QUANTILE REGRESSION

Lingxin Hao
The Johns Hopkins University

Daniel Q. Naiman
The Johns Hopkins University

1. INTRODUCTION

The purpose of regression analysis is to expose the relationship between a response variable and predictor variables. In real applications, the response variable cannot be predicted exactly from the predictor variables. Instead, the response for a fixed value of each predictor variable is a random variable. For this reason, we often summarize the behavior of the response for fixed values of the predictors using measures of central tendency. Typical measures of central tendency are the average value (mean), the middle value (median), or the most likely value (mode).

Traditional regression analysis is focused on the mean; that is, we summarize the relationship between the response variable and predictor variables by describing the mean of the response for each fixed value of the predictors, using a function we refer to as the *conditional mean* of the response. The idea of modeling and fitting the *conditional-mean function* is at the core of a broad family of regression-modeling approaches, including the familiar simple linear-regression model, multiple regression, models with heteroscedastic errors using weighted least squares, and nonlinear-regression models.

Conditional-mean models have certain attractive properties. Under ideal conditions, they are capable of providing a complete and parsimonious description of the relationship between the covariates and the response distribution. In addition, using conditional-mean models leads to estimators (least squares and maximum likelihood) that possess attractive statistical properties, are easy to calculate, and are straightforward to interpret. Such models have been generalized in various ways to allow for heteroscedastic errors so that given the predictors, modeling of the conditional mean and conditional scale of the response can be carried out simultaneously.

Conditional-mean modeling has been applied widely in the social sciences, particularly in the past half century, and regression modeling of the relationship between a continuous response and covariates via least squares and its generalization is now seen as an essential tool. More recently, models for binary response data, such as logistic and probit models and Poisson regression models for count data, have become increasingly popular in social-science research. These approaches fit naturally within the conditional-mean modeling framework. While quantitative social-science researchers have applied advanced methods to relax some basic modeling assumptions under the conditional-mean framework, this framework itself is seldom questioned.

The conditional-mean framework has inherent limitations. First, when summarizing the response for fixed values of predictor variables, the conditional-mean model cannot be readily extended to noncentral locations, which is precisely where the interests of social-science research often reside. For instance, studies of economic inequality and mobility have intrinsic interest in the poor (lower tail) and the rich (upper tail). Educational researchers seek to understand and reduce group gaps at preestablished achievement levels (e.g., the three criterion-referenced levels: basic, proficient, and advanced). Thus, the focus on the central location has long distracted researchers from using appropriate and relevant techniques to address research questions regarding noncentral locations on the response distribution. Using conditional-mean models to address these questions may be inefficient or even miss the point of the research altogether.

Second, the model assumptions are not always met in the real world. In particular, the homoscedasticity assumption frequently fails, and focusing exclusively on central tendencies can fail to capture informative trends in the response distribution. Also, heavy-tailed distributions commonly occur in social phenomena, leading to a preponderance of outliers. The conditional mean can then become an inappropriate and misleading measure of central location because it is heavily influenced by outliers.

Third, the focal point of central location has long steered researchers' attention away from the properties of the whole distribution. It is quite natural to go beyond location and scale effects of predictor variables on the response and ask how changes in the predictor variables affect the underlying *shape* of the distribution of the response. For example, much social-science research focuses on social stratification and inequality, areas that require close examination of the properties of a distribution. The central location, the scale, the skewness, and other higher-order properties—not central location alone—characterize a distribution. Thus, conditional-mean models are inherently ill equipped to characterize the relationship between a response distribution and predictor variables. Examples of inequality topics include economic inequality in wages, income, and wealth; educational inequality in academic achievement; health inequality in height, weight, incidence of disease, drug addiction, treatment, and life expectancy; and inequality in well-being induced by social policies. These topics have often been studied under the conditional-mean framework, while other, more relevant distributional properties have been ignored.

An alternative to conditional-mean modeling has roots that can be traced to the mid-18th century. This approach can be referred to as conditional-median modeling, or simply median regression. It addresses some of the issues mentioned above regarding the choice of a measure of central tendency. The method replaces least-squares estimation with least-absolute-distance estimation. While the least-squares method is simple to implement without high-powered computing capabilities, least-absolute-distance estimation demands significantly greater computing power. It was not until the late 1970s, when computing technology was combined with algorithmic developments such as linear programming, that median-regression modeling via least-absolute-distance estimation became practical.

The median-regression model can be used to achieve the same goal as conditional-mean-regression modeling: to represent the relationship between the central location of the response and a set of covariates. However, when the distribution is highly skewed, the mean can be challenging to interpret while the median remains highly informative. As a consequence, conditional-median modeling has the potential to be more useful.

The median is a special *quantile,* one that describes the central location of a distribution. Conditional-median regression is a special case of quantile regression in which the conditional .5th quantile is modeled as a function of covariates. More generally, other quantiles can be used to describe noncentral positions of a distribution. The *quantile* notion generalizes specific terms like *quartile, quintile, decile,* and *percentile.* The *p*th quantile denotes that value of the response below which the proportion of the population is *p.* Thus, quantiles can specify any position of a distribution. For example, 2.5% of the population lies below the .025th quantile.

Koenker and Bassett (1978) introduced *quantile regression,* which models conditional quantiles as functions of predictors. The quantile-regression model is a natural extension of the linear-regression model. While the linear-regression model specifies the change in the conditional mean of the dependent variable associated with a change in the covariates, the quantile-regression model specifies changes in the conditional quantile. Since any quantile can be used, it is possible to model any predetermined position of the distribution. Thus, researchers can choose positions that are tailored to their specific inquiries. Poverty studies concern the low-income population; for example, the bottom 11.3% of the population lived in poverty in 2000 (U.S. Census Bureau, 2001). Tax-policy studies concern the rich, for example, the top 4% of the population (Shapiro & Friedman, 2001). Conditional-quantile models offer the flexibility to focus on these population segments whereas conditional-mean models do not.

Since multiple quantiles can be modeled, it is possible to achieve a more complete understanding of how the response distribution is affected by predictors, including information about shape change. A set of equally spaced conditional quantiles (e.g., every 5% or 1% of the population) can characterize the shape of the conditional distribution in addition to its central location. The ability to model shape change provides a significant methodological leap forward in social research on inequality. Traditionally, inequality studies are non-model based; approaches include the Lorenz curve, the Gini coefficient, Theil's measure of entropy, the coefficient of variation, and the standard deviation of the log-transformed distribution. In another book for the Sage QASS series, we will develop conditional Lorenz and Gini coefficients, as well as other inequality measures based on quantile-regression models.

Quantile-regression models can be easily fit by minimizing a generalized measure of distance using algorithms based on linear programming. As a result, quantile regression is now a practical tool for researchers. Software packages familiar to social scientists offer readily accessed commands for fitting quantile-regression models.

A decade and a half after Koenker and Bassett first introduced quantile regression, empirical applications of quantile regression started to grow rapidly. Empirical researchers took advantage of quantile regression's ability to examine the impact of predictor variables on the response distribution. Two of the earliest empirical papers by economists (Buchinsky, 1994; Chamberlain, 1994) provided practical examples of how to apply quantile regression to the study of wages. Quantile regression allowed them to examine the entire conditional distribution of wages and determine if the returns to schooling, and experience and the effects of union membership differed across wage quantiles. The use of quantile regression to analyze wages increased and expanded to address additional topics such as changes in wage distribution (Machado & Mata, 2005; Melly, 2005), wage distributions within specific industries (Budd & McCall, 2001), wage gaps between whites and minorities (Chay & Honore, 1998) and between men and women (Fortin & Lemieux, 1998), educational attainment and wage inequality (Lemieux, 2006), and the intergenerational transfer of earnings (Eide & Showalter, 1999). The use of quantile regression also expanded to address the quality of schooling (Bedi & Edwards, 2002; Eide, Showalter, & Sims, 2002) and demographics' impact on infant birth weight (Abreveya, 2001). Quantile regression also spread to other fields, notably sociology (Hao, 2005, 2006a, 2006b), ecology and environmental sciences (Cade, Terrell, & Schroeder, 1999; Scharf, Juanes, & Sutherland, 1989), and medicine and public health (Austin et al., 2005; Wei et al., 2006).

This book aims to introduce the quantile-regression model to a broad audience of social scientists who are interested in modeling both the location and shape of the distribution they wish to study. It is also written for readers who are concerned about the sensitivity of linear-regression models to skewed distributions and outliers. The book builds on the basic literature of Koenker and his colleagues (e.g., Koenker, 1994; Koenker, 2005; Koenker & Bassett, 1978; Koenker & d'Orey, 1987; Koenker & Hallock, 2001; Koenker & Machado, 1999) and makes two further contributions. We develop conditional-quantile-based shape-shift measures based on quantile-regression estimates. These measures provide direct answers to research questions about a covariate's impact on the shape of the response distribution. In addition, inequality research often uses log transformation of right-skewed responses to create a better model fit even though "inequality" in this case refers to raw-scale distributions. Therefore, we develop methods to obtain a covariate's effect on the location and shape of conditional-quantile functions in absolute terms from log-scale coefficients.

Drawn from our own research experience, this book is oriented toward those involved with empirical research. We take a didactic approach, using language and procedures familiar to social scientists. These include clearly defined terms, simplified equations, illustrative graphs, tables and graphs based on empirical data, and computational codes using statistical software popular among social scientists. Throughout the book, we draw examples from our own research on household income distribution. In order to provide a gentle introduction to quantile regression, we use simplified model specifications wherein the conditional-quantile functions for the raw or log responses are linear and additive in the covariates. As in linear regression, the methodology we present is easily adapted to more complex model specifications, including, for example, interaction terms and polynomial or spline functions of covariates.

Quantile-regression modeling provides a natural complement to modeling approaches dealt with extensively in the QASS series: *Understanding Regression Assumptions* (Berry, 1993), *Understanding Regression Analysis* (Schroeder, 1986), and *Multiple Regression in Practice* (Berry & Feldman, 1985). Other books in the series can be used as references to some of the techniques discussed in this book, e.g., *Bootstrapping* (Mooney, 1993) and *Linear Programming* (Feiring, 1986).

The book is organized as follows. Chapter 2 defines quantiles and quantile functions in two ways—using the cumulative distribution function and solving a minimization problem. It also develops quantile-based measures of location and shape of a distribution in comparison with distributional moments (e.g., mean, standard deviation). Chapter 3 introduces the basics of the quantile-regression model (QRM) in comparison with the linearregression model, including the model setup, the estimator, and properties. The numerous quantile-regression equations with quantile-specific parameters are a unique feature of the quantile-regression model. We describe how quantile-regression fits are obtained by making use of the minimum distance principle. The QRM possesses properties such as monotonic equivariance and robustness to distributional assumptions, which produce flexible, nonsensitive estimates, properties that are absent in the linear-regression model. Chapter 4 discusses inferences for the quantile-regression model. In addition to introducing the asymptotic inference for quantile-regression coefficients, the chapter emphasizes the utility and feasibility of the bootstrap method. In addition, this chapter briefly discusses goodness of fit for quantile-regression models, analogous to that for linear-regression models. Chapter 5 develops various ways to interpret estimates from the quantile-regression model. Going beyond the traditional examination of the effect of a covariate on specific conditional quantiles, such as the median or off-central quantiles, Chapter 5 focuses on a distributional interpretation. It illustrates graphical interpretations of quantile-regression estimates and quantitative measures of shape changes from quantile-regression estimates, including location shifts, scale shifts, and skewness shifts. Chapter 6 considers issues related to monotonically transformed response variables. We develop two ways to obtain a covariate's effect on the location and shape of conditional-quantile functions in absolute terms from log-scale coefficients. Chapter 7 presents a systematic application of the techniques introduced and developed in the book. This chapter analyzes the sources of the persistent and widening income inequality in the United States between 1991 and 2001. Finally, the Appendix provides Stata codes for performing the analytic tasks described in Chapter 7.