CHAPTER 1

Introduction

1.1 MEANS AND ENDS

Much of applied statistics may be viewed as an elaboration of the linear regression model and associated estimation methods of least squares. In beginning to describe these techniques, Mosteller and Tukey (1977), in their influential text, remark:

> What the regression curve does is give a grand summary for the averages of the distributions corresponding to the set of xs. We could go further and compute several different regression curves corresponding to the various percentage points of the distributions and thus get a more complete picture of the set. Ordinarily this is not done, and so regression often gives a rather incomplete picture. Just as the mean gives an incomplete picture of a single distribution, so the regression curve gives a correspondingly incomplete picture for a set of distributions.

My objective in the following pages is to describe explicitly how to "go further." Quantile regression is intended to offer a comprehensive strategy for completing the regression picture.

Why does least-squares estimation of the linear regression model so pervade applied statistics? What makes it such a successful tool? Three possible answers suggest themselves. One should not discount the obvious fact that the computational tractability of linear estimators is extremely appealing. Surely this was the initial impetus for their success. Second, if observational noise is normally distributed (i.e., Gaussian), least-squares methods are known to enjoy a certain optimality. But, as it was for Gauss himself, this answer often appears to be an *ex post* rationalization designed to replace the first response. More compelling is the relatively recent observation that least-squares methods provide a general approach to estimating conditional mean functions.

And yet, as Mosteller and Tukey suggest, the mean is rarely a satisfactory end in itself, even for statistical analysis of a single sample. Measures of spread, skewness, kurtosis, boxplots, histograms, and more sophisticated density estimation are all frequently employed to gain further insight. Can something similar be done in regression? A natural starting place for this would be to supplement the conditional mean surfaces estimated by least squares with several estimated conditional quantile surfaces. In the chapters that follow, methods are described to accomplish this task. The basic ideas go back to the earliest work on regression by Boscovich in the mid-18th century to Edgeworth at the end of the 19th century.

1.2 THE FIRST REGRESSION: A HISTORICAL PRELUDE

It is ironic that the first faltering attempts to *do* regression are so closely tied to the notions of quantile regression. Indeed, as I have written on a previous occasion, the present enterprise might be viewed as an attempt to set statistics back 200 years, to the idyllic period before the discovery of least squares.

If least squares can be dated to 1805 by the publication of Legendre's work on the subject, then Boscovich's initial work on regression was half a century prior. The problem that interested Boscovich was the ellipticity of the earth. Newton and others had suggested that the earth's rotation could be expected to make it bulge at the equator with a corresponding flattening at the poles, making it an oblate spheroid, more like a grapefruit than a lemon. On the early history of regression and the contribution of Boscovich in particular, Stigler (1986) is the definitive introduction. Smith (1987) gives a detailed account of the development of geodesy, focusing attention on the efforts that culminated in the data appearing in Table 1.1.

To estimate the extent of this effect, the five measurements appearing in Table 1.1 had been made. Each represented a rather arduous direct measurement of the arc-length of 1° of latitude at five quite dispersed points – from Quito on the equator to a site in Lapland at 66° 19′N. It was clear from these measurements that arc length was increasing as one moved toward the pole from the equator, thus qualitatively confirming Newton's conjecture. But how the five measurements should be combined to produce one estimate of the earth's ellipticity was unclear.

For short arcs, the approximation

$$y = a + b \sin^2 \lambda, \tag{1.1}$$

Table 1.1. *Boscovich ellipticity data*

Location	Latitude	$\sin^2$ (Latitude)	Arc Length
Quito	0° 0′	0	56,751
Cape of Good Hope	33° 18′	0.2987	57,037
Rome	42° 59′	0.4648	56,979
Paris	49° 23′	0.5762	57,074
Lapland	66° 19′	0.8386	57,422

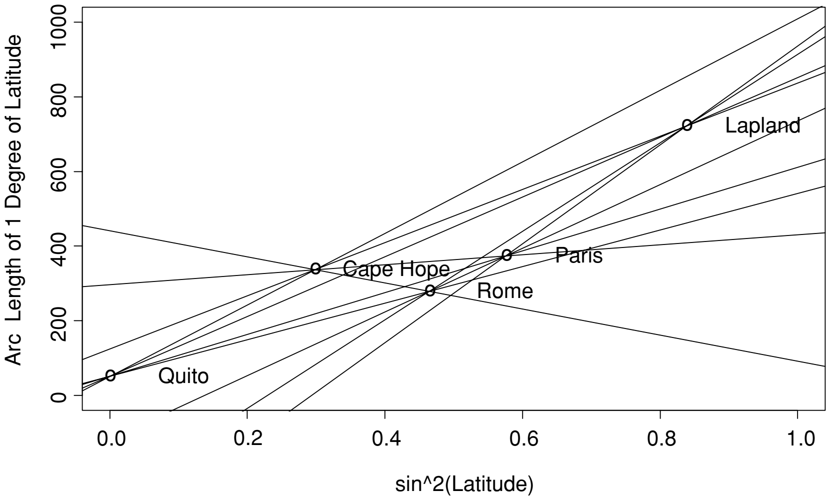

Figure 1.1. Boscovich ellipticity example. Boscovich computed all the pairwise slopes and initially reported a trimmed mean of the pairwise slopes as a point estimate of the earth's ellipticity. Arc length is measured as the excess over 56,700 toise per degree where one toise ≈ 6.39 feet, or 1.95 meters.

where y is the length of the arc and λ is the latitude, was known to be satisfactory. The parameter a could be interpreted as the length of a degree of arc at the equator and b as the exceedence of a degree of arc at the pole over its value at the equator. Ellipticity could then be computed as $1/\text{ellipticity} = \eta = 3a/b$. Boscovich, noting that any pair of observations could be used to compute an estimate of a and b, hence of η, began by computing all ten such estimates. These lines are illustrated in Figure 1.1. Some of these lines seemed quite implausible, especially perhaps the downward-sloping one through Rome and the Cape of Good Hope. Boscovich reported two final estimates: one based on averaging all ten distinct estimates of b, the other based on averaging all but two of the pairwise slopes with the smallest implied exceedence. In both cases the estimate of a was taken directly from the measured length of the arc at Quito. These gave ellipticities of 1/155 and 1/198, respectively. A modern variant on this idea is the median of pairwise slopes suggested by Theil (1950), which yields the somewhat lower estimate of 1/255.

It is a curiosity worth noting that the least-squares estimator of (a, b) may also be expressed as a weighted average of the pairwise slope estimates. Let h index the ten pairs, and write

$$b(h) = X(h)^{-1} y(h), \tag{1.2}$$

where, for the simple bivariate model and $h = (i, j)$,

$$X(h) = \begin{pmatrix} 1 & x_i \\ 1 & x_j \end{pmatrix} \quad y(h) = \begin{pmatrix} y_i \\ y_j \end{pmatrix}; \tag{1.3}$$

then we may write the least-squares estimator as

$$\hat{b} = \sum_h w(h) b(h), \tag{1.4}$$

where $w(h) = |X(h)|^2 / \sum_h |X(h)|^2$. As shown by Subrahmanyam (1972) and elaborated by Wu (1986), this representation of the least-squares estimator extends immediately to the general p-parameter linear regression model. In the bivariate example the weights are obviously proportional to the distance between each pair of design points, a fact that, in itself, portends the fragility of least squares to outliers in either x or y observations.

Boscovich's second attack on the ellipticity problem formulated only two years later brings us yet closer to quantile regression. In effect, he suggests estimating (a, b) in (1.1) by minimizing the sum of absolute errors subject to the constraint that the errors sum to zero. The constraint requires that the fitted line pass through the centroid of the observations, $(\bar{x}, \bar{y})$. Boscovich provided a geometric algorithm, which was remarkably simple, to compute the estimator. Having reduced the problem to regression through the origin with the aid of the constraint, we may imagine rotating a line through the new origin at $(\bar{x}, \bar{y})$ until the sum of absolute residuals is minimized. This may be viewed algebraically, as noted later by Laplace, as the computation of a *weighted median*. For each point we may compute

$$b_i = \frac{y_i - \bar{y}}{x_i - \bar{x}} \tag{1.5}$$

and associate with each slope the weight $w_i = |x_i - \bar{x}|$. Now let $b_{(i)}$ be the ordered slopes and $w_{(i)}$ the associated weights, and find the smallest j, say j^*, such that

$$\sum_{i=1}^{j} w_{(i)} > \frac{1}{2} \sum_{i=1}^{n} w_{(i)} \tag{1.6}$$

The Boscovich estimator, $\hat{\beta} = b_{(j^*)}$, was studied in detail by Laplace in 1789 and later in his monumental *Traite de Méchanique Céleste*. Boscovich's proposal, which Laplace later called the "method of situation," is a curious blend of mean and median ideas; in effect, the slope parameter b is estimated as a median, while the intercept parameter a is estimated as a mean.

This was clearly recognized by Edgeworth, who revived these ideas in 1888 after nearly a century of neglect. In his early work on index numbers and weighted averages, Edgeworth had emphasized that the putative optimality of the sample mean as an estimator of location was crucially dependent on the assumption that the observations came from a common normal distribution. If the observations were "discordant," say from normals with different variances, the median, he argued, could easily be superior to the mean. Indeed, anticipating the work of Tukey in the 1940s, Edgeworth compares the asymptotic variances of the median and mean for observations from scale mixtures of normals, concluding that, for equally weighted mixtures with relative scale greater than 2.25, the median had smaller asymptotic variance than the mean.

Edgeworth's work on median methods for linear regression brings us directly to quantile regression. Edgeworth (1888) discards the Boscovich–Laplace constraint that the residuals sum to zero and proposes to minimize the sum of absolute residuals in both intercept and slope parameters, calling it a "double median" method and noting that it could be extended, in principle, to a "plural median" method. A geometric algorithm was given for the bivariate case, and a discussion of conditions under which one would prefer to minimize absolute error rather than the by-then well-established squared error is provided. Unfortunately, the geometric approach to computing Edgeworth's new median regression estimator was rather awkward, requiring, as he admitted later, "the attention of a mathematician; and in the case of many unknowns, some power of hypergeometrical conception." Only considerably later did the advent of linear programming provide a conceptually simple and efficient computational approach.

Once we have a median regression estimator it is natural to ask, "are there analogs for regression of the other quantiles?" The answer to this question is explored in the next section.

1.3 QUANTILES, RANKS, AND OPTIMIZATION

Any real-valued random variable X may be characterized by its (right-continuous) distribution function

$$F(x) = P(X \leq x), \tag{1.7}$$

whereas for any $0 < \tau < 1$,

$$F^{-1}(\tau) = \inf\{x : F(x) \geq \tau\} \tag{1.8}$$

is called the τth quantile of X. The median, $F^{-1}(1/2)$, plays the central role.

The quantiles arise from a simple optimization problem that is fundamental to all that follows. Consider a simple decision theoretic problem: a point estimate is required for a random variable with (posterior) distribution function F. If loss is described by the piecewise linear function illustrated in Figure 1.2

$$\rho_\tau(u) = u(\tau - I(u < 0)) \tag{1.9}$$

for some $\tau \in (0, 1)$, find $\hat{x}$ to minimize expected loss. This is a standard exercise in decision theory texts (e.g., Ferguson, 1967, p. 51). The earliest reference that I am aware of is Fox and Rubin (1964), who studied the admissibility of the quantile estimator under this loss function. We seek to minimize

$$E\rho_\tau(X - \hat{x}) = (\tau - 1)\int_{-\infty}^{\hat{x}} (x - \hat{x})dF(x) + \tau \int_{\hat{x}}^{\infty} (x - \hat{x})dF(x). \tag{1.10}$$

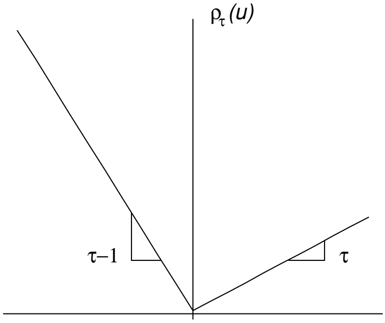

Figure 1.2. Quantile regression ρ function.

Differentiating with respect to $\hat{x}$, we have

$$0 = (1-\tau)\int_{-\infty}^{\hat{x}} dF(x) - \tau \int_{\hat{x}}^{\infty} dF(x) = F(\hat{x}) - \tau. \tag{1.11}$$

Since F is monotone, any element of $\{x : F(x) = \tau\}$ minimizes expected loss. When the solution is unique, $\hat{x} = F^{-1}(\tau)$; otherwise, we have an "interval of τth quantiles" from which the smallest element must be chosen – to adhere to the convention that the empirical quantile function be left-continuous.

It is natural that an optimal point estimator for asymmetric linear loss should lead us to the quantiles. In the symmetric case of absolute value loss it is well known to yield the median. When loss is linear and asymmetric, we prefer a point estimate more likely to leave us on the flatter of the two branches of marginal loss. Thus, for example, if an underestimate is *marginally* three times more costly than an overestimate, we will choose $\hat{x}$ so that $P(X \leq \hat{x})$ is three times greater than $P(X > \hat{x})$ to compensate. That is, we will choose $\hat{x}$ to be the 75th percentile of F.

When F is replaced by the empirical distribution function

$$F_n(x) = n^{-1} \sum_{i=1}^{n} I(X_i \leq x), \tag{1.12}$$

we may still choose $\hat{x}$ to minimize expected loss:

$$\int \rho_\tau(x - \hat{x}) dF_n(x) = n^{-1} \sum_{i=1}^{n} \rho_\tau(x_i - \hat{x}) \tag{1.13}$$

and doing so now yields the τth *sample* quantile. When τn is an integer there is again some ambiguity in the solution, because we really have an interval of solutions, $\{x : F_n(x) = \tau\}$, but we shall see that this is of little practical consequence.

Much more important is the fact that we have expressed the problem of finding the τth sample quantile, a problem that might seem inherently tied to the notion of an ordering of the sample observations, as the solution to a simple optimization problem. In effect we have replaced *sorting* by *optimizing*. This will prove to be the key idea in generalizing the quantiles to a much richer class of models in subsequent chapters. Before doing this, though, it is worth examining the simple case of the ordinary sample quantiles in a bit more detail.

The problem of finding the τth sample quantile, which may be written as

$$\min_{\xi \in \mathbf{R}} \sum_{i=1}^{n} \rho_\tau(y_i - \xi), \tag{1.14}$$

may be reformulated as a linear program by introducing $2n$ artificial, or “slack,” variables $\{u_i, v_i : 1, \ldots, n\}$ to represent the positive and negative parts of the vector of residuals. This yields the new problem

$$\min_{(\xi,u,v) \in \mathbb{R} \times \mathbb{R}_+^{2n}} \left\{ \tau 1_n^\top u + (1-\tau) 1_n^\top v \,|\, 1_n \xi + u - v = y \right\}, \tag{1.15}$$

where 1_n denotes an n-vector of 1. Clearly, in (1.15) we are minimizing a linear function on a polyhedral constraint set consisting of the intersection of the $(2n+1)$-dimensional hyperplane determined by the linear equality constraints and the set $\mathbb{R} \times \mathbb{R}_+^{2n}$.

Figure 1.3 illustrates the most elementary possible version of the median linear programming problem. We have only one observation, at $y = 1$, and we wish to solve

$$\min_{(\xi,u,v) \in \mathbb{R} \times \mathbb{R}_+^{2}} \{u + v \,|\, \xi + u - v = y\}.$$

The constraint set is the triangular region representing the intersection of the plane $\{(\xi, u, v) | \xi + u - v = 1\}$ with the cone $\{(\xi, u, v) \in \mathbb{R}^3 | u \geq 0, v \geq 0\}$. The long edge of this triangle extends off into the deeper regions of the figure. The objective function is represented by a series of vertical planes perpendicular to the 45° line in the (u, v) (horizontal) plane. Moving toward the origin reduces $u + v$, thus improving the objective function. It is apparent that any feasible point (ξ, u, v) that has both u and v strictly positive can be improved by reducing v and increasing u to compensate. But with only one observation we can move further. Reducing u and increasing ξ to compensate – that is, moving along the interior edge of the constraint set – allows us to reduce the objective function to zero, setting $\xi = 1$, coming to rest at the upper-left corner of the constraint set. Now, if we try to imagine increasing the number of observations, we have contributions to the objective function from each observation like the one illustrated in Figure 1.3. Given a trial value of the parameter ξ, we can consider a feasible point that sets each u_i equal to the positive part of the residual $y_i - \xi$ and v_i equal to the negative part of the ith residual. But, as in Figure 1.3, such solutions can always be improved by moving ξ closer to one of the sample observations.

Many features of the solution are immediately apparent from these simple observations. To summarize, $\min\{u_i, v_i\}$ must be zero for all i, because otherwise the objective function may be reduced without violating the constraint

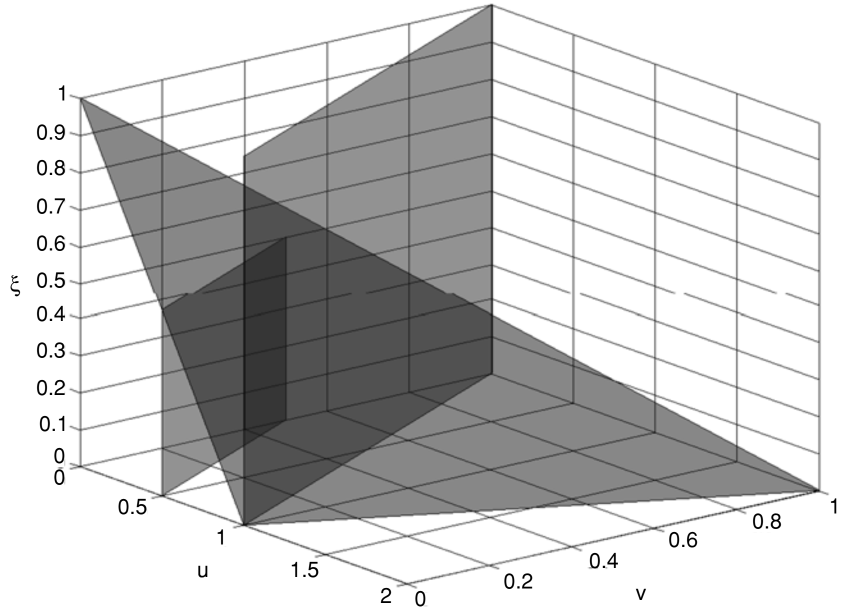

Figure 1.3. Computing the median with one observation. The figure illustrates the linear programming formulation of the median problem. The triangular region represents the constraint set; the vertical planes represent two different contours of the objective function, which decreases as ones moves toward the origin in the (u, v)-plane.

by shrinking such a pair toward zero. This is usually called complementary slackness in the terminology of linear programming. Indeed, for essentially the same reason we can restrict attention to "basic solutions" of the form $\xi = y_i$ for some observation i. Figure 1.4 depicts objective function (1.14) for three different random samples of varying sizes. The graph of the objective function is convex and piecewise linear with kinks at the observed y_is. When ξ passes through one of these y_is, the slope of the objective function changes by exactly 1 since a contribution of $\tau - 1$ is replaced by τ or vice versa.

Optimality holds at a point $\hat{\xi}$ if the objective function

$$R(\xi) = \sum_{i=1}^{n} \rho_\tau(y_i - \xi)$$

is increasing as one moves away from $\hat{\xi}$ to either the right or the left. This requires that the right and left derivatives of R are both nonnegative at the point $\hat{\xi}$. Thus,

$$R'(\xi+) \equiv \lim_{h \to 0}(R(\xi + h) - R(\xi))/h = \sum_{i=1}^{n}(I(y_i < \xi + 0) - \tau)$$

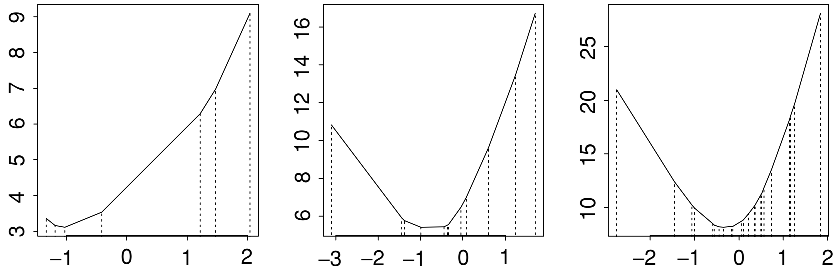

Figure 1.4. Quantile objective function with random data. The figure illustrates the objective function for the optimization problem defining the ordinary $\tau = 1/3$ quantile for three different random problems with y_is drawn from the standard normal distribution and sample sizes 7, 12, and 23. The vertical dotted lines indicate the position of the observations in each sample. Note that because 12 is divisible by 3, the objective function is flat at its minimum in the middle figure, and we have an interval of solutions between the fourth- and fifth-largest observations.

and

$$R'(\xi-) \equiv \lim_{h\to 0}(R(\xi - h) - R(\xi))/h = \sum_{i=1}^{n}(\tau - I(y_i < \xi - 0))$$

must both be nonnegative, and so $n\tau$ lies in the closed interval $[N^-, N^+]$, where N^+ denotes the number of y_i less than or equal to ξ and N^- denotes the number of y_i strictly less than ξ. When $n\tau$ is not an integer, there is a unique value of ξ that satisfies this condition. Barring ties in the y_is, this value corresponds to a unique order statistic. When there are ties, ξ is still unique, but there may be several y_i equal to ξ. If $n\tau$ *is* an integer then $\hat{\xi}_\tau$ lies between two adjacent order statistics. It is unique only when these order statistics coalesce at a single value. Usually, we can dismiss the occurrence of such ties as events of probability zero.

The duality connecting the sample quantiles and the ranks of the order statistics is further clarified through the formal duality of linear programming. While the primal problem, (1.15), may be viewed as generating the sample quantiles, the corresponding dual problem may be seen to generate the order statistics, or perhaps more precisely the *ranks* of the observations. This approach to ranks generalizes naturally to the linear model, yielding an elegant generalization of rank tests for the linear model.

1.4 PREVIEW OF QUANTILE REGRESSION

The observation developed in Section 1.3 that the quantiles may be expressed as the solution to a simple optimization problem leads, naturally, to more general methods of estimating models of conditional quantile functions. Least squares offers a template for this development. Knowing that the sample mean solves the problem

$$\min_{\mu\in\mathbb{R}} \sum_{i=1}^{n} (y_i - \mu)^2 \tag{1.16}$$

suggests that, if we are willing to express the *conditional* mean of y given x as $\mu(x) = x^\top \beta$, then β may be estimated by solving

$$\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} \left(y_i - x_i^\top \beta\right)^2 . \tag{1.17}$$

Similarly, since the τth sample quantile, $\hat{\alpha}(\tau)$, solves

$$\min_{\alpha\in\mathbb{R}} \sum_{i=1}^{n} \rho_\tau(y_i - \alpha), \tag{1.18}$$

we are led to specifying the τth *conditional* quantile function as $Q_y(\tau|x) = x^\top \beta(\tau)$, and to consideration of $\hat{\beta}(\tau)$ solving

$$\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau \left(y_i - x_i^\top \beta\right) . \tag{1.19}$$

This is the germ of the idea elaborated by Koenker and Bassett (1978).

Quantile regression problem (1.19) may be reformulated as a linear program as in (1.15):

$$\min_{(\beta,u,v)\in\mathbb{R}^p\times\mathbb{R}_+^{2n}} \left\{\tau 1_n^\top u + (1-\tau)1_n^\top v | X\beta + u - v = y\right\}, \tag{1.20}$$

where X now denotes the usual n by p regression design matrix. Again, we have split the residual vector $y - X\beta$ into its positive and negative parts, and so we are minimizing a linear function on a polyhedral constraint set, and most of the important properties of the solutions, $\hat{\beta}(\tau)$, which we call "regression quantiles," again follow immediately from well-known properties of solutions of linear programs.

We can illustrate the regression quantiles in a very simple bivariate example by reconsidering the Boscovich data. In Figure 1.5 we illustrate all of the *distinct* regression quantile solutions for this data. Of the ten lines passing through pairs of points in Figure 1.1, quantile regression selects only four. Solving (1.19) for any τ in the interval (0, 0.21) yields as a unique solution the line passing through Quito and Rome. At $\tau = 0.21$, the solution jumps, and throughout the interval (0.21, 0.48) we have the solution characterized by the line passing through Quito and Paris. The process continues until we get to $\tau = 0.78$, where the solution through Lapland and the Cape of Good Hope prevails up to $\tau = 1$.

In contrast to the ordinary sample quantiles that are equally spaced on the interval [0,1], with each distinct order statistic occupying an interval of length exactly $1/n$, the lengths of the regression quantile solution intervals for $\tau \in [0, 1]$ are irregular and depend on the configuration of the design as well as the realized values of the response variable. *Pairs of points now play the role*

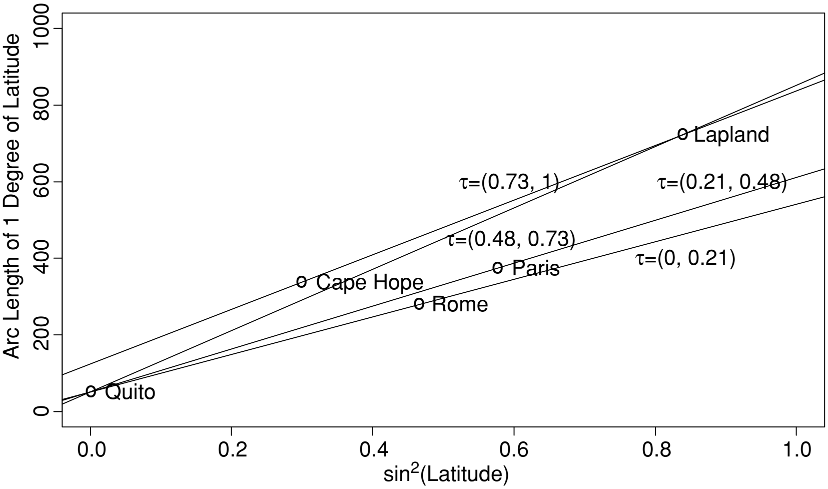

Figure 1.5. Regression quantiles for Boscovich ellipticity example. Only four of the full ten pairs of points form quantile regression solutions. The subintervals of (0, 1) for which each pair solves (1.19) are given in the figure.

of order statistics and serve to define the estimated linear conditional quantile functions. Again, in the terminology of linear programming, such solutions are "basic" and constitute extreme points of the polyhedral constraint set. If we imagine the plane represented by the objective function of (1.19) rotating as τ increases, we may visualize the solutions of (1.19) as passing from one vertex of the constraint set to another. Each vertex represents an exact fit of a line to a pair of sample observations. At a few isolated τ points, the plane will make contact with an entire edge of the constraint set and we will have a set-valued solution. It is easy to see, even in these cases, that the solution is characterized as the convex hull of its "basic" solutions.

One occasionally encounters the view that quantile regression estimators must "ignore sample information" since they are inherently determined by a small subset of the observations. This view neglects the obvious fact that all the observations participate in which "basic" observations are selected as basic.

We shall see that quantile regression does preserve an important robustness aspect of the ordinary sample quantiles: if we perturb the order statistics above (or below) the median in such a way that they *remain* above (or below) the median, the position of the median is unchanged. Similarly, for example, if we were to perturb the position of the Lapland observation upward, this would not affect the solutions illustrated in the figure for any τ in the interval (0, 0.48).

The Boscovich example is a bit too small to convey the full flavor of quantile regression even in the bivariate setting, so I will conclude this section with two other examples that exhibit various aspects of quantile regression in the bivariate context where pictures are easily available to illustrate the results.

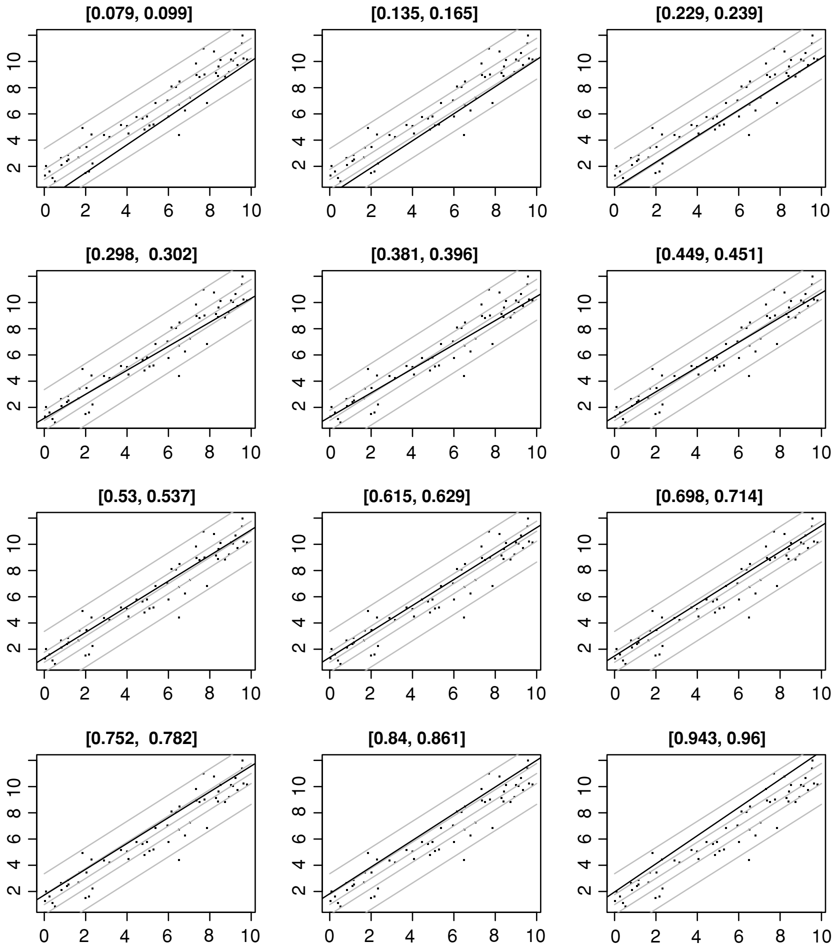

Figure 1.6. Regression quantiles for iid-error bivariate regression.

Consider an artificial sample in which we have a simple bivariate regression model with independent and identically distributed (iid) errors:

$$y_i = \beta_0 + x_i\beta_1 + u_i$$

and so the conditional quantile functions of y are

$$Q_y(\tau|x) = \beta_0 + x\beta_1 + F_u^{-1}(\tau),$$

where F_u denotes the common distribution function of the errors. In this simple case the quantile functions are simply a vertical displacement of one another and $\hat{\beta}(\tau)$ estimates the population parameters, $(\beta_0 + F^{-1}(\tau), \beta_1)^\top$.

In Figure 1.6 we illustrate data and several fitted regression quantile lines from such a model. The dots indicate 60 observations generated from the iid

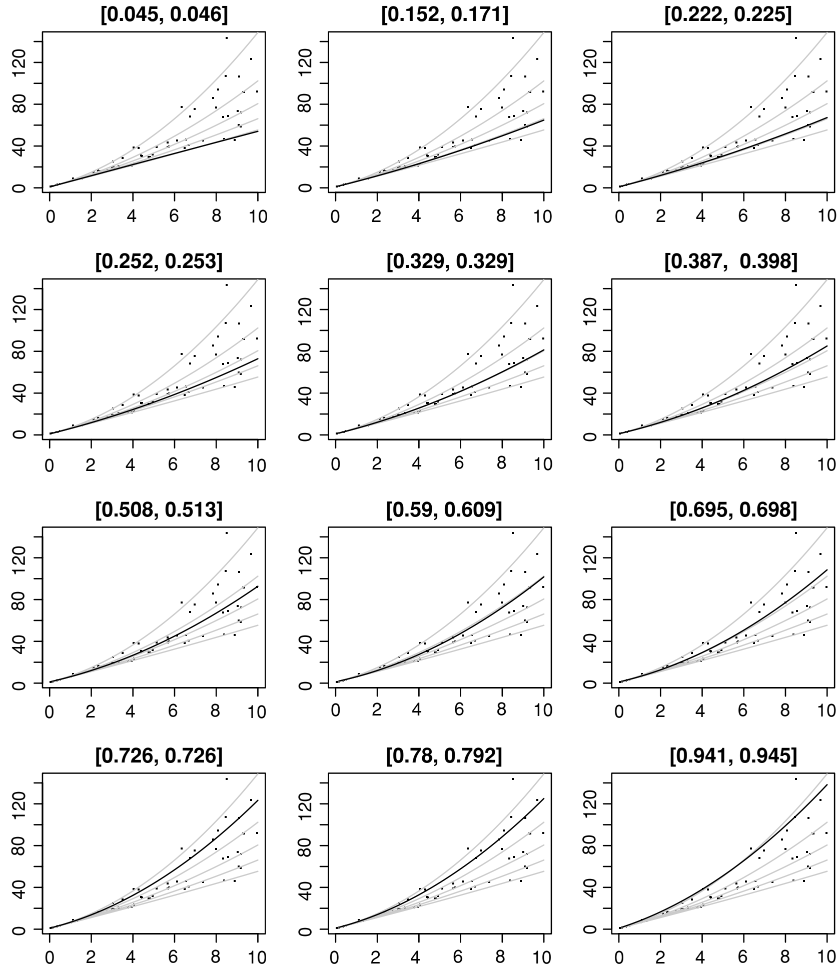

Figure 1.7. Regression quantiles for heteroscedastic bivariate regression.

error model with F selected to be Gaussian. The gray lines represent the *true* $\{0.05, 0.25, 0.50, 0.75, 0.95\}$ conditional quantile lines. The black line in each panel depicts the estimated conditional quantile line for the τ interval indicated above the panel. As τ increases, we see that these estimated lines move up through the data, retaining in most cases a slope reasonably close to that of the family of true conditional quantile functions. In this example there are 66 distinct regression quantile solutions. Rather than illustrate *all* of them, we have chosen to illustrate only 12, spaced roughly evenly over the interval [0, 1]. Above each panel we indicate the τ interval for which the illustrated solution is optimal.

If real data analysis were always as well behaved as the iid linear model depicted in Figure 1.6, there would be little need for quantile regression. The least-squares estimate of the conditional mean function and some associated measure of dispersion would (usually) suffice. Robust alternatives to least squares could be used to accommodate situations in which errors exhibited long tails.

In Figure 1.7 we illustrate a somewhat more complicated situation. The model now takes the heteroscedastic form,

$$y_i = \beta_0 + x_i\beta_1 + \sigma(x_i)u_i,$$

where $\sigma(x) = \gamma x^2$ and the $\{u_i\}$ are again iid. The conditional quantile functions of y are now

$$Q_y(\tau|x) = \beta_0 + x\beta_1 + \sigma(x)F^{-1}(\tau)$$

and can be consistently estimated by minimizing

$$\sum \rho_\tau(y_i - \beta_0 - x_i\beta_1 - x_i^2\beta_2)$$

so that $\hat{\beta}(\tau)$ converges to $(\beta_0, \beta_1, \gamma F^{-1}(\tau))^\top$. Figure 1.7 illustrates an example of this form. Again, the *population* conditional quantile functions are shown as gray lines with the observed sample of 60 points superimposed and a sequence of estimated quantile regression curves appearing as black lines. The estimated quantile regression curves provide a direct empirical analog for the family of conditional quantile functions in the population.

1.5 THREE EXAMPLES

A simple bivariate example and two somewhat more elaborate multivariate examples are used to motivate quantile regression methods.

1.5.1 Salaries versus Experience

In Figure 1.8 we illustrate a p-sample version of the basic quantile regression problem with results of the 1995 survey of academic salaries in statistics conducted by the American Statistical Association (ASA). The figure is based on data from 99 departments in U.S. colleges and universities on 370 full professors of statistics. The data are grouped into three-year experience categories defined as years since promotion to the rank of full professor. The boxes appearing in the figure represent the interquartile range of salaries for each experience group. The upper limit of the box represents the 75th percentile of the salary distribution in each experience group from the survey, while the lower limit represents the 25th percentile. Thus, the central half of the surveyed salaries would fall within the boxes. Median salary for each group is depicted by the horizontal line drawn in each box. The width of the boxes is proportional to the square root of the respective sample sizes of the groups.

What can we conclude from the boxes? There clearly seems to be a tendency for salary to increase at a decreasing rate with "years in rank," with some suggestion that salary may actually decline for the oldest group. There

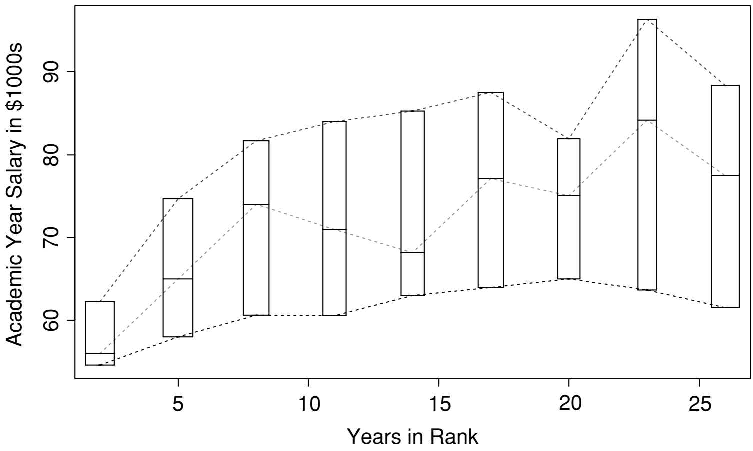

Figure 1.8. Boxplots of 1995 ASA academic salary survey for full professors of statistics in U.S. colleges and universities.

is also a pronounced tendency for the dispersion of the salary distribution to increase with experience. None of these findings is particularly surprising, but taken together they constitute a much more complete description than would be available from conventional least-squares regression analysis. The boxplot takes us much further than we are able to go with only the conditional mean function. Of course we would like to go still further: to estimate more quantiles of the distribution, to introduce additional covariates, to disaggregate the experience groups, and so forth. However, each of these steps diminishes the viability of the boxplot approach, which relies on adequate sample sizes for each of the groups, or cells, represented by the boxes. What could we do if the sample size of the salary survey were only 96 points, as it was in 1973–74, rather than the 370 observations of 1995?

Hogg (1975) provides an answer to this question, an answer that constituted an elaboration of a much earlier proposal by Brown and Mood (1951) for median regression. Hogg suggested dividing the observations (x_i, y_i) into two groups according to whether $x_i \leq \text{median}\{x_j\}$ or $x_i > \text{median}\{x_j\}$ and then estimating linear conditional quantile functions,

$$Q_Y(\tau|x) = \alpha + \beta x,$$

by choosing $(\hat{\alpha}, \hat{\beta})$ so that the number of observations in both groups had (approximately) the same proportion, τ, of their observations below the line. This can be accomplished relatively easily "by eye" for small data sets using a method Hogg describes. A more formal version of Hogg's proposal may be cast as a quantile regression version of the Wald (1940) instrumental variables estimator for the errors-in-variable model. This connection is developed more fully in Section 8.8. Based on the 1973–74 ASA data for full professor salaries, he obtains the estimates reported in Table 1.2. Since the estimated slope parameters

Table 1.2. *Hogg's (1975) linear quantile regression results for the 1973–74 ASA academic salary survey of full professors of statistics in U.S. colleges and universities. The monotone relation of the slope estimates indicates heteroscedasticity (i.e., increasing salary dispersion with experience)*

Quantile τ	Initial Professorial Salary $\hat{\alpha}$	Annual Increment $\hat{\beta}$
0.75	21,500	625
0.50	20,000	485
0.25	18,800	300

$\hat{\beta}$ increase with the quantile, these estimates reflect the same increasing dispersion, or heteroscedasticity, that we saw in the boxplots of Figure 1.8 for the more recent salary data. In this case, with so few data, it does not seem prudent to venture an opinion about the curvature of the salary profile.

We could probably agree that the dotted curves connecting the boxplot salary quartiles of Figure 1.8 appear somewhat undersmoothed. A parametric model for the conditional quartiles might improve the appearance of the plot, if we could agree on a transformation that would adequately capture the curvature of the salary profile. One attempt to do this is illustrated in Figure 1.9, where we have chosen the parametric model

$$Q_{\log(y)}(\tau|x) = \alpha + \beta \log x$$

for each of the quartiles, $\tau \in \{1/4, 1/2, 3/4\}$. The curves shown in Figure 1.9 have been estimated by median (ℓ_1) regression using only the respective grouped

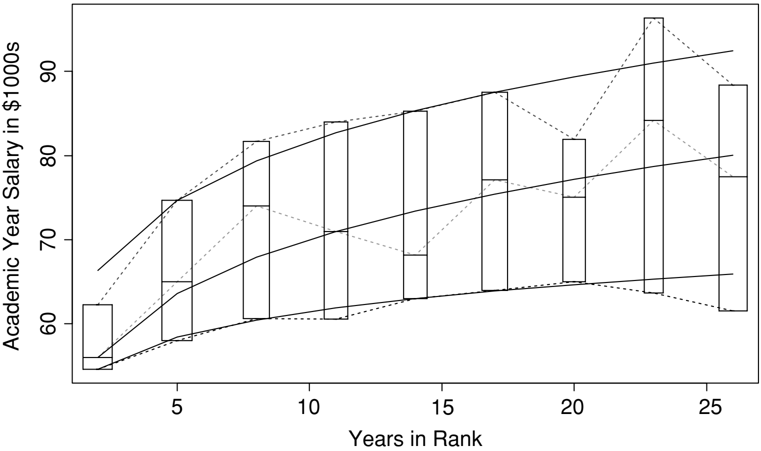

Figure 1.9. Boxplots of 1995 ASA academic salary survey for full professors of statistics in U.S. colleges and universities.

quartile data. (The individual data collected by the ASA are protected by confidentiality assurances.) These curves, and the parameters that characterize them, have a straightforward interpretation. The slope parameter in the log-linear quantile regression is simply a rate of growth of salary with respect to experience. In this example, the first quartile of the salary distribution has an estimated growth rate of 7.3% per year of tenure, whereas the median and the upper quartile grow at 14 and 13% respectively. As for Hogg's linear specification, higher coefficients at the higher quantiles imply increasing dispersion in salaries for more experienced faculty. However, in this case, the tendency is pronounced only in the left tail of the distribution and there is actually a slight narrowing of the gap between the median and the upper quartile for older faculty.

As this example illustrates, the specification and interpretation of quantile regression models is very much like that of ordinary regression. However, unlike ordinary regression, we now have a family of curves to interpret, and we can focus attention on particular segments of the conditional distribution, thus obtaining a more complete view of the relationship between the variables. If the slope parameters of the family of estimated quantile regression models seem to fluctuate randomly around a constant level, with only the intercept parameter systematically increasing with τ, we have evidence for the iid error hypothesis of classical linear regression. If, however, some of the slope coefficients are changing with τ, then this is indicative of some form of heteroscedasticity. The simplest example of this kind of heteroscedasticity is what we have called the linear location-scale model,

$$y_i = x_i^\top \beta + \left(x_i^\top \gamma\right) u_i,$$

with $\{u_i\}$ iid from F. In this case the coefficients of the τth quantile regression, $\hat{\beta}(\tau)$, converge to $\beta + \gamma F_u^{-1}(\tau)$, and so all of the parameters would share the same monotone behavior in τ, governed by the quantile function of the errors $F_u^{-1}(\tau)$. Clearly, this too is an extremely restrictive model, and we often find very different behavior (in τ) across slope coefficients. Such findings should remind us that the theory of the linear statistical model and its reliance on the hypothesis of a scalar iid error process is only a convenient fiction; life can be stranger, and more interesting.

1.5.2 Student Course Evaluations and Class Size

Our second example illustrates several advantages of the optimization approach to quantile regression. The data consist of mean course evaluation scores for 1482 courses offered by a large public university over the period 1980–94. We are primarily concerned with the effect of class size on course evaluation questionnaire (CEQ) score, but also of interest is the possibility of a time trend in the scores and any special effects due to the nature of particular types of courses.

In Figure 1.10 we illustrate the data for this example and plot five estimated quantile regression curves. These curves are specified as quadratic in the number

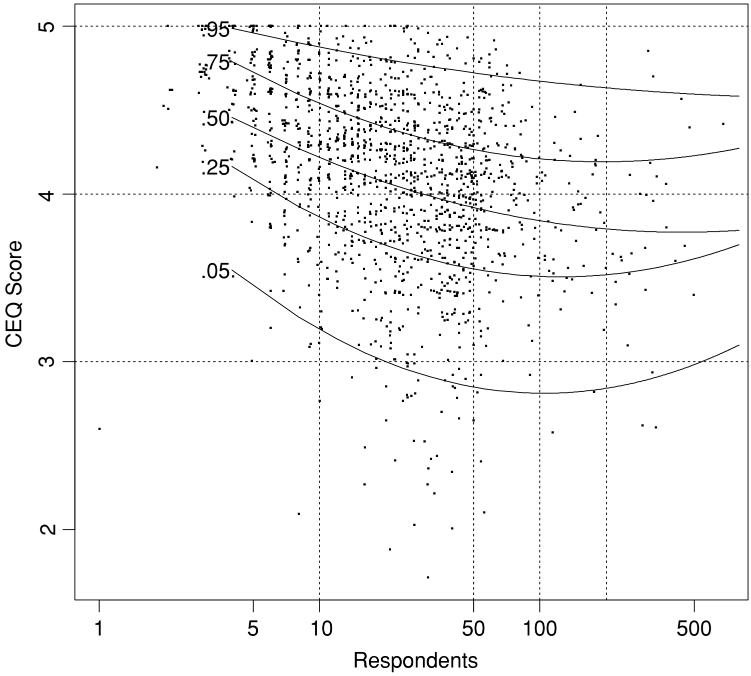

Figure 1.10. Course evaluation scores. Solid lines indicate estimated quantiles of CEQ response for an undergraduate course in 1992 as a function of the class size measured by number of CEQ respondents.

of CEQ respondents, which can be taken as the relevant measure of class size. In addition to the class size effect, we have included a linear time trend and an indicator variable that takes the value 1 for graduate courses and 0 for undergraduate courses. The model may thus be written as

$$Q_Y(\tau|x) = \beta_0(\tau) + \text{Trend}\,\beta_1(\tau) + \text{Grad}\,\beta_2(\tau) + \text{Size}\,\beta_3(\tau) + \text{Size}^2\beta_4(\tau)$$

and can be estimated for any $\tau \in (0, 1)$ by solving the problem

$$\min_{b\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^\top b). \tag{1.21}$$

The estimated quantile regression parameters and their confidence intervals are given in Table 1.3. Details on the construction of the confidence intervals appear in the next chapter.

From Table 1.3 it can be seen that there is some evidence for a downward trend in CEQ scores for the lower quantiles, on the order of 0.01 to 0.02 rating points per year, but no evidence of a trend in the upper tail of the ratings distribution. One tentative conclusion from this is that ornery students are getting ornerier. Graduate courses have a fairly consistent tendency to be rated higher by about 0.10 rating points than undergraduate courses.

Table 1.3. *Quantile regression estimates for a model of student course evaluation scores. Numbers in parentheses give a 95% confidence interval for each reported coefficient.*

τ	Intercept	Trend	Graduate	Size	Size^2
0.050	4.749 (4.123,5.207)	−0.032 (−0.041,−0.016)	0.054 (−0.065,0.169)	−0.642 (−0.930,−0.233)	0.069 (0.013,0.104)
0.250	5.003 (4.732,5.206)	−0.014 (−0.023,−0.008)	0.132 (0.054,0.193)	−0.537 (−0.604,−0.393)	0.056 (0.034,0.066)
0.500	5.110 (4.934,5.260)	−0.014 (−0.018,−0.008)	0.095 (0.043,0.157)	−0.377 (−0.484,−0.274)	0.031 (0.014,0.050)
0.750	5.301 (5.059,5.379)	−0.001 (−0.005,0.005)	0.111 (0.027,0.152)	−0.418 (−0.462,−0.262)	0.040 (0.015,0.050)
0.950	5.169 (5.026,5.395)	0.001 (−0.004,0.006)	0.054 (−0.001,0.099)	−0.159 (−0.323,−0.085)	0.010 (−0.005,0.035)

In order to plot the curves illustrated in Figure 1.10, we have set the indicator variable to zero to represent an undergraduate course and the trend variable to represent the last year in the sample, 1994. The curves clearly show a tendency for larger classes to receive lower ratings by students, with this decline occurring at a decreasing rate. The apparent tendency for scores to increase slightly for courses with more than 100 respondents may be entirely an artifact of the quadratic specification of the curves, but it may also be partially attributed to a departmental policy of trying to allocate its best teachers to the larger courses.

In the course evaluation example we have seen that the downward time trend in student evaluations is apparent at the median and lower quantiles but there is essentially no trend in the upper conditional quantile estimates. In contrast, the estimated disparity between graduate and undergraduate course ratings is positive and quite large (0.1 rating points) for the central quantiles, but negligible in the tails. This $\cap$-shape for $\hat{\beta}_j(\tau)$ may seem strange at first, but it is easily reconciled by considering a very simple two-sample quantile regression problem.

Suppose, to continue the course evaluation example, that sample one of undergraduate scores, supported on the interval $[1, 5]$, was quite symmetric around its median, while sample two of graduate ratings was skewed toward the upper bound of 5. If the two distributions have similar tail behavior, then the quantile regressions, which in the two-sample case simply connect the corresponding quantiles of the two distributions, would also display a $\cap$-shaped pattern – central quantiles with a significant positive slope, extreme quantiles with negligible slope. The effect of class size on the quantile regressions for CEQ scores is illustrated in Figure 1.10. There is some tendency for these curves to be initially more steeply sloped and to exhibit more curvature at lower quantiles.

Taken together, it is difficult to reconcile these observations with a conventional scalar-error linear model, but they do offer a much richer view of the data than the one provided by a least-squares analysis.

1.5.3 Infant Birth Weight

The third example reconsiders an investigation by Abreveya (2001) of the impact of various demographic characteristics and maternal behavior on the birth weight of infants born in the United States. Low birth weight is known to be associated with a wide range of subsequent health problems and has even been linked to educational attainment and labor market outcomes. Consequently, there has been considerable interest in factors influencing birth weight and public policy initiatives that might prove effective in reducing the incidence of low-birth-weight infants.

Although most of the analysis of birth weight has employed conventional least-squares regression methods, it has been recognized that the resulting estimates of various effects on the conditional mean of birth weights were not necessarily indicative of the size and nature of these effects on the lower tail of the birth-weight distribution. In an effort to focus attention more directly on the lower tail, several studies have recently explored binary response (e.g., probit) models for the occurrence of low birth weights – conventionally defined to be infants weighing less than 2500 grams. Quantile regression offers a natural complement to these prior modes of analysis. A more complete picture of covariate effects can be provided by estimating a family of conditional quantile functions.

The analysis will be based on the June 1997 Detailed Natality Data published by the National Center for Health Statistics. Like Abreveya's study, the sample is restricted to singleton births, with mothers recorded as either black or white, between the ages of 18 and 45, resident in the United States. Observations with missing data for any of the variables described in the following were also dropped from the analysis. This process yielded a sample of 198,377 babies. Education of the mother is divided into four categories: less than high school, high school, some college, and college graduate. The omitted category is "less than high school," so coefficients must be interpreted relative to this category. The prenatal medical care of the mother is also divided into four categories: those with no prenatal visit, those whose first prenatal visit was in the first trimester of the pregnancy, those with the first visit in the second trimester, and those with the first visit in the last trimester. The omitted category is the group with a first visit in the first trimester; they constitute almost 85 percent of the sample. The other variables are, hopefully, self-explanatory.

Figure 1.11 presents a concise summary of the quantile regression results for this example. Each plot depicts one coefficient in the quantile regression model. The solid line with filled dots represents the point estimates, $\{\hat{\beta}_j(\tau) : j = 1, \ldots, 16\}$, with the shaded gray area depicting a 90% pointwise confidence band. Superimposed on the plot is a dashed line representing the ordinary least-squares estimate of the mean effect, with two dotted lines again representing a 90% confidence interval for this coefficient.

In the first panel of the figure, the intercept of the model may be interpreted as the estimated conditional quantile function of the birth-weight distribution of

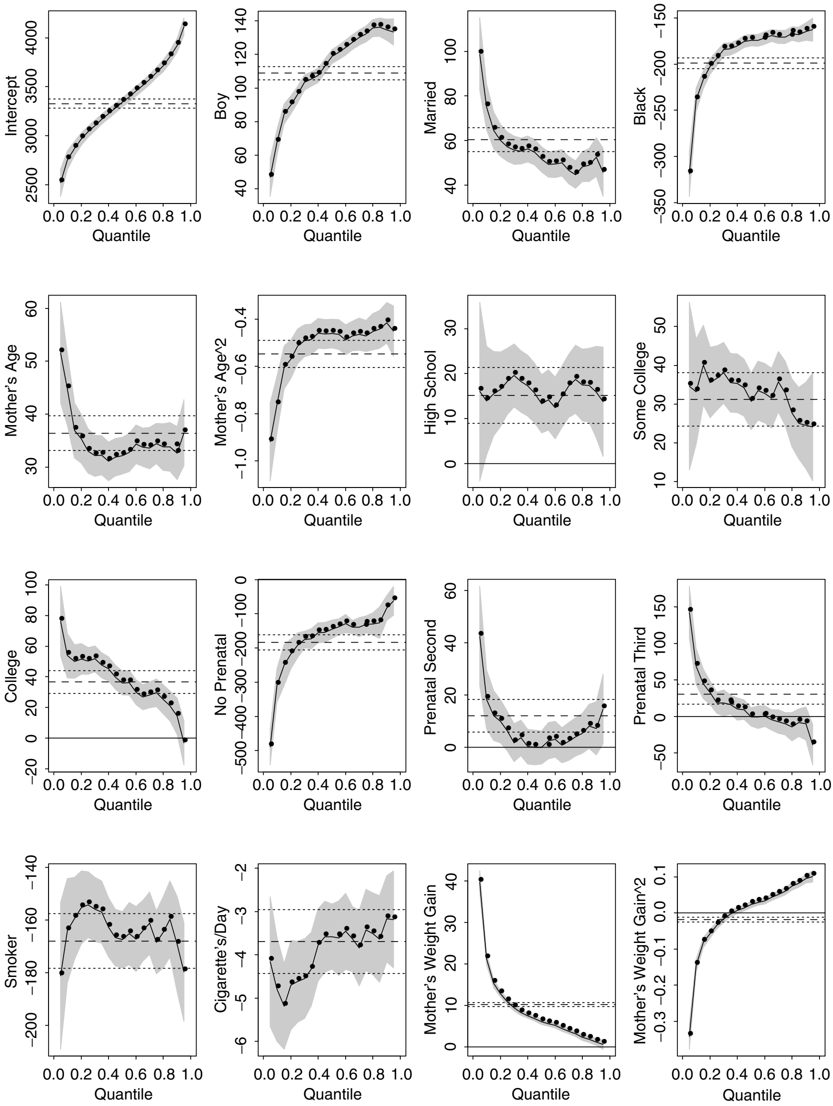

Figure 1.11. Quantile regression for birth weight.

a girl born to an unmarried, white mother with less than a high school education, who is 27 years old and had a weight gain of 30 pounds, did not smoke, and had her first prenatal visit in the first trimester of the pregnancy. The mother's age and weight gain are chosen to reflect the means of these variables in the sample. Note that the $\tau = 0.05$ quantile of this distribution is just at the margin of the conventional definition of a low-birth-weight baby.

Boys are obviously bigger than girls, about 100 grams bigger according to the ordinary least-squares (OLS) estimates of the mean effect, but, as is clear from the quantile regression results, the disparity is much smaller in the lower quantiles of the distribution and somewhat larger than 100 grams in the upper tail of the distribution. At any chosen quantile we can ask how different the corresponding weights of boys and girls are, given a specification of the other conditioning variables. The second panel answers this question.

Perhaps surprisingly, the marital status of the mother seems to be associated with a rather large positive effect on birth weight, especially in the lower tail of the distribution. The public health implications of this finding should, of course, be viewed with caution, however.

The disparity between birth weights of infants born to black and white mothers is very large, particularly at the left tail of the distribution. The difference in birth weight between a baby born to a black mother and a white mother at the 5th percentile of the conditional distribution is roughly one-third of a kilogram.

Mother's age enters the model as a quadratic. At the lower quantiles the mother's age tends to be more concave, increasing birth weight from age 18 to about age 30, but tending to decrease birth weight when the mother's age is beyond 30. At higher quantiles there is also this optimal age, but it becomes gradually older. At the third quantile it is about 36, and at $\tau = 0.9$ it is almost 40. This is illustrated in Figure 1.12.

Education beyond high school is associated with a modest increase in birth weight. High school graduation has a quite uniform effect over the whole range of the distribution of about 15 grams. This is a rare example of an effect that really does appear to exert a pure location shift effect on the conditional distribution. Some college education has a somewhat more positive effect in the lower tail than in the upper tail, varying from about 35 grams in the lower tail to 25 grams in the upper tail. A college degree has an even more substantial positive effect, but again much larger in the lower tail and declining to a negligible effect in the upper tail.

The effect of prenatal care is of obvious public health policy interest. Since individuals self-select into prenatal care, results must be interpreted with considerable caution. Those receiving no prenatal care are likely to be at risk in other dimensions as well. Nevertheless, the effects are sufficiently large to warrant considerable further investigation. Babies born to mothers who received no prenatal care were on average about 150 grams lighter than those who had a prenatal visit in the first trimester. In the lower tail of the distribution this effect is considerably larger – at the 5th percentile it is nearly half a kilogram! In contrast, mothers who delayed prenatal visits until the second or third trimester have substantially *higher* birth weights in the lower tail than mothers who had a visit in the first trimester. This might be interpreted as the self-selection effect of mothers confident about favorable outcomes. In the upper three quarters of the distribution there seems to be no significant effect.

Smoking has a clearly deleterious effect. The indicator of whether the mother smoked during the pregnancy is associated with a decrease of about 175 grams in birth weight. In addition, there is an effect of about 4 to 5 grams per cigarette per day. Thus, a mother smoking a pack per day appears to induce a birth-weight

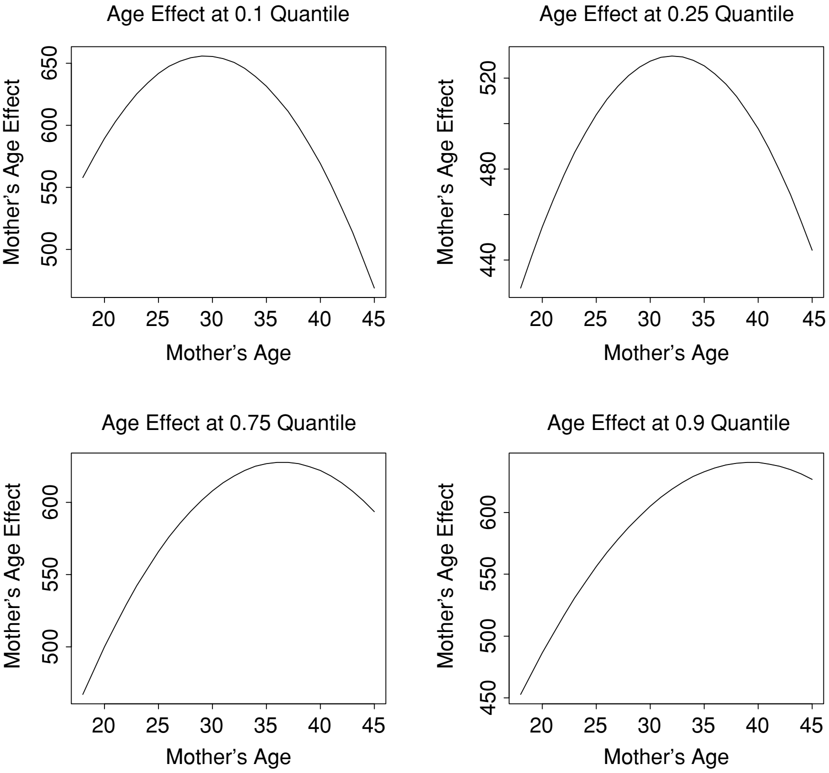

Figure 1.12. Mother's age effect on Birth weight. The estimated quadratic effect of mother's age on infant birth weight is illustrated at four different quantiles of the conditional birth-weight distribution. In the lower tail of the conditional distribution, mothers who are roughly 30 years of age have the largest children, but in the upper tail it is mothers who are 35–40 who have the largest children.

reduction of about 250 to 300 grams, or from about one-half to two-thirds of a pound. In contrast to some of the other effects, the effect of smoking is quite stable over the entire distribution, as indicated by the fact that the least-squares point estimates of the two smoking effects are (nearly) covered by the quantile regression confidence band.

Lest this smoking effect be thought to be attributable to some associated reduction in the mother's weight gain, one should hasten to point out that the weight gain effect is explicitly accounted for with a quadratic specification. Not surprisingly, the mother's weight gain has a very strong influence on birth weight, and this is reflected in the very narrow confidence band for both linear and quadratic coefficients. Figure 1.13 illustrates the marginal effect of weight gain by evaluating over the entire range of quantiles for four different levels of weight gain. At low weight gains by the mother, the marginal effect of another

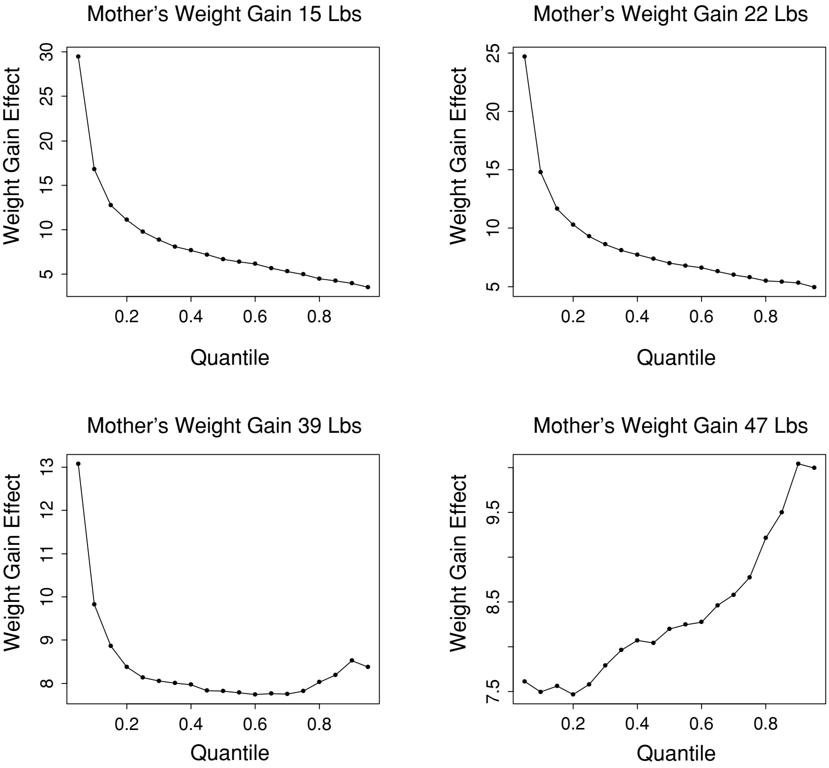

Figure 1.13. Mother's weight gain marginal effect. The marginal effect of the mother's weight gain, again parameterized as a quadratic effect, tends to decrease over the entire range of the conditional distribution of birth weight. Thus, incremental weight gain is most influential in increasing the weight of low-birth-weight infants. But for mothers with unusually large weight gains, this pattern is reversed and the effect is largest in the upper tail of the conditional birth-weight distribution.

pound gained is about 30 grams at the lowest quantiles and declines to only about 5 grams at the upper quantiles. This pattern of declining marginal effects is maintained for large weight gains, until we begin to consider extremely large weight gains, at which point the effect is reversed. For example, another pound gained by the mother who has already gained 50 pounds has only a 7-gram effect in the lower tail of the birth-weight distribution, and this increases to about 10 grams at the upper quantiles. The quadratic specification of the effect of mother's weight gain offers a striking example of how misleading the OLS estimates can be. Note that the OLS estimates strongly suggest that the effect is linear with an essentially negligible quadratic effect. However, the quantile regression estimates give a very different picture, one in which the quadratic effect of the weight gain is very significant except where it crosses the zero axis at about $\tau = 0.33$.

1.6 CONCLUSION

Although much more could be drawn out of the foregoing analyses, it may suffice to conclude here with the comment that the quantile regression results offer a much richer, more focused view of the applications than could be achieved by looking exclusively at conditional mean models. In particular, it provides a way to explore sources of heterogeneity in the response that are associated with the covariates. The next chapter delves more deeply into the interpretation of quantile regression models and their estimation.