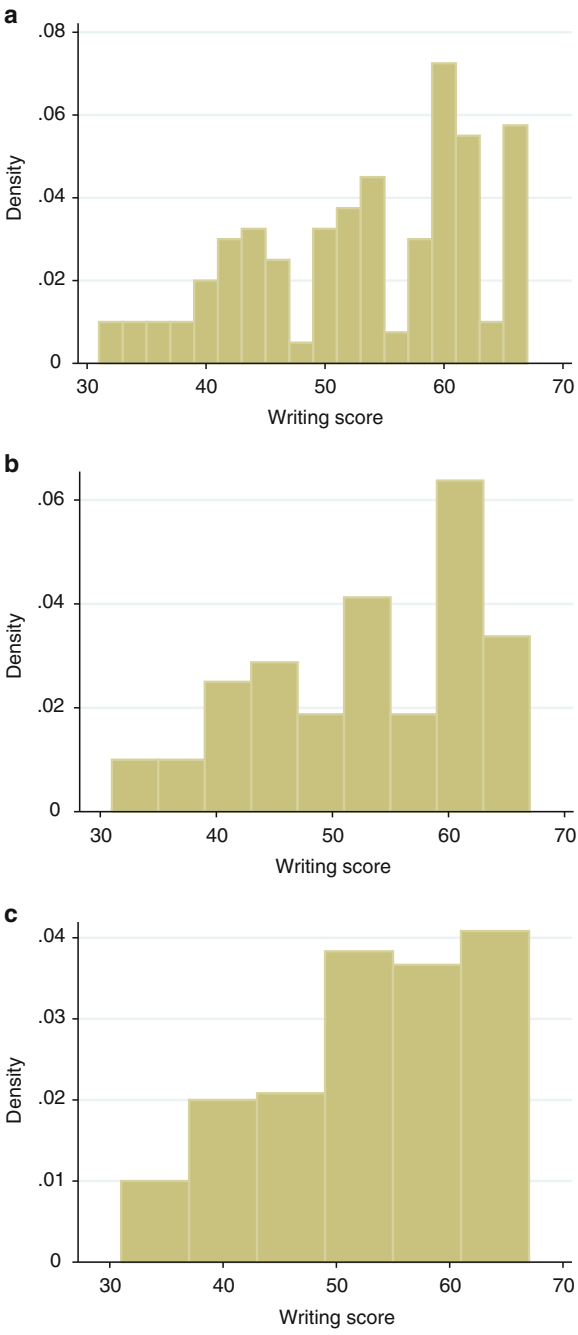

Fig. 8.3 Distribution of writing test scores using different bin sizes. (**a**) Bin = 2. (**b**) Bin = 4. (**c**) Bin = 6

the size of h determines how smooth or spiky the estimated density curve is, much as the width of the bins for a histogram determine the smoothness of its shape.

The kernel function specifies a distribution to be used when estimating the density. The standard normal density function can be used to draw a normal curve, and should be familiar from any basic statistics class,

$$k(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} \tag{8.6}$$

and using this as the kernel, we can rewrite Eq. 8.5 as

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{\left(\frac{x-x_i}{h}\right)^2}{2}\right)}. \tag{8.7}$$

While this equation may look complex, it provides an elegant solution for plotting the density of x. A simple example illustrates how the kernel density function works.

Table 8.3 provides a dataset consisting of a single variable x with five observations, and Fig. 8.4 plots the density of this variable using a bandwidth of 1. From the

Table 8.3 Calculating the density of x at 3 with a Gaussian kernel density estimator

x_i	(1) $\frac{(x-x_i)^2}{2}$	(2) $e^{-(\text{column 1})}$	(3) $\frac{1}{\sqrt{2\pi}} * (\text{column 2})$
1	2.0	0.1353	0.0540
1	2.0	0.1353	0.0540
3	0.0	1.0000	0.3989
4	0.5	0.6065	0.2420
4	0.5	0.6065	0.2420
			$\frac{\sum_{i=1}^{5} \text{column 3}}{5} = 0.1982$

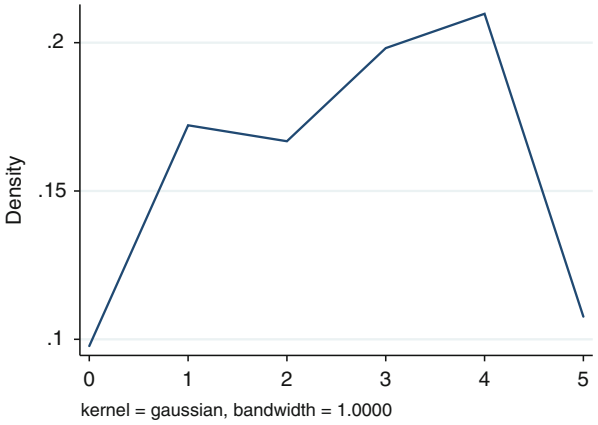

Fig. 8.4 Estimated density using Gaussian kernel and a bandwidth of 1

graph, the density of x when $x = 3$ is approximately .2, and we can use Eq. 8.7 to calculate this directly. In the first column of Table 8.3, we subtract each observation from 3, square it, and divide by 2 (because the bandwidth is 1 in this example, we can ignore the h's in Eq. 8.7). In the next column, we multiply this quantity by -1 and exponentiate it. Finally, we divide this quantity by the square root of 2π. Column 3 thus contains the quantity to the right of the summation sign in Eq. 8.7, and we sum these over the entire dataset and divide by the sample size to determine the density of x when $x = 3$, .198, matching what is shown in Fig. 8.4.

As the table demonstrates, the algorithm places greater weight on observations closest to the chosen value of x. More importantly, in Eq. 8.7 the differenced quantity, $x - x_i$, is divided by the bandwidth parameter. The size of this parameter will greatly determine the quantity for each observation before summing, thus determining what the final density will look like. Because the exact value of the RIF is determined by the density of y (the term $f_Y(q_\tau)$ in Eq. 8.4), determining the bandwidth is an important choice. Unfortunately, the literature does not provide much advice as to determining the appropriate kernel function and bandwidth for unconditional quantile regression, and in practice, researchers appear to be using the defaults of their particular software package.

Returning to the distribution of the writing test score variable, we can calculate its density using a variety of kernels and bandwidths. Figures 8.5 and 8.6 provide the estimated density of test scores using bandwidths of 1, 2, and 3, with two commonly-used kernels, the Gaussian (standard normal distribution) and the Epanechnikov. Both figures demonstrate that the estimated density at a specific test score can vary greatly depending on the bandwidth used, with smaller differences due to the choice of kernel.

While the help file for the Stata command `rifreg` suggests that

> The RIF for quantiles may be sensitive to the choice of bandwidth. It is advisable to graph the density and explore alternative choices of bandwidth for appropriate smoothness using the options in [the Stata command] vkdensity, for example.

kernel and bandwidth choices appear to be rarely discussed in papers applying unconditional quantile regression in education. I describe several ways to determine the optimal bandwidth in the empirical example below.

Presentation of Results

As should be clear at this point, quantile regression models yield numerous sets of results, depending on the quantiles of interest. As Davino et al. (2014) demonstrate, the number of distinct quantiles that can be estimated increases with the sample size, so that it is possible to estimates hundreds of different quantiles. In practice, such a vast quantity of output is unnecessary. Instead, authors adopt one of two approaches to the presentation of results, and sometimes both.

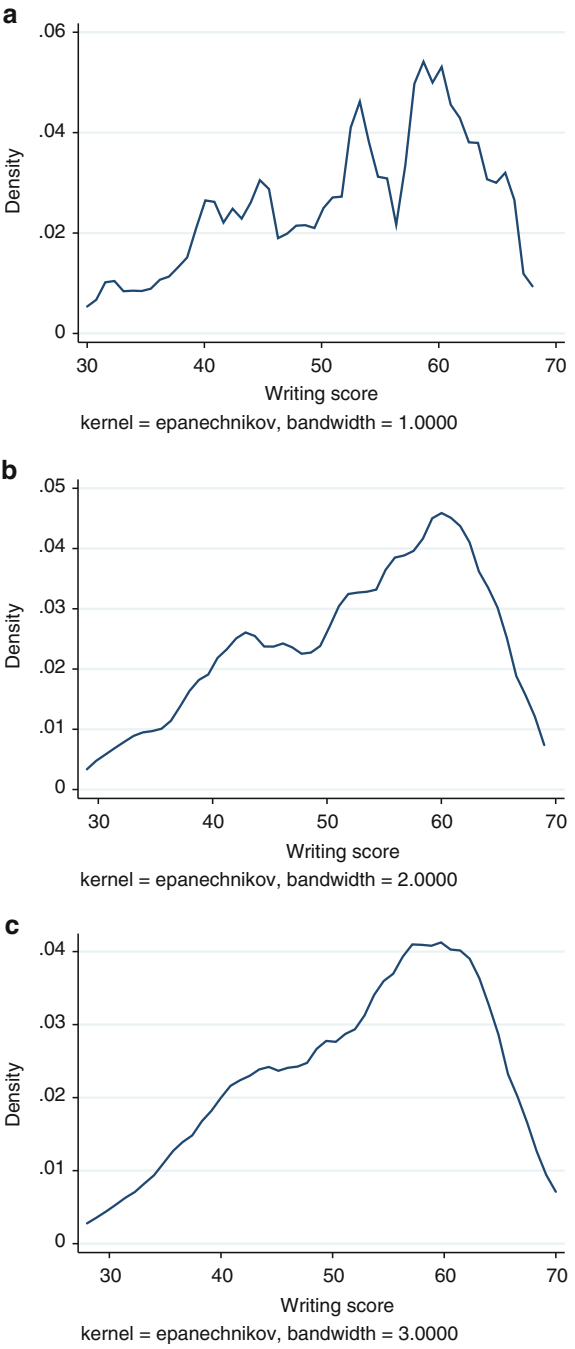

Fig. 8.5 Distribution of writing test scores using Epanechnikov kernel and different bandwidths. (**a**) Bandwidth = 1. (**b**) Bandwidth = 2. (**c**) Bandwidth = 3

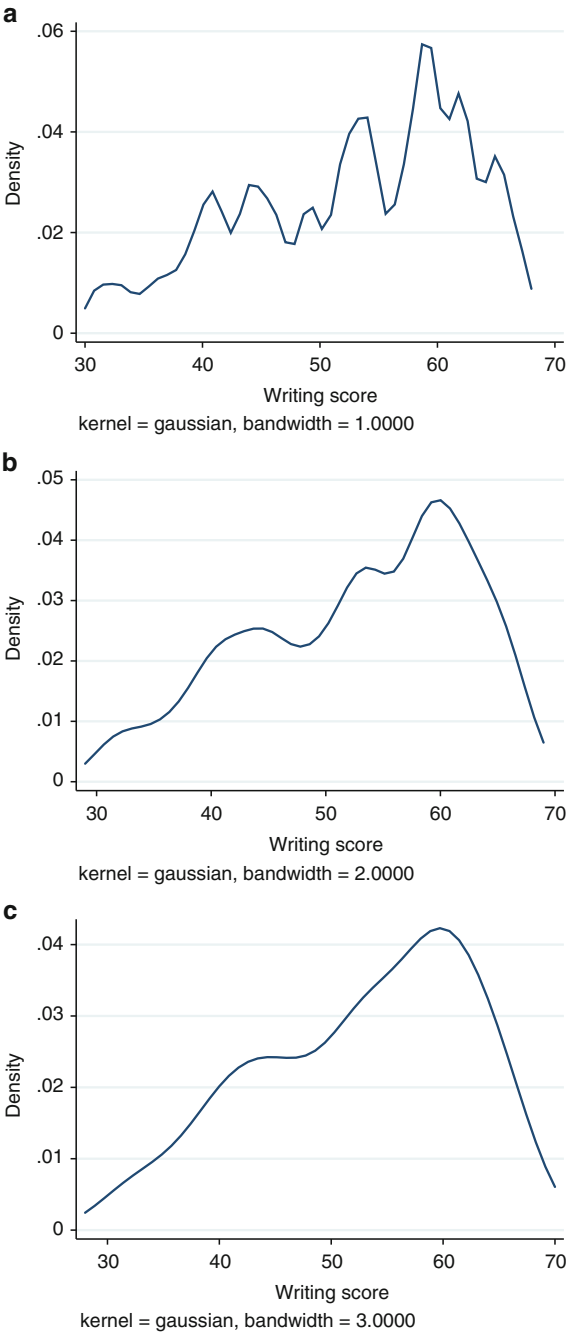

Fig. 8.6 Distribution of writing test scores using Gaussian kernel and different bandwidths. (**a**) Bandwidth = 1. (**b**) Bandwidth = 2. (**c**) Bandwidth = 3

In the first approach, specific quantiles are chosen and a table is created, with each column corresponding to the quantile regression results for a given quantile. Typical quantiles displayed in a table are the 10th, 25th, 50th (median), 75th, and 90th. This approach has the advantage of providing the reader with complete model results, and by following coefficients across columns, examine how the effect of an independent variable differs across the distribution of y.

A second approach to presenting results calculates the effect for each independent variable for each quantile in .01 increments, from .01 to .99. That is, the effect is calculated for the 1st quantile, the 2nd quantile, and so forth. Obviously a table of 99 results is not feasible, nor likely to be comprehensible, so the results are instead graphed, with the quantiles along the x-axis and the size of the quantile regression coefficient on the y-axis. Superior graphs also include 95 % confidence intervals for each quantile, so that the reader can understand at what points along the distribution of y the effect is not statistically significant (the intervals bracket 0). Both of these approaches will be shown below.

Empirical Example

To illustrate the use of unconditional quantile regression in postsecondary research, I use data from the 2004 National Survey of Postsecondary Faculty (NSOPF) to understand the impact of gender and other covariates on the distribution of faculty compensation. Faculty compensation and its determinants have long been a topic of study within higher education, as researchers have striven to understand why racial and gender differentials exist in faculty pay.

The first example estimates the male-female differential in faculty compensation.[2] The dependent variable is the amount of base salary received during the calendar year from the faculty member's institution, excluding other sources of compensation from within the institution (such as summer salary and payment for overload courses and administrative duties) as well as outside the institution (such as consulting fees and honoraria).

As noted earlier in the chapter, the unconditional quantile regression estimator as implemented by Firpo et al. (2009) relies an the estimated density of y, and this quantity varies depending on the kernel function and bandwidth chosen for estimation. This in turn raises the crucial question, how should one choose the kernel and bandwidth? Applied researchers appear to rely on software defaults for these choices, which may not always be the best strategy for estimation. Instead, the researcher should investigate the distribution of y, determine the optimal bandwidth, and run a sensitivity analysis by altering the kernel and bandwidth and reestimating the quantile regression model for different values of the bandwidth

[2]To simplify the analysis, no survey weights or adjustments of the standard errors for the complex sampling design of the NSOPF are used, and the dependent variable is not logged.

and different choices of kernel. This approach is similar to many regression-discontinuity applications, which estimate the regression model multiple times using varying bandwidths around the cutoff score.[3] Given the amount of output that an unconditional quantile regression model produces (one set of model results for each quantile), it is not feasible to include all of these sensitivity analyses in the typical journal article. However, the results should be summarized in the text, and a web appendix that details the analyses should be provided.

Several commands are available in Stata for estimating unconditional quantile regression models, as well as for determining the optimal bandwidth for a given application. Firpo et al. (2009) have developed the Stata command `rifreg` to implement their unconditional quantile regression estimator.[4] It relies on Stata's `kdensity` command to estimate the density of y at the specified quantile using the Gaussian kernel as the default, and the Stata manual explains how this is accomplished. The `kdensity` command estimates the "optimal" bandwidth by using "the width that would minimize the mean integrated squared error if the data were Gaussian and a Gaussian kernel were used, so it is not optimal in any global sense. In fact, for multimodal and highly skewed densities, this width is usually too wide and oversmooths the density" (StataCorp LP, 2013, p. 1003). Such language is not reassuring, and highlights the risks of relying on software defaults for modeling choices.

Another user-written command, `vkdensity` (Fiorio, 2004), allows the user to use three different approaches to determining the optimal bandwidth. The field of density estimation is fairly extensive, so the following description is only a brief overview. Each of the three approaches takes some measure of the spread in the distribution of y, combined with the sample size and a numerical adjustment, to determine the optimal bandwidth h. The default in Stata, for example, is the approach proposed by Silverman (1992)

$$h = \frac{.9m}{n^{1/5}} \tag{8.8}$$

where m is the smaller of either the standard deviation of y or the interquartile range (75th percentile−25th percentile) divided by 1.349 (StataCorp LP, 2013, p. 1010). Härdle (1991) proposes a similar formula, using 1.06 in the numerator instead of .9. Not surprisingly, these two approaches tend to yield similar h's. Finally, Scott (1992) proposes a more complex approach, combining measures of the "roughness" $R(K)$ and variance σ_K of the kernel with the standard deviation and sample size of y

$$h = 3\left[\frac{R(K)}{35\sigma_K^4}\right]^{1/5} \sigma_y n^{-1/5}. \tag{8.9}$$

[3]Indeed, one of the co-authors of the Firpo et al. (2009) paper has done this in their discussion papers, but omitted the sensitivity analyses from their published papers (Fortin, June 2 2014, Personal communication).

[4]While their estimator is easily programmed by hand, the ado files for this command can be found at http://faculty.arts.ubc.ca/nfortin/datahead.html

Table 8.4 Optimal bandwidth estimation

s_y	28,604	Silverman	$\frac{.9(22{,}606)}{9{,}949^{1/5}} = 3{,}228$
25th percentile	47,500		
75th percentile	77,996	Härdle	$\frac{1.06(22{,}606)}{9{,}949^{1/5}} = 3{,}802$
IQR	30,496		
IQR/1.349	22,606	Scott	$\frac{1.144(28{,}604)}{9{,}949^{1/5}} = 5{,}192$
n	9,949		

For the Gaussian kernel, $R(K) = .5/\sqrt{\pi}$ and $\sigma_K = 1$ (Salgado-Ugarte, Shimizu, & Taniuchi, 1995), so that Eq. 8.9 simplifies to

$$h = 1.144\sigma_y n^{-1/5}. \tag{8.10}$$

These different approaches to determining the optimal bandwidth differ in two ways. First, the factor used to adjust the standard deviation to determine the bandwidth h varies. Second, either the standard deviation or the interquartile range is used as a measure of the spread of the distribution. Table 8.4 estimates h, using the three approaches and the base salary data from the NSOPF. The Scott estimate is larger not only due to the larger factor (1.144), but also because in this application, the standard deviation (28,604) is larger than the interquartile range divided by 1.349 (22,606). The Silverman optimal bandwidth is the default of the `kdensity` command, and the other two optimal bandwidths can easily be estimated with the `vkdensity` command using the `hardle` and `scott` options.

In practice, it can be difficult to determine which approach is optimal, so I recommend using all three to determine the sensitivity of your results and reporting the results using the Silverman formula in your tables (simply because this is the default, and your results will be comparable to other researchers who rely on the software defaults). As Figs. 8.5 and 8.6 demonstrate, bandwidth choice has a much larger impact on the shape of the density than does kernel choice. Nevertheless, it is very easy to construct code that runs your model using all of the eight kernels that can be used with `kdensity` as a sensitivity check, and then using `rifreg`'s default Gaussian kernel for reporting your main model results.

Finally, one could always argue that for a given application, it makes sense to use a bandwidth that differs from the Silverman, Härdle, and Scott optimal bandwidths. Such an approach would require (a) a detailed explanation of why the particular bandwidth is better suited for the distribution of y than one of the optimal bandwidth calculations listed here (e.g., a distributional argument) and, (b) a summary of results based on the Silverman, Härdle, and Scott optimal bandwidths, as these are some of the more common approaches to the knotty issue of which bandwidth to use in kernel density estimation. The worry here is that one could play around with the bandwidth until the desired results are found, much as researchers can run multiple linear models with different specifications until they find what they are seeking (Ho, Imai, King, & Stuart, 2007). Without such an explanation, the reader will be left wondering how robust one's results really are.

Table 8.5 Male-female salary differentials, OLS and unconditional quantile regression results

		Quantiles of y				
	OLS	.10	.25	.50	.75	.90
Female	−5,441	−979	−1,562	−3,662	−7,228	−12,991
	(516)***	(420)	(412)***	(534)***	(810)***	(1,372)***
Asian	684	1,224	2,372	2,917	457	−2,909
	(891)	(610)	(659)***	(936)**	(1,528)	(2,685)
Black	−50	−259	246	606	−233	284
	(992)	(848)	(820)	(1,014)	(1,465)	(2,584)
Latino	550	2,053	1,164	724	−2,079	762
	(1,057)	(788)**	(888)	(1,101)	(1,580)	(2,865)
Nat. Amer.	−4,513	490	−377	−5,482	−8,172	−10,867
	(1,678)**	(1,306)	(1,419)	(1,846)**	(2,474)***	(3,593)**
Full	24,506	8,935	16,545	25,265	34,129	35,600
	(578)***	(476)***	(450)***	(583)***	(963)***	(1,692)***
Associate	8,040	7,560	11,120	10,925	6,941	489
	(581)***	(500)***	(498)***	(601)***	(795)***	(1,266)
Articles	1,901	332	533	1,078	2,310	4,823
	(61)***	(33)***	(41)***	(67)***	(119)***	(268)***
Books	573	257	459	754	457	562
	(170)***	(89)**	(115)***	(194)***	(302)	(547)
Constant	46,739	36,509	40,150	50,409	55,026	51,223
	(1,966)***	(1,061)***	(1,326)***	(1,906)***	(3,107)***	(3,745)***

Note: Cell entries are coefficients, with robust standard errors in parentheses. Models include 31 discipline-specific fixed effects. Unweighted n equals 9,949
** $p < 0.05$; *** $p < 0.01$

As with OLS, researchers usually display unconditional quantile regression results in a tabular format, often with the OLS results as a comparison to illustrate how conclusions can differ when understanding effects across the entire distribution. Table 8.5 presents one approach to displaying the faculty compensation results, with the OLS coefficients in the first column, and results for selected quantiles in the other columns. Note that with unconditional quantile regression, a separate regression model is estimated for every specific quantile, so to produce the results in Table 8.5, I estimated five different unconditional quantile regression models using the `rifreg` command.

Substantively, both the OLS and unconditional quantile regression results are in line with the literature, suggesting a negative male-female differential. The OLS results indicate that female faculty make, on average, over $5,000 less than male faculty with the same demographic and professional profile. With OLS, this estimate is the differential at the mean of the salary distribution. The unconditional quantile regression results, however, tell a different story. At the low end of the distribution, the male-female differential is about $1,000, increasing to almost $4,000 at the median and then to $13,000 at the 90th percentile. In other words, the results suggest

a male-female differential that is small when compensation is low, but much larger when compensation is high. This trend is masked when using OLS to estimate the male-female differential.

Another way to conceptualize the quantile regression results is with a thought experiment, in which females suddenly become males. In the case of OLS, if this occurred, mean compensation for females would increase over $5,000. In the case of the quantile regression results, we should think of the entire distribution of compensation shifting, as females become males. If this occurred, the distribution would shift to the right (in a positive direction), with small shifts at the low end of the distribution, and much larger shifts at the higher end of the distribution.

Graphical presentation is very helpful in presenting conditional and unconditional quantile regression model results, as the results for every .01 quantile can be summarized in a single graphic. In Fig. 8.7, the x-axis consists of quantiles ordered from .01 to .99, and the y-axis is the size of the female dummy variable coefficient. In other words, the figure displays the male-female differential for the 1st through the 99th quantiles, plotted as the thick line, summarizing the results from 99 different unconditional quantile regression models. They are different in that they are each estimated at a different quantile; the set of independent variables is the same for each model. The dotted lines above and below the thick line plot the 95 % confidence intervals for each coefficient, and the horizontal dashed line plots the OLS estimate of the differential (it is constant across the quantiles because OLS yields only one estimate of the differential).

Figure 8.7 adds two additional details to the story of male-female salary differences that are not apparent in Table 8.5. First, the confidence intervals for the coefficients below the 8th quantile bracket zero, indicating that the differential is not statistically significant at the very bottom of the distribution. Salary equity, it would seem, is achieved at the lowest end of the distribution. Second, the size of

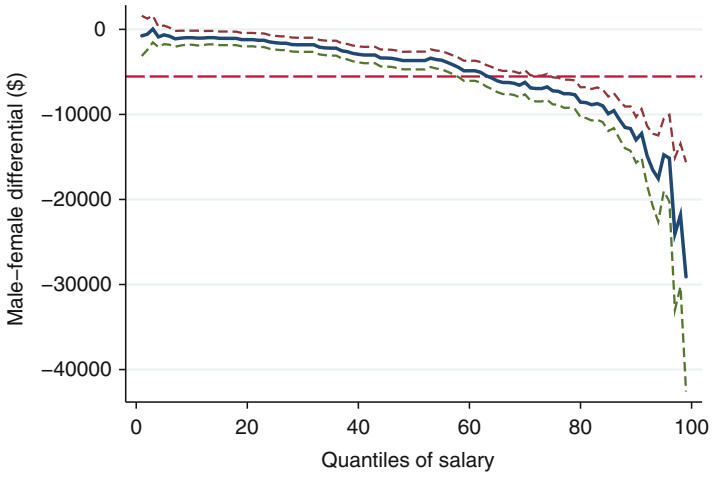

Fig. 8.7 Male-female differential in faculty compensation, summary of quantile regression results

Fig. 8.8 Effect of one additional publication on faculty compensation, summary of quantile regression results. (**a**) Books. (**b**) Articles

the salary differential increases rapidly above the 90th quantile, increasing to almost $30,000 at the 99th quantile, although the confidence intervals for this part of the distribution are wide.[5]

Graphics such as Fig. 8.7 are also useful when reviewing results for a large set of independent variables. Figure 8.8, for example, summarizes the results for the effect

[5]Please note that for expository purposes I am assuming selection on observables, but this clearly does not hold here. There are many differences between male and female faculty that are not taken into account by the simple model estimated here, so the results should not be interpreted as the "true" male-female salary differential.

of the number of books and articles published in the previous 2 years (not career publications) on compensation. An additional book yields a small, modest increase in salary along all parts of the distribution. An additional article, however, yields a small payoff at the low end of the distribution, with an increasingly larger yield at the higher end of the salary distribution. An additional article results in a $1,900 increase in compensation at the mean, but a $4,800 increase in compensation at the 90th quartile (see Table 8.5).

Inference

Correctly estimated standard errors are crucial for most analyses, as they are used to calculate test statistics for hypotheses, such as whether β is different from 0, as well as for confidence intervals. Yet besides dealing with non-independence of observations (such as the clustering of students within colleges), higher education researchers have tended to ignore this issue in much of their applied work. For example, robust standard errors are widely acknowledged as more appropriate for most applications using OLS and related models, but relatively few published papers in our field use robust standard errors, and instead use default standard errors that assume homoskedasticity.

For unconditional quantile regression, researchers face the choice of using standard errors derived from formulas assuming asymptotic normality (the default for most linear models such as OLS, logistic regression, and HLM) or standard errors derived from bootstrapping. Asymptotic normality refers to the idea that as the sample size for a random variable increases, its probability density function more closely approximates the standard normal distribution. In formal proofs, the sample size is taken to infinity, which raises the question of how large does a sample have to be for the assumption of asymptotic normality to hold? Unfortunately, there is no simple answer to this question, and most researchers simply assume it holds when estimating their regression models and standard errors.

Rather than using derived formulas to estimate the standard error of a regression coefficient β, bootstrapping uses the data at hand to estimate the standard errors. Assuming that the sample at hand is representative of the population, repeated subsamples of the sample are drawn, and the parameter of interest (in this case, β) is estimated. The variance and standard deviation of β is estimated, and because we are viewing the distribution of β in the bootstrapped samples as a sampling distribution, this standard deviation is the standard error for β. While bootstrapped standard errors are appealing because we do not need to rely on distributional assumptions, one drawback is that they do change as the model is reestimated, due to the drawing of random samples to estimate the standard error. As with multiple imputation, this can be avoided by choosing a seed number that starts the random process, so that results can be replicated.

As an example, Fig. 8.9 presents the 100 unconditional quantile regression coefficients at the median for the female dummy variable that are produced when

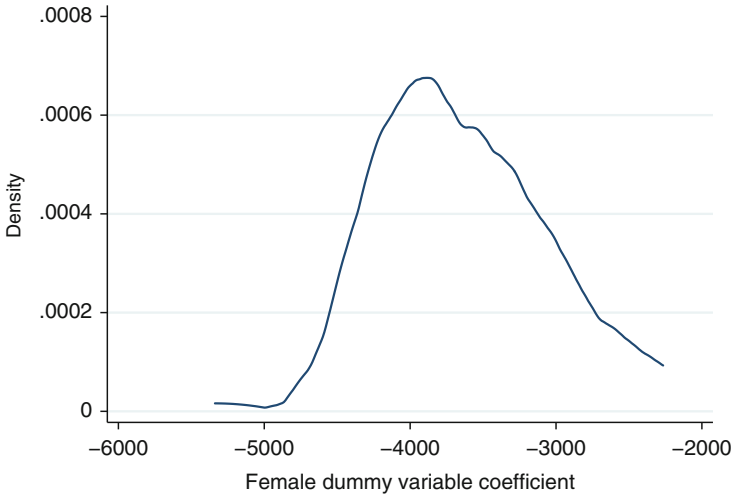

Fig. 8.9 Distribution of bootstrapped regression coefficients, male-female differential

estimating standard errors via the bootstrap. They are the result of drawing 100 subsamples from the data and then estimating the faculty compensation model at the 50th percentile using unconditional quantile regression on each sample. The estimate for the male-female differential reported in Table 8.5 is −3,662; note that the distribution of the 100 bootstrapped coefficients is centered just to the right of −4,000, and the mean of the coefficients equals −3,635, close to our original quantile regression estimate. The standard deviation of the coefficients is 579, which is close to the asymptotic standard error of 534 reported in Table 8.5.

Table 8.6 compares the two sets of standard errors that can be estimated for any unconditional quantile regression model, formula-based versus bootstrapped. The last set of numbers are ratios of the bootstrapped standard errors to the asymptotic standard errors (the default option for `rifreg` and most statistical software), such that the ratios can be interpreted as percentage differences. For example, at the 10th quantile, the bootstrapped standard errors for the female dummy variable coefficient are 7 % larger than the asymptotic standard errors. On average, the bootstrapped standard errors are about 5 % larger, with some much larger differences, especially for the 90th quantile. Such differences naturally raise the question of which set should be used when reporting results. Like many areas of statistics, partisans can be found on both sides of the issues. Given its lack of distributional assumptions, I tend to favor the bootstrapping standard errors, with two caveats. First, a specific seed for the random number generator should always be used, otherwise you (and other scholars) will not be able to exactly replicate your results. Second, the traditional standard errors should also be estimated and compared to the bootstrapped standard errors, as a sensitivity analysis.

Table 8.6 Comparison of standard errors

	Asymptotic					Bootstrapped					Ratio				
	.10	.25	.50	.75	.90	.10	.25	.50	.75	.90	.10	.25	.50	.75	.90
Female	420	412	534	810	1,372	450	458	579	820	1,374	1.07	1.11	1.08	1.01	1.00
Asian	610	659	936	1,528	2,685	711	690	980	1,527	2,878	1.17	1.05	1.05	1.00	1.07
Black	848	820	1,014	1,465	2,584	881	794	973	1,423	2,712	1.04	0.97	0.96	0.97	1.05
Latino	788	888	1,101	1,580	2,865	766	823	1,148	1,592	2,755	0.97	0.93	1.04	1.01	0.96
Native Amer.	1,306	1,419	1,846	2,474	3,593	1,439	1,634	1,941	2,245	3,577	1.10	1.15	1.05	0.91	1.00
Full	476	450	583	963	1,692	523	525	683	1,331	2,534	1.10	1.17	1.17	1.38	1.50
Associate	500	498	601	795	1,266	536	468	635	788	1,114	1.07	0.94	1.06	0.99	0.88
Articles	33	41	67	119	268	32	43	69	122	370	0.97	1.05	1.03	1.03	1.38
Books	89	115	194	302	547	84	118	152	277	526	0.94	1.03	0.78	0.92	0.96
Constant	1,061	1,326	1,906	3,107	3,745	1,099	1,386	1,743	3,217	5,028	1.04	1.05	0.91	1.04	1.34

Sensitivity of Results

The unconditional quantile regression model results discussed so far are based on the defaults for the `rifreg` command; that is, they use the Gaussian kernel and the optimal bandwidth calculated with the Silverman (1992) method. As a sensitivity analysis, the unconditional regression models were reestimated using the Gaussian, Epanechnikov, and uniform kernels, and then for each kernel using the optimal bandwidth formulas of Silverman, Härdle, and Scott, as outlined above. Table 8.7 presents the results for the female dummy variable coefficient only. Comparing the results using different bandwidths within each of the three kernels, the largest dollar difference between the estimates is less than $2,000, and the other differences are much smaller. Comparing the results using the different kernels in the table, the differences are even smaller, which is not surprising given that kernel density estimators are generally more sensitive to choice of bandwidth than choice of kernel. For this application, the results appear relatively insensitive regardless of whether the Gaussian, Epanechnikov, or uniform kernel is used, as well as to how the optimal bandwidth is calculated.

Table 8.7 Sensitivity of results to kernel selection and bandwidth calculation

Kernel	Bandwidth calculation	Quantiles of y				
		.10	.25	.50	.75	.90
Gaussian	Silverman	−979	−1,562	−3,662	−7,228	−12,991
		(420)**	(412)***	(534)***	(810)***	(1,372)***
	Härdle	−1,024	−1,651	−3,719	−7,207	−13,254
		(439)**	(435)***	(542)***	(808)***	(1,400)***
	Scott	−993	−1,582	−3,689	−7,232	−13,187
		(426)**	(417)***	(538)***	(811)***	(1,393)***
Epanechnikov	Silverman	−981	−1,557	−3,685	−7,300	−13,251
		(420)**	(410)***	(537)***	(818)***	(1,400)***
	Härdle	−1,032	−1,664	−3,717	−7,212	−13,320
		(442)**	(439)***	(542)***	(808)***	(1,407)***
	Scott	−1,001	−1,573	−3,717	−7,262	−13,483
		(429)**	(415)***	(542)***	(814)***	(1,424)***
Uniform	Silverman	−904	−1,486	−3,486	−6,262	−12,492
		(387)**	(392)***	(508)***	(702)***	(1,320)***
	Härdle	−909	−1,577	−3,374	−7,135	−12,323
		(390)**	(416)***	(492)***	(800)***	(1,302)***
	Scott	−1,023	−1,488	−3,957	−7,227	−14,236
		(438)**	(392)***	(577)***	(810)***	(1,504)***

Note: Cell entries are coefficients, with standard errors in parentheses
** $p < 0.05$; *** $p < 0.01$

Comparison to Conditional Quantile Regression

As noted previously, conditional quantile regression estimates are not only difficult to interpret compared to unconditional quantile regression, but the substantive sizes of the coefficients often differ. Table 8.8 presents the conditional quantile regression results for the exact same faculty compensation model from Table 8.5, using the `qreg` command. Similar trends are evident for the male-female differential and the effect of publications on compensation, although the effects are smaller in the conditional quantile regression model than the unconditional model. Figure 8.10 plots the coefficients and confidence intervals for the unconditional and conditional regression coefficients. In each graph the solid dark line plots the unconditional results and the lighter, dashed line plots the conditional results. The male-female differential is relatively the same for both estimators until about the 40th quantile, after which the conditional estimates suggest a smaller effect for gender (panel *a*). The estimated effect of one additional book is about the same for both estimators (panel *b*). The results for articles, however diverge, with larger coefficients for the conditional estimates at lower quantiles, and then reversing at the 80th quantile, exhibiting much smaller estimates than the unconditional results.

Table 8.8 Male-female salary differentials, conditional quantile regression results

	Quantiles of *y*				
Variable	.10	.25	.50	.75	.90
Female	−1,077	−1,370	−2,597	−4,181	−7,600
	(526)**	(482)***	(469)***	(645)***	(1,185)***
Asian	613	2,055	1,648	149	373
	(906)	(832)**	(809)**	(1,112)	(2,043)
Black	−1,503	−320	108	430	774
	(1,010)	(927)	(901)	(1,238)	(2,276)
Latino	428	−105	352	−221	3,026
	(1,076)	(988)	(960)	(1,319)	(2,425)
Native Amer.	−3,477	−3,361	−2,597	−3,788	−4,385
	(1,707)**	(1,568)**	(1,524)	(2,094)	(3,848)
Full	12,636	16,932	23,371	29,612	34,933
	(588)***	(540)***	(525)***	(721)***	(1,325)***
Associate	6,026	6,895	8,352	8,888	9,733
	(591)***	(542)***	(527)***	(725)***	(1,332)***
Articles	1,010	1,425	1,824	2,537	3,240
	(62)***	(57)***	(55)***	(76)***	(139)***
Books	410	321	503	369	1,233
	(173)**	(159)**	(154)***	(212)	(389)***
Constant	37,597	44,586	47,173	51,112	52,547
	(2,001)***	(1,837)***	(1,786)***	(2,454)***	(4,510)***

Note: Cell entries are coefficients, with standard errors in parentheses. Models include 31 discipline-specific fixed effects. Unweighted n equals 9,949
** $p < 0.05$; *** $p < 0.01$

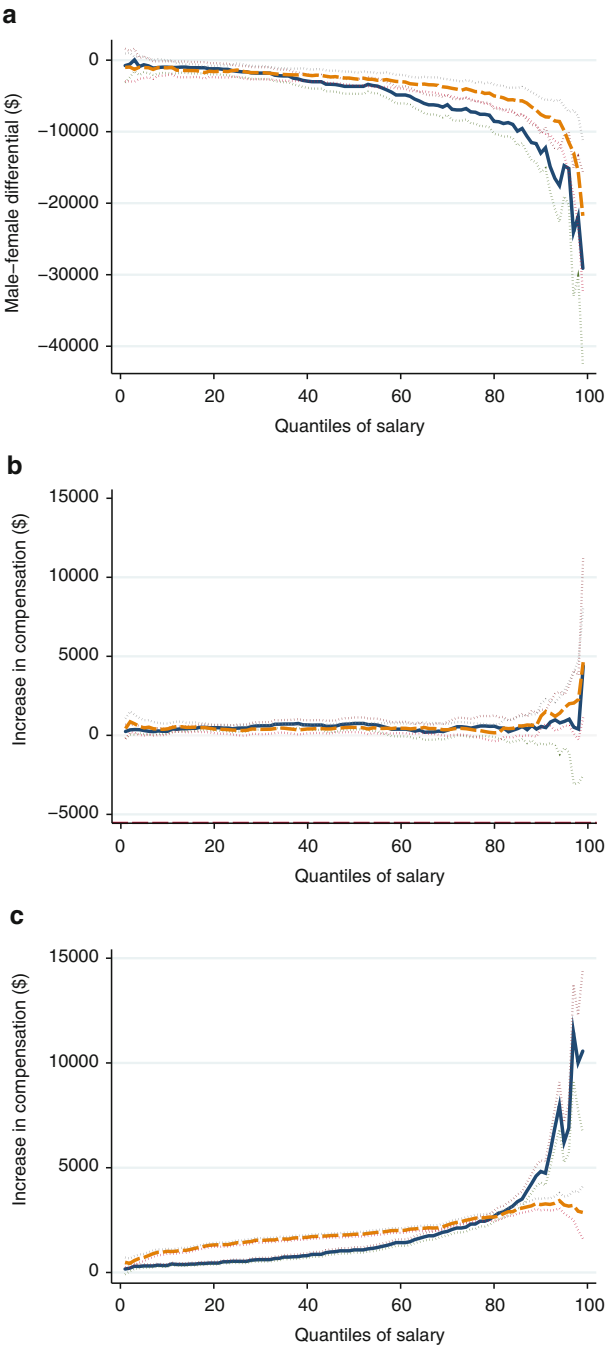

Fig. 8.10 Comparison of unconditional and conditional regression models. (**a**) Male-female differential. (**b**) Books. (**c**) Articles

Unconditional Quantile Regression with Endogenous Treatment

The unconditional quantile regression estimator described above simply uses the recentered influence function (RIF) to transform y before employing OLS to estimate the coefficients. OLS relies on several assumptions so that we can use the results to infer the effect of an independent variable on y. Most of these assumptions can be easily dealt with in one way or another. Heteroskedastic errors, for example, can be handled with robust standard errors, while severe multicollinearity can be addressed through data reduction or an increase in the number of observations. The single most important assumption underlying OLS, however, also turns out to be the most difficult to address.

In the context of educational data, in which students, families, and institutions make a variety of choices that we can only observe in our data (as opposed to experimentally manipulate), OLS assumes *exogeneity*, *conditional independence*, or *selection on observables*. Exogeneity means that the independent variables in the model are uncorrelated with the error term u

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + u_i \tag{8.11}$$

where D is a dummy variable indicating participation in a policy, program, or behavior of interest, and X represents a set of control variables.

Given our interest in the effect of D, unbiased estimation of β_1 is crucial. However, we can only conclude that β_1 is unbiased if D is uncorrelated with u. Given that u represents the variables that affect Y but are not included in the model, this assumption is unrealistic in most areas of higher education research. Consider a simple example that should be familiar to all postsecondary researchers: student outcomes. A few of the factors that drive student decision-making and affect student outcomes are the quality and culture of the primary and secondary schools attended, how much the family emphasizes education and how supportive they are of postsecondary educational pursuits, the attitudes of friends and peers towards educational choices and appropriate aspirations for life, as well as myriad other sources of social and cultural capital, student academic ability, psychological makeup such as conscientiousness and grit, their physical health, and other sources of human capital, and the financial resources available through family connections, postsecondary institutions, and other sources, such as state and federal agencies.

The central issue is that these factors drive decisions about outcomes of interest, such as college access and persistence, and many of these factors also drive decisions to participate in programs and behaviors of interest (D's), such as remediation, first-year initiatives, and student engagement. We can only credibly claim exogeneity if none of the factors in u are correlated with D. Clearly, this is a high hurdle to jump, which is why alternative forms of OLS that can credibly claim exogeneity of treatment variables (such as instrumental variables, regression discontinuity, and panel models) are becoming more popular with educational researchers (Murnane & Willett, 2011).

Instrumental Variables

Instrumental variables is a simple yet powerful approach to the problem of endogeneity of treatment. Given

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + u_i \tag{8.12}$$

we seek an alternative form of D that is not correlated with u. If a variable exists that is highly correlated with D but not with u, we can use a two-step process to "purge" D of its correlation with u.

First, if D and Z are correlated, we can estimate the following model in which D is driven in part by Z

$$D_i = \theta_0 + \theta_1 Z_i + \theta_2 X_i + v_i \tag{8.13}$$

and then create predicted values from this model

$$\hat{D}_i = \hat{\theta}_0 + \hat{\theta}_1 Z_i + \hat{\theta}_2 X_i. \tag{8.14}$$

Second, we use these predicted values in place of D in our original treatment effect model

$$Y_i = \beta_0 + \beta_1 \hat{D}_i + \beta_2 X_i + u_i \tag{8.15}$$

because if Z and X are uncorrelated with u, then this new estimate of D must also be uncorrelated with u. See Porter (2014) for an explanation of the assumptions underlying IV, and Bielby, House, Flaster, and DesJardins (2013) for an overall review.

Building on the work of Abadie, Angrist, and Imbens (2002) and Abadie (2003), Fröhlich and Melly (2010, 2013) propose an IV estimator for unconditional quantile regression when the main focus of interest is the effect of a binary treatment variable, and a credible binary instrument for the treatment exists.[6] Similar to the conditional quantile regression approach, their estimator is formulated as an optimization problem with weights, such that

$$\arg\min \sum_{i=1}^{N} \rho_\tau (Y_i - \alpha - \beta D_i) W_i \tag{8.16}$$

where $\rho_\tau (Y_i - \alpha - \beta D_i)$ is again an absolute value function for the linear model of $Y_i = \alpha + \beta D_i$, such that $\rho_\tau (u) = u \cdot (\tau - \mathbf{1}(u < 0))$, and W_i represent the weights for the IV estimator.

[6] Continuous instruments can be dichotomized to satisfy this requirement.

Equation 8.16 appears very similar to Eq. 8.3, and like the conditional quantile regression estimator is solved through optimization. There are, however, two major differences between the conditional quantile regression estimator and the instrumental variables unconditional quantile regression estimator. First, note that the covariates X (other than the treatment variable D) do not appear in Eq. 8.16, as they do in the formula for conditional quantile regression. This results in unconditional versus conditional quantile estimates. Second, Eq. 8.16 includes the set of IV weights W, which are used to identify the effect of D for compliers in the population (see Porter (2014) for an explanation of compliers, defiers, always-takes and never-takers).

The weights W are derived from the treatment variable D, the instrument Z, and an estimate of the probability that $Z = 1$ (notated by $\pi(Z = 1|X)$)

$$W_i = \frac{Z_i - \pi(Z_i = 1|X_i)}{\pi(Z_i = 1|X_i)(1 - \pi(Z_i = 1|X_i))}(2D_i - 1). \tag{8.17}$$

The weights are the crucial part if this estimator, and can be thought of as "complier weights." They weight the data in order to estimate the effect of D for compliers, relying on the relationship between the instrument and the endogenous regressor.

Recall that with IV, we can only estimate the effect of D for units whose behavior is actually affected by the instrument Z. With binary instruments and treatments, we can partition units into four cells (assuming monotonicity, i.e., the absence of defiers). Table 8.9 illustrates these cells based on the values of the instrument and the treatment for units. Compliers fall across the diagonal, because they decline treatment when $Z = 0$, and agree to treatment when $Z = 1$. We cannot identify them individually, because always-takers and never-takers also appear in these cells as well ($Z = 0, D = 0; Z = 1, D = 1$). For example, never-takers always decline treatment regardless of the value of the instrument, so they are units whose $D = 0$ for both $Z = 0$ and $Z = 1$.

We can, however, estimate the treatment effect for the compliers across the entire dataset, even if we cannot identify them individually, and the weights W achieve this, as well as balancing the distribution of covariates between treated and untreated compliers (Fröhlich & Melly, 2010). This allows the estimated treatment effects to be considered unconditional even with the inclusion of covariates, similar to the

Table 8.9 Compliance behavior of units

	Treatment D_i	
Instrument Z_i	0	1
0	Compliers and never-takers	Always-takers
1	Never-takers	Compliers and always-takers

Table 8.10 Weights for Fröhlich and Melly (2013) IV estimator

	Z	D	$\pi(Z)$	W
Compliers and always-takers	1	1	0.5	2
Compliers and never-takers	0	0	0.5	2
Always takers	0	1	0.5	−2
Never-takers	1	0	0.5	−2

effects estimated by the Firpo et al. (2009) recentered influence function approach that assumes exogeneity of treatment.

In the simplest example, suppose we estimate an IV unconditional quantile regression model with no covariates. In this case, $\pi(Z = 1|X) = \pi(Z)$, or the mean of Z. Suppose further that for half of the sample, the instrument takes the value of 1, so $\pi(Z) = .5$. Using Eq. 8.17 and the four groups from Table 8.9, we can estimate the weights as shown in Table 8.10. The two groups that contain compliers always receive positive weights, while the always-takers and never-takers always receive negative weights. The size of the weight is determined by the propensity score, $\pi(Z = 1|X)$; the weights are equal among the groups only when $\pi(Z = 1|X) = .5$.

Estimation

This estimator has been implemented in Stata via the user-created `ivqte` command. The main issue in using the IV unconditional quantile regression estimator is generating W, specifically, estimating $\pi(Z = 1|X)$. With no covariates in the model, $\pi(Z = 1|X)$ is the mean of Z. With covariates, the estimated probability of Z becomes a type of propensity score, and there are different ways of estimating it.

First, because Z is a binary variable, we can use either logistic regression (the `ivqte` default) or a linear probability model (an OLS regression with a binary dependent variable). Typically logistic regression is preferred, because it yields predicted probabilities bounded within 0 and 1.

Second, we can use either global or local models. Global models use the entire sample to estimate $\pi(Z = 1|X)$; for example, $\pi(Z = 1|X)$ is estimated using a logistic regression model with Z as the dependent variable and X as the covariate(s). Local models use a kernel and weighted subsets of the data to estimate $\pi(Z_i)$, somewhat similar to the kernel density estimator. As with the kernel density estimator, some choice must be made as to how much of the data should be used. With local logistic regression, two smoothing parameters must be set to determine the bandwidth used: h for continuous predictors of Z, which varies between 0 and ∞, and λ for discrete predictors of Z, which varies between 0 and 1. When h is set to infinity and λ to 1, the entire dataset is used and a global model is estimated. In addition, a kernel must be chosen for local logistic regression; the Epanechikov kernel is the default in `ivqte`.

As with kernel density estimators, the researcher faces choices as to how smooth the estimates should be (h and λ), as well as which kernel to use. Because the literature indicates that kernel choice has little practical impact on results (Fröhlich & Melly, 2010), the primary issue is choosing the optimal values of h and λ. Fröhlich and Melly (2010) have developed a related command, `locreg`, which provides the researcher with these optimal values.

Empirical Example

To illustrate the use of unconditional quantile regression with instrumental variables, I continue with the NSOPF data to understand the impact of faculty unions on faculty compensation. The major concern with literature in this area is the endogeneity of unionization. The literature suggests two reasons why a variable measuring the presence of a faculty union at an institution is endogenous in a model with faculty compensation as the dependent variable. First, there may be omitted variables from the model. Faculty at unionized institutions may differ from faculty at non-unionized institutions in ways that are not easily measured. Unionized institutions may tend to attract faculty who prefer to teach rather than conduct research; faculty who do not conduct much research tend to earn less compensation than faculty who do. Even if we tried to control for teaching and research emphasis between campuses, our measures will be crude (such as the Carnegie classification), and their inclusion in the model will not sufficiently remove the correlation between the unionization variable and the error term, which leads to bias in the estimate of the effect of unionization.

Second, OLS assumes that the causal chain of events runs from x to y. Yet the literature on why faculty choose to unionize indicates that one of the primary drivers is low compensation. So while we might expect faculty unions to raise faculty salaries through collective bargaining, a strong case can be made that faculty compensation also drives unionization. Such simultaneity between the dependent and independent variables results in endogeneity, just as in the case of omitted variables.

Porter (2013) has argued that state public employee unionization laws can be considered a valid instrument for faculty unions at an institution, because faculty at public institutions are public employees. These laws vary in strength across the country, in terms of the ease in which faculty can form a union and the institution is required to collectively bargain with the union. Conditional on two covariates, state political ideology and the strength of state oversight over higher education, these laws should have a strong, direct effect on unionization and should not affect faculty compensation other than through unionization. He also demonstrates that this is a strong instrument (correlation of .58 between state ideology and campus unionization).

Figure 8.11 compares the results of the two approaches to understanding the effects of unions on faculty compensation. Two sets of models are estimated. The first set assumes unionization is exogenous, and uses the `rifreg` command to

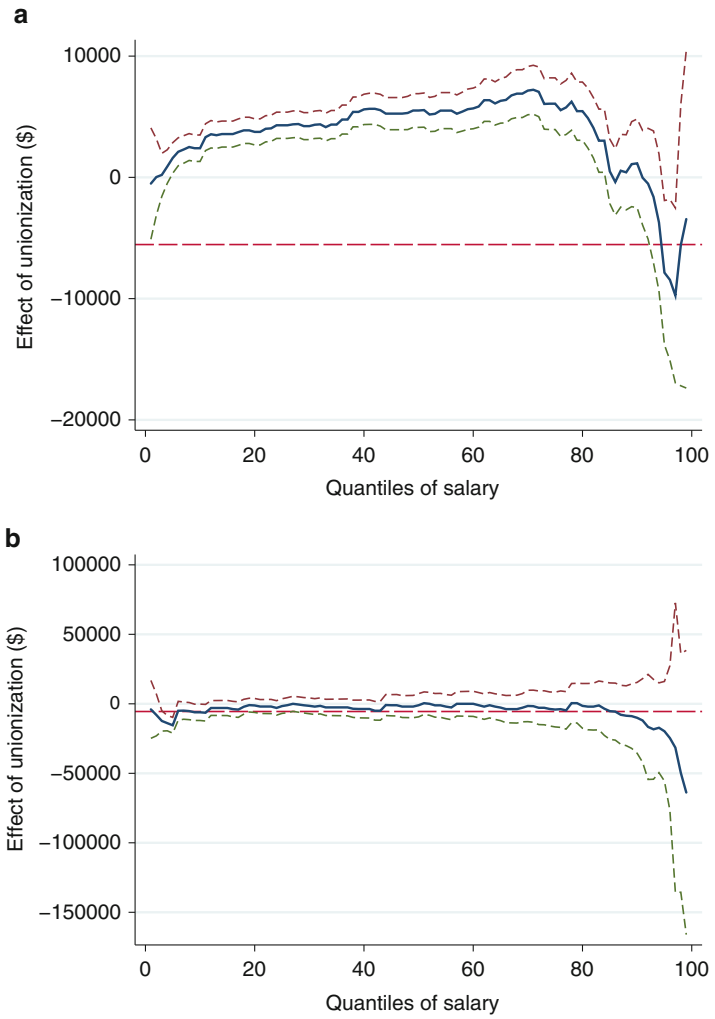

Fig. 8.11 Unionization and compensation, exogenous and endogenous (IV) quantile regression. (**a**) RIF-OLS. (**b**) Quantile IV

estimate the effect of unionization. Besides a dummy variable indicating unionization of the individual faculty member's campus, the model includes the individual-level faculty covariates used in the previous models, as well as logged student enrollment at the institution, logged expenditures per student, Barron's college selectivity index, and dummy variables for Carnegie classification as control variables. The second set of results assumes unionization is endogenous, and uses the `ivqte` command, with a binary indicator for weak/strong state public employee collective bargaining rights as the instrument. The model also includes two covariates, state political ideology and whether the state had a consolidated governing board.

Panel *a* of the figure contains the RIF-OLS results, which suggests unionization has a strong, positive effect on faculty compensation. The effects increase until about the 75th percentile, and then rapidly drop off to zero (no difference between faculty at unionized and non-unionized institutions). Panel *b* shows the quantile IV results using state laws as an instrument. The 95 % confidence intervals bracket almost the entire distribution, leading to the conclusion that unionization has no effect on faculty compensation.

As with any IV estimate, care must be made in interpretation of the results. In an OLS model with a truly exogenous treatment variable (e.g., analysis of an experiment where college students were randomly assigned to a treatment and control condition, with perfect compliance), the regression coefficient β can be interpreted as the average treatment effect – the estimated effect if we randomly selected students from the population and then administered the treatment. IV estimates, however, do not have the same interpretation. Instead, they produce what are known as *local average treatment effects*, where *local* refers to the subset of the population on which the treatment effect is estimated. In the context of IV, this is the group of units, known as compliers, whose assignment to treatment is determined by the instrument. In the current example, this means we should not conclude that unionization has a null effect. Instead, we can conclude that for the group of institutions whose faculty decide to unionize based on the strength of state laws, unionization has no effect. For example, we can say little about the effect of unionization on colleges that would never unionize despite how easy state public employee union laws may make the collective bargaining process. A very conservative, religious college may be hostile to unions, for example, and would always remain non-unionized regardless of state law.

Sensitivity Analysis

The previous analysis used global logistic regression to estimate $\pi(Z)$ when creating the IV weights; in other words, it assumed the default smoothing parameters $h = \infty$ and $\lambda = 1$. The `locreg` command (Fröhlich & Melly, 2010) allows users to find the optimal smoothing values for the IV estimator, using a leave-one-out cross-validation approach that seeks the smallest mean squared error. In leave-one-out cross-validation, values are first chosen for h and λ. The sample is then split into N datasets (the training sets), on which the local logistic regression model is estimated using all of the sample except for one observation, and the coefficients from the model are used with values of the independent variables from the remaining observation (the validation dataset) to make a prediction for Y. The predicted value is compared to the actual value, and the mean squared error (MSE) is calculated across the datasets for this pair of smoothing values. The user tries out different sets of values for the smoothing parameters to find the pair that yields the lowest mean squared error; these are the optimal values.

Table 8.11 Search process for optimal values of h and λ

First iteration			Second iteration			Third iteration			Fourth iteration		
h	λ	MSE	h	λ	MSE	h	λ	MSE	h	λ	MSE
0.2	**0.2**	**0.084307**	0.05	0.05	0.073809	0.06	0.01	0.073854	**0.1**	**0**	**0.069696**
0.2	0.5	0.084712	0.05	0.1	0.073813	0.06	0.02	0.073854	0.1	0.0025	0.070971
0.2	0.8	0.084924	0.05	0.15	0.073818	0.06	0.03	0.073855	0.1	0.005	0.071008
1	0.2	0.126978	0.05	0.2	0.073823	0.06	0.04	0.073855	0.1	0.0075	0.071069
1	0.5	0.13043	0.05	0.25	0.073827	0.06	0.05	0.073856	0.1	0.01	0.071155
1	0.8	0.131862	**0.1**	**0.05**	**0.073693**	0.08	0.01	0.073844			
∞	0.2	0.151678	0.1	0.1	0.073979	0.08	0.02	0.073844			
∞	0.5	0.152310	0.1	0.15	0.073985	0.08	0.03	0.073844			
∞	0.8	0.152905	0.1	0.2	0.073992	0.08	0.04	0.073845			
			0.1	0.25	0.073999	0.08	0.05	0.073845			
			0.15	0.05	0.082831	**0.1**	**0.01**	**0.071155**			
			0.15	0.1	0.082858	0.1	0.02	0.071741			
			0.15	0.15	0.082896	0.1	0.03	0.072719			
			0.15	0.2	0.082943	0.1	0.04	0.073595			
			0.15	0.25	0.082997	0.1	0.05	0.073693			
			0.2	0.05	0.08369	0.12	0.01	0.081329			
			0.2	0.1	0.083944	0.12	0.02	0.081334			
			0.2	0.15	0.084158	0.12	0.03	0.081337			
			0.2	0.2	0.084307	0.12	0.04	0.08134			
			0.2	0.25	0.084414	0.12	0.05	0.081343			
			0.25	0.05	0.086121						
			0.25	0.1	0.086493						
			0.25	0.15	0.086727						
			0.25	0.2	0.086883						
			0.25	0.25	0.086994						

This procedure is computationally intensive due to the cross-validation, and even more so given the large number of pairs of values to be tested. Rather than testing many values at once, I recommend using smaller sets of values in an iterative process to narrow down the choices and find the optimal values. For the faculty union example, I first tested the values .2, 1 and ∞ for h and .2, .5, and .8 for λ. MSEs were calculated for each possible pair that could be created from the six values, and as noted in Table 8.11, the pair (.2,.2) had the lowest MSE (shown in bold). Next, MSEs were calculated for the values .05 to .25 for both, resulting in the optimal values of .1 and .05 for this set of numbers. The process was repeated for .06, .08, .1, and .12 for h and .01, .02, .03, .04, and .05 for λ, and once more with .1 for h and 0, .0025, .005, .0075, and .01 for λ, yielding final values of .1 for h and 0 for λ.

Table 8.12 shows the results for the faculty union model with the default smoothing values of ∞ for h and 1 for λ compared to the optimal values of .1 for h and 0 for λ. For the estimates using the default settings, we would conclude that

Table 8.12 Effect of faculty unionization on compensation: IV unconditional quantile regression estimates

	Quantiles of y				
	.10	.25	.50	.75	.90
$h = \infty$ and $\lambda = 1$	−6,095	−1,622	−1,035	−4,114	−10,060
	(3,008)**	(2,684)	(4,271)	(6,377)	(13,008)
$h = .1$ and $\lambda = 0$	−12,850	−14,578	−11,600	−7,500	−11,400
	(46,586)	(29,946)	(18,926)	(97,642)	(75,227)

Note: Cell entries are coefficients, with standard errors in parentheses
** $p < 0.05$

unions decrease compensation at the 10th percentile, with no statistically significant differences along the rest of the distribution. For the estimates using the optimal bandwidths, the coefficients have the same sign as the default estimates, and while some are much larger in value, none are statistically significant. In this example, both approaches yield substantively similar results, and we would conclude that unionization has no effect on faculty compensation.

Discussion

The preceding review of the literature on quantile regression demonstrates the potential of analyzing distributions instead of means in postsecondary research. The main drawback to using OLS in applied research is that it only shows us the effect of independent variables on the mean of y. Quantile regression allows the researcher to estimate how the entire distribution of an outcome changes given a unit change in x, rather than just the change in the mean of y. While quantile regression models generate a much larger set of empirical results compared with OLS, careful use of tables and graphical presentation of results can easily illustrate how an independent variables affects the distribution of y.

For researchers seeking to use quantile regression, the first modeling choice they face is conditional versus unconditional quantile regression. For most postsecondary applications, conditional quantile regression does not seem to be a useful approach. It estimates the effect of an independent variable on the conditional distribution of y, such that the coefficient must be interpreted as a within-group effect, where the groups are defined by the independent variables used in the model. In general, this is not the effect that is useful for most evaluation and policy discussions. Instead, unconditional quantile regression would appear to be the best choice, because it tells us the effect of x on the unconditional distribution of y. In other words, if x increases by one unit, how much does the distribution of y change? This interpretation is similar to how we interpret OLS regression results.

Next, some assumptions must be made about the density of y. While the choice of kernel typically does not matter, the bandwidth does, and researchers

should investigate different bandwidths to determine the robustness of their results. Bootstrapping the standard errors, rather than relying on the default standard errors estimated by the software, is also recommended.

Whether researchers should use the recentered influence function approach that assumes exogeneity, or an alternative approach that assumes endogeneity, will depend on the particular research question. Given the ubiquity of unobservable selection processes in higher education, not just on the part students and their families, but also faculty and institutions, endogeneity is a more realistic assumption than exogeneity. Besides instrumental variables, econometricians have been devising other quantile regression approaches that can handle endogeneity, such as regression discontinuity quantile regression (Frandsen, Fröhlich, & Melley, 2012). Work in this area is changing rapidly, so readers are advised to conduct a thorough literature review before using these techniques.

Further Resources

Conditional Quantile Regression

For readers seeking a short introduction, Koenker and Hallock (2001) provide an accessible overview of these models and their application in economics while Buchinsky (1998) goes into more depth, especially regarding estimation issues. Davino et al. (2014) is a very recent, book-length treatment of conditional quantile regression and is probably the single-best source for anyone interested in these models.

Unconditional Quantile Regression

Firpo et al. (2009) describe the RIF-OLS approach in their seminal paper on unconditional quantile regression, and is required reading for anyone using these models. Their supplement provides proofs for their main paper, and it is probably not very useful for most researchers.

Firpo (2007) has proposed another estimator for unconditional quantile treatment effects under exogeneity. This estimator has been implemented in the `ivqte` command, but does not seem to be widely used.

For endogenous regressors, two approaches have been proposed. Fröhlich and Melly (2013) have developed an IV approach to quantile treatment effects, and have implemented their estimator in the Stata command `ivqte` (Fröhlich & Melly, 2010). This command is somewhat complicated, in that it will produce four different estimators, depending on the syntax used. Their paper in the *Stata Journal* requires close reading to correctly use this command.

Kernel Density Estimators

Kernel density estimators play an important role not only in quantile regression estimators, but also in the visual display of data, as well as propensity score matching. Cox (2007) provides an accessible introduction to these estimators, as does Salgado-Ugarte, Shimizu, and Taniuchi (1993).

Appendix

Below is the Stata syntax used to generate the results in this chapter.

```
global figures directory

*** Example in Table 8.1 ***
use http://www.ats.ucla.edu/stat/stata/notes/hsb2, clear
sum write, detail
sum write if female==0, detail
sum write if female==1, detail
reg write female
qreg write female
qreg write female, quantile(.25)
replace write=1000 if id==192
reg write female
qreg write female

*** Graphing densities for Fig. 8.6, panel (a) ***
kdensity write, bwidth(1) kernel(gau) legend(off)
  graphregion(color(white) lwidth(large)) xtitle
  ("Writing score") title("")
graph export $figures\kernel1gau.eps, replace
!epstopdf $figures\kernel1gau.eps

*** Input faculty salary data ***
use nsopfdata.dta, clear
* Define faculty group for analysis
keep if q1==1 & q2==1 & q3==1 & q5==1 // only instr.
 duties, faculty status, full-time
keep if q4==1 | q4==2 // principal activity is teaching
 or research
keep if q10==1 | q10==2 | q10==3 // rank of prof, assoc
 or asst
* Code independent variables
recode q17a1 (1=1) (0 2/7=0), gen(phd)
```

```
recode q71 (2=1) (1=0), gen(female)
gen age=2003-q72
recode q10 (1=1) (2 3=0) (0 4 5 6=.), gen(full)
recode q10 (2=1) (1 3=0) (0 4 5 6=.), gen(assoc)
rename q74b asian
rename q74c black
gen native=0
replace native=1 if q74a==1 | q74d==1
rename q73 latino
rename q52ba articles
rename q52bd books
rename q16cd2 disc
xi i.disc // discipline dummy vars
* Dependent variable
rename q66a basesalary
drop if basesalary<20000 // seems odd to be FT prof and
 making less than 20K
* Create analytic sample
reg basesalary female asian black latino native full
 assoc articles books _Idisc_2-_Idisc_32
keep if e(sample)

*** OLS-RIF results for Table 8.5 ***
reg basesalary female asian black latino native full
 assoc articles books _Idisc_2-_Idisc_32
estimate store ols
foreach i in 10 25 50 75 90
rifreg basesalary female asian black latino native full
 assoc articles books _Idisc_2-_Idisc_32,
 quantile(.`i')
estimates store q`i'
estimates table ols q10 q25 q50 q75 q90,
  drop(_Idisc_2-_Idisc_32) b(%9.0f) se se(%9.0f)

*** bootstrapping SEs for Table 8.6 ***
foreach i in 10 25 50 75 90
bootstrap, reps(100) seed(642014): rifreg basesalary
 female asian black latino native full assoc articles
 books _Idisc_2-_Idisc_32, quantile(.`i')
estimates store q`i'
estimates table q10 q25 q50 q75 q90,
  drop(_Idisc_2-_Idisc_32) b(%9.0f) se se(%9.0f)

*** Testing sensitivity of results in Table 8.7 ***
* Gaussian
```

```
foreach i in 10 25 50 75 90
rifreg basesalary female asian black latino native full
 assoc articles books _Idisc_2-_Idisc_32, quantile(.`i')
estimates store silverq`i'
foreach i in 10 25 50 75 90
rifreg basesalary female asian black latino native full
 assoc articles books _Idisc_2-_Idisc_32, quantile(.`i')
 width(5192)
estimates store hardleq`i'
foreach i in 10 25 50 75 90
rifreg basesalary female asian black latino native full
 assoc articles books _Idisc_2-_Idisc_32, quantile(.`i')
 width(3802)
estimates store scottq`i'
estimates table silverq10 silverq25 silverq50 silverq75
 silverq90, drop(_Idisc_2-_Idisc_32) b(%9.0f)
 se se(%9.0f)
estimates table hardleq10 hardleq25 hardleq50 hardleq75
 hardleq90, drop(_Idisc_2-_Idisc_32) b(%9.0f)
 se se(%9.0f)
estimates table scottq10 scottq25 scottq50
  scottq75 scottq90, drop(_Idisc_2-_Idisc_32) b(%9.0f)
  se se(%9.0f)
* to see results with Epanechnikov and uniform
  distributions, just add kernop(ep) or kernop(rec) as
  options

*** Conditional QR results for Table 8.8 ***
foreach i in 10 25 50 75 90
qreg basesalary female asian black latino native full
 assoc articles books _Idisc_2-_Idisc_32, quantile(.`i')
estimates store q`i'
estimates table q10 q25 q50 q75 q90,
  drop(_Idisc_2-_Idisc_32) b(%9.0f) se se(%9.0f)

*** Graph unconditional QR results for gender (Fig. 8.7)
 ***
* This set of code can be used to create the other
 figures in the chapter
matrix quantiles = J(1,3,.) // create blank matrix to
 add model results to
matrix colnames quantiles = B SE Q
matrix identity=J(1,1,1) // to add to counter matrix
 per loop
matrix counter=J(1,1,0) // will save quatiles for
```

```
  qraphing
forvalues i=.01(.01)1
matrix counter=counter+identity
qui:rifreg basesalary female asian black latino native
 full assoc articles books _Idisc_2-_Idisc_32,
 quantile(‘i’)
matrix table=r(table) // create a matrix of results for
 each rd (have to rename matrix)
matrix b_se=table[1..2,1..1]’ // grab B and SE and
  transpose so they are in column format rather than row
matrix temp=b_se,counter // add quantile as a column
matrix quantiles=quantiles\temp //add most recent set
 of model results to matrix
matrix quantiles2=quantiles[2..100,1..3] // drop missing
 first row
clear svmat quantiles2, names(col) // converts matrix
 of results to dataset for graphing
gen ciplus=B+1.96*SE
gen cineg=B-1.96*SE
graph twoway connected B Q, msymbol(none) legend(off)
 graphregion(color(white)) yline(-5540, lpattern
 (longdash)) lwidth(medthick) xtitle("Quantiles of
 salary") ytitle(Male-female differential ($)) ||
 connected ciplus Q, msymbol(none) lpattern(dash) ||
 connected cineg Q, msymbol(none) lpattern(dash)
graph export $figures\gender.eps, replace
!epstopdf "$figures\gender.eps

*** Finding optimal bandwidths for ivqte command (Table
 8.11) ***
locreg facultyunion, logit bandwidth(.2 1 .) lambda(.2
 .5 .8) continuous(citi6008) dummy(gov_cons)
locreg facultyunion, logit bandwidth(.05 .1 .15 .2 .25)
 lambda(.05 .1 .15 .2 .25) continuous(citi6008)
 dummy(gov_cons)
locreg facultyunion, logit bandwidth(.06 .08 .1 .12)
 lambda(.01 .02 .03 .04 .05) continuous(citi6008)
 dummy(gov_cons)
locreg facultyunion, logit bandwidth(.1) lambda(0 .0025
 .005 .0075 .01) continuous(citi6008) dummy(gov_cons)

*** IV QR estimates for Table 8.12 **
foreach i in 10 25 50 75 90
ivqte basesalary (facultyunion = statelaws) , variance
 quantiles(.‘i’) continuous(citi6008) dummy(gov_cons)
```

```
foreach i in 10 25 50 75 90
ivqte basesalary (facultyunion = statelaws) , variance
 quantiles(.‘i’) continuous(citi6008) dummy(gov_cons)
 bandwidth(.1) lambda(0)
```

References

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response variables. *Journal of Econometrics, 113*, 231–263.

Abadie, A., Angrist, J., & Imbens, G. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica, 70*(1), 91–117.

Bielby, R. M., House, E., Flaster, A., & DesJardins, S. L. (2013). Instrumental variables: Conceptual issues and an application considering high school coursetaking. In M. B. Paulsen (Ed.), *Higher education: Handbook of theory and Research*. New York: Springer.

Buchinsky, M. (1998). Recent advances in quantile regression models: A practical guideline for empirical research. *Journal of Human Resources, 33*(1), 88–126.

Budig, M. J., & Hodges, M. J. (2010). Differences in disadvantage: Variation in the motherhood penalty across white women's earnings distribution. *American Sociological Review, 75*, 705–728.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications*. New York: Cambridge University Press.

Castellano, K. E., & Ho, A. D. (2013a). Contrasting OLS and quantile regression approaches to student “growth” percentiles. *Journal of Educational and Behavioral Statistics, 38*(2), 190–215.

Castellano, K. E., & Ho, A. D. (2013b). *A practioners guide to growth models*. Technical report, Council of Chief State School Officers.

Chen, C. L. (2005). An introduction to quantile regression and the quantreg procedure. In *SUGI 30 proceedings*, Philadelphia.

Cox, N. J. (2007). Kernel estimation as a basic tool for geomorphological data analysis. *Earth Surface Processes and Landforms, 32*, 1902–1912.

Davino, C., Furno, M., & Vistocco, D. (2014). *Quantile regression: Theory and applications*. Chichester, UK: Wiley.

Fiorio, C. V. (2004). Confidence intervals for kernel density estimation. *The Stata Journal, 4*(2), 168–179.

Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica, 75*(1), 259–276.

Firpo, S., Fortin, N. M., & Lemieux, T. (2009). Unconditional quantile regressions. *Econometrica, 77*(3), 953–973.

Frandsen, B. R., Fröhlich, M., & Melley, B. (2012). Quantile treatment effects in the regression discontinuity design. *Journal of Econometrics, 168*, 382–395.

Fröhlich, M., & Melly, B. (2010). Estimation of quantile treatment effects with Stata. *The Stata Journal, 10*(3), 423–457.

Fröhlich, M., & Melly, B. (2013). Unconditional quantile treatment effects under endogeneity. *Journal of Business & Economic Statistics, 31*(3), 346–357.

Härdle, W. (1991). *Smoothing techniques*. New York: Springer.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis, 15*(3), 199–236.

Killewald, A., & Bearak, J. (2014). Is the motherhood penalty larger for low-wage women? A comment on quantile regression. *American Sociological Review, 79*(2), 350–357.

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

Maclean, J. C., Webber, D. A., & Marti, J. (2014). An application of unconditional quantile regression to cigarette taxes. *Journal of Policy Analysis and Management, 33*(1), 188–210.

Mueller, S. (2013). Teacher experience and the class size effect – Experimental evidence. *Journal of Public Economics, 98*, 44–52.

Murnane, R. J., & Willett, J. B. (2011). *Methods matter: Improving causal inference in educational and social science*. New York: Oxford University Press.

Porter, S. R. (2013). The causal effect of faculty union on institutional decision-making. *Industrial & Labor Relations Review, 66*(5), 1192–1211.

Porter, S. R. (2014). *Understanding the modern approach to instrumental variables*. North Carolina State University. Raleigh: NC.

Salgado-Ugarte, I. H., Shimizu, M., & Taniuchi, T. (1993). Exploring the shape of univariate data using kernel density estimators. *Stata Technical Bulletin, 16*, 8–19.

Salgado-Ugarte, I. H., Shimizu, M., & Taniuchi, T. (1995). Practical rules for bandwidth selection in univariate density estimation. *Stata Technical Bulletin, 27*, 5–19.

Scott, D. (1992). *Multivariate density estimation*. New York: Wiley.

Silverman, B. W. (1992). *Density estimation for statistics and data analysis*. London: Chapman & Hall.

StataCorp LP (2013). *Stata base reference manual release 13*. College Station, TX: Stata Press.

Webber, D. A., & Ehrenberg, R. G. (2010). Do expenditures other than instructional expenditures affect graduation and persistence rates in American higher education? *Economics of Education Review, 29*, 947–958.