2

Quantile Regression

This primer focuses on quantile regression. We assume you have read the section on quantile regression in Chapter 11, but we repeat parts of it here to set context. We begin with a brief review of the logic of quantile regression. We then discuss coefficient interpretation and indices of model fit. We then contrast quantile analysis with another common form of analysis that is based on breakpoints in a distribution. Finally, we consider the practice of data jittering to make quantile regression more widely applicable.

As noted in the main text, traditional multiple regression is fundamentally an analysis of conditional means. Consider a simple bivariate case where the outcome is the amount of time per week that an adolescent spends with his or her mother (Y) and the predictor is the age of the adolescent (X), which ranges from 12 to 17. Each age is a different predictor "profile" or population "segment." Consider the following data where we calculate the mean of the outcome for each segment:

Age	Mean Hours Spent Together
12	32
13	30
14	28
15	26
16	24
17	22

For any given profile or segment, we can characterize the mean of the outcome for that profile. This is called a *conditional mean*. The mean amount of time spent with one's mother given adolescents are 12 years old is 32 hours. The mean amount of time spent with one's mother given adolescents are 13 years old is 30 hours. And so on.

If we have multiple predictors, then a "profile" refers to a specific combination of scores across the predictors. If we predict time spent with one's mother from age and gender, then one predictor profile is "12 year old males," another predictor profile is "12 year old females," another predictor profile is "13 year old males," and so on.

In addition to describing the mean of Y at any given predictor profile, multiple

regression also describes how the outcome means change as we move across values of X from one predictor profile to another predictor profile. For example, how does the mean amount of time spent with one's mother change as an adolescent's age increases or decreases? From the data above, we can see that for every one unit that age increases, the mean amount of time spent together decreases by 2 hours. Thus, the unstandardized regression coefficient is -2.0. It provides perspectives on mean changes across X values.

Quantile regression is an extension of multiple regression. Instead of estimating Y means for different predictor profiles and how the means vary across profiles, it seeks instead to characterize the values of different quantiles of Y for predictor profiles and how these values vary across profiles.

A *quantile* is analogous to the concept of percentiles. Informally, a percentile is a score or point in a distribution that a specified percentage of scores are less than or equal to. If we tell you a GRE score of 161 defines the 80th percentile, this means that 80% of individuals score 161 or less on it. If we tell you a GRE score of 153 defines the 50th percentile, this means that 50% of individuals score 153 or less on it. Statisticians typically use the term quantile instead of percentile, they refer to it with the letter q, and they state it in probability or proportion terms rather than as percents. For example, the 0.80 quantile (q = 0.80) for the GRE is a score of 161.

One of the most well-known quantiles is q = 0.50, which is the median. When we use quantile regression to analyze q = 0.50 on the outcome variable, we are applying the spirit of traditional regression analysis, but to medians instead of means. Medians are more outlier resistant than means.

Interestingly, we also can analyze other quantiles, so we can characterize predictors of what is happening at the lower end of the outcome distribution or the upper end of the outcome distribution. This is important because sometimes there will not be group/profile differences in means or medians in the middle of an outcome distribution, but there will be group/profile differences in the lower or upper parts of the distribution. Quantile regression allows us to explore such possibilities.

As an example, suppose we want to compare males and females on their reported depression levels as reported on the classic CES-D scale. When the analysis is conducted using q=0.50 (the median), we might find a statistically significant difference, such that females have a higher median than males (12.0 versus 9.0). When we conduct the same contrast at q = 0.20, we might find the quantile values do not differ (males = 5.0, females = 5.0). Finally, we might find at a quantile of 0.80, the gender difference is even more exaggerated than what it was at the median (males = 19.0, females = 25.0). Stated another way, the regression coefficient for the dummy variable of gender is 0.0 at q = 0.20, 3.0 at q = 0.50, and 6.0 at q = 0.80. Koenker (2005) describes methods that can be used for

significance testing of differences in coefficients for different quantiles. For example, in the depression analysis, is the coefficient for gender of 6.0 when q = 0.80 significantly different from the coefficient of 3.0 when q = 0.50?

As another example, gender gaps in math achievement tend to be more pronounced at the upper end of the achievement dimension as contrasted with the middle of the distribution, and they are almost non-existent at the lower end of the distribution (Reeves and Lowe, 2009). One explanation of this result is that females often choose not to take higher level math courses, thereby promoting a larger gender difference at the higher levels of math achievement. Quantile regression has identified health disparities between groups that went undetected when the focus was on group differences in means (e.g., Gebregziabher et al., 2011; Juarez, Tan, Davis, et al., 2014). For introductory discussions of quantile regression, see Cade and Noon (2003) and Hao and Naiman (2007).

Quantile regression analysis has several useful properties. One of these is that it does not make some of the strong assumptions about error terms that traditional regression makes (e.g., it does not make the assumption of normally distributed errors). Quantiles also are outlier resistant (but see below for qualifications).

Just as we do not literally calculate the mean of the outcome for each predictor profile in OLS, the same is true in quantile regression – the target quantile for each predictor profile is not literally computed. In traditional OLS, we make the assumption that means are a linear function of the predictor(s) and then estimate parameters by making reference to a regression surface. Quantile regression uses analogous methods, but applied to quantiles. There are different algorithms for coefficient estimation in quantile regression, with the most popular one based on the simplex methods proposed by Barrodale and Roberts (1974). For a description of this approach, see Koenker (2005). Another approach to coefficient estimation is based on interior point methods (Portnoy & Koenker, 1997), with a third option being Frisch-Newton methods based on sparse linear algebra (Koenker, 2005). There also are multiple strategies for estimating standard errors and confidence intervals. These are discussed in Koenker (2005) and we do not delve into them here. A popular method for estimating confidence intervals uses a rank inversion method. Another method uses an asymptotic covariance approach based on Huber-like sandwich estimators. Powell (1990) suggests a kernel estimation approach while others prefer bootstrapping. See Koenker (2005) for details.

OUTLIERS AND QUANTILE REGRESSION

In traditional regression models, distinctions are sometimes made between outliers and leverage points. Outliers are extreme observations on the outcome variable that either distort the basic trend in the data or adversely affect standard errors and statistical power.

Distorting cases also can occur in the predictor space independent of the outcome, i.e., they represent highly unusual predictor profiles. These are often called leverage points instead of outliers (because they "leverage" the regression plane). Quantile regression is generally robust to outliers, but not necessarily to large leverage points. For this reason, it is not uncommon for researchers to conduct leverage analyses of the predictor space and to pursue corrective methods for observations with high leverages. For a discussion of robust methods for accomplishing this task, see Wilcox (2017).

In general, larger sample sizes make it harder for any one case to affect regression results. One (imperfect) way to reduce the impact of leveraged cases is to use large N.

INTERPRETATION OF COEFFICIENTS

The interpretation of coefficients in quantile regression follows the same basic logic as traditional multiple regression except instead of means, the parameter of interest is the targeted quantile. Dummy variables, product terms, and polynomials all are interpreted in ways directly analogous to multiple regression. If we perform a quantile analysis using q = 0.50, and our predictor X is a dummy variable for gender scored 0 = females and 1 = males, then the regression coefficient for X is the outcome median difference between males and females, holding constant all other predictors in the equation. If a predictor, Z, is the number of years of education, then the regression coefficient for Z indicates how much the predicted median of the outcome changes for every one unit increase in the number of years of education, holding constant the other predictors.

OVERALL MODEL FIT

Traditional multiple regression generates a squared multiple correlation to provide a sense of the degree of predictive accuracy of the regression model. There is no such statistic in quantile regression. Some analysts calculate an analog to it, called a *pseudo R squared*, that uses the statistical concept of log-likelihoods. A log-likelihood, stated informally, is an index of fit that posits a population model and then estimates the probability of the sample data occurring given that model. The log-likelihood is the natural log of this probability. The pseudo R squared calculates a log-likelihood for an intercept only model with no predictors and then calculates how much this log-likelihood improves if that model is compared to a model that includes the predictors of interest. For example, a pseudo R squared of 0.10 indicates that the log-likelihood of the model that does not include the predictors was improved by 10% when the predictors were added to the intercept only model. The use of pseudo R squares is somewhat controversial. For a discussion of pseudo R squares and log likelihoods, see Long (1997). Log likelihoods

also are discussed in the primer on mixture regression.

BREAKPOINTS VERSUS QUANTILES

Quantile analysis works with cumulative frequency distributions of the outcome variable. You likely have encountered cumulative frequency distributions in your use of popular statistical packages. For example, consider in Figure 2.1 the output from an SPSS frequency analysis of adult marijuana users who reported their age at first use. The column on the extreme right provides the cumulative percents and allows us to state the percent of individuals who first tried marijuana at a given age or lower. For example, 11.2% of the sample tried marijuana at age 13 or younger. These percentages, of course, can be reframed as probabilities by dividing them by 100, yielding what is known as a cumulative probability distribution.

agemarijuana age of first marijuana use

		Frequency	Percent	Cumulative Percent
Valid	8	10	.1	.2
	9	13	.1	.5
	10	20	.2	.9
	11	30	.3	1.6
	12	160	1.8	5.0
	13	285	3.1	11.2
	14	445	4.9	20.9
	15	707	7.7	36.2
	16	882	9.7	55.3
	17	571	6.2	67.7
	18	638	7.0	81.5
	19	236	2.6	86.6
	20	192	2.1	90.8
	21	145	1.6	93.9
	22	95	1.0	96.0
	23	60	.7	97.3
	24	40	.4	98.2
	25	38	.4	99.0
	26	17	.2	99.3
	27	18	.2	99.7
	28	5	.1	99.8
	29	4	.0	99.9
	30	3	.0	100.0
	Total	4614	50.5	

FIGURE 2.1. Age at First Marijuana Use

One approach to comparing groups using cumulative probability distribution concepts is called *breakpoint analysis*. Consider an example where we want to compare males and females on the CES-D depression scale (hence, depression is our outcome and gender is our predictor). In Figure 2.2, we plot a cumulative probability distribution for depression scores for males and females, separately. The CES-D scale ranges from 0 to 60, with higher scores indicating greater levels of depressive symptoms. In the general population, most people score near 0. A score of 16 on the scale is said to be "clinically meaningful" and is a cutpoint (or breakpoint) that is a flag for potential depression associated problems.

In breakpoint analysis, one *a priori* identifies a score on the outcome that is of theoretical interest (e.g., a score of 16 on the CES-D) and uses the cumulative distribution to compare the proportion of cases above and below that score for the two groups. In Figure 2.3, we can draw a (solid) vertical line upward from the breakpoint on the X axis and then examine the proportion on the vertical axis where that line intersects the cumulative distribution for each group (see the arrows with dashed lines) As seen in Figure 2.3, about twenty percent of males have a CES-D score above 16, with 80% being at or below it. By contrast, 32% of females have a CES-D score above 16, with 68% being at or below it. Females thus are more likely to be above the breakpoint than males. Since 16 is an important "marker" for troublesome depression, this result is of import.

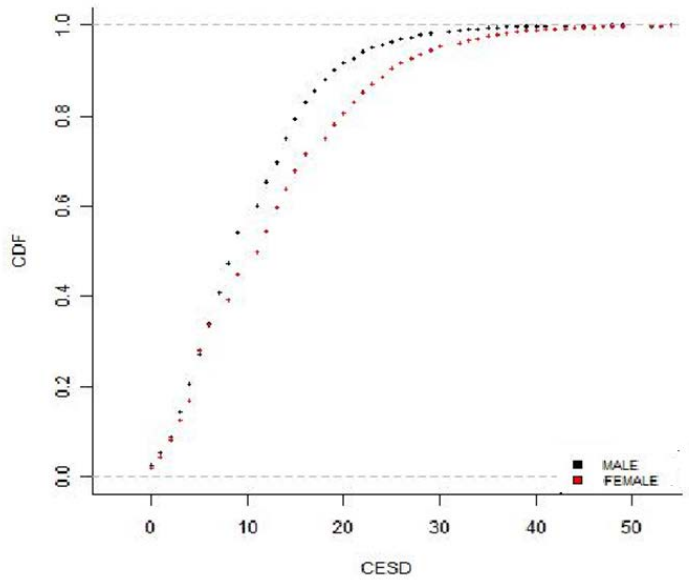

FIGURE 2.2. Cumulative CES-D Distribution for Males and Females

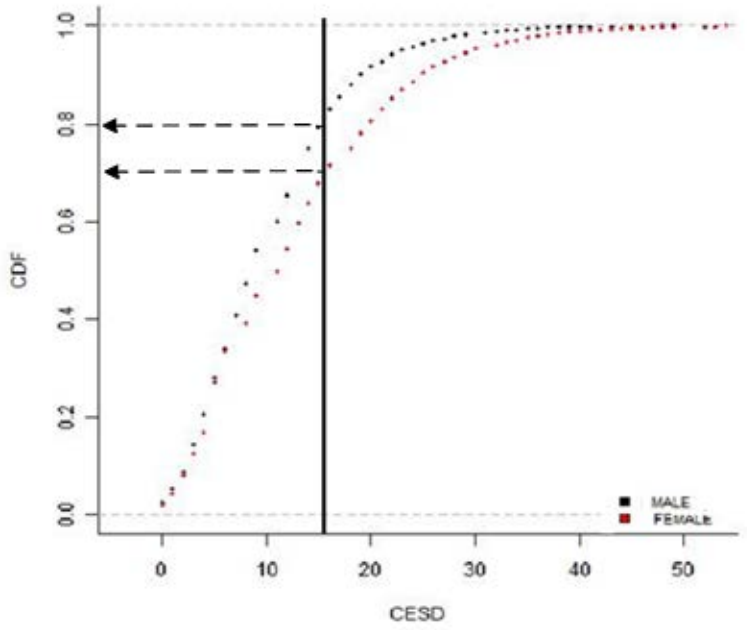

FIGURE 2.3. Breakpoint Analysis

Breakpoints can be subjected to multivariate analysis using binary regression modeling. For example, each person in the sample is given a dichotomous-based score for the CES-D based on the breakpoint value of 16, such that 0 = person is at or below the breakpoint, and 1 = person is above the breakpoint. Then, logistic/probit regression or a modified linear probability model is used to probe the relations between this dichotomous outcome and a set of predictors, such as gender, income, ethnicity, and so on. The choice of breakpoint values is theoretically or substantively driven and must be carefully justified. Dichotomizing continuous variables can be problematic, so one does so only in cases where there is theoretical or practical justification for it.

In quantile regression, the same cumulative distribution plot is of interest, but we examine it from a different perspective than breakpoint analysis. In quantile regression, we start by *a priori* specifying a quantile, say q=0.80, rather than a breakpoint score to focus on. We then use the cumulative distribution to identify the score that maps onto that quantile for each group. The dynamics are shown in Figure 2.4, where we draw a (solid) horizontal line across the plot at the quantile of interest. We then identify the scores on the horizontal CES-D dimension where the line intersects the cumulative distribution for each group (see dashed lines). If we are interested in q= 0.80, for example, the quantile value is 15 for males and 20 for females: Twenty percent of males have a CES-D score above 15, whereas 20% of females have a CES-D score above 20.

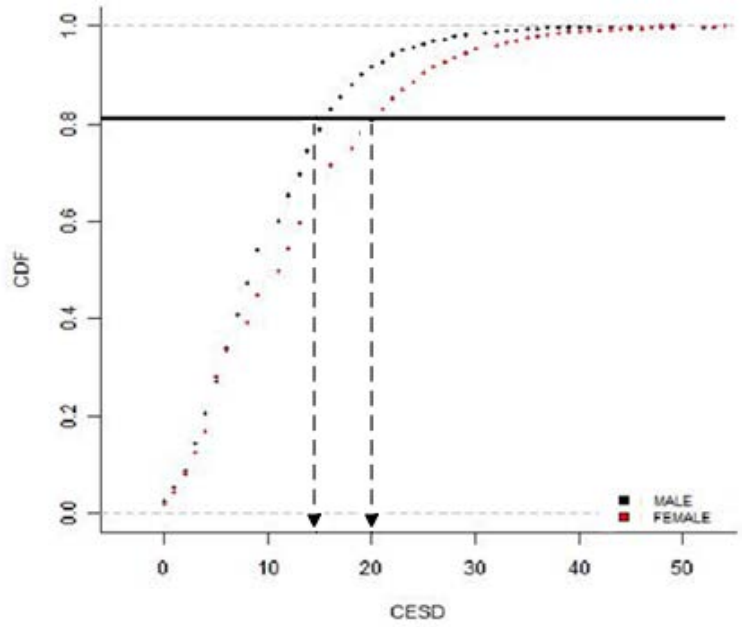

FIGURE 2.4. Quantile Analysis

The choice of the particular quantiles to focus on, ideally, is theoretically or substantively driven, but the choice also can be informed by preliminary examination of the distances between curves for the groups on their respective cumulative distributions (ASA offers program for plotting cumulative distributions for two groups on the same graph). What matters most from the perspective of quantile analysis is the degree of horizontal separation between the curves (from left to right).

Which analysis should you conduct, breakpoint-based binary regression analyses or quantile regression analyses? It depends on what questions you are trying to answer. If you have identified a specific score on the outcome that is of theoretical or practical significance and you want to identify factors that push individuals above or below that score, then binary regression is appropriate. You do not really care how far below or how far above the breakpoint that people are; you just want to understand factors that make people above or below it. Our experience is that in practice, such breakpoints are rare and one must be careful about invoking them. For example, with the benchmark value of 16 for the CES-D, do we really not care how much above the score of 16 people are? Do we really want to treat someone with a score of 17 as being the same as someone who has a score of 50 (both are above the breakpoint, hence both get a score of 1 on the dichotomously defined outcome)? And do we really want to treat someone with a score of 0 as being the same as someone with a score of 15 (both are below the breakpoint)? We can think of some cases where this might be useful, but generally speaking, it is a

rather crude approach to building a science of depression. If instead your interest is in comparing scores of groups of individuals at different portions of the outcome dimension (at the low end of the outcome dimension, in the middle of the outcome dimension, and at the high end of the outcome dimension), then quantile regression is an appropriate tool for doing so.

JITTERING

Sometimes we treat variables as continuous, but we only have coarse, discrete measures of them. Quantile regression assumes a continuous outcome, but sometimes we operationalize such outcomes with, say, only a small number (e.g., 5 to 7) of measurement categories. This can produce degenerate solutions in quantile regression as a result. There are different ways of handling such scenarios. A somewhat crude but often workable solution is to smooth the outcome variable by adding some "jitter" to it (Machado and Santos Silva, 2005). Jittering adds a very small amount of random perturbance to each score – not enough to affect substantive results but enough to allow the statistical algorithms to estimate the parameters of interest.

The most common way of jittering integer measures (such as a 1 to 7 Likert-scale) is to generate random numbers from a uniform distribution whose span is defined by the lower and upper real limit of each number. For example, let V be the span around the number that you wish to cover. For the score of 6, the real limits are 5.5 to 6.5, so the span is 1. For the score of 1, the real limits are 0.5 to 1.5, so again the span is 1. Thus, V = 1 for each score on the variable. Random noise is then added to a given score using a uniform distribution between –V/2 and +V/2. This is called *symmetric jittering*. The span is called the *dequantization value*. It assumes that people with a score of 5, for example, actually scored somewhere between 4.5 and 5.5 on the underlying continuous dimension and that these individuals are uniformly distributed across this span. For some scenarios, we might use instead an algorithm called *to-the-right jittering*, such as for cases involving counts (see Machado & Santos Silva, 2005). In this case, random noise is added to a score based on a uniform distribution from 0 to V (where V is commonly set to 1) instead of –V/2 to +V/2. When using jittering, we usually specify a random seed in the computer program so results can be replicated by repeating the seed. Random seeds are the starting point that algorithms for generating random numbers use.

Jittering must be used with caution. We must admit, it feels a bit perverse to purposely add random noise to data. But the underlying assumptions of jittering often are not unreasonable and if its use ultimately allows us to apply more powerful methods of analysis, then this is desirable. See Machado and Santos-Silva (2005) for an example that applies jittering to quantile regression for counts.

CONCLUDING COMMENTS

Traditional ordinary least squares (OLS) regression is the analytic stalwart of the social and health sciences. It can be thought of as a method that describes how the central tendency of an outcome as represented by the mean varies across different predictor profiles. However, we might be interested in using alternative measures of central tendency for the outcome (e.g., the median) or we might want to focus our analysis on the lower or upper end of the outcome distribution, not just the middle. Quantile regression is a tool for doing so.

Quantile regression does not make assumptions about normality and it is outlier resistant, which are desirable properties. However, it can yield distorted characterizations of data trends due to unusual leverage points in the predictor space (although this is less likely for large N). Care must be taken in applying quantile regression, accordingly. Quantile regression also requires outcome measures not be too coarse, but such coarseness often can be addressed through judicious use of jittering.

We make it a common practice to explore data using quantile regression and to gain an appreciation for what is going on at the lower, middle and upper end of outcome distributions.

REFERENCES

Cade, B. and Noon, B. (2003). A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment*, 1, 412-420.

Gebregziabher, M., Lynch, C., Mueller, M. et al. (2011). Using quantile regression to investigate racial disparities in medication non-adherence. *BMC Medical Research Methodology*, 11, 88-95.

Hao, L. & Naiman, D. (2007). *Quantile regression*. Newbury Park: Sage.

Juarez, D, Tan, C., Davis, J. et al. (2014). Using quantile regression to assess disparities in medication non-adherence. *American Journal of Health*, 38, 53-62

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

Koenker, R. & Portnoy, S. (1987). L-Estimation for linear models. *Journal of the American Statistical Association*, 82, 851-857.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks: Sage.

Reeves, E. B. & Lowe, J (2009). Quantile regression: An education policy research tool. *South Rural Sociology*, 24, 175-199.

Wilcox, R. (2017). *Introduction to robust estimation and hypothesis testing*. New York: Academic Press (fourth edition).