Quantile Regression†

Roger Koenker
Department of Economics
University of Illinois, Urbana-Champaign
Champaign, IL 61820 USA

and

Stephen Portnoy
Department of Statistics
University of Illinois, Urbana-Champaign
Champaign, IL 61820

June 1999

† This material constitutes very preliminary versions of four chapters from a proposed monograph on quantile regression. Portions of the present version were prepared for a minicourse offered by Roger Koenker in February 1997 at Campos dos Jordão, Sao Paulo, Brazil. All rights are reserved by the above authors. Comments on any aspect of the chapters would be most welcome and could be sent to `rkoenker@uiuc.edu` or to the postal addresses provided above.

Contents

Preface

Francis Galton in a famous passage defending the "charms of statistics" against its many detractors, chided his statistical colleagues

> [who] limited their inquiries to Averages, and do not seem to revel in more comprehensive views. Their souls seem as dull to the charm of variety as that of a native of one of our flat English counties, whose retrospect of Switzerland was that, if the mountains could be thrown into its lakes, two nuisances would be got rid of at once. [*Natural Inheritance,* p. 62]

It is the fundamental task of statistics to bring order out of the diversity, at times the apparent chaos, of scientific observation. And this task is often very effectively accomplished by exploring how *averages* of certain variables depend upon the values of other "conditioning" variables. The method of least squares which pervades statistics is admirably suited for this purpose. And yet, like Galton, one may question whether the exclusive focus on conditional mean relations among variables, ignores some "charm of variety" in matters statistical.

As residents of one of the flattest American counties, our recollections of Switzerland and its attractive nuisances are quite different from the ones described by Galton. Not only the Swiss landscape, but many of its distinguished statisticians have in recent years made us more aware of the charms and perils of the diversity of observations, and the consequences of too-blindly limiting our inquiry to averages.

Quantile regression offers the opportunity for a more complete view of the statistical landscape and the relationships among stochastic variables. The simple expedient of replacing the familiar notions of sorting and ranking observations in the most elementary one-sample context by *optimization* enables us to extend these ideas to a much broader class of statistical models. Just as minimizing sums of squares permits us to estimate a wide variety of models for conditional mean functions, minimizing a simple asymmetric version of absolute errors yields estimates for conditional quantile functions. For linear parametric models computation is greatly facilitated by the reformulation of our optimization problem as a parametric linear program. Formal duality results for linear programs yields a new approach to rank statistics and rank-based inference for linear models.

We hope that this book can provide a comprehensive introduction to quantile regression methods, and that it will serve to stimulate others to explore and further develop these ideas in their own research. Since ultimately the test of any statistical method must be its success in applications, we have sought to illustrate the application of quantile regression methods throughout the book wherever possible. Formal mathematical development, which in our view plays an indispensable role in clarifying precise conditions under which statistical methods can be expected to perform reliably and efficiently, are generally downplayed, but Chapter 4 is devoted to an exposition of the basic asymptotic theory of quantile regression, and other chapters include technical appendices which provide further mathematical details.

Statistical software for quantile regression is now widely available in many well-known statistical packages including S, SAS, Shazam, and Stata. Fellow S users will undoubtedly recognize by our graphics that we are S-ophiles and much of the software described here may be obtained for S from `http://www.econ.uiuc.edu`.

We are grateful to many colleagues who have, over the years, collaborated with us on various aspects of the work described here. Gib Bassett whose Ph.d. thesis on l_1-regression served as a springboard for much of the subsequent work in this area has been a continuing source of insight and enthusiastic support. Jana Jurečková , who took an early interest in this line of research, has made an enormous contribution to the subject especially in developing the close connection between quantile regression ideas and rank statistics in work with Cornelius Guttenbrunner. Independent work by David Ruppert, Ray Carroll, Alan Welsh, Jim Powell, Gary Chamberlain, Probal Chaudhuri, and Moshe Buchinsky among others, has also played a crucial role in the development of these ideas. We have also collaborated in recent years with a number of our students who have contributed significantly to the development of these ideas including: José Machado, Pin Ng, Lin-An Chen, Liji Shen, Qing Zhou, Quanshui Zhao, Beum-Jo Park, and M. N. Hasan. Research support by the NSF and the Center for Advanced Study at the University of Illinois has also been deeply appreciated.

Urbana, June, 1999

CHAPTER 1

Introduction

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 of x's. 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.

Our 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. We 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. Secondly, 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 our 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 we describe methods to accomplish this. The basic ideas go back to the earliest work on regression by Boscovich in the mid-18th century, and to Edgeworth at the end of the 19th century.

2. The first regression: an 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 we 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 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. To estimate the extent of this effect the five measurements appearing in Table **??** had been made. Each represented a rather arduous direct measurement of the arc-length of 1° of latitude at 5 quite dispersed points – from Quito on the equator to Lapland at $66^\circ 19'$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.

Location	latitude	$\sin^2$ (latitude)	arc-length
Quito	$0^o\ 0'$	0	56751
Cape of Good Hope	$33^o\ 18'$	0.2987	57037
Rome	$42^o\ 59'$	0.4648	56979
Paris	$49^o\ 23'$	0.5762	57074
Lapland	$66^o\ 19'$	0.8386	57422

TABLE 1.1. Boscovich Ellipticity Data

For short arcs the approximation

$$y = a + b \sin^2 \lambda \tag{1.2.1}$$

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 the excedence 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 10 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 10 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 1/255.

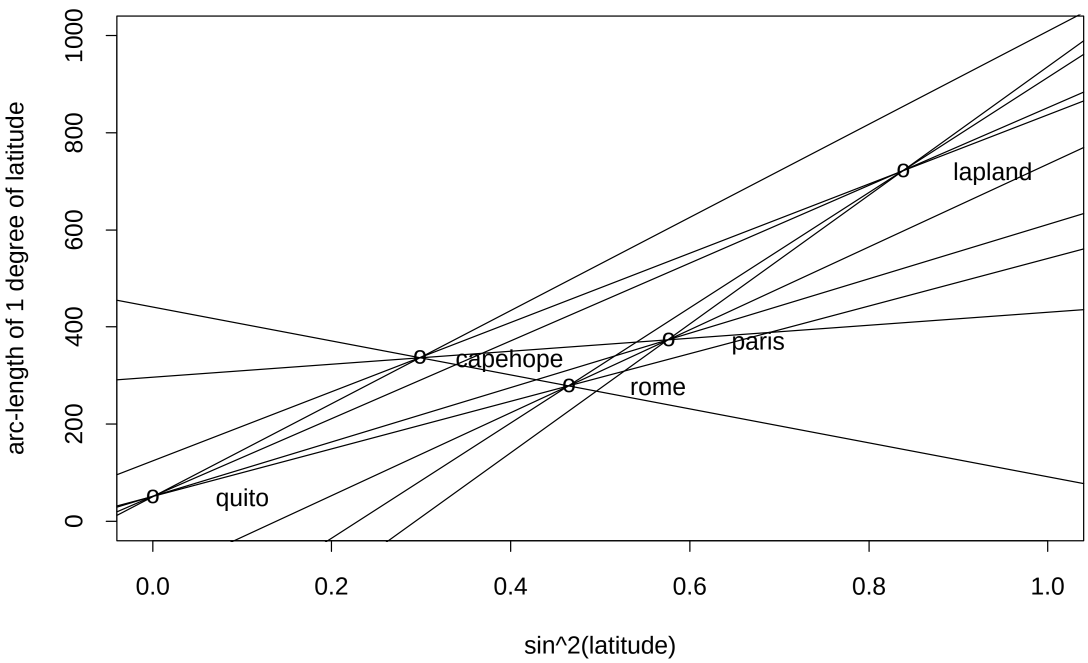

FIGURE 1.1. Boscovich Ellipticity Example. Boscovich computed all the pairwise slopes and initially reported a trimmed mean of the pairwise slopes as a point estimate of the earth's ellipticity. Arc length is measured as the excess over 56,700 toise per degree where one toise $\approx$ 6.39 feet, or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

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

10 pairs, and write

$$b(h) = X(h)^{-1}y(h) \tag{1.2.2}$$

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

$$X(h) = \begin{pmatrix} 1 & x_i \\ 1 & x_j \end{pmatrix} \quad y(h) = \begin{pmatrix} y_i \\ y_j \end{pmatrix}; \tag{1.2.3}$$

then we may write the least squares estimator as

$$\hat{b} = \sum_h w(h)b(h) \tag{1.2.4}$$

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 1.2.1 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 to compute the estimator which was remarkably simple. Having reduced the problem to regression through the origin with the aid of the 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

$$b_i = \frac{y_i - \bar{y}}{x_i - \bar{x}} \tag{1.2.5}$$

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

$$\sum_{i=1}^{j} w_{(i)} > \frac{1}{2} \sum_{i=1}^{n} w_{(i)} \tag{1.2.6}$$

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 F.Y. 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 work of Tukey in the 1940's, 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 analogues for regression of the other quantiles?" We begin to explore the answer to this question in the next section.

3. Quantiles, Ranks, and Optimization

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

$$F(x) = P(X \leq x) \tag{1.3.1}$$

while for any $0 < \tau < 1$

$$F^{-1}(\tau) = inf\{x : F(x) \geq \tau\} \tag{1.3.2}$$

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 which 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

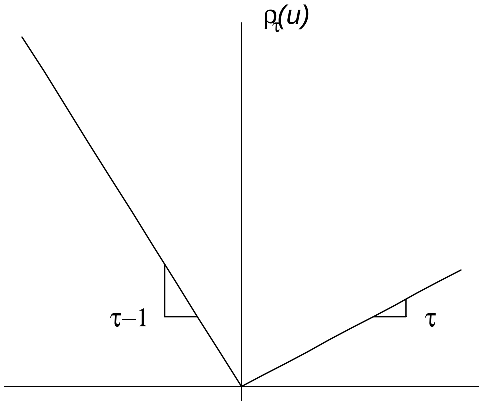

FIGURE 1.2. Quantile Regression ρ Function

described by the function

$$\rho_\tau(u) = u(\tau - I(u < 0)) \tag{1.3.3}$$

for some $\tau \in (0, 1)$, find $\hat{x}$ to minimize expected loss. This is a standard exercise in many decision theory texts e.g. Ferguson (1967, p. 51). We seek to minimize

$$E\rho_\tau(X - \hat{x}) = (\tau - 1)\int_{-\infty}^{\hat{x}}(x - \hat{x})dF(x) + \tau\int_{\hat{x}}^{\infty}(x - \hat{x})dF(x). \tag{1.3.4}$$

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

$$0 = (1 - \tau)\int_{-\infty}^{\hat{x}} dF(x) - \tau\int_{\hat{x}}^{\infty} dF(x) = F(\hat{x}) - \tau. \tag{1.3.5}$$

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 we may choose the smallest element – to adhere to the convention that the empirical quantile function be left-continuous.

It is natural that our 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,

$$F_n(x) = n^{-1} \sum_{i=1}^{n} I(X_i \leq x) \tag{1.3.6}$$

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

$$\int \rho_\tau(x - \hat{x}) dF_n(x) = n^{-1} \sum_{i=1}^{n} \rho_\tau(x_i - \hat{x}) = \min! \tag{1.3.7}$$

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, which 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 we do 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 we may now write as,

$$\min_{\xi \in \mathbf{R}} \sum_{i=1}^{n} \rho_\tau(y_i - \xi), \tag{1.3.8}$$

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

$$\min_{(\xi, u, v) \in \mathbb{R} \times \mathbb{R}_+^{2n}} \{\tau 1_n' u + (1 - \tau) 1_n' v | 1_n' \xi + u - v = y\} \tag{1.3.9}$$

where 1_n denotes an n-vector of ones. Clearly, in (1.3.9) 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 $\mathbb{R} \times \mathbb{R}_+^{2n}$ Many features of the solution are immediately apparent from this simple fact. For example, $\min\{u_i, v_i\}$ must be zero for all i, since otherwise, the objective function may be reduced without violating the constraint by shrinking such a pair toward zero. This is usually called complementary slackness in the terminology of linear programming. Indeed, for this same reason we can restrict attention to "basic solutions" of the form $\xi = y_i$ for some observation i. To see this consider Figure 1.3 which depicts the objective function (1.3.8) 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_i's. When ξ passes through one of these y_i's, the slope of the objective function changes by exactly 1 since a contribution of $\tau - 1$ is replaced by τ or vice-versa.

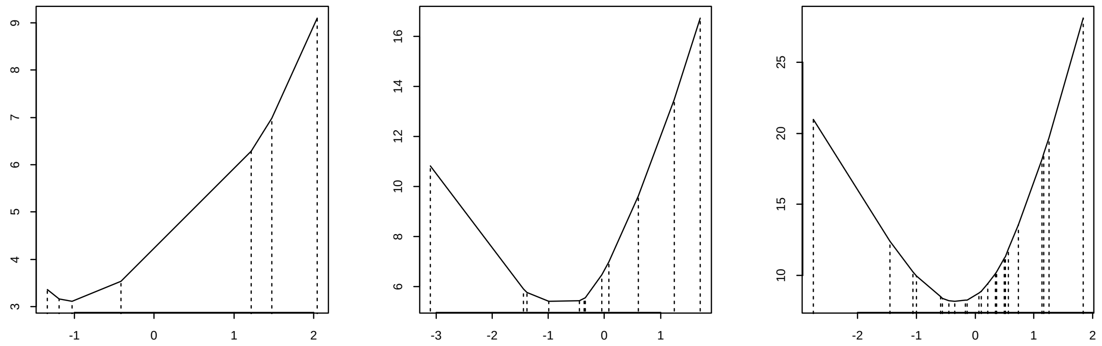

FIGURE 1.3. Quantile objective function with random data. The figure illustrates the objective function for the optimizaiton problem defining the ordinary $\tau = 1/3$ quantile, for three different random problems with y_i's drawn from the standard normal distribution, and sample sizes 7, 12, and 23. The vertical dotted lines indicate the position of the observations in each sample. Note that since 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 order statistics.

Optimality holds if the right and left derivatives,

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

and

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

are both nonnegative, that is if $n\tau$ lie in the closed interval $[N^-, N^+]$ where

$$N^{\pm} = \sharp\{y_i < \xi \pm 0\}.$$

When $n\tau$ is not an integer there is a unique value of ξ which satisfies this condition. Barring ties in the y_i's, 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.

The duality connecting the sample quantiles and the ranks of the order statistics is further clarified through the formal duality of linear programming. The dual program to (1.3.9) is

$$\max\{y'a | 1_n' a = (1-\tau)n, \quad a \in [0,1]^n\} \tag{1.3.10}$$

While the former (primal) problem may be viewed as generating the sample quantiles, the dual problem may be seen to generate the order statistics or perhaps more precisely the *ranks* of the observations. What does the solution $\hat{a}(\tau)$ look like for the simple dual problem (1.3.10)?

Clearly at $\tau = 0$ feasibility requires that all of the $\hat{a}_i(0) = 1$, and similarly at $\tau = 1$, $\hat{a}_i(1) = 0$ for all $i = 1, ..., n$. Starting at $\tau = 0$, consider increasing τ. How should we modify the $\hat{a}_i$'s? Initially we should focus on $y_{(1)} = \min\{y_1, ..., y_n\}$ since decreasing its weight has the least impact on the sum $y'a$. Thus, if $y_{(1)} = y_j$ then as τ increases a_j must decrease linearly to satisfy the equality constraint. This is fine until τ reaches $1/n$, but at this point a_j has been driven to 0 and it is allowed to go no further. Now $y_{(2)}$, being the smallest available response, is gradually downweighted, and the process continues until all the observations have achieved weight $a_i = 0$, at which point $\tau = 1$. The functions $\hat{a}_i(\tau)$ take the form

$$\hat{a}_i(\tau) = \begin{cases} 1 & \tau \le (R_i - 1)/n \\ R_i - \tau n & (R_i - 1)/n < \tau \le R_i/n \\ 0 & R_i/n < \tau \end{cases} \tag{1.3.11}$$

where R_i is the rank of y_i among $\{y_1, ..., y_n\}$. These functions coincide exactly with the rankscore generating functions introduced by Hájek and Šidák (1967, V.3.5). They provide a natural approach to the construction of the ranks and test statistics based upon the ranks. Note for example that integrating with respect to the Wilcoxon score function $\varphi(t) = 1$ we have

$$\int_0^1 \hat{a}_i(t) d\varphi(t) = (R_i - 1/2)/n.$$

We will see that this approach to ranks generalizes naturally to the linear model, yielding an elegant generalization of rank tests for the linear model.

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

$$\min_{\mu \in \mathbb{R}} \sum_{i=1}^{n} (y_i - \mu)^2 \tag{1.4.1}$$

suggests that if we are willing to express the *conditional* mean of y given x as $\mu(x) = x'\beta$ then we might estimate β by solving

$$\min_{\beta \in \mathbb{R}} \sum_{i=1}^{n} (y_i - x_i'\beta)^2. \tag{1.4.2}$$

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

$$\min_{\alpha \in \mathbb{R}} \sum_{i=1}^{n} \rho_\tau(y_i - \alpha), \tag{1.4.3}$$

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

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - x_i'\beta). \tag{1.4.4}$$

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

The quantile regression problem (1.4.4) may be reformulated as a linear program as in (1.3.9)

$$\min_{(\beta,u,v) \in \mathbb{R}^p \times \mathbb{R}^{2n}_+} \{\tau 1_n' u + (1-\tau) 1_n' v | X\beta + u - v = y\} \tag{1.4.5}$$

where X now denotes the usual n by p regression design matrix. Again, we see that 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 1.4 we illustrate all of the *distinct* regression quantile solutions for this data. Of the 10 lines passing through pairs of points in Figure 1.1, quantile regression selects only 4. Solving (1.4.4) for any τ in the interval $(0, .21)$ yields as a unique solution the line passing through Quito and Rome. At $\tau = .21$ the solution jumps and throughout the interval $(.21, .48)$ we have the solution characterized by the line passing through Quito and Paris. The process continues until we get to $\tau = .78$, where the solution through Lapland and the Cape of Good Hope prevails up to $\tau = 1$.

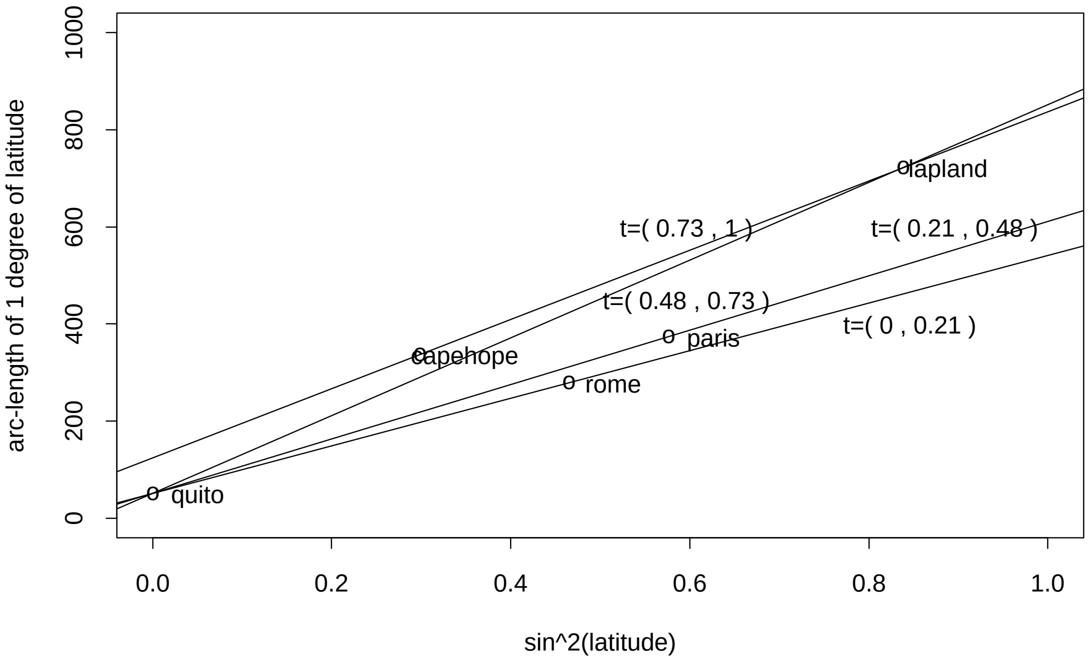

FIGURE 1.4. Regression Quantiles for Boscovich Ellipticity Example. Only 4 of the full 10 pairs of points form quantile regression solutions. The subintervals of $(0, 1)$ for which each pair solves (1.4.4) are given in the figure.

In contrast to the ordinary sample quantiles which 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 upon the configuration of the design as well as the realized values of the response variable. *Pairs of points now play the role 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 (1.4.4) rotating as τ increases, we may visualize the solutions of (1.4.4) 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 the position of the Lapland observation upwards this would not affect the solutions illustrated in the figure for any τ in the interval $(0, .48)$.

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

Consider an artificial sample in which we have a simple bivariate regression model with independent and identically distributed errors,

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

so the quantile functions of y_i are

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

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, $(\beta_0 + F^{-1}(\tau), \beta_1)'$

In Figure 1.5 we illustrate data and several fitted regression quantile lines from such a model. The dots indicate 60 observations generated from the iid error model with F selected to be Gaussian. The dotted lines represent the *true* $\{.05, .25, .50, .75, .95\}$ conditional quantile lines. The solid 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 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 1.5 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.

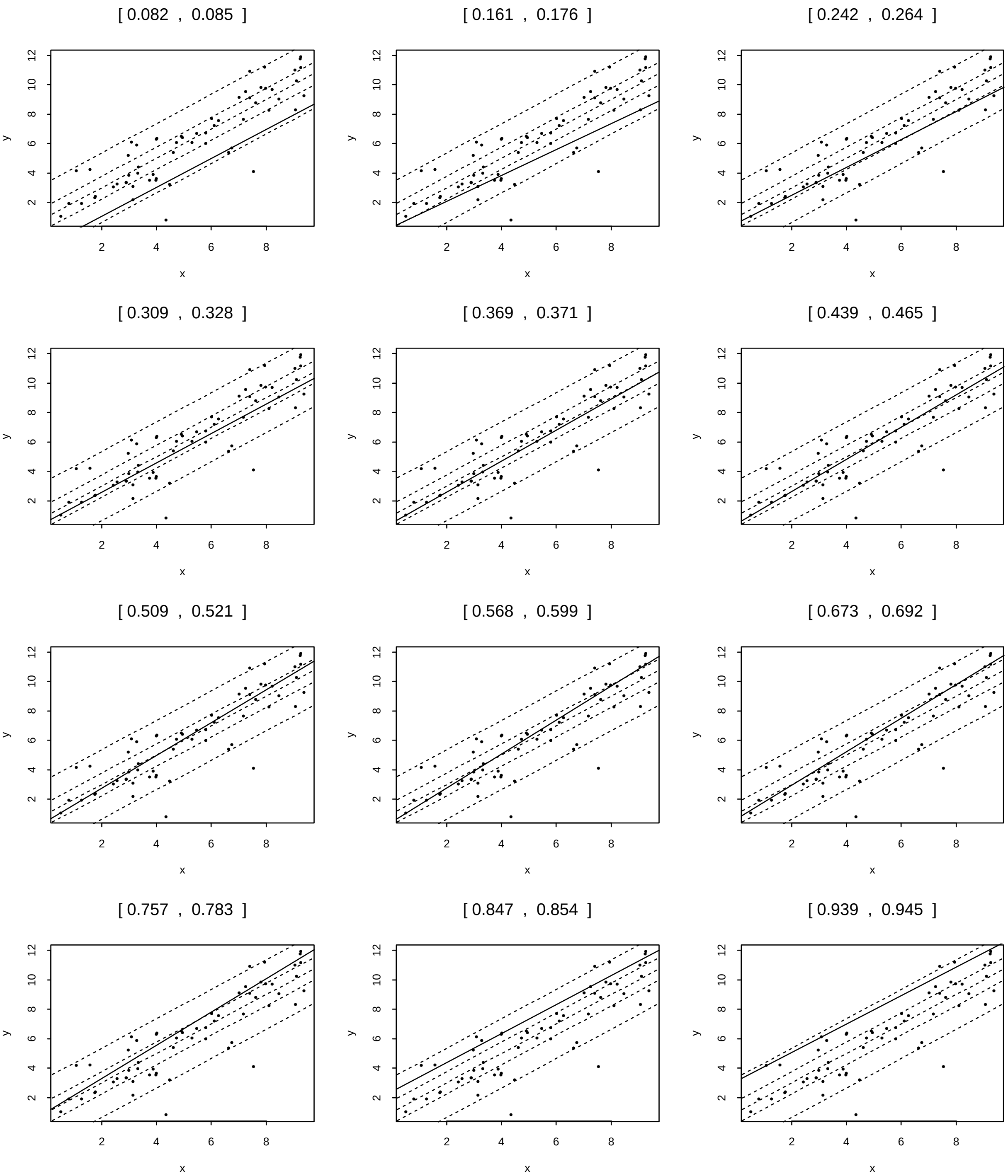

FIGURE 1.5. Regression Quantiles for iid-Error Bivariate Regression

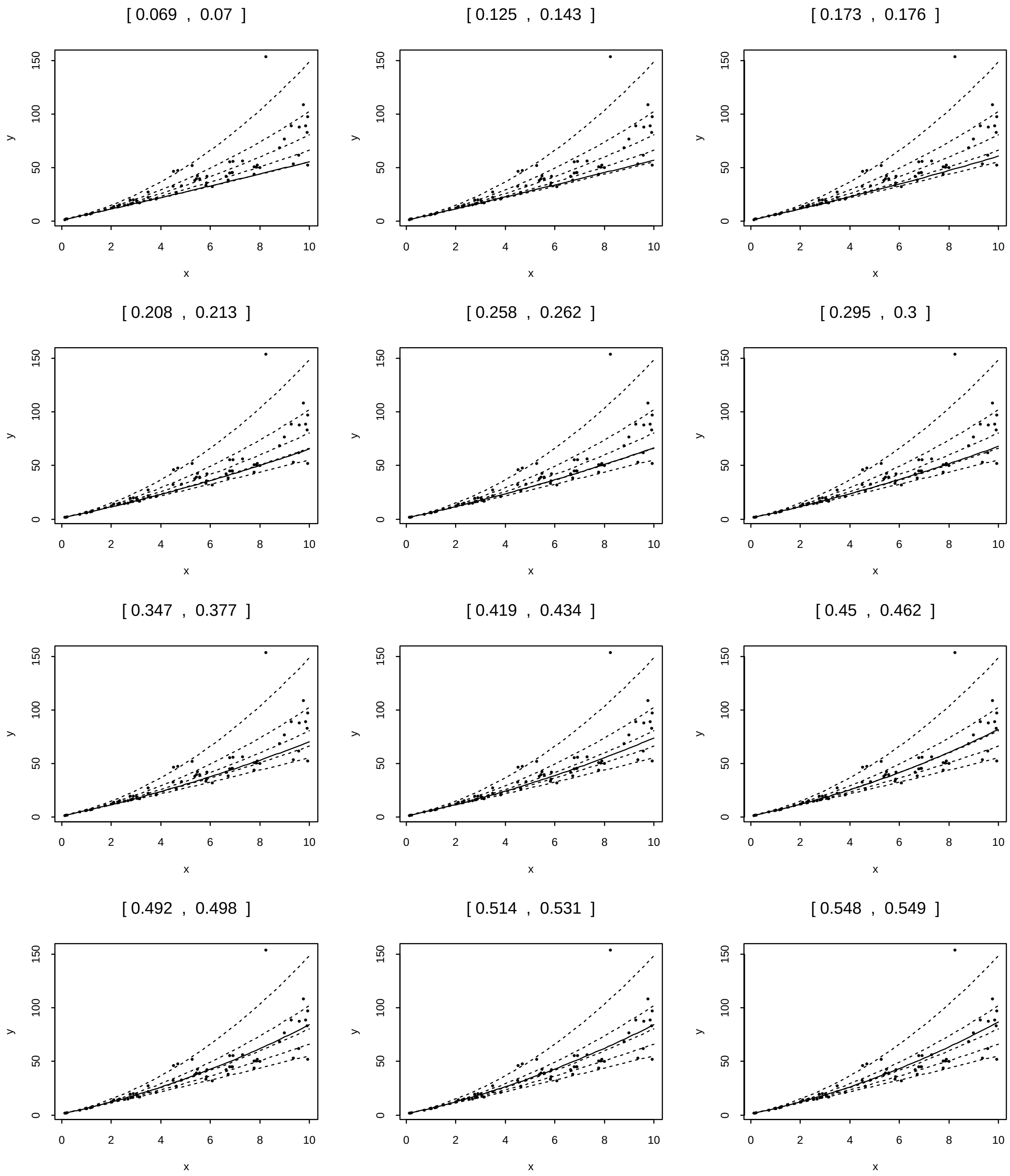

FIGURE 1.6. Regression Quantiles for Heteroscedastic Bivariate Regression

In the next figure 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 quantile functions of y_i are now easily seen to be

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

and can be consistently estimated by minimizing

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

so $\hat{\beta}(\tau)$ converges to $(\beta_0, \beta_1, \gamma F^{-1}(\tau)))$. Figure 1.6 illustrates an example of this form. Again, the *population* conditional quantile functions are shown as dotted lines with the observed sample of 60 points superimposed and a sequence of estimated quantile regression curves appearing as the solid lines. The estimated quantile regression curves provide a direct empirical analogue for the family of conditional quantile functions in the population.

5. Bibliographic Notes

On the early history of regression and the contribution of Boscovitch in particular, Stigler(1986) is the definitive introduction. Smith(1986) contains a detailed account of the development of geodesy, focusing attention on the efforts which culminated in the data appearing in Table **??**. Sheynin(1973) and Harter(1974) also offer useful accounts of the early history of regression. Edgeworth's (1887,1888) contributions to the development of median regression were crucial to the continuing interest in these methods in economics. Only with the emergence of the simplex algorithm for linear programming in the late 1940's did ℓ_1 methods become practical on a large scale. Papers by Charnes, Cooper and Ferguson (1955), Wagner (1959) and others provided a foundation for modern implementations, such as Barrodale and Roberts (1974) and Bartels and Conn (1980).

CHAPTER 2

Fundamentals of Quantile Regression

In this Chapter we seek to provide a basic conceptual guide to quantile regression, illustrating the ideas with a number of examples and stressing various aspects of the interpretation of quantile regression. We will begin by illustrating the methods in two simple examples. The first example involves earnings-experience profiles for academic salaries in statistics. The second is a somewhat more complicated analysis of determinants of student course evaluation responses. The bivariate nature the first example allows us to develop a close link between quantile regression methods and the well-known boxplot. The basic objective of Tukey's boxplot is to provide an efficient graphical summary of the main features of an entire univariate distribution. By aligning boxplots for neighboring conditional distributions, one can achieve – for bivariate regression – the "completed picture" of regression alluded to by Mosteller and Tukey, quoted in our introductory paragraph.

Quantile regression permits us to extend this link to more complicated situations such as our second example where there are several covariates and offers an extensive menu of possible formal inference strategies. The basic insight is exceedingly simple: the notions of ordering, sorting and ranking traditionally associated with univariate statistical analysis can be extended to regression by viewing these procedures as the outcome of an elementary optimization process. The resulting optimization problems not only yield a convenient computational strategy for quantile regression, but they illuminate many important properties of the methods in a particularly convenient way. In subsequent chapters these themes will be developed more fully.

1. Quantile Regression Treatment Effects

The simplest formulation of regression is the classical two-sample treatment-control model. We will begin by reconsidering a general model of two-sample treatment response introduced by Lehmann and Doksum in the 1970's. This model provides a natural introduction to the interpretation of quantile regression models in more general settings.

Lehmann (1974) proposed the following model of treatment response:

> "Suppose the treatment adds the amount $\Delta(x)$ when the response of the untreated subject would be x. Then the distribution G of the treatment

> responses is that of the random variable $X+\Delta(X)$ where X is distributed according to F."

Special cases obviously include the location shift model $\Delta(X) = \Delta_0$, and the scale shift model $\Delta(x) = \Delta_0 X$, If the treatment is beneficial in the sense that,

$$\Delta(x) \geq 0 \quad \text{for all } x$$

then the distribution of treatment responses, G, is stochastically larger than the distribution of control responses, F.

Doksum (1974) shows that if we define $\Delta(x)$ as the "horizontal distance" between F and G at x, so

$$F(x) = G(x + \Delta(x))$$

then $\Delta(x)$ is uniquely defined and can be expressed as

$$\Delta(x) = G^{-1}(F(x)) - x. \tag{2.1.1}$$

Thus, on changing variables so $\tau = F(x)$ we have the *quantile treatment effect*,

$$\delta(\tau) = \Delta(F^{-1}(\tau)) = G^{-1}(\tau) - F^{-1}(\tau).$$

Doksum provides a thorough axiomatic analysis of this formulation of treatment response.

In the two sample setting the quantile treatment effect is naturally estimable by

$$\hat{\delta}(\tau) = \hat{G}_n^{-1}(\tau) - \hat{F}_m^{-1}(\tau)$$

where G_n and F_m denote the empirical distribution functions of the treatment and control observations, based on n and m observations respectively. If we formulate the quantile regression model for the binary treatment problem as,

$$Q_{Y_i}(\tau|D_i) = \alpha(\tau) + \delta(\tau)D_i \tag{2.1.2}$$

where D_i denotes the treatment indicator, with $D_i = 1$ indicating treatment, $D_i = 0$, control, then we may estimate the quantile treatment effect directly.

To illustrate, Doksum (1974) reconsiders a 1960 study by Bjerkedal of the effect of injections of tubercle bacilli on guinea pigs. Survival times, following injection, were recorded (in days) for 107 control subjects and 60 treatments subjects. Of the control subjects, 42 lived longer than the experimental censoring threshold of 736 days. None of the treatment subjects survived more than 600 days. In Figure **??** we plot the estimated functions $\hat{\alpha}(\tau)$ and $\hat{\delta}(\tau)$. The plots are "censored" beyond $\tau = .6$ due to the censoring of the survival times of the control subjects. Confidence bands are indicated by the lightly shaded regions. The treatment effect in this example, depicted in right panel, is evidently neither a location shift, which would appear as a horizontal line, or a scale shift, which would appear as a proportional dilation of the "control effect" depicted in the left (intercept) panel. Here, animals receiving the treatment injection of bacilli appear to benefit from the treatment in the lower tail of the distribution, while the upper tail the treatment shows a strongly significantly adverse effect on survival. The treatment thus appears to have an advantageous effect on survival in the short-run, but seems very disadvantageous in the longer run.

Doksum suggests that we may wish interpret control subjects in terms of a latent characteristic. Control subjects may be called frail if they are prone to die at an early age, and robust if prone to die at an advanced age. This characteristic is thus implicitly indexed by τ, the quantile of the survival distribution at which the subject would appear if untreated, i.e., $(Y_i|D_i = 0) = \alpha(\tau)$. And the treatment, under the Lehmann-Doksum model, is assumed to alter the subjects control response, $\alpha(\tau)$, making it $\alpha(\tau)+\delta(\tau)$ under the treatment. If the latent characteristic, say, propensity for longevity, were observable *ex ante*, then we might view the treatment effect $\delta(\tau)$ as an explicit interaction with this observable variable. However, in the absence of such an observable variable, the quantile treatment effect may be regarded as a natural measure of the treatment response. Of course, there is no way of knowing whether the treatment actually operates in the manner proscribed by $\delta(\tau)$. In fact, the treatment may miraculously make weak subject especially robust, and turn the strong into jello. All we can observe from experimental evidence, however, is the difference in the two survival distributions, and it is natural to associate the treatment effect with the difference in the corresponding quantiles of the two distributions. This is what the quantile treatment effect does.

When the treatment variable takes more than two values, this interpretation requires only slight adaptation. In the case of p distinct treatments, we can write

$$Q_{Y_i}(\tau|D_{ij}) = \alpha(\tau) + \sum_{j=1}^{p} \delta_j(\tau) D_{ij}$$

where $D_{ij} = 1$ if observation i received the jth treatment and $D_{ij} = 0$ otherwise. Here $\delta_j(\tau)$ constitutes the quantile treatment effect connecting the distribution of control responses to the responses of subjects under treatment j. If the treatment is continuous as, for example, in dose-response studies, then it is natural to consider the assumption that the effect is linear, and write,

$$Q_{Y_i}(\tau|x_i) = \alpha(\tau) + \beta(\tau)x_i.$$

We assume thereby that the treatment effect, $\beta(\tau)$, of changing x from x_0 to x_0+1 is the same as the treatment effect of an alteration of x from x_1 to $x_1 + 1$. Interpreted in this fashion the quantile treatment effect offers a natural extension to continuously varying treatments of the Lehmann-Doksum formulation for the discrete case.

In economics, a common application of this type involves investigations of the effect of years of schooling on observed wages. In this literature, it is common to identify latent components of wage determination with unobserved characteristics such as "spunk" or "ability" and thus these terms play the same role as "propensity

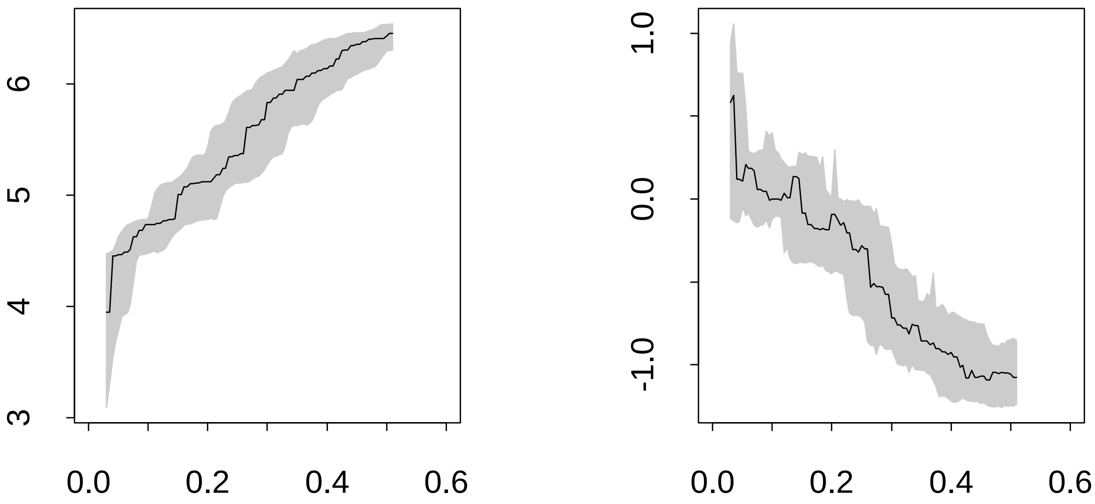

FIGURE 2.1. This figure illustrates quantile regression results for the guinea pig example analyzed in Doksum (1974), and taken from Bjerkedal (1960). The fitted model (2.1.2) for log survival times is based on a sample of 107 controls and 60 treatment subjects injected with the tubercle bacilli. In the left panel we plot the function $\hat{\alpha}(\tau)$ representing the empirical quantiles of the log survival time distribution for the control sample. In the right panel we depict, $\hat{\delta}(\tau)$, the estimated quantile treatment effect. In this simple two sample setting, the quantile treatment effect, $\hat{\delta}(\tau)$, is just the horizontal distance between the empirical cdfs of the control and treatment samples. Note that the treatment has a positive effect on survival in the left tail, thus improving survival prospects for the weakest subjects. But the treatment has a very adverse effect on survival times in the right tail, dramatically reducing survival times for the stronger subjects. The lightly shaded region illustrates a 90% confidence band for the estimated effects.

for longevity" in survival examples. The quantile treatment effect, $\beta(\tau)$, may be interpreted as an interaction effect between unobserved "ability" and the level of education. This interpretation has been recently explored in work of Arias, Hallock and Sosa (1999) in a study of the earnings of identical twins.

Finally, it may be noted that the quantile treatment effect (2.1.1), is intimately tied to the traditional two-sample QQ-plot which has a long history as a graphical diagnostic device. Note that the function $\hat{\Delta}(x) = G_n^{-1}(F_m(x)) - x$ is exactly what is plotted in the traditional two sample QQ-plot. The connection between the Lehmann-Doksum treatment effect and the QQ-plot is explored in Doksum and Sievers (1976), Nair (1982) for the the p-sample problem. Quantile regression may be as a means of extending the two-sample QQ plot and related methods to general regression settings with continuous covariates. We will return to this observation and its implications for inference in Chapter 3.

2. Two Examples

2.1. Salaries vs Experience. In Figure 2.1 we illustrate this with results of the 1995 survey of academic salaries in statistics conducted by the American Statistical Association. The figure is based on data from 99 departments in U.S. colleges and universities on 370 full professors of statistics. The data is grouped into 3 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 is also a pronounced tendency for the dispersion of the salary distribution to increase with experience. None of these findings are 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 diminish the viability of the boxplot approach which relies upon adequate sample sizes for each of the groups, or cells, represented by the boxes. What could we do if

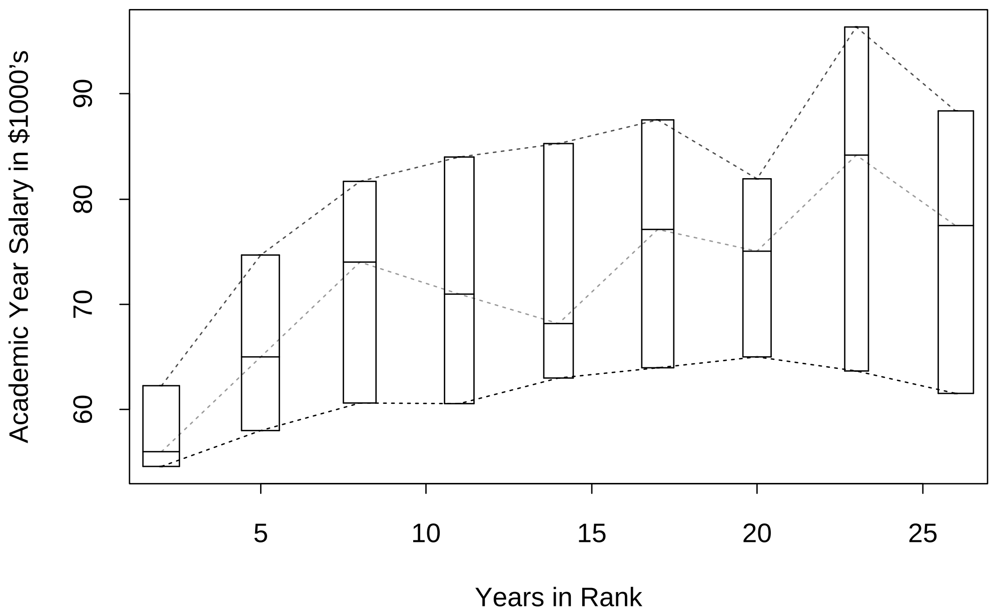

FIGURE 2.2. Boxplots of 1995 ASA Academic Salary Survey for Full Professors of Statistics in U.S. Colleges and Universities.

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 which 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 $x_i \le \text{median}\{x_j\}$ 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 8.2. Based on the the 1973-74 ASA data for full professor salaries, he obtains the estimates reported in

Table 2.1 Since the estimated slope parameters, $\hat{\beta}$ increase with the quantile, these

Quantile τ	Initial Professorial Salary $\hat{\alpha}$	Annual Increment $\hat{\beta}$
0.75	21500	625
0.50	20000	485
0.25	18800	300

TABLE 2.1. 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.

estimates reflect the same increasing dispersion, or heteroscedasticity, that we saw in the boxplots of Figure 2.1 for the more recent salary data. In this case, with so little data, it does not seem prudent to venture an opinion about the curvature of the salary profile.

2.2. Student Course Evaluations and Class Size. Our second example illustrates several advantages of the optimization approach to quantile regression introduced in the previous chapter. The data consists of mean course evaluation scores for 1482 courses offered by large U.S. 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 the nature of particular types of courses.

In Figure 2.3 we illustrate the data for this example and plot five estimated quantile regression curves. These curves are specified as quadratic in the number of CEQ respondents which we take 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 which takes the value 1 for graduate courses, and 0 for undergraduate courses. The model may thus be written as,

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

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

$$\min_{b \in \mathbb{R}} \sum_{i=1}^{n} \rho_\tau(y_i - x_i'b). \tag{2.2.1}$$

The estimated quantile regression parameters and their confidence intervals are given in Table 2.2 Details on the construction of the confidence intervals appear in in the next chapter

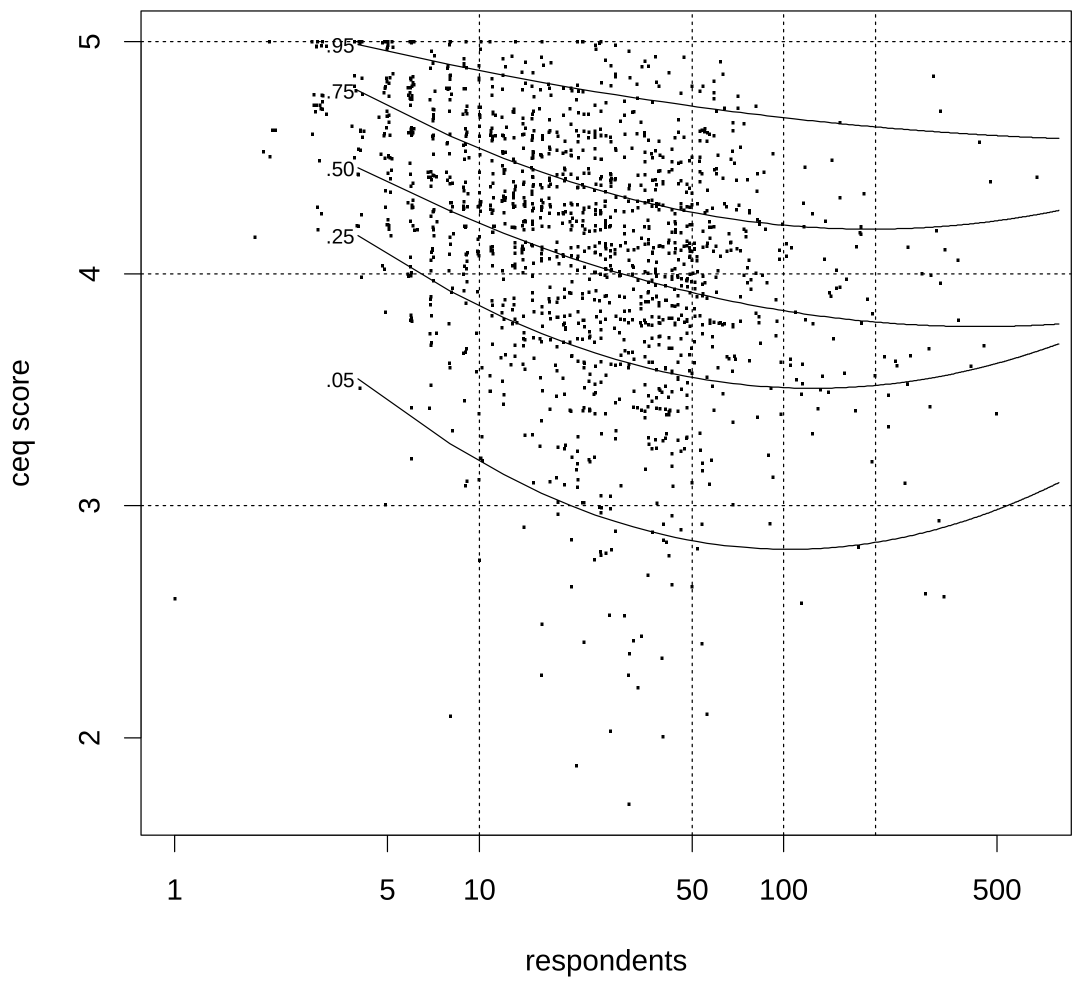

FIGURE 2.3. Course Evaluation Scores: Solid lines indicate estimated quantiles of CEQ response for an undergraduate course in 1992 as a function the class size measured by number of CEQ respondents.

From the table it can be seen that there is some evidence for a downward trend in CEQ scores for the lower quantiles, on the order of .01 to .02 rating points per year, but no evidence of a trend in the upper tail of the ratings distribution. Our tentative conclusion from this is that ornery students are getting ornerier. Graduate courses

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)

TABLE 2.2. Quantile regression estimates for a model of student course evaluation scores. Numbers in parentheses give a 95% confidence interval for each reported coeficient.

have a fairly consistent tendency to be rated higher by about .10 rating points than undergraduate courses.

In order to plot the curves illustrated in Figure 2.3 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 may also be partially attributed to a departmental policy of trying to allocate its best teachers to the larger courses.

We could probably agree that the dotted curves connecting the boxplot salary quartiles of Figure 2.1 appear somewhat undersmoothed. A parametric model for the conditional quartiles might improve the appearance of the plot, if we could agree on a transformation which would adequately capture the curvature of the salary profile. One attempt to do this is illustrated in Figure 2.2 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 2.2 have been estimated by median (ℓ_1) regression using only the respective grouped quartile data. (The individual data collected by the ASA is 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 our example, the first quartile of the salary distribution has an estimated growth rate of 7.3% per year of tenure, while the median and the upper quartile grow at 14 and 13 percent respectively. As for

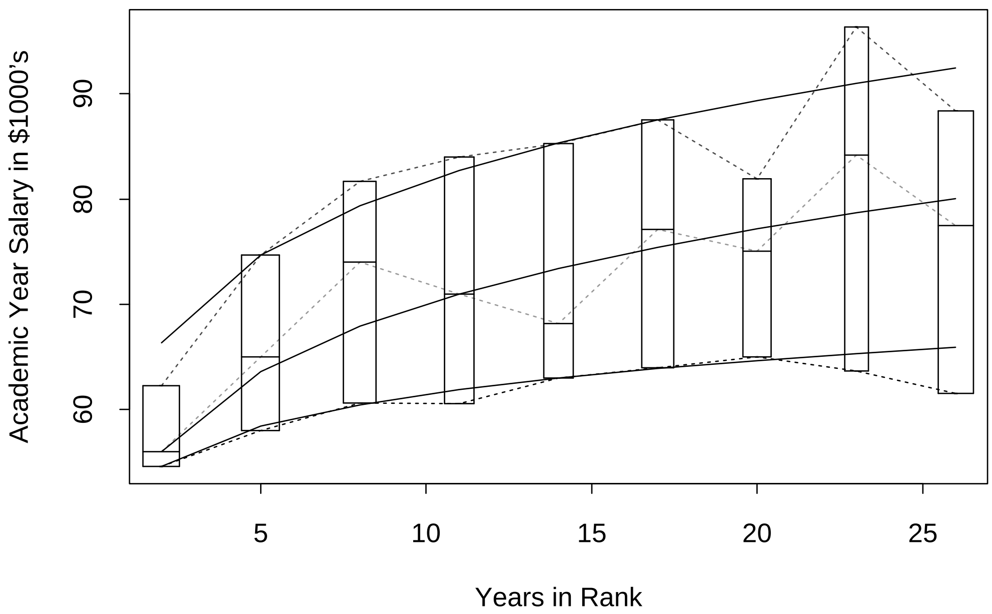

FIGURE 2.4. Boxplots of 1995 ASA Academic Salary Survey for Full Professors of Statistics in U.S. Colleges and Universities.

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 also 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,

$$y_i = x_i\beta + (x_i'\gamma)u_i$$

with $\{u_i\}$ iid from F. In this case the coefficients of the τth quantile regression, $\hat{\beta}(\tau)$ converge to $\beta + \gamma F_u^{-1}(\tau)$, 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.

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, .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 $[1, 5]$, were 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 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 2.3. 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.

3. How does quantile regression work?

Much of our intuition about how ordinary regression "works" comes from the geometry of least squares projection. The idea of minimizing the Euclidean distance $\| y - \hat{y} \|$ over all $\hat{y}$ in the linear span of the columns of X is very appealing. We may just imagine blowing up a beach ball centered at y until it touches the subspace spanned by X. Replacing Euclidean beach balls by polyhedral diamonds of the ρ_τ-distance,

$$d_\tau(y, \hat{y}) = \sum_{i=1}^{n} \rho_\tau(y_i - \hat{y}_i),$$

raises some new problems, but many nice features and insights persist. We do not obtain an elegant "closed form" solution like

$$\hat{y} = X(X'X)^{-1}X'y,$$

but the algorithm which leads us to the quantile regression estimates is really no more esoteric than say, the Householder transformations which gives a QR decomposition of X, and lead eventually to the "closed form" least squares estimate.

To minimize

$$\| y - \hat{y}(\beta) \|^2 = (y - X\beta)'(y - X\beta)$$

we differentiate to obtain the "normal equations"

$$\nabla_\beta ||y - \hat{y}(\beta)|| = X'(y - X\beta) = 0$$

and solve for $\hat{\beta}$. In quantile regression we proceed likewise. We differentiate,

$$R(\beta) = d_\tau(y, \hat{y}(\beta)) = \sum_{i=1}^{n} \rho_\tau(y_i - x_i\beta),$$

but recognizing that these derivatives may depend upon the *direction*, when some residuals are zero, we consider the directional derivatives,

$$\begin{aligned}
\nabla R(\beta, w) &\equiv \frac{d}{dt} R(\beta + tw) \mid_{t=0} \\
&= \frac{d}{dt} \sum_{i=1}^{n} u_i(\beta + tw)[\tau - I(u_i(\beta + tw) < 0)] \mid_{t=0} \\
&= -\sum \psi^*(y_i - x_i'\beta, -x_i'w)x_i'w
\end{aligned}$$

where

$$\psi^*(u, v) = \begin{cases} \tau - I(u < 0) & \text{if } u \neq 0 \\ \tau - I(v < 0) & \text{if } u = 0. \end{cases}$$

If $\nabla R(\hat{\beta}, w) \geq 0$ for all $w \in \mathbb{R}^p$ with $\| w \| = 1$, then $\hat{\beta}$ minimizes $R(\beta)$. This is a natural generalization of simply setting $\nabla R(\beta) = 0$ when R is smooth.

One important feature of $\hat{\beta}(\tau)$ that is immediately apparent from the geometry of the problem of finding the point $\hat{y} = X\hat{\beta}(\tau)$ closest to y in d_τ-distance is that the choice should keep as many coordinates of the residual vector $u(\beta) = y - X\beta$ equal to zero as possible. If we are estimating p parameters, i.e., $\beta \in \mathbb{R}^p$, then we usually can not hope to have more than p zero $u(\beta)$-coordinates, but there is no reason to tolerate fewer than p zero coordinates either. This is an immediate consequence of the piecewise linearity of the objective function in the residuals, and the polyhedral nature of the constraint set. Just as our search for the median leads us to a unique middle observation or a pair of adjacent middle observations either of which serve to minimize the sum of absolute residuals, in quantile regression we are led to seek subsets of p-observations which will serve to characterize the solution. We have already commented on this feature in our discussion of Figure 1.1. Remarkably, this feature was already noted by Gauss (1809) in his commentary on Boscovich's estimator.

In the terminology of linear programming the class of these p-element subsets are called basic solutions. They may be seen as extreme points of the polyhedral constraint set, vertices of the polyhedron which constitutes the constraint set. Minimizing a linear function with respect to a constraint set of this form *is* the task of linear programming. It is clear from the geometry that solutions must either occur uniquely, when the plane representing the objective function touches only a single vertex of the constraint set, or occur multiply when the objective function happens to come to rest on an edge or on entire facet of the constraint set. We will have more to say about non-uniqueness later, for now it will suffice to observe that even when it occurs the basic solutions play a fundamental role since any element of the solution set can be constructed as a linear combination of solution of this form. They necessarily constitute the vertices of the full solution set and thus must constitute a polyhedral, convex set themselves. This is already familiar from the elementary case of the median.

To facilitate our ability to consider these p-element subsets of observations we will introduce a bit more notation. Let $h \in \mathcal{H}$ index p-element subsets of the first n integers, $\mathcal{N} = \{1, 2, \ldots, n\}$, and $X(h)$ denote the submatrix of X with rows $\{x_i : i \in h\}$. Likewise, let $y(h)$ be a p-vector with coordinates $\{y_i : i \in h\}$. The complement of h with respect to $\mathcal{N}$, will be written as $\bar{h}$ and $X(\bar{h})$ and $y(\bar{h})$ may be defined accordingly.

With this notation in mind we can express any basic solution which passes through the points $\{(x_i, y_i), i \in h\}$ as

$$\beta(h) = X(h)^{-1} y(h)$$

presuming, of course, that the matrix $X(h)$ is nonsingular. There are obviously too many of these basic solutions, $\binom{n}{p} = O(n^p)$, in fact, to simply search through them like a drawer of old socks. What the simplex algorithm of linear programming finally provided was an efficient way to conduct this search, essentially by traversing from vertex to vertex of the constraint set always taking the direction of steepest descent.

3.1. The subgradient condition. We are now ready to introduce the basic optimality condition which characterizes the regression quantiles. We have seen that we can restrict attention to candidate solutions of the "basic" form

$$b(h) = X(h)^{-1} y(h).$$

For some $h, X(h)$ may be singular. This needn't worry us, we can restrict attention to $b(h)$ with $h \in \mathcal{H}^* = \{h \in \mathcal{H} : |X(h)| \neq 0\}$. We have also seen that our optimality condition entails verifying that the directional derivatives are non-negative in all directions. To check this at $b(h)$ we must consider

$$\nabla R(b(h), w) = -\sum_{i=1}^{n} \psi_\tau^*(y_i - x_i'b(h), -x_i'w)x_i'w$$

Reparameterizing the directions, so $v = X(h)w$, we have optimality if and only if,

$$0 \leq -\sum_{i=1}^{n} \psi_\tau^*(y_i - x_i'b(h), -x_iX(h)^{-1}v)x_i'X(h)^{-1}v$$

for all $v \in \mathbb{R}^p$. Now note that for $i \in h$, we have $e_i' = x_i'X(h)^{-1}$, the ith unit basis vector of $\mathbb{R}^p$, so we may rewrite this as

$$0 \leq -\sum_{i \in h} \psi_\tau^*(0, v_i)v_i - \xi'v = -\sum_{i \in h}(\tau - I(v_i < 0))v_i - \xi'v$$

where

$$\xi(v_i) = \sum_{i \in \bar{h}} \psi_\tau^*(y_i - x_ib(h), -x_iX(h)^{-1}v_i)x_i'X(h)^{-1}.$$

Finally, note that the space of "directions", $v \in \mathbb{R}^p$, are spanned by $v = \pm e_k, k = 1, \ldots, p$. That is the directional derivative condition holds for all $v \in \mathbb{R}^p$ if and only if holds for the $2p$ canonical directions $\{\pm e_i : i = 1, \ldots, p\}$. Thus for $v = e_i$ we have the p-inequalities

$$0 < -(\tau - 1) + \xi_i(e_i) \qquad i = 1, \ldots, p$$

while for $v = -e_i$ we have,

$$0 < \tau - \xi_i(-e_i) \qquad i = 1, \ldots, p.$$

Combining these inequalities we have our optimality condition in its full generality. If none of the residuals of the non-basic observations, $i \in \bar{h}$, are zero, as would be the case with probability one if the y's had a density with respect to Lesbesgue measure, then the dependence of ξ on v disappears and we may combine the two sets of inequalities to yield,

$$(\tau - 1)1_p \leq \xi_h \leq \tau 1_p.$$

Summarizing the foregoing discussion we may reformulate Theorem 3.3 of Koenker and Bassett (1978) with the aid of the following definition introduced by Rousseeuw and Leroy (1987).

DEFINITION 3.1. *We will say that the regression observations (y, X) are in general position if any p of them yield a unique exact fit, that is for any $h \in \mathcal{H}^*$,*

$$y_i - x_i b(h) \neq 0 \quad \text{for any } i \notin h.$$

Note that if the y_i's have a density with respect to Lesbesgue measure then the observations (y, X) will be in general position with probability one.

THEOREM 2.1. *If (y, X) are in general position, then there exists a solution to the quantile regression problem (1.1) of the form $b(h) = X(h)^{-1}y(h)$ if and only if for some $h \in \mathcal{H}^*$*

$$(\tau - 1)1_p \leq \xi_h \leq \tau 1_p \tag{2.3.1}$$

where $\xi_h = \sum_{i \in \bar{h}} \psi_\tau(y_i - x_i' b(h)) x_i' X(h)^{-1}$ and $\psi_\tau = \tau - I(u < 0)$. Furthermore, $b(h)$ is the unique solution if and only if the inequalities are strict, otherwise the solution set is the convex hull of several solutions of the form $b(h)$.

Remark: Several comments on degeneracy and multiple optimal solutions may be useful at this point. Primal degeneracy in the quantile regression problem refers to circumstances in which (y, X) are not in general position and therefore we have more than p zero residuals – either at a solution, or more generally in exterior point algorithms like simplex on the path to a solution. This is unusual, unless the y_i's are discrete. On the other hand multiple optimal solutions occur when the inequalities (2.3.1) are satisfied only weakly. This occurs, typically, when the x's are discrete, so that sums of the x_i's, weighted by τ or $(\tau - 1)$, sum exactly to τ or $\tau - 1$. If the x's have a component that has a density with respect to Lesbesgue measure, then for any given τ this occurs with probability zero. In the dual problem the roles of degeneracy and multiple optimal solutions are reversed, degeneracy arising from discrete x's and MOS from discrete y's.

It might be thought that such *inequalities* could not offer the same essential analytical services provided by the more conventional gradient conditions of smooth (quasi-) maximum likelihood theory. Fortunately, as we shall see, that pessimism is not justified. Indeed, as we have already seen in Figure 1.3 the graph of the objective function actually appears quite smooth as long as n is moderately large, relative to p.

An important finite sample implication of the optimality condition (2.3.1) is the following result that shows, provided the design matrix "contains an intercept" that there will be roughly $n\tau$ negative residuals and $n(1-\tau)$ positive ones.

THEOREM 2.2. *Let N^+, N^-, N^0 denote the number of positive, negative, and zero elements of the residual vector $y - X'\hat{\beta}(\tau)$. If X contains an intercept, i.e., if there exists, $\alpha \in \mathbb{R}^p$, such that $X\alpha = 1_n$, then for any $\hat{\beta}(\tau)$ solving (1.4.4) we have*

$$N^- \leq n\tau \leq N^- + N^0$$

and

$$N^+ \leq n(1-\tau) \leq N^+ + N^0$$

Proof: We have optimality of $\hat{\beta}(\tau)$ if and only if

$$-\sum_{i=1}^{n} \psi_\tau^*(y_i - x_i'\hat{\beta}(\tau), -x'w)x_i'w \geq 0$$

for all directions $w \in \mathbb{R}^p$. For $w = \alpha$, such that $X\alpha = 1_n$ we have

$$\sum \psi_\tau^*(y_i - x_i'\hat{\beta}(\tau), -1) \geq 0$$

which yields

$$\tau N^+ - (1-\tau)N^- - (1-\tau)N^0 \geq 0$$

similarly for $w = \alpha$, we obtain

$$-\tau N^+ + (1-\tau)N^- - \tau N^0 \geq 0.$$

Combining these inequalities and using the fact that $n = N^- + N^+ + N^0$ completes the proof. ■

COROLLARY 2.1. *As a consequence, if $N^0 = p$, which occurs whenever there is no degeneracy, then the proportion of negative residuals is approximately τ,*

$$\frac{N^-}{n} \leq \tau \leq \frac{N^- + p}{n}$$

and the number of positive residuals is approximately $(1-\tau)$,

$$\frac{N^+}{n} \leq 1 - \tau \leq \frac{N^+ + p}{n}$$

Remark: In the special case that $X \equiv 1_n$, this result fully characterizes the τth sample quantile. If τn is an integer, then we will have only weak satisfaction of the inequalities, and consequently there will be an interval of τth sample quantiles between two adjacent order statistics. If τn isn't an integer, then the τth sample quantile is unique.

The foregoing remark can be extended to the two sample problem in the following manner.

COROLLARY 2.2. *Consider the two sample model where X takes the form*

$$X = \begin{bmatrix} 1_{n_1} & 0 \\ 0 & 1_{n_2} \end{bmatrix}$$

and write $y = (y_1', y_2')'$ to conform to X. Denote any τth sample quantile of the subsample y_i by $\hat{\beta}_i(\tau)$, then any regression quantile solution for this problem takes form

$$\hat{\beta}(\tau) = (\hat{\beta}_1(\tau), \hat{\beta}_2(\tau))',$$

that is, the line characterizing a τth regression quantile solution in the two sample problem simply connects two corresponding ordinary sample quantiles from the two samples.

Proof: The result follows immediately by noting that the optimality condition

$$-\sum_{i=1}^{n} \psi^*(y_i - b, -x_i'w) \quad x_i'w) \leq 0, \quad j = 1, 2,$$

for $b \in \mathbb{R}^2$ and $w \in \mathbb{R}^2$ separates into two independent conditions,

$$-\sum_{i=1}^{n_j} \psi^*(y_{i_j} - b_j, -w_j) \quad w_j \leq 0, \quad j = 1, 2.$$

where y_{ij} denotes the ith element of y_i. ■

Our formulation of the optimality conditions for quantile regression in this section is fully equivalent to the approach based on the subgradient introduced in Rockafellar(1970) and developed by Clark(1983). To make this connection more explicit, recall that the subgradient of a function f: $X \to \mathbb{R}$, at x, denoted $\partial f(x)$ is the subset of the dual space X^* given by

$$\partial f(x) = \{\xi \in X^* | \nabla f(x, v) \geq \xi' v \text{ for all } v \in X\}.$$

It is then clear that $\nabla f(x, v) \geq 0$ for all $v \in \mathbb{R}^p$ if and only if $0 \in \partial f(x)$.

3.2. Equivariance. Several important features of the least squares regression estimator are sometimes taken for granted in elementary treatments of regression, but play an important role in enabling a coherent interpretation of regression results. Suppose we have a model for the temperature of a liquid, y, but we decide to alter the scale of our measurements from Fahrenheit to Centigrade. Or we decide to reparametrize the effect of two covariates to investigate the effect of their sum and their difference. We expect such changes to have no fundamental effect on our estimates. When the data is altered in one of these entirely predictable ways we expect the regression estimates also to change in a way that leaves our interpretation of the results *invariant*. We group several such properties of quantile regression estimators together under the heading of *equivariance* and treat them quite explicitly since they are often an important aid in careful interpretation of statistical results. To facilitate this treatment we will explicitly denote a τ-th regression quantile based on observations (y, X) by $\hat{\beta}(\tau; y, X)$. Four basic equivariance properties of $\hat{\beta}(\tau; y, X)$ are collected in the following result.

THEOREM 2.3. *(Koenker and Bassett (1978)) Let A be any $p \times p$ nonsingular matrix, $\gamma \in \mathbb{R}^p$, and $a > 0$. Then for any $\tau \in [0, 1]$,*

$$\begin{array}{ll} (i) & \hat{\beta}(\tau; ay, X) = a\hat{\beta}(\tau; y, X) \\ (ii) & \hat{\beta}(\tau; -ay, X) = a\hat{\beta}(1-\tau; y, X) \\ (iii) & \hat{\beta}(\tau; y + X\gamma, X) = \hat{\beta}(\tau; y, X) + \gamma \\ (iv) & \hat{\beta}(\tau; y, XA) = A^{-1}\hat{\beta}(\tau; y, X) \end{array}$$

Remark: Properties (i) and (ii) imply a form of scale equivariance, (iii) is usually called shift or regression equivariance, and (iv) is called equivariance to reparameterization of design.

Presuming that X "contains an intercept" i.e., there exists a $\gamma \in \mathbb{R}^p$ such that $X\gamma = 1_n$, the effect of our temperature scale change is simply to shift $\hat{\beta}(\tau; y, X)$ to $\frac{5}{9}(\hat{\beta}(\tau; y, X) - 32\gamma)$. Typically in this example γ would be the first unit basis vector e_1 so the first column of X would be 1_n. The first coordinate of $\hat{\beta}$ would be shifted by 32 and all the coordinates would be then rescaled by the factor $\frac{5}{9}$. In the second example, the situation is even simpler. The result of reparameterizing the x's is that the new coefficients are now one half the sum and one half the difference of the old pair of coefficients, respectively. These equivariance properties are shared by the least squares estimator but this is not universally true for other regression estimators.

Quantiles enjoy another equivariance property, one much stronger than those already discussed. This property which we may term *equivariance to monotone transformations* is critical to an understanding of the full potential of quantile regression. Let $h(\cdot)$ be a nondecreasing function on $\mathbb{R}$, then for any random variable Y,

$$Q_{h(Y)}(\tau) = h(Q_Y(\tau)), \tag{2.3.2}$$

that is the quantiles of the transformed random variable $h(Y)$ are simply the transformed quantiles of the original Y. Of course, the mean does not share this property:

$$Eh(Y) \neq h(E(Y)),$$

except for affine h as we have considered above, or other exceptional circumstances. Condition (2.3.2) follows immediately from the elementary fact that for any monotone h,

$$P(Y \leq y) = P(h(Y) \leq h(y)),$$

but the property has many important implications.

It is common in considering least-squares regression to posit a model of the form

$$h(y_i, \lambda) = x_i'\beta + u_i$$

where $h(y, \lambda)$ denotes a transformation of the original response variable, y, which (*mirable dictu!*) achieves three objectives simultaneously:

(i) makes $E(h(y_i, \lambda)|x)$ linear in the covariates, x,
(ii) makes $V(h(y_i, \lambda)|x)$ independent of x, i.e., homoscedastic, and
(iii) makes $u_i = h(y_i, \lambda) - x_i\beta$ Gaussian.

Frequently, in practice however, these objectives are conflicting, and we need a more sophisticated strategy. There is certainly no *a priori* reason to expect that a single transformation, even the celebrated Box-Cox transformation

$$h(y, \lambda) = (y^\lambda - 1)/\lambda$$

which is the archetypical choice in this context would be capable of so much. There is also an associated difficulty that, having built a model for $E(h(y, \lambda)|x)$, we may still wish to predict or interpret the model as if were constructed for $E(y|x)$. One often sees $h^{-1}(x'\hat{\beta})$ used in place of $E(y|x)$ in such circumstances, $\exp(x'\hat{\beta})$ when the model has been specified as $\log(y) = x'\beta$, for example, but this is difficult to justify formally.

Transformations are rather more straightforward to interpret in the context of quantile regression than they are for ordinary, mean regression. Because of the equivariance property, having estimated a linear model, $x'\hat{\beta}$, for the conditional median of $h(y)$ given x we are perfectly justified in interpreting $h^{-1}(x'\hat{\beta})$ as an appropriate estimate of the conditional median of y given x.

Furthermore, because we have focused on estimating a local feature of the conditional distribution rather than a global feature like the conditional mean we may concentrate on the primary objective of the transformation – achieving linearity of the conditional quantile function – and leave the other objectives aside for the moment.

3.3. Censoring. A particularly instructive application of the foregoing equivariance results, and one which has proven extremely influential in the econometric application of quantile regression, involves censoring of the observed response variable. The simplest model of censoring may be formulated as follows. Let y_i^* denote a latent (unobservable) response assumed to be generated from the linear model

$$y_i^* = x_i'\beta + u_i \quad i = 1, \ldots, n \tag{2.3.3}$$

with $\{u_i\}$ iid from distribution function F. Due to censoring, we do not observe the y_i^*'s directly, but instead we see

$$y_i = \max\{0, y_i^*\}.$$

This model may be estimated by maximum likelihood

$$\hat{\beta} = \text{argmin}_{\beta}\{\prod_{i=1}^{n}(1 - F(x_i'\beta))^{\Delta_i} f(x_i'\beta)^{1-\Delta_i}\}$$

where Δ_i denotes the censoring indicator, $\Delta_i = 1$ if the ith observation is censored, $\Delta_i = 0$ otherwise. For F Gaussian, this leads to an estimate of the conditional mean function and has received intense scrutiny by Heckman (1979) and many subsequent authors. However, another F yields another functional in place of the conditional mean and consequently leads to a specification bias for the Gaussian maximum likelihood estimator. See Goldberger (1983) for an discussion of this bias in some typical cases.

Powell (1986) observed that the equivariance of the quantiles to monotone transformations implied that in this model the conditional quantile functions of the response depended only on the censoring point, but were independent of F. Formally, we may express the τ^{th} conditional quantile function of the observed response, y_i, in the model (2.3.3) as

$$Q_i(\tau|x_i) = \max\{0, x_i'\beta + F_u^{-1}(\tau)\} \tag{2.3.4}$$

The censoring transformation, by the prior equivariance result becomes, transparently, the new conditional quantile function. The parameters of the conditional quantile functions may now be estimated by replacing

$$\min_b \sum_{i=1}^{n} \rho_\tau(y_i - x_i'b)$$

by

$$\min_b \sum_{i=1}^{n} \rho_\tau(y_i - \max\{0, x_i'b\}) \tag{2.3.5}$$

where we assume, as usual, that the design vectors x_i, contain an intercept to absorb the additive effect of $F_u^{-1}(\tau)$.

Generalizing the model (2.3.4) slightly to accommodate a linear scale (heteroscedasticity) effect

$$y_i^* = x_i'\beta + (x_i'\gamma)u_i \quad i = 1, \ldots, n \tag{2.3.6}$$

with u_i iid F, it is clear that the new conditional quantile functions

$$Q_i(\tau|x_i) = \max\{0, x_i'\beta + x_i'\gamma F_u^{-1}(\tau)\} \tag{2.3.7}$$

can also be estimated by solving (2.3.5). Since heteroscedasticity of this form is also a source of specification bias for the iid error maximum likelihood estimator, even in the Gaussian case, its straightforward accommodation within the conditional quantile formulation must be counted as a significant advantage.

A constant censoring point is typical of many econometric applications where 0 is a natural lower bound, or institutional arrangements dictate, for example, top-coding of a specified amount. However, it is also straightforward to accommodate observation specific censoring from the right and left. Suppose, we observe

$$y_i = \begin{cases} \bar{y}_i & \text{if } y_i^* \geq \bar{y}_i \\ y_i^* & \text{otherwise} \\ \underline{y}_i & y_i^* < \underline{y}_i \end{cases}$$

then, by the same argument that led to (2.3.5), as in Fitzenberger (1996), we would now have

$$\min_b \sum_{i=1}^{n} \rho_\tau(y_i - \max\{\underline{y}_i, \min\{\bar{y}_i, x_i'b\}\}) \tag{2.3.8}$$

This framework provides a quite general treatment of fixed censoring for linear model applications. We will defer the discussion of computational aspects of solving problems (2.3.5) and (2.3.8) until Chapter X. For computational purposes, the nonlinear "kinks" in the response function created by the censoring require careful attention, since they take us out of the strict linear programming formulation of the original quantile regression problem. The linear equality constraints become, under censoring, nonlinear equality constraints.

Censoring is also typical in survival analysis applications. Random censoring, in which the censoring points are only observed for the censored observations, has recently been considered within the quantile regression framework by Ying, Jung and Wei (1991) and Powell (1994). Powell (1986) deals with the truncated regression situation in which only the uncensored observations are available to the investigator. It is an elementary point that censoring beyond a fixed threshold has no effect on the uncensored quantiles but the extension of this idea to regression has proven to be one of the most compelling rationales for the use of quantile regression in applied work.

[Perhaps more should be said about this here or elsewhere. For the moment, this subsection is intended only as a simple illustration of the monotone equivariance result. Further development of these ideas especially to more complicated Heckmanesque sample selection models and random, but not independent, censoring remains a challenging problem. It might also be reasonable to comment on some applications of these methods like Fitzenberger, Chamberlain, Buchinsky, Chay, Conley and Galenson, etc.]

3.4. Robustness. The comparison of the relative merits of the mean and median in statistical applications has a long, illustrious history. Since Gauss it has been recognized that the mean enjoys a strong optimality if the "law of errors" happens to be proportional to e^{-x^2}. On the other hand, if there are occasional, very large errors, as was commonly the case in early astronomical calculations, for example, the performance of the median can be superior; a point stressed by Laplace and many subsequent authors including, remarkably, Kolmogorov (1931).

The modern view of this, strongly influenced by Tukey, see e.g. Andrews, et al (1972), is framed by the sensitivity curve, or influence function of the estimators, and perhaps to a lesser degree, by their finite sample breakdown points. The influence function, introduced by Hampel (1974) is a population analogue of Tukey's empirical sensitivity curve. It offers a concise description of how an estimator, $\hat{\theta}$, evaluated at a distribution F is affected by "contaminating" F. Formally, we may view $\hat{\theta}$ as a functional of F and write $\hat{\theta}(F)$, and consider contaminating F by replacing a small amount of mass ε from F by an equivalent mass concentrated at y, allowing us to write the contaminated distribution function as

$$F_\varepsilon = \varepsilon\delta_y + (1-\varepsilon)F$$

where δ_y denotes the *df* which assigns mass 1 to the point y. Now we may express the influence function of $\hat{\theta}$ at F as

$$IF_{\hat{\theta}}(y, F) = \lim_{\varepsilon \to 0} \frac{\hat{\theta}(F_\varepsilon) + \hat{\theta}(F)}{\varepsilon}$$

For the mean

$$\hat{\theta}(F_\varepsilon) = \int y dF_\varepsilon = \varepsilon y + (1-\varepsilon)\hat{\theta}(F)$$

so

$$IF_{\hat{\theta}}(y, F) = y - \hat{\theta}(F),$$

whereas for the sample median, see Problem 2.5,

$$\tilde{\theta}(F_\varepsilon) = F_\varepsilon^{-1}(1/2)$$

$$IF_{\tilde{\theta}}(y, F) = \operatorname{sgn}\,(y - \tilde{\theta}(F))/f(F^{-1}(1/2)) \tag{2.3.9}$$

presuming, of course, the existence, and positivity, of the density term in the denominator.

There is a dramatic difference between the two influence functions. In the case of the mean, the influence of contaminating F at y, is simply proportional to y implying that a little contamination, *however small* at a point y sufficiently far from $\theta(F)$ can take the mean arbitrarily far away from its initial value at F. In contrast, the influence of contamination at y on the median is *bounded* by the constant $s(1/2) = 1/f(F^{-1}(1/2))$ which we will, following Tukey, call the "sparsity" at the median, since it is simply the reciprocal of the density function evaluated at the median. The sparsity is low where the density is high and vice-versa.

The comparison of the influence functions of the mean and median graphically illustrates the fragility of the mean and the robustness of the median in withstanding the contamination of outlying observations. Much of what we have already said extends immediately to the quantiles generally, and from there to quantile regression. The influence function of the τth quantile is obtained simply by replacing the 1/2 in (2.3.9) by τ. The boundedness of the quantile influence function is obviously maintained provided that the sparsity at τ is finite. Extending the IF to regression is straightforward, but we now need F to represent the joint distribution of the pairs (x, y). Writing dF in the conditional form,

$$dF = dG(x)f(y|x)dy$$

and again assuming that f is continuous and strictly positive when needed we have,

$$IF_{\hat{\beta}_F(\tau)}((y,x),F) = Q^{-1}x \text{ sgn } (y - x'\hat{\beta}_F(\tau))$$

where

$$Q = \int xx' f(x'\hat{\beta}_F(x))dG(x)$$

Again we see that the estimator has bounded influence in y since y appears only clothed by the protective sgn $(\cdot)$ function. However, the naked x appearing in IF should be a cause of some concern. It implies that introducing contamination at (x, y) with x sufficiently deviant can have extremely deleterious consequences. We could illustrate this effect with an example in which we gradually move a single outlier further and further from the mass of the data until eventually all of the quantile regression lines are forced to pass through this same offending point. There is nothing surprising or unusual here; similar behavior of the least squares estimator is illustrated in the lower panels. We will consider several proposals to robustify the behavior of quantile regression to influential design observations in Section X.x where we deal with the breakdown point of quantile regression estimators.

The robustness of the quantile regression estimator to outlying y's can be seen clearly in the following thought-experiment. Imagine a data cloud with the fitted τth quantile regression plane slicing through it. Now consider taking any point, say y_i, above that plane and moving it further way from the plane *in the y direction*. How is the position of the fitted plane affected? A moment's reflection on the subgradient condition reveals that the contribution of the point to the subgradient is independent of y_i as long as sgn $(y_i - x_i'\hat{\beta}(\tau))$ does not change. In other words, we are free to move y_i up and down at will *provided we do not cross the fitted plane* without altering the fit. This clarifies somewhat our earlier remarks that (i) the influence function is constant above the fitted quantile and (ii) observations are never "neglected", rather they participate equally in electing the representative points. Unlike the sample mean where influence is increasing in the discrepancy, $y - \hat{\theta}_F$, quantile influence depends upon y only through the sign of this discrepancy.

This feature of quantile regression can be restated more formally as follows.

THEOREM 2.4. *Let D be a diagonal matrix with nonnegative elements d_i, for $i = 1, \dots, n$, then*

$$\hat{\beta}(\tau; y, X) = \hat{\beta}(\tau; X\hat{\beta}(\tau; y, X) + D(y - X\hat{\beta}(\tau; y, X)), X)$$

As long as we don't alter the sign of the residuals *any* of the y observations may be altered without altering the initial solution. While this may, at first thought, appear astonishing, even bizarre, a second thought assures us that without it we couldn't have a quantile. It is a crucial aspect of interpreting quantile regression. When a mean dog wags its tail even its essential center moves. When the kinder, median dog wags its tail its soul remains at rest.

The influence function is an indispensable tool, exquisitely designed to measure the sensitivity of estimators to infinitesimal perturbations of the nominal model. But procedures can be infinitesimally robust, but still highly sensitive to small, finite perturbations. Take, for example, the α-trimmed mean, which is capable of withstanding a proportion $0 < \epsilon < \alpha$ of contamination, but also capable of breaking down completely when $\epsilon > \alpha$.

The finite sample breakdown point of Donoho and Huber (1983) has emerged as the most successful notion of *global* robustness of estimators. Essentially, it measures the smallest fraction of contamination of an initial sample that can cause an estimator to take values arbitrarily far from its value at the initial sample. This concept has played a crucial role in recent work on robust estimation and inference. It offers an appealing, yet tractable, global quantification of robustness, complementing the local assessment captured by the influence function. Indeed a primary goal of recent research in robustness has been the construction of so-called "high-breakdown" methods exemplified by Rousseeuw's (1984) least-median-of-squares estimator for the linear regression model which achieves asymptotic breakdown point one-half. Despite the attention lavished on the breakdown point of estimators in recent years, it remains a rather elusive concept. In particular, its non-probabilistic formulation poses certain inherent difficulties. In HJKP (1990) it is shown that the breakdown point of regression estimators is closely tied to a measure of tail-performance introduced by Jurečková (1981) for location estimators.

Let $T_n = T_n(X_1, ..., X_n)$ be an estimator of a location parameter θ_0, where $X_1, ..., X_n$ are independent and identically distributed with common, symmetric about zero, distribution function $F(x)$. Jurečková considered the measure of performance,

$$B(a, T_n) = \frac{-\log P_\theta(|T_n - \theta| > a)}{-\log(1 - F(z))}$$

for fixed n as $a \to \infty$, and she showed that this rate is controlled by the tail behavior of F. For any (reasonable) translation equivariant T_n, she showed that,

$$1 \leq \liminf_{a \to \infty} B(a, T_n) \leq \limsup_{a \to \infty} B(a, T_n) \leq n.$$

For the sample mean, $T_n = \bar{X}_n$, and F with exponential tails, so

$$\lim_{a\to\infty} \frac{-\log(1-F(z))}{ca^r} = 1$$

for some $c > 0$ and $r > 0$, $\bar{X}_n$ attains optimal tail performance with log of the probability of a large error tending to zero n times faster than the log of the probability that a single observation exceeding the bound a. While, on the contrary, for F with algebraic tails, so

$$\lim_{a\to\infty} \frac{-\log(1-F(z))}{m\log a} = 1$$

for some $m > 0$, this ratio tends to one. In contrast the sample median has much better tail behavior with the $\log P(|T_n - \theta| > a)$ tending to zero as $a \to \infty$ at least $n/2$-times faster than the tails of the underlying error distribution, for either exponential or algebraic tailed errors.

For location equivariant estimators, $T_n(X_1, ..., X_n)$ that are monotone in each argument, it can be shown (Theorem 2.1 of HJKP(1990)) that T_n has a universal breakdown point, m^*, independent of the initial sample, and for any symmetric absolutely continuous, F, having density, $f(z) = f(-z) > 0$, for $z \in \mathbb{R}$, and such that $\lim_{z\to\infty} \log(1 - F(z + c))/\log(1 - F(z)) = 1$ for any $c > 0$,

$$m^* \leq \liminf B(a, T_n) \leq \limsup B(a, T_n) \leq n - m^* + 1.$$

This close link between breakdown and tail performance extends to regression, where the least squares estimator is found to have $\lim(B(a, T_n) = \bar{h}^{-1}$, with $\bar{h} = \max_i h_{ii}$ and $h_{ii}x_i'(X'X)^{-1}x_i$, for iid Gaussian errors, but again $\lim B(a, T_n) = 1$ for F's with algebraic tails. For quantile regression estimators a trivial upper bound on tail performance and breakdown is given by $\lim B(a, \hat{\beta}(\tau)) \leq [\min\{\tau, 1-\tau\}n] + 1$. But the corresponding lower bound is more challenging.

Of course, $\hat{\beta}(\tau)$ has breakdown point, $m^* = 1$, if we consider contamination of (x, y)-pairs; a single observation judiciously pulled to infinity in both x and y directions can force *all* of the quantile regression hyperplanes to pass through it. This sensitivity to contamination of design observations is a well known defect of the entire class of M-estimators. Before addressing this issue directly, it is revealing to consider briefly the question of breakdown and tail performance in the context of fixed design observations.

For the regression median, $\hat{\beta}(1/2)$, the quantities,

$$g_i = \sup_{||b||=1} \frac{|x_i'b|}{\sum_{i\in N} |x_i'b|}$$

play the role of influence diagnostics analogous to the $h_{ii} = x_i'(X'X)^{-1}x_i$ in conventional least squares theory. Define m_* to be the largest integer m such that for any subset M of $N = \{1, 2, ..., n\}$ of size m,

$$\inf_{||b||=1} \frac{\sum_{i \in N \setminus M} |x_i' b|}{\sum_{i \in N} |x_i' b|} > 1/2.$$

Then $\lim B(a, \hat{\beta}(1/2)) \geq m_* + 1$ for algebraic tailed F, and the breakdown point, m^* of $\hat{\beta}(1/2)$ satisfies $m_* + 1 \leq m^* \leq m_* + 2$. Although it is somewhat difficult to compute precisely the value of m_* for designs in higher dimensions, for scalar, regression through the origin it is quite easy. In this case, with x_i iid $U[0, 1]$, for example, m_*/n tends to $1 - 1/\sqrt{2} \approx .29$, a quite respectable breakdown point. Clearly, for regression quantiles other than the median breakdown is determined by similar considerations.

There have been several proposals for "robustifying" quantile regression with respect to outlying design observations. Both DeJongh, DeWet and Welsh (1987) and Antoch and Jurečková (1985?) have proposed bounded influence versions of the quantile regression objective function, but unfortunately, there is little experience with these approaches in applications. Recently, Rousseeuw and Hubert (1999) have proposed a new, highly design robust variant of quantile regression based on the concept of regression depth. We will briefly describe this approach in the context of bivariate regression.

Suppose we have data $Z_n = \{(x_i, y_i) : i = 1, ..., n\} \in \mathbb{R}^2$ and the model

$$y_i = \theta_1 x_i + \theta_2 + u_i \tag{2.3.10}$$

Rousseeuw and Hubert introduce the following definitions:

DEFINITION 3.2. *A candidate fit $\theta = (\theta_1, \theta_2)$ to Z_n is called a nonfit iff there exists a real number, $v_\theta = v$ which does not coincide with any x_i and such that*

$$r_i(\theta) < 0 \text{ for all } x_i < v \text{ and } r_i(\theta) > 0 \text{ for all } x_i > v$$

or

$$r_i(\theta) > 0 \text{ for all } x_i < v \text{ and } r_i(\theta) < 0 \text{ for all } x_i > v$$

where $r_i(\theta) = y_i - \theta_1 x_i - \theta_2$.

DEFINITION 3.3. *The regression depth of a fit $\theta = (\theta_1, \theta_2)$ relative to a data set $Z_n \in \mathbb{R}^2$ is the smallest number of observations that need to be removed to make θ a nonfit.*

A mechanical description of regression depth in the "primal" or data-space plot is also provided by Rousseeuw and Hubert: the existence of v_θ for any nonfit θ, corresponds to a point on the line $y = \theta_1 x + \theta_2$ about which one could rotate the line to the vertical without encountering any observations. However, the geometric notion of "depth" is more clearly brought out by the fundamental concept of the dual plot.

In the bivariate regression version of the dual plot, each point, (x_i, y_i) appears as a line in parameter space, that is, all the points on the line

$$\theta_2 = y_i - \theta_1 x_i$$

in (θ_1, θ_2)-space have ith residual zero, and intersections of such lines correspond to points which have two zero residuals. Rousseeuw and Hubert observe,

> The [regression depth] of a fit θ is (in dual space) the smallest number of lines L_i that need to be removed to set θ free, i.e. so that it lies in the exterior of the remaining arrangement of lines.

In fact, this view brings us very close to several fascinating papers by F.Y. Edgeworth on median regression, or what he called the "plural median." Edgeworth (1888) contains an almost prescient description of the simplex algorithm for linear programming:

> The method may be illustrated thus:–Let $C - R$ (where C is a constant, [and R denotes the objective function]) represent the height of a surface, which will resemble the roof of an irregularly built slated house. Get on this roof somewhere near the top, and moving continually upwards along some one of the edges, or *arrétes*, climb up to the top. The highest position will in general consist of a solitary pinnacle. But occasionally there will be, instead of a single point, a horizontal ridge, or even a flat surface.

Supplemented by a more explicit rule for choosing the edges at each vertex, this description would fit nicely into modern textbooks of linear programming. In terms of the dual plot this strategy can be described as starting from an arbitrary intersection corresponding to a basic feasible solution, finding the directional derivatives corresponding to all of the possible directions emanating from this point, choosing the most favorable direction, and going in this direction until the objective function stops decreasing. This turns out to be a concise description of the most commonly used algorithm for quantile regression originally developed for the median case by Barrodale and Roberts (1973) and modified by Koenker and d'Orey(1983) for general quantile regression. A more detailed discussion of this approach, and its alternatives is provided in Chapter 6.

It is a curious irony that Edgeworth's long time collaborator A.L. Bowley in trying to describe Edgeworth's geometric method for computing the "plural median", came very close to the formulation of the maximum depth regression estimator of Rousseeuw and Hubert. Bowley (1902) speaking of the dual plot, suggests,

> ... we may with advantage apply Prof. Edgeworth's "double median" method and find the point, line or small area, such that, whether we proceed from it to the left, or right, or up, or down, we always intersect the same number of lines before we are clear of the network.

This is clearly not the same as finding the "deepest point" in the network, as formulated by Rousseeuw and Hubert, but if we interpret it a bit generously to include *all* possible directions not just the canonical ones, we obtain something akin to their "deepest point" and this "point" corresponds to the "deepest regression line".

Unlike the conventional median regression estimator which has a breakdown point of $1/n$ in the (x, y)-contamination model, and only marginally better breakdown properties in the fixed-x, y-contamination model, as discussed in He, Jurečková , Koenker, and Portnoy (1993) and Mizera and Muller (1997), the deepest line estimator has breakdown point 1/3. It shares the equivariance properties of the ℓ_1 estimator, but exhibits a somewhat greater tendency toward non-uniqueness. It is worth remarking in this connection that one of Theil's (1950) earliest papers also deal with a variant of this type which is usually described as the "median of pairwise slopes" and may be viewed geometrically in the dual plot by projecting all the the intersections onto the axis of the "slope" parameter and then choosing the median of these projected values.

The contrast between the deepest line estimator and the usual median regression estimator is, perhaps, most clearly seen in their asymptotic behavior, which has been recently studied by He and Portnoy (1998). It is well known that

$$\hat{\theta} = \arg \min_{\theta \in \mathbb{R}^2} \sum |y_i - \theta_1 x_i - \theta_2|$$

satisfies, under mild conditions given in Bassett and Koenker (1978) and related work by numerous subsequent authors,

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightsquigarrow \mathcal{N}(0, \omega^2 D^{-1})$$

where $\theta_0 = (\theta_1, \theta_2)'$, $\omega^2 = 1/(4f^2(0))$ and

$$\lim_{n \to \infty} n^{-1} X'X \to D$$

with $X = (x_i, 1)_{i=1}^n$.

In contrast, the deepest line estimator may be formulated as

$$\tilde{\beta}_n = \text{ argmin } \max_{x_{(1)} \leq a \leq x_{(n)}} |D_n(b, a)|$$

where

$$D_n(b, a) = \sum \text{ sgn } \{(y_i - \theta_1 x_i - \theta_2)(x_i - a)\}.$$

To formulate an asymptotic theory for the maximum regression depth estimator He and Portnoy (1997) assume that the sequence $\{x_i\}$ satisfies the conditions:

A1.) $\sum x_i^2 = O(n)$

A2.) $n^{-1} \sum x_i \text{ sgn } (x_i - x_{[tn]}) \to g_1(t)$ uniformly from $t \in (0, 1)$, with $g_1''(t) < 0$ for all t.

In addition, they assume, A3.) The $\{u_i\}$'s are iid random variables with median zero, bounded density f, $f(0) > 0$, and that f is Lipschitz in a neighborhood of zero.

When the $\{x_i\}$'s are iid from distribution function G with positive density on its entire support, they note that

$$g_1(t) = \int_t^1 G^{-1}(u)du - \int_0^t G^{-1}(u)du$$

so $g_1'(t) = -2G^{-1}(t)$ and therefore, (A2) follows immediately from the Kolmogorov strong law and the monotonicity of G^{-1}. Now let $g_0(t) = 1 - 2t$ denote the limit of $n^{-1} \sum$ sgn $(z_i - z_{[nt]})$ and set $g(t) = (g_0(t), g_1(t))'$. He and Portnoy prove the following theorem.

THEOREM 2.5. *Under conditions A1 - 3,* $\sqrt{n}(\hat{\theta} - \theta)$ *converges in distribution to a random variable whose distribution is that of the unique minimizer of the function*

$$h(\delta) = \max_t |2B(t) - B(1) + 2f(0)g(t)'\delta|$$

where $B(t)$ *is standard Brownian motion.*

Unfortunately, it is rather difficult to compare the asymptotic performance of the maximal depth estimator with the more familiar median regression estimator even in this simple iid-error bivariate setting. Even under non-iid error conditions, as long as the conditional median function is linear in parameters, both approaches can be shown to be $\sqrt{n}$-consistent for the same parameter; this is in itself quite remarkable. We would expect that the improved robustness of the maximal depth estimator would come at the price of some efficiency loss under the idealized conditions A1-3 where influential design observations are highly desirable. He and Portnoy provide a very limited evaluation of the asymptotic relative efficiency of the two estimates which is reported in Table 2.1.

Given that the maximal depth estimator consistently estimates the linear conditional median function under essentially similar conditions to those required by the ℓ_1-estimator, it is natural to ask whether it is possible to estimate the parameters of other linear conditional quantile models using similar methods. A simple reweighting of the maximal depth objective function allows us to answer this question affirmatively.

Asymmetrically reweighting positive and negative residuals suggests the quantile regression depth function

$$d_\tau(\theta) = \min_t\{\min\{\tau L^+(t) + (1-\tau)R^-(t), \tau R^+(t) + (1-\tau)L^-(t)\}\}$$

and essentially the same asymptotic analysis of He and Portnoy shows that the minimizer

$$\hat{\theta}_n(\tau) = \text{argmin } d_\tau(\theta)$$

Design	Intercept	Slope
Uniform	.90	.95
Normal	.82	.87
$t(3)$	.86	.62

TABLE 2.3. Asymptotic Relative Efficiencies of the Maximal Depth and Median Regression Estimators: The Table reports He and Portnoy's (1997) estimates of the relative asymptotic variances of the median (ℓ_1) estimator to Rousseeuw and Hubert's (1998) maximal depth estimator for three design distributions: uniform, standard normal, and Student's t on 3 degrees of freedom. In all cases the y_i's were standard normal. In all cases there is a non-trivial efficiency loss which is accentuated in the case of the slope estimator in the t model.

is a $\sqrt{n}$ consistent estimator of the parameters of the linear τ^{th} conditional quantile function.

Thus, regression depth provides an alternative "influence robust" approach to quantile regression estimation which could be compared to the earlier GM-type weighting proposals of Antoch and Jurečková (1985) and DeJongh, DeWet and Welsh (1988). Extending the regression depth idea beyond the bivariate model poses some challenges particularly on the asymptotic and algorithmic fronts, but the basic conceptual apparatus is already provided by Rousseeuw and Hubert (1998), and Rousseeuw and Struyf (1998), Mizera(1999) and He and Bai (1999).

4. Interpreting Quantile Regression Models

In the classical linear regression model where,

$$E(Y|X = x) = x'\beta,$$

we are used to interpreting the coefficients, β, in terms of the partial derivatives,

$$\frac{\partial E(Y|X = x)}{\partial x_j} = \beta_j.$$

Of course there are many *caveats* that must accompany this interpretation. For instance, we may have several coefficients associated with a single covariate in a model with quadratic effects or interaction terms. In this case changes in a single covariate induce changes in several coordinates of the vector, x, and derivatives must be computed accordingly. For example, if we have

$$E(Y|Z = z) = \beta_0 + \beta_1 z + \beta_2 z^2,$$

it is clear that

$$\frac{\partial E(Y|Z=z)}{\partial z} = \beta_1 + 2\beta_2 z$$

and therefore the "effect" of a change in z on the conditional expectation of y now depends upon both β_1 and β_2 and perhaps more significantly the effect depends upon the value of z, we choose to evaluate the derivative at, as well.

In the transformation model,

$$E(h(Y)|X=x) = x'\beta,$$

there is a strong temptation to write,

$$\frac{\partial E(Y|X=x)}{\partial x_j} = \frac{\partial h^{-1}(x'\beta)}{\partial x_j}$$

This is a common practice in logarithmic models, *i.e.* where $h(Y) = \log(Y)$, but this practice is subject to the famous Nixon dictum, "You can do it, but it would be wrong." The difficulty is obviously that $Eh(Y)$ is not the same as $h(EY)$ except in very exceptional circumstances, and this makes interpretation of mean regression models somewhat trickier in practice than one might gather from some applied accounts.

As we have already noted, the situation is somewhat simpler in this respect, in the case of quantile regression. Since, as we have already noted,

$$Q_{h(Y)}(\tau|X=x) = h(Q_Y(\tau|X=x))$$

for any monotone transformation, $h(\cdot)$, we have immediately that, if

$$Q_{h(Y)}(\tau|X=x) = x'\beta(\tau)$$

then

$$\frac{\partial Q_Y(\tau|X=x)}{\partial x_j} = \frac{\partial h^{-1}(x'\beta)}{\partial x_j}.$$

So, for example, if we specify,

$$Q_{\log(Y)}(\tau|X=x) = x'\beta(\tau)$$

then it follows that

$$\frac{\partial Q_Y(\tau|X=x)}{\partial x_j} = \exp(x'\beta)\beta_j,$$

subject, of course to our initial qualifications about the possible interdependence among the components of x.

The interpretation of the partial derivative, $\partial Q_Y(\tau|X=x)/\partial x_j$, itself, often requires considerable care. We have emphasized earlier in the context of the two sample problem that the Lehmann-Doksum quantile treatment effect is simply the response necessary to keep a respondent at the same quantile under both control and treatment regimes. Of course, this is not to say that for a particular subject who happens to fall at the τth quantile initially, and then receives an increment, Δx_j, say another year of education, will necessarily fall on the τth conditional quantile function following the increment. Indeed as much of the recent literature on treatment effects has stressed, see, e.g. Angrist, Imbens, and Rubin (1997), we are typically unable to identify features of the joint distribution of control and treatment responses since we don't observe responses under both regimes for the same subjects. With longitudinal data one may be able to explore in more detail the dynamics of response, but in many applications this will prove impossible. This is certainly also the case in conventional mean regression, where we are able to estimate the average response to treatment, but its dynamics remain hidden.

4.1. Some Examples. At this stage it is useful to consider some examples in an effort to clarify certain issues of interpretation.

4.1.1. *The Union Wage Premium.* Chamberlain (1994) considers the union wage premium, that is the percentage wage premium that union workers receive over comparable non-union employees. Based on 1987 data from the U.S. Current Population Survey, Chamberlain estimated a model of log hourly wages for 5338 men with 20-29 years of work experience. In addition to union status the model included several other covariates that are conventionally included in earnings models of this type: years of schooling, years of potential work experience, indicators of whether the respondent was married, or living in a metropolitan area, and indicators of regional, occupational and industrial categories.

Sector	0.1	0.25	0.5	0.75	0.9	OLS
Manufacturing	0.281 (0.12)	0.249 (0.12)	0.169 (0.11)	0.075 (0.1)	−0.003 (0.11)	0.158 (0.14)
Non-manufacturing	0.47 (0.14)	0.406 (0.14)	0.333 (0.13)	0.248 (0.16)	0.184 (0.18)	0.327 (0.16)

TABLE 2.4. The Union Wage Premium: Quantile regression estimates of the union wage premium in the US as estimated by Chamberlain (1994) based on 5358 observations from the 1987 CPS data on workers with 20-29 years experience.

The results for the union wage effect are summarized in Table X.i, for manufacturing and non-manufacturing employees separately. In the last column of the Table the conditional mean effect estimated by least squares is reported. It shows nearly a 16% wage premium for union workers in manufacturing and almost a 33% premium for non-manufacturing employees. But is important to ask, how is this premium distributed? Is the union wage premium shared equally by all strata of workers, as would be the case if union membership induced a pure location shift in the distribution of log wages, or do some strata benefit more than others from union status.

The results clearly indicate that conditional on other labor market characteristics, it is the lowest wage workers that benefit most from union membership. If there were a pure location shift effect, as we implicitly assume in the mean regression model, we would expect to see that the coefficients at each of the five estimated quantiles would be the same as the 15.8% mean effect for manufacturing. Instead, we see that workers at the first decile of the conditional wage distribution receive a 28% boost in wages from union membership, and this figure declines steadily as one moves up through the conditional wage distribution until, at the upper decile, the union wage premium has vanished. For non-manufacturing workers the picture is quite similar; the mean shift of 32.7% is strongest at the lower quantiles, and essentially disappears in the upper tail of the conditional wage distribution.

These findings should not, as Chamberlain comments, surprise students of unionism. Prior work had shown that the dispersion of wages conditional on covariates similar to those used by Chamberlain was considerably smaller for union workers than for non-union workers. And the pattern of observed quantile regression union effects can be roughly anticipated from this dispersion effect. But the precise nature of the pattern, its asymmetry, and the effect of other covariates on aspects of the conditional distribution other than its location are all revealed more clearly by the quantile regression analysis.

An important aspect of the union wage premium problem, one that is quite explicitly neglected in Chamberlain's work involves the causal interpretation of the estimated model. There is a large econometric literature on this aspect of the interpretation, which stresses the endogoneity of union status. Individuals are obviously not randomly assigned to union, or non-union status, they are selected in a rather complicated procedure that makes causal interpretation of estimated union effects fraught with difficulties. We shall return to this important issue in Section 8.2.

4.1.2. *Demand for Alcohol.* Manning, Blumberg, and Moulton (1995) estimate a model for the demand for alcohol based on a sample of 18,844 observations from the U.S. National Helath Interview Survey. The model is a conventional log linear demand equation,

$$\log q_i = \beta_0 + \beta_1 \log p_i + \beta_2 \log x_i + u_i$$

where q_i denotes annual alcohol consumption as reported by individual i, $\log p_i$ is a price index for alcohol computed on the basis of the place of residence of individual i, and x_i is the annual income of the ith individual. Roughly 40 percent of the respondents reported zero consumption so for quantiles with $\tau < .4$, we have no demand response to either price or income. Results for $\tau > .4$ are illustrated in Figure X.ii. The income elasticity is fairly constant at about $\hat{\beta} \approx .25$, with some evidence of a somewhat less elastic response near $\tau = .4$ and $\tau = 1$. More interesting is the pattern

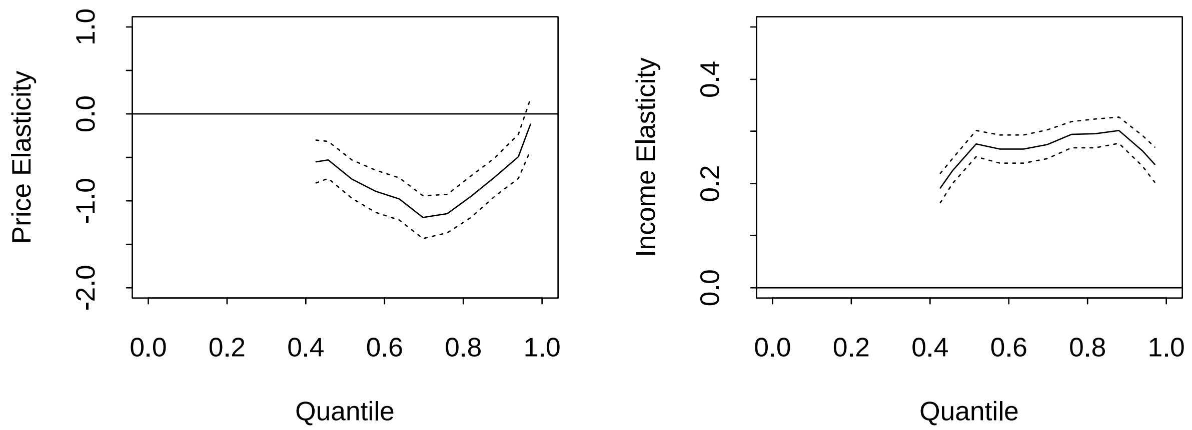

FIGURE 2.5. Price and Income Elasticities of Alcohol Demand: Quantile regression estimates of the demand for alcohol taken from Manning, Blumberg, and Moulton (1995), and based 18844 observations from the National Health Interview Survey (NHIS). Price and income elasticities are plotted as a function of the quantile of the demand distribution conditional on several other covariates. A pointwise .70 level confidence band is indicated by the dotted curves. The plotted curves begin at $\tau = .4$ because approximately 40 percent of the sample reports zero consumption."

of the price elasticity, $\beta_1(\tau)$, which is most elastic at moderate consumption levels with $\tau \approx .7$, and becomes very inelastic (unresponsive to price changes) for individuals with either very low levels of consumption, $\tau = .4$, or very high levels of consumption, $\tau = 1$. This seems quite consistent with prior expectations. Given income, individuals with very low levels of demand could be expected to be quite insensitive to price, as would those with very high levels of demand – those for whom demand is dictated more by physiological considerations. Again, the presumption that price and income act as a pure location shift effect on log consumption appears to be a very inadequate representation of the actual state of affairs. Certainly, from a policy standpoint it is important to have a clear indication of how the mean response to changes in prices is "allocated" to the various segments of the conditional distribution of dmeand, and this is what the quantile regression analysis provides.

4.1.3. *Glacier Lilies, Gophers, and Rocks.* Cade, Terrell and Schroeder (1999) consider a model of the viability of the glacier lily (*Erythronium grandiflorum*) as a function of several ecological covariates. They argue generally that in ecology it is often of interest to formulate models for maximum sustainable population densities, and they suggest that it may therefore be more informative to estimate the effect of certain covariates on upper quantiles of the response, rather than focus on models of conditional central tendency. Cade *et al* explore several models for the prevalence of lily seedlings as a function of the number of flowers observed in 256 contiguous 2 $\times$ 2 m quadrats of subalpine meadow in western Colorado. An index of rockiness of the terrain and an index of gopher burrowing activity are also used as explanatory variables.

As in the alcohol demand example there is a preponderance of observations with zero response, making conventional least squares estimation of mean regression models problematic. In a simple bivariate model in which the number of seedlings depends solely on the number of flowers observed, we illustrate several fitted log linear quantile regression models in Figure X.iii. As can be seen in these figures, the relationship is very weak until we reach the upper tail. Only the .95 and .99 quantile regression estimates exhibit a significant slope. Note that in fitting the log linear model it was necessary to deal with the fact that nearly half of the response observations were zero. In mean regression it is occassionally suggested that one transform by $\log(y+\epsilon)$ to account for this, but it is clear that the least squares fit can be quite sensitive to the choice of epsilon. In contrast for the quantile regression model, as long as we are interested in quantiles such that all the zero response observations fall below the fitted relationship, the choice of ϵ has no effect.

Regarding the strong *negative* relationship between the number of seedlings and the number of observed flowers in the upper tail of the conditional distribution, Cade *et al*, comment,

> "Negative slopes for upper regression quantiles were consistent with the explanation provided by Thompson *et al* that sites where flowers were most numerous because of lack of pocket gophers (which eat lilies), were rocky sites that provided poor moisture conditions for seed germination; hence seedling numbers were lower."

Here we risk missing the primary relationship of interest by focusing too much attention on the conditional central tendency. Fitting the upper quantile regressions reveals a strong relationship posited in prior work. Cade *et al* go on to explore the effect of other covariates and find that their measure of the rockiness of the terrain plays a significant role. After the inclusion of the index of rockiness, the number of observed flowers exert a more natural, statistically significant, *positive* effect on the presence of seedlings at the upper quantiles of the conditional distribution. This reversal of sign for the flower effect further supports the view of Thompson *et al* cited above.

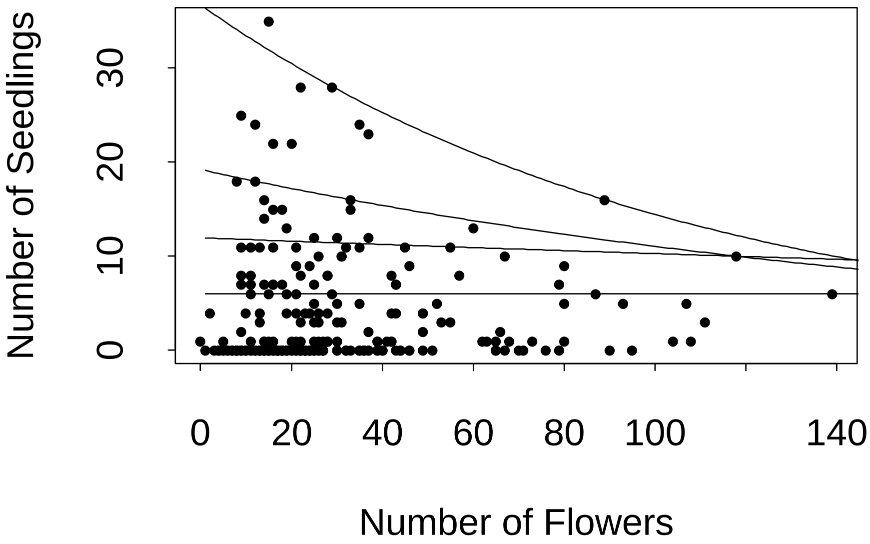

FIGURE 2.6. Glacial Lily Seedling Counts: The figure plots observations on flower and seedling counts for 256 contiguous 2 by 2 meter quadrats of subalpine meadow in western Colorado. As in Cade *et al* one outlying count of 72 seedlings in a region with 16 flowers was omitted from the plot, but included in the fitting. The four plotted curves are estimates of the $\tau \in \{.75, .9, .95, .99\}$ conditional quantile functions. Note that almost half, 127 of 256, of the observations have zero seedling counts.

It is common in many disciplines that theory offers predictions about upper or lower bounds of stochastic processes conditional on observable covariates. In these cases it is valuable to be able to estimate these extreme regression quantiles directly as we have suggested the foregoing example. Of course the theory of the most extreme regression quantiles is considerably more complicated than the theory for more central quantile regression, and we must balance considerations of robustness and efficiency. In Section 8.4 we offer a more extensive review of the literature on extreme quantile regression estimation.

4.1.4. *Daily Melbourne Temperatures.* As a final example we will reconsider a semi-parametric AR(1) model for daily temperature in Melbourne, Australia. Hyndman, Bashtannyk, and Grunwald (1996) have recently analyzed these data using the modal regression approach of Scott(1992). The quantile regression approach is strongly complementary and offers a somewhat more complete view of the data. In Figure X.iv we provide an AR(1) scatter plot of 10 years of daily temperature data. Today's maximum daily temperature is plotted against yesterday's maximum. Not surprisingly one's first impression from the plot suggests that a "unit-root" model in which today's forecasted maximum is simply yesterday's maximum. But closer examination of the plot reveals that this impression is based primarily on the left side of the plot where the central tendency of the scatter follows the 45 degree line quite closely. On the right side, however, corresponding to summer conditions, the pattern is more complicated. There, it appears that *either* there is another hot day, falling again along the 45 degree line, *or* there is a dramatic cooling off. But a mild cooling off appears to be quite rare. In the language of conditional densities, if today is hot, tomorrow's temperature appears to be bimodal with one mode roughly centered at today's maximum, and the other mode centered at about 20°.

In Figure X.v we have superimposed 19 estimated quantile regression curves. Each curve is specified as a linear B-spline of the form,

$$Q_{Y_t}(\tau|Y_{t-1}) = \sum_{i=1}^{p} \phi_i(Y_{t-1})\beta_i(\tau)$$

where $\{\phi_i(\cdot) : i = 1, ..., p\}$ denote the basis functions of the spline. Having selected the knot positions of the spline such models are linear in parameters and thus can be easily estimated by the methods already introduced. Related smoothing spline methods are discussed later in Chapter 7.

Given a family of estimated conditional quantile functions, it is straightforward to estimate the conditional density of the response at various values of the conditioning covariate. In Figure X.vi we illustrate this approach with several of density estimates based on the Melbourne data. In the last panel of this Figure we see clearly the bimodal form of the conditional density for the case in which we are conditioning on a high value of yesterday's temperature.

The particular form of mean reversion illustrated in this example has a natural meteorological explanation as high pressure systems bringing hot weather from the interior of the continent, must eventually terminate with a cold front generated over the Tasman Sea, generating a rapid drop in temperature. This sort of dynamic does not seem entirely implausible in other time-series settings, including those in economics and finance, and yet the conventional time series models that we usually consider are incapable of accommodating behavior of this type. Clearly, models in which the conditioning covariates affect only the location of the response distribution

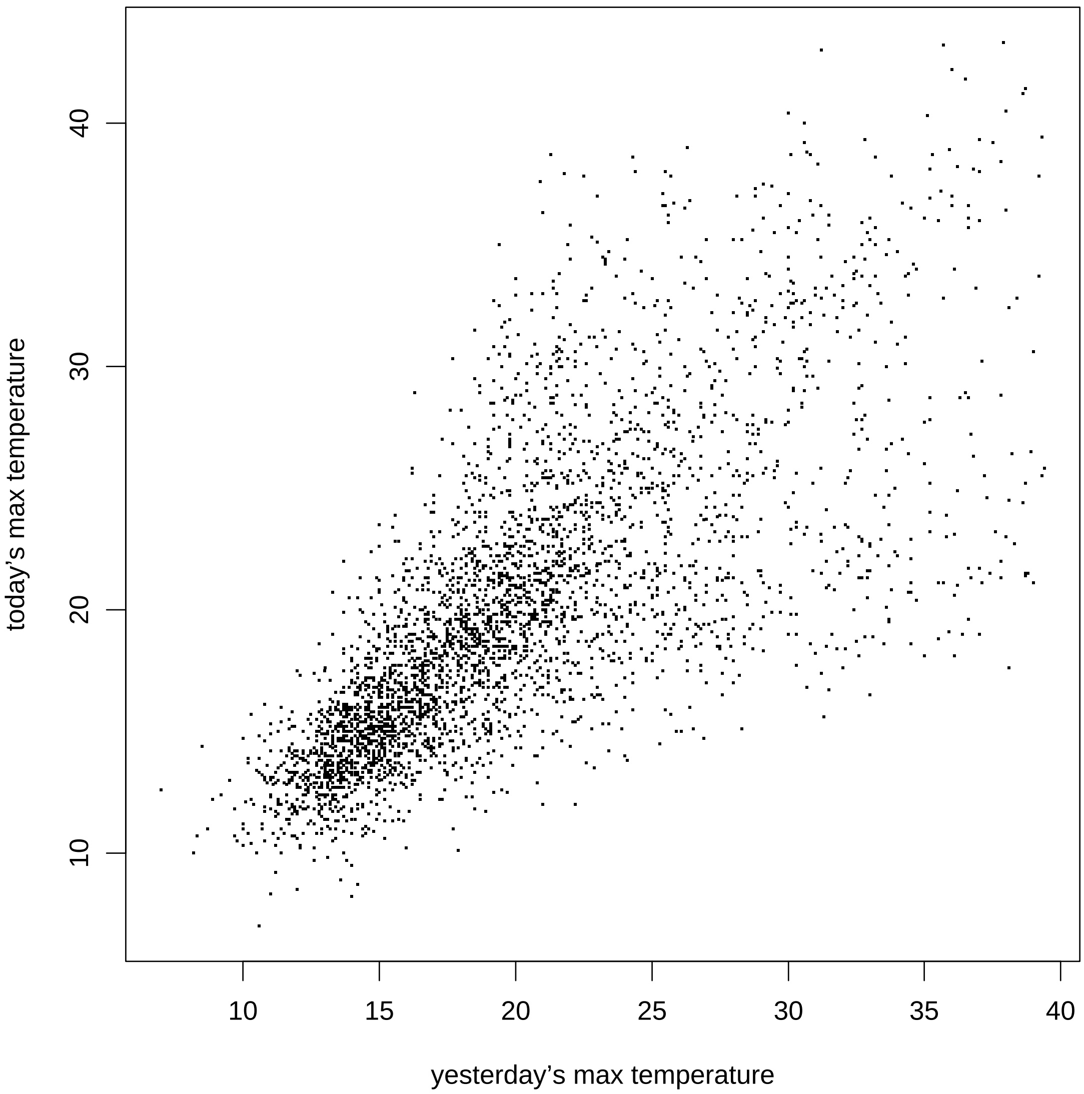

Figure 2.7. Melbourne Maximum Daily Temperature: The plot illustrates 10 years of daily maximum (centigrade) temperature data for Melbourne, Australia as an AR(1) scatterplot. Note that conditional on hot weather on the prior day, the distribution of maximum temperature on the following day appears to be bimodal.

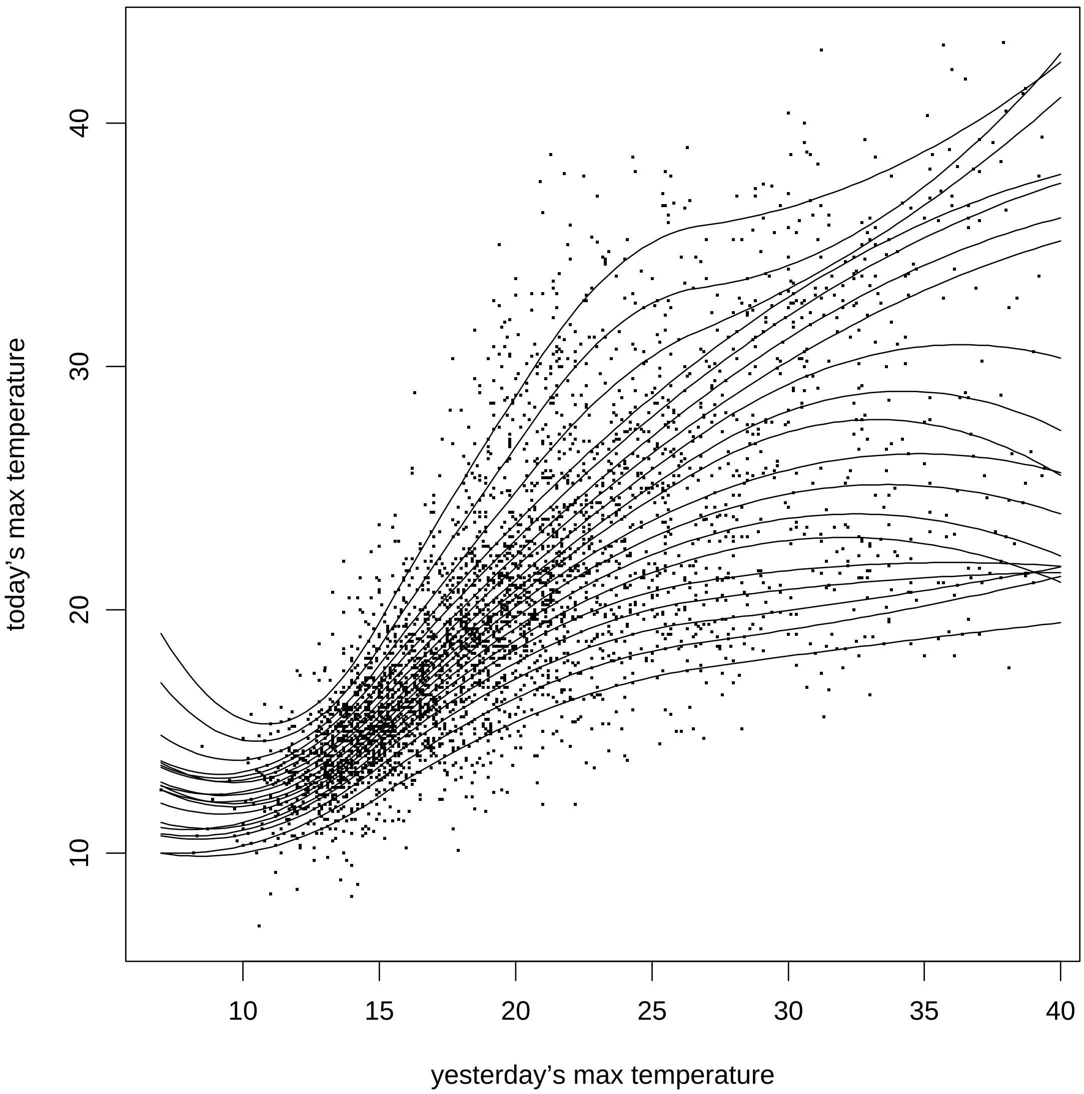

FIGURE 2.8. Melbourne Daily Maximum Temperature: Superimposed on the AR(1) scatterplot of daily maximum temperatures are 12 estimated conditional quantile functions. These functions support the view that the conditional density of maximum temperature conditional on prior warm weather is bimodal.

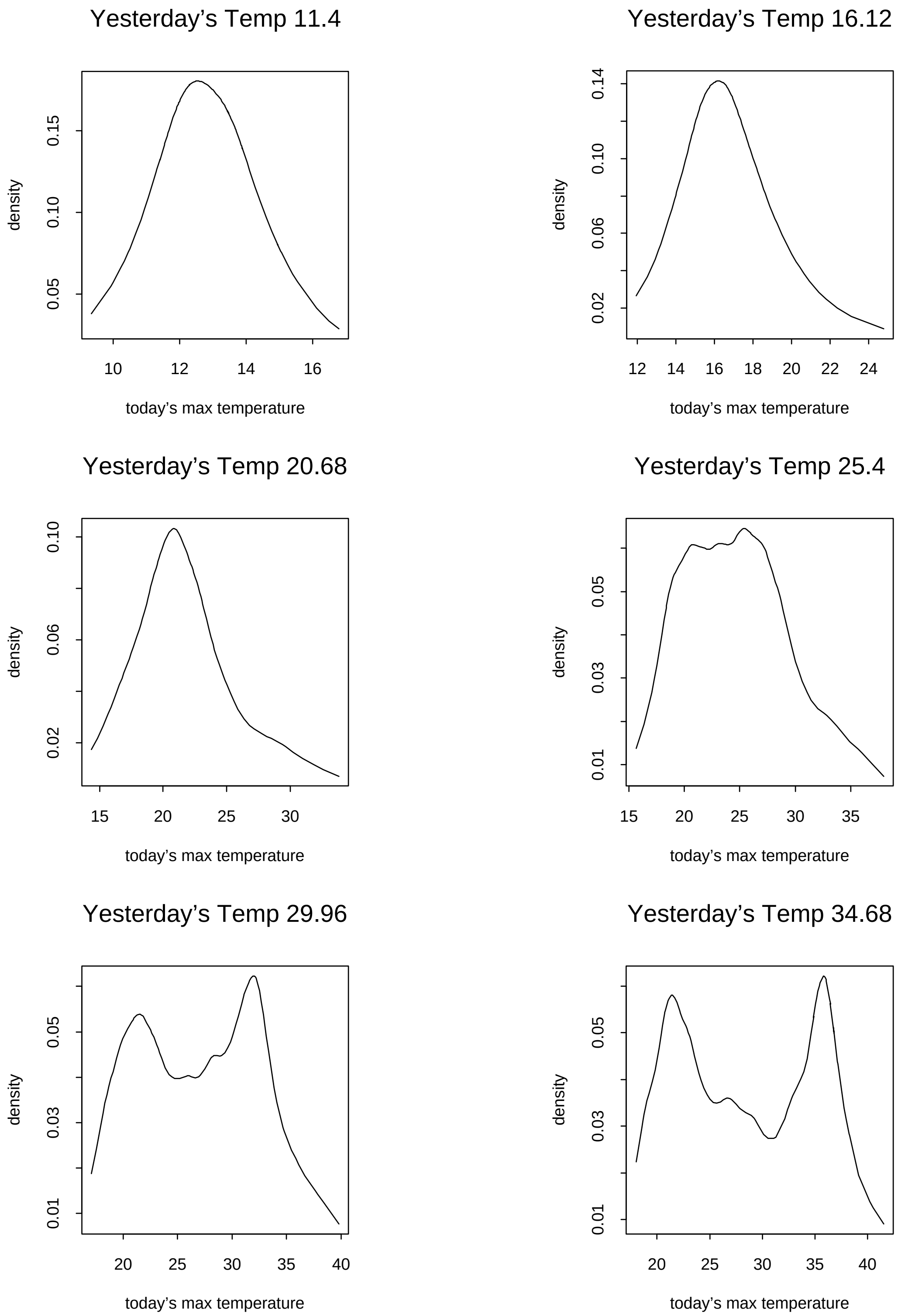

FIGURE 2.9. Melbourne Daily Maximum Temperature: Conditional density estimates of today's maximum temperature for several values of yesterday's maximum temperature, based on the Melbourne data. Note that today's temperature is bimodal when yesterday was hot.

are inadequate, and the recent wave of models for conditional scale, variance, etc. also are unsatisfactory. We must allow the entire shape of the conditional density to change with x, and this is readily done within the scope of the quantile regression formulation.

5. Interpreting Misspecified Quantile Regression Models

In the classical least squares setting if we consider the model

$$y_i = \theta(z_i) + u_i \tag{2.5.11}$$

with the $\{u_i\}$ iid from the distribution F, and $Eu_1 = 0$; but, mistakenly, we estimate the linear model,

$$y_i = \beta_0 + \beta_1 z_i + v_i, \tag{2.5.12}$$

then the least squares estimator, $\hat{\beta} = (X'X)^{-1}X'y$ with $X = (x_i) = (1, z_i)$ has the property that,

$$\hat{\beta} = (X'X)^{-1}X'\theta$$

where $\theta = (\theta(z_i))_{i=1}^n$. Thus, in effect, we can view the misspecified least squares projection as occurring in two steps: in the first, the response, y is projected to obtain its correct conditional mean vector θ, and, in the second step, θ is projected into the linear subspace spanned by the columns of X. We may thus interpret the least squares estimator, $\hat{\beta}$, as an estimator of the best $\mathcal{L}_2$ approximation of the true conditional mean vector, θ by a vector lying in the column space of X. This approximation is clearly dependent on design points, $\{x_i\}$, since it minimizes the quantity $\sum(\theta(x_i) - x_i'b)^2$.

In quantile regression the analysis of the consequences of misspecification is somewhat more complicated due to the fact that we cannot decompose the analogous "projection" in the same way. To see this, note that,

$$\hat{\beta}(\tau) = \text{argmin} \sum \rho_\tau(y_i - x_i'b)$$

solves, asymptotically, the equations,

$$\Psi(b) = n^{-1} \sum \psi_\tau(y_i - x_i'b)x_i = 0.$$

Write,

$$\begin{aligned} \Psi(b) &= n^{-1} \sum \psi_\tau(u_i + \theta(x_i) - x_i'b)x_i \\ &= n^{-1} \sum (\tau - I(u_i + \theta(x_i) - x_i'b < 0)x_i \end{aligned}$$

so

$$E\Psi(b) = n^{-1} \sum (\tau - F(x_i'b - \theta(x_i)))x_i.$$

And thus, we see that the solution, $\hat{\beta}(\tau)$, that makes $E\Psi(b) = 0$, depends not only upon the function, $\theta(\cdot)$, and the observed, $\{x_i\}$, as in the least squares case, but also upon the form of the distribution function, F.

An interesting, albeit rather implausible, special case is that of the uniform distribution for the $\{u_i\}$. In this case, we have simply,

$$E\Psi(b) = n^{-1} \sum (\tau + \theta(x_i) - x_i' b) x_i.$$

so we can write the solution, $\beta(\tau)$, explicitly as,

$$\beta(\tau) = (\sum x_i x_i')^{-1} \sum x_i'(\theta(x_i) + \tau) = (X'X)^{-1} X'(\theta + \tau)$$

Since, X, explicitly contains an intercept the effect of the τ term appears only in the intercept component of $\beta(\tau)$ and for the slope parameters we have the same projection of the conditional mean function as we found for the least squares case. In general, of course, the distribution function is *nonlinear* and thus enters the determination of $\beta(\tau)$ in a more complicated manner.

6. Problems

1. *Extend Corollary 2.2 to the p sample problem with design matrix*

$$X = \begin{bmatrix} 1_{n_1} & 0 & \dots & 0 \\ 0 & 1_{n_2} & & \\ \vdots & & \ddots & \\ 0 & & & 1_{n_p} \end{bmatrix}.$$

2. *Suppose we have the reformulated p sample design matrix*

$$X = \begin{bmatrix} 1_{n_1}, & 0 & \dots & 0 \\ 1_{n_2} & 1_{n_2} & \vdots & \\ \vdots & \vdots & & 0 \\ 1_{n_p} & 0 & & 1_{n_p} \end{bmatrix}$$

express the regression quantile estimator $\hat{\beta}(\tau)$ in this case as,

$$\hat{\beta}(\tau) = (\hat{\beta}_1(\tau), \hat{\delta}_2(\tau), \dots, \hat{\delta}_p(\tau))'$$

where $\hat{\delta}_i(\tau) = \hat{\beta}_i(\tau) - \hat{\beta}_1(\tau)$, and interpret.

3. *Bofinger (1975). Show using standard density estimation techniques that the optimal bandwidth for minimum mean-squared error estimation of the sparsity function at τ, is,*

$$h_n = n^{-1/5} \left(\frac{4.5 s^2(t)}{(s''(t))^2} \right)^{1/5}$$

Of course, if we knew $s(t)$ and $s''(t)$ we wouldn't need h_n, but fortunately $s(t)/s''(t)$ is not very sensitive to F. Show that h_n is invariant to location and scale of F, for example. Compare h_n for some typical distributional shapes - say, the Gaussian, Student, and lognormal. Show

$$\frac{s(t)}{s''(t)} = \frac{f^2}{2(f'/f)^2 + [(f'/f)^2 - f''/f]}$$

and, for example, if f is Gaussian, $(f'/f)(F^{-1}(t)) = -\Phi^{-1}(t)$ so the term in square brackets is 1 , and the optimal bandwidth becomes,

$$h_n = n^{-1/5}\left(\frac{4.5\phi^4(\Phi^{-1}(t))}{(2\Phi^{-1}(t)^2+1)^2}\right)^{1/5}.$$

Plot this bandwidth, as a function of n for several quantiles, and compare the plots across the distributions. Sheather and Maritz(1983) discuss preliminary estimation of s and s'' as a means of estimating a plug-in h_n.

4. *Compare the Sheather-Hall bandwidth rule given in the text with the Bofinger bandwidth of the previous problem.*

5. *Let $X_1,\ldots,X_n$ be a random sample from a df $F(x-\theta_0)$. Consider L-estimators of location of the form*

$$\hat{\theta}_0 = \int_0^1 J(u)F_n^{-1}(u)du$$

where $F_n(\cdot)$ denotes the empirical distribution function constructed from the X_i's. If F has finite Fisher information $I(F)$ and a twice continuously differentiable log density, then the optimal L-estimator has weight function of the form

$$J^*(F(x)) = -\frac{(\log\ f(x))''}{I(F)}$$

1. *Explain the observation that $\hat{\theta}_n$ is location equivariant, since,*

$$\int_{-\infty}^{\infty} J^*(F(y))dF(y) = \int_0^1 J^*(u)du = 1$$

2. *The optimality of $\hat{\theta}_n$ may be seen by computing the influence function of the general L-estimator as follows:*

   I.: *The IF of the u^{th} sample quantile is*

$$IF(x, F^{-1}(u), F) = \frac{d}{d\varepsilon}F_\varepsilon^{-1}(u) = \frac{u - \delta_x(F^{-1}(u))}{f(F^{-1}(u))}$$

   which may be shown by differentiating the identity

$$F_\varepsilon(F_\varepsilon^{-1}(u)) = u$$

where $F_\varepsilon(y) = (1-\varepsilon)F(y) + \varepsilon\delta_x(y)$ *to obtain*

$$0 = -F(F_\varepsilon^{-1}(u)) + \delta_x(F_\varepsilon^{-1}(y)) + f_\varepsilon(F_\varepsilon^{-1}(u))\frac{d}{d\varepsilon}F_\varepsilon^{-1}(u)$$

and evaluating at $\varepsilon = 0$.

II.: *Thus*

$$\begin{aligned} IF(x,\hat{\theta}_n,F) &= \int_0^1 (J^*(u)(u-\delta_x(F^{-1}(u)))/f(F^{-1}(u))du \\ &= \int_{-\infty}^{\infty} J^*(F(y))(F(y)-\delta_x(y))dy \\ &= \int_{-\infty}^{x} J^*(F(y))dy - \int_{-\infty}^{\infty}(1-F(y))J^*(F(y))dy \\ &= -I(F)^{-1}\int_{-\infty}^{x}(\log f)''(y)dy \\ &= -I(F)^{-1}(\log f)'(x) \end{aligned}$$

III.: *Setting* $\psi(x) = -(\log f)'(x) = -f'(x)/f(x)$ *we conclude that* $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow \mathcal{N}(0, EIF^2)$ *where*

$$\begin{aligned} EIF^2 &= \int (\psi^2(x)/I(F)^2)dF(x) \\ &= I(F)^{-1} \end{aligned}$$

Explain briefly the foregoing result. Focus on the following aspects

(i) *How to compute* $\hat{\theta}_n$.
(ii) *How does* $\hat{\theta}_n$ *differ from the mle.*
(iii) *What does the IF tell us about* $\hat{\theta}_n$.

6. *Consider the mean squared error of the kernel smoothed quantile estimator,* $\tilde{\beta}_n(\tau)$, *of* (5.2.1). *Show that the bias my be expressed as,*

$$Bias(\tilde{\beta}_n(\tau)) = \frac{1}{2}h_n^2\sigma_k^2 + o(h^2) + O(n^{-1}),$$

and the variance as,

$$V(\tilde{\beta}_n(\tau)) = n^{-1}\tau(1-\tau)s^2(\tau) - n^{-1}h_n s^2(\tau)\eta_k + o(h/n),$$

where $\sigma_k^2 = \int t^2 k(t)dt$, *and* $\eta_k = \int tk(t)K(t)dt$. *Conclude from this that the optimal (mean-squared-error minimizing) bandwidth is of the form*

$$h^* = (\kappa/n)^{1/3}[s(\tau)/s'(\tau)]^{2/3}$$

where $\kappa = 2\eta_k/(\sigma_k^2)^2$. *For the Epanechnikov kernel, show that* $\kappa = .287494$, *and illustrate them for few representative n's for the standard normal distribution. Even* *for quite large sample sizes these bandwidths are quite wide and we would be reluctant to recommend such aggressive smoothing. See Sheather and Marron (1990) and Falk (1984) for assistance on this problem.*

CHAPTER 3

Inference for Quantile Regression

In this chapter we will try to provide a practical guide to statistical inference for quantile regression applications. There are a number of competing approaches in the literature and we will offer some guidance on their advantages and disadvantages. Ideally, of course, we would aspire to provide a finite-sample apparatus for statistical inference about quantile regression like the elegant classical theory of least squares inference under iid Gaussian errors. But we must recognize that even in the least squares theory it is necessary to resort to asymptotic approximations as soon as we depart significantly from idealized Gaussian conditions.

Nevertheless, we will briefly describe what is known about the finite sample theory of the quantile regression estimator and its connection to the classical theory of inference for the univariate quantiles. We then introduce the asymptotic theory of inference with a heuristic discussion of the scalar, regression-through-the-origin model; a more detailed treatment of the asymptotic theory of quantile regression is deferred to Chapter 4. We then describe several approaches to inference: Wald tests and related problems of direct estimation of the asymptotic covariance matrix, rank tests based on the dual quantile regression process, likelihood ratio type tests based on the value of the objective function under null and alternative models and finally several resampling methods are introduced. The chapter concludes with a description of a small Monte Carlo experiment designed to evaluate and compare the foregoing methods.

1. Some Finite Sample Distribution Theory

Suppose $Y_1, ..., Y_n$ are independent and identically distributed (iid) random variables with common distribution function F, and assume that F has a continuous density, f, in a neighborhood of $\xi_\tau = F^{-1}(\tau)$ with $f(\xi_\tau) > 0$. The objective function of τth sample quantile,

$$\hat{\xi}_\tau \equiv \inf_{\xi}\{\xi \in \mathbb{R} | \sum \rho_\tau(Y_i - \xi) = \min!\}$$

is the sum of convex functions, hence is itself convex. Consequently by the monotonicity of the gradient,

$$g_n(\xi) = \sum_{i=1}^{n} (I(Y_i < \xi) - \tau),$$

we have,

$$\begin{aligned} P\{\hat{\xi}_\tau > \xi\} &= P\{g_n(\xi) < 0\} \\ &= P\{\sum I(Y_i < \xi) < n\tau\} \\ &= P\{B(n, F(\xi)) < n\tau\}, \end{aligned}$$

where $B(n,p)$ denotes a binomial random variable with parameters (n,p). Thus, letting $m = \lceil n\tau \rceil$ denote the smallest integer $\geq n\tau$, we may express the distribution function, $G(\xi) \equiv P\{\hat{\xi}_\tau \leq \xi\}$, of $\hat{\xi}(\tau)$, using the incomplete beta function, as,

$$\begin{aligned} G(\xi) &= 1 - \sum_{k=m}^{n} \binom{n}{k} F(\xi)^k (1 - F(\xi))^{n-k} \\ &= n \binom{n-1}{m-1} \int_0^{F(\xi)} t^{m-1}(1-t)^{n-m} dt. \end{aligned}$$

Differentiating, yields the density function for $\hat{\xi}(\tau)$,

$$g(\xi) = n\binom{n-1}{m-1} F(\xi)^{m-1}(1 - F(\xi))^{n-m} f(\xi). \tag{3.1.1}$$

This form of the density can be deduced directly, by noting that the event $\{x < Y_{(m)} < x + \delta\}$ requires that $m-1$ observations lie below x, $n-m$ lie above $x+\delta$ and one lies in the interval $(x, x+\delta)$. The number of ways that this arrangement can occur is

$$\frac{n!}{(m-1)!1!(n-m)!} = n\binom{n-1}{m-1},$$

and each arrangement has the probability, $F(\xi)^{m-1}(1-F(\xi))^{n-m}[F(\xi+\delta) - F(\xi))]$. Thus,

$$P\{x < Y_{(m)} < x + \delta\} = n\binom{n-1}{m-1} F(\xi)^{m-1}(1-F(\xi))^{n-m} f(\xi)\delta + o(\delta^2),$$

and we obtain (3.1.1) by dividing both sides by δ and letting it tend to zero.

This approach may also be used to construct confidence intervals for ξ_τ of the form,

$$P\{\hat{\xi}_{\tau_1} < \xi_\tau < \hat{\xi}_{\tau_2}\} = 1 - \alpha$$

where τ_1 and τ_2 are chosen to satisfy

$$P\{n\tau_1 < B(n,\tau) < n\tau_2\} = 1 - \alpha.$$

In the case of continuous F these intervals have the remarkable property that they are distribution-free, that is, they hold irrespective of F. In the case of discrete F, closely related distribution-free *bounds* for $P\{\hat{\xi}_{\tau_1} < \xi_\tau < \hat{\xi}_{\tau_2}\}$ and $P\{\hat{\xi}_{\tau_1} \le \xi_\tau \le \hat{\xi}_{\tau_2}\}$ may be constructed.

The finite sample density of the regression quantile estimator, $\hat{\beta}(\tau)$, under iid errors may be derived from the subgradient condition (2.2.1) introduced in the previous chapter in a manner which much like the derivation of the density in the one-sample case given above.

THEOREM 3.1 (Bassett and Koenker (1978)). *Consider the linear model*

$$Y_i = x_i'\beta + u_i \quad i = 1, \ldots, n,$$

with iid errors $\{u_i\}$ having common distribution function F and strictly positive density f at $F^{-1}(\tau)$. Then the density of $\hat{\beta}(\tau)$ takes the form,

$$g(b) = \sum_{h \in \mathcal{H}} P\{\xi_h(b) \in \mathcal{C}\} \big|X(h)\big| \prod_{i \in h} f(x_i'(b - \beta(\tau)) + F^{-1}(\tau)) \tag{3.1.2}$$

where $\xi_h(b) = \sum_{i \in \bar{h}} \psi_\tau(y_i - x_i b) x_i' X(h)^{-1}$ and $\mathcal{C}$ denotes the cube $[\tau - 1, \tau]^p$.

Proof: From Theorem 2.1, $\hat{\beta}(\tau) = b(h) \equiv \big(X(h)\big)^{-1} Y(h)$ if and only if $\xi_h(b(h)) \in \mathcal{C}$. For any $b \in \mathbb{R}^p$, let $B(b,\delta) = b + [-\delta/2, \delta/2]^p$ denote the cube centered at b with edges of length δ, and write,

$$\begin{aligned} P\{\beta(\tau) \in B(b,\delta)\} &= \sum_{h \in \mathcal{H}} P\{b(h) \in B(b,\delta),\ \xi_h(b(h)) \in \mathcal{C}\} \\ &= \sum_{h \in \mathcal{H}} EI(b(h) \in B(b,\delta)) P\{\xi_h(b(h)) \in \mathcal{C} | Y(h)\} \end{aligned} \tag{3.1.3}$$

where the expectation is taken with respect to the vector $Y(h)$. The conditional probability above is defined from the distribution of $\xi_h(b(h))$ which is a discrete random variable (taking on 2^{n-p} values for each $h \in \mathcal{H}$). As $\delta \to 0$, this conditional probability tends to $P\{\xi_h(b) \in \mathcal{C}\}$ this probability is independent of $Y(h)$.

Now, divide both sides by Volume$(B(b,\delta)) = \delta^p$, and let $\delta \to 0$ to put things in density form. The conditional probability tends to $P\{\xi_h(b) \in \mathcal{C}\}$ which no longer depends on $Y(h)$. The other factor tends to the joint density of the vector $X(h)^{-1}Y(h)$ which since

$$f_{Y(h)}(y) = \prod_{i \in h} f(y_i - x_i'\beta).$$

can be written as,

$$
\begin{aligned}
(3.1.4) \qquad f_{\left(X(h)\right)^{-1}Y(h)}(b) &= \left|X(h)\right| \prod_{i \in h} f\Big(\big(X(h)b\big)_i - x_i'\beta(\tau) + F^{-1}(\tau)\Big) \\
&= \left|X(h)\right| \prod_{i \in h} f\Big(x_i'(b - \beta(\tau)) + F^{-1}(\tau)\Big).
\end{aligned}
$$

Note that h for which $X(h)$ is singular contribute nothing to density. The result now follows by reassembling the pieces. ■

Unfortunately, from a practical standpoint the $\binom{n}{p}$ summands of (3.1.2) are not very tractable in most applications and, as in the least squares theory, we must resort to asymptotic approximations for a distribution theory adaptable to practical statistical inference. In the next section we try to provide a heuristic introduction to the asymptotic theory of quantile regression. A more detailed formal treatment of the asymptotic theory of quantile regression may be found in Chapter 4.

2. Some Asymptotic Heuristics

The optimization view of the sample quantiles also affords an elementary approach to their asymptotic theory. Suppose $Y_1, \ldots, Y_n$ are independent and identically distributed (iid) from the distribution F and for some quantile $\xi_\tau = F^{-1}(\tau)$ assume that F has a continuous density, f, at ξ_τ with $f(\xi_\tau) > 0$. The objective function of τth sample quantile,

$$\hat{\xi}_\tau \equiv \inf_{\xi}\{\xi \in \mathbb{R} | \sum \rho_\tau(Y_i - \xi) = \min!\}$$

as we have already noted, this objective function is the sum of convex functions, hence is itself convex. Consequently its gradient,

$$g_n(\xi) = n^{-1}\sum_{i=1}^{n}(I(Y_i < \xi) - \tau)$$

is monotone in ξ. Of course, when ξ equals one of the Y_i then this "gradient" needs the subgradient interpretation discussed above, but this is not crucial to the argument that follows. By monotonicity, $\hat{\xi}_\tau$ is greater than ξ if and only if $g_n(\xi) < 0$. so

$$
\begin{aligned}
P\{\sqrt{n}(\hat{\xi}_\tau - \xi_\tau) > \delta\} &= P\{g_n(\xi_\tau + \delta/\sqrt{n}) < 0\} \\
&= P\{n^{-1}\sum(I(Y_i < \xi_\tau + \delta/\sqrt{n}) - \tau) < 0\}.
\end{aligned}
$$

Thus, we have reduced the behavior of $\hat{\xi}_\tau$ to a DeMoivre-Laplace central limit theorem problem in which we have a triangular array of Bernoulli random variables. The summands take the values $(1-\tau)$ and $-\tau$ with probabilities $F(\xi_\tau+\delta/\sqrt{n})$ and $1-F(\xi_\tau+\delta/\sqrt{n})$. Since

$$Eg_n(\xi_\tau+\delta/\sqrt{n}) = (F(\xi_\tau+\delta/\sqrt{n})-\tau) \to f(\xi_\tau)\delta/\sqrt{n},$$

and

$$V(n^{-1}g_n(\xi_\tau+\delta/\sqrt{n})) = F(\xi_\tau+\delta/\sqrt{n})(1-F(\xi_\tau+\delta/\sqrt{n}))/n \to \tau(1-\tau)/n.$$

we may set $\omega^2 = \tau(1-\tau)/f^2(\xi_\tau)$ and write,

$$\begin{aligned} P\{\sqrt{n}(\hat{\xi}_\tau-\xi_\tau)>\delta\} &= P\left\{\frac{g_n(\xi_\tau+\delta/\sqrt{n})-f(\xi_\tau)\delta/\sqrt{n}}{\sqrt{\tau(1-\tau)/n}} < -\omega^{-1}\delta\right\} \\ &\to 1-\Phi(\omega^{-1}\delta) \end{aligned}$$

and therefore

$$\sqrt{n}(\hat{\xi}_\tau-\xi_\tau) \rightsquigarrow \mathcal{N}(0,\omega^2). \tag{3.2.1}$$

The $\tau(1-\tau)$ effect tends to make $\hat{\xi}_\tau$ more precise in the tails, but this would be typically dominated by the effect of the density term which tends to make $\hat{\xi}_\tau$ less precise in regions of low density.

Extending the foregoing argument to consider the limiting form of the joint distribution of several quantiles, set $\hat{\zeta}_n = (\hat{\xi}_{\tau_1},\dots,\hat{\xi}_{\tau_m})$ with $\zeta_n = (\xi_{\tau_1},\dots,\xi_{\tau_m})$ and we obtain, see Problem 3.1,

$$\sqrt{n}(\hat{\zeta}_n-\zeta) \rightsquigarrow \mathcal{N}(0,\Omega) \tag{3.2.2}$$

where $\Omega = (\omega_{ij}) = (\tau_i\wedge\tau_j - \tau_i\tau_j)/(f(F^{-1}(\tau_i)f(F^{-1}(\tau_j))$. This result is the starting point of the large sample theory for finite linear combinations of order statistics (L-statistics) which we have introduced in the previous chapter.

These results for the ordinary sample quantiles in the one-sample model generalize nicely to the classical linear regression model

$$y_i = x_i'\beta + u_i$$

with iid errors $\{u_i\}$. Suppose that the $\{u_i\}$ have common distribution function F with associated density f, with $f(F^{-1}(\tau_i)) > 0$ for $i=1,\dots,m$, and $n^{-1}\sum x_i x_i' \equiv Q_n$ converges to a positive definite matrix, Q_0. Then the joint asymptotic distribution of the m p-variate quantile regression estimators $\hat{\zeta}_n = (\hat{\beta}_n(\tau_1)',\dots,\hat{\beta}_n(\tau_m)')'$ takes the form,

$$\sqrt{n}(\hat{\zeta}_n-\zeta) = (\sqrt{n}(\hat{\beta}_n(\tau_j)-\beta(\tau_j)))_{j=1}^m = \mathcal{N}(0,\Omega\otimes Q_0^{-1}). \tag{3.2.3}$$

In the iid error regression setting, the form of the $\beta(\tau_j)$ is particularly simple as we have seen. The conditional quantile planes of $y|x$ are parallel so presuming that the first coordinate of β corresponds to the "intercept" parameter we have, $\beta(\tau) = \beta + \xi_\tau e_1$ where $\xi_\tau = F^{-1}(\tau)$ and e_1 is the first unit basis vector of $\mathbb{R}^p$. Since Ω takes the same form as in the one sample setting many of the classical results on L-statistics can be directly carried forward to iid error regression using this result.

This result, which is essentially the content of Theorem 4.1 of KB(1978) affords considerable scope for Wald-type inference in the quantile regression setting. Hypotheses which may be formulated as linear restrictions of the vector ζ, are immediately subject to test using the limiting normal theory and its chi-square adaptations. We now turn to the problem of estimating the asymptotic covariance matrix required for these tests.

3. Wald Tests

The classical theory of linear regression *assumes* that the conditional quantile functions of the response variable, y, given covariates, x, are all parallel to one another, implying that the slope coefficients of distinct quantile regressions will be identical. In applications, however, as we have seen, quantile regression slope estimates often vary considerably across quantiles, so an immediate and fundamental problem of inference in quantile regression involves testing for equality of slope parameters across quantiles.

Some simple tests designed for this purpose were suggested in Koenker and Bassett(1982). For the two-sample problem they correspond to tests of equality between the interquantile ranges of the two samples. Thus, they may be considered to be tests of homogeneity of scale, or tests for heteroscedasticity. Consider the two-sample model

$$Y_i = \alpha_1 + \alpha_2 x_i + u_i$$

where $x_i = 0$ for n_1 observations in the first sample and $x_i = 1$ for n_2 observations in the second sample. The τth regression quantile estimate of the "slope" parameter α_2 in this model is, simply the difference between the τth sample quantiles of the two samples. See Problem 2.2. Thus a test of the equality of the slope parameters across quantiles τ_1 and τ_2 is just a test of the hypothesis

$$\begin{aligned}\alpha_2(\tau_2) - \alpha_1(\tau_1) &= (Q_2(\tau_2) - Q_1(\tau_2)) - (Q_2(\tau_1) - Q_1(\tau_1))\\ &= (Q_2(\tau_2) - Q_2(\tau_1)) - (Q_1(\tau_2) - Q_1(\tau_1))\\ &= 0,\end{aligned}$$

i.e. that the $(\tau_2 - \tau_1)$-interquantile ranges are identical for the two samples. By (3.2.3) the asymptotic variance of $\hat{\alpha}_2(\tau_2) - \hat{\alpha}_1(\tau_1)$ is given by

$$\sigma^2(\tau_1,\tau_2) = \left[\frac{\tau_1(1-\tau_1)}{f^2(F^{-1}(\tau_1))} - 2\frac{\tau_1(1-\tau_2)}{f(F^{-1}(\tau_1))f(F^{-1}(\tau_2))} + \frac{\tau_2(1-\tau_2)}{f^2(F^{-1}(\tau_2))}\right]\left[\frac{n}{nn_1 - n_1^2}\right]$$

and a test of the null hypothesis can be based on the asymptotic normality of the statistic,

$$T_n = (\hat{\alpha}_2(\tau_2) - \hat{\alpha}_1(\tau_1))/\hat{\sigma}(\tau_1, \tau_2)$$

Of course, it is obvious from the form of $\sigma^2(\tau_1, \tau_2)$ that it is necessary to estimate the nuisance parameters, $1/f(F^{-1}(\tau_1))$ and $1/f(F^{-1}(\tau_2))$, a topic which is taken up in the next subsection.

3.1. Sparsity Estimation. It is a somewhat unhappy fact of life that the asymptotic precision of quantile estimates in general, and quantile regression estimates in particular, depend upon the reciprocal of a density function evaluated at the quantile of interest – a quantity Tukey(1965) has termed the "sparsity function" and Parzen(1979) calls the quantile-density function. It is perfectly natural that the precision of quantile estimates should depend on this quantity since it reflects the density of observations near the quantile of interest, if the data is very sparse at the quantile of interest it will be difficult to estimate. On the other hand, when the sparsity is low, so observations are very dense, then the quantile will be more precisely estimated. Thus, to estimate the precision of the τth quantile regression estimate directly, the nuisance quantity

$$s(\tau) = [f(F^{-1}(\tau))]^{-1}$$

must be estimated, and this leads us into the realm of density estimation and smoothing. In fact, we shall see that it may be possible to pull oneself out of this swamp by the bootstraps, or other statistical necromancy, but we defer the exploration of these strategies a bit, and begin by exploring the direct approach to estimating the asymptotic covariance matrix.

Luckily, there is a large literature on estimating $s(\tau)$ in the one-sample model, including Siddiqui (1960), Bofinger (1975), Sheather and Maritz (1983), Welsh (1986) and Hall and Sheather (1988). Siddiqui's idea is simplest and has received the most attention in the literature so we will focus on it. Differentiating the identity, $F(F^{-1}(t)) = t$ we find that the sparsity function is simply the derivative of the quantile function, i.e.,

$$\frac{d}{dt} F^{-1}(t) = s(t).$$

So, just as differentiating the distribution function, F, yields he density function, f, differentiating the quantile function, F^{-1}, yields the sparsity function s. It is therefore natural, following Siddiqui, to estimate $s(t)$ by using a simple difference quotient of the empirical quantile function, i.e.,

$$\hat{s}_n(t) = [\hat{F}_n^{-1}(t + h_n) \; - \; \hat{F}_n^{-1}(t - h_n)]/2h_n$$

where $\hat{F}^{-1}$ is an estimate of F^{-1} and h_n is a bandwidth which tends to zero as $n \to \infty$. A bandwidth rule suggested Hall and Sheather (1988) based on Edgeworth expansions for Studentized quantiles is

$$h_n = n^{-1/3} z_\alpha^{2/3} [1.5 s(t)/s''(t)]^{1/3}$$

where z_α satisfies $\Phi(z_\alpha) = 1 - \alpha/2$ for the construction of $1-\alpha$ confidence intervals. In the absence of other information about the form of $s(\cdot)$ we may use the Gaussian model to select the bandwidth h_n, which yields

$$h_n = n^{-1/3} z_\alpha^{2/3} [1.5\phi^2(\Phi^{-1}(t))/(2(\Phi^{-1}(t))^2 + 1)]^{1/3}$$

Having chosen a bandwidth h_n the next question is: how should we compute $\hat{F}^{-1}$? The simplest approach seems to be to use the residuals from the quantile regression fit. Let $r_i : i = 1, ..., n$ be these residuals, and $r_{(i)} : i = 1, ..., n$, be the corresponding order statistics. One may define the usual empirical quantile function, $\hat{F}^{-1}(t) = r_{(j)}$ for $t \in [(j-1)/n,\ j/n\ )$. Alternatively, one may wish to interpolate to get a piecewise linear version

$$\tilde{F}^{-1}(t) = \begin{cases} r_{(1)} & \text{if } t \in [0, 1/2n) \\ \lambda r_{(j+1)} + (1-\lambda) r_{(j)} & \text{if } t \in [(2j-1)/2n, (2j+1)/2n)\ j = 1, ..., n-1 \\ r_{(n)} & \text{if } t \in [(2n-1)/2n, 1] \end{cases}$$

where $\lambda = tn - j + 1/2$. Other schemes are obviously possible. A possible pitfall of the residual-based estimates of F^{-1} is that if the number of parameters estimated, say p, is large relative to n, then since there must be p residuals equal to zero at the fitted model, we must make sure that the bandwidth is large enough to avoid these zero residuals. The simplest approach seems to be to ignore the zero residuals entirely in the construction of $\hat{F}^{-1}$ and $\tilde{F}^{-1}$ and treat the effective sample size as $n-p$. We may think of the deletion of these zero residuals as a direct analogue of the usual degrees freedom adjustment made when computing an estimate of the regression variance in least squares regression. In applications with discrete y's it is sometimes possible to have more than p zero residuals, in the terminology of linear programming this phenomenon is called degeneracy. It is prudent to delete all such residuals before proceeding with the estimation of $s(\tau)$.

An alternative, perhaps less obvious, approach to estimating F^{-1} is to employ the empirical quantile function suggested in Bassett and Koenker (1982). In effect this amounts to using $\hat{F}_Y^{-1}(t) = \bar{x}'\hat{\beta}(t)$ where $\hat{\beta}(\cdot)$ is the usual regression quantile process. As we have emphasized already, the functions

$$\hat{Q}_Y(\tau|x) = x'\hat{\beta}(\tau)$$

constitute a family of conditional quantile functions for the response variable Y. At any fixed x we can regard $\hat{Q}_Y(\tau|x)$ as a viable estimate of the conditional quantile function of Y given x. Of course, the precision of this estimate depends upon the x

at which we evaluate the expression, but the precision is maximized at $x = \bar{x}$. This makes $\hat{F}_Y^{-1}(t) = \bar{x}'\hat{\beta}(t)$ an attractive choice, but we should verify that as a function of τ this function satisfies the fundamental monotonicity requirement of a quantile function. It is clear from the examples that we have already seen that the estimated conditional quantile functions fitted by quantile regression may cross – indeed this is inevitable since the estimated slope coefficients are not identical and therefore the functions are not parallel. One might hope, and expect, that this crossing occurred only in the remote regions of design space – near the centroid of the design, $\bar{x}$, crossing should not occur. This "wishful thinking" is supported by the following result.

THEOREM 3.2. *The sample paths of $\hat{Q}_Y(\tau|\bar{x})$ are non-decreasing in τ on $[0, 1]$.*

Proof: We will show that

$$\tau_1 < \tau_2 \quad \Rightarrow \quad \bar{x}'\hat{\beta}(\tau_1) \leq \bar{x}'\hat{\beta}(\tau_2). \tag{3.3.1}$$

We first note a simple property of the quantile regression objective function. For any $b \in \boldsymbol{R}^p$,

$$\sum_{i=1}^{n} \Big[\rho_{\tau_2}(Y_i - x_i'b) - \rho_{\tau_1}(Y_i - x_i'b) \Big] = n(\tau_2 - \tau_1)(\bar{Y} - \bar{x}'b) \tag{3.3.2}$$

This equation follows directly from the definition of ρ_τ:

$$\begin{aligned} \rho_{\tau_1}(Y_i - x_i't) - \rho_{\tau_2}(Y_i - x_i't) & \\ = (\tau_2 - \tau_1)(Y_i - x_i't)^+ + \big[(1-\tau_2) - (1-\tau_1)\big](Y_i - x_i't)^- & \\ = (\tau_2 - \tau_1)\Big[(Y_i - x_i't)^+ - (Y_i - x_i't)^-\Big] & \\ = (\tau_2 - \tau_1)(Y_i - x_i't)\,. & \end{aligned} \tag{3.3.3}$$

Now, using the definition of $\hat{\beta}(\tau)$ as a minimizer of ρ_τ, and applying (3.3.2) with $b = \bar{x}'\hat{\beta}(\tau_k)$ for $k = 1, 2$, we have

$$\begin{aligned}
&\sum_{i=1}^{n} \rho_{\tau_1}\big(Y_i - x_i'\hat{\beta}(\tau_1)\big) + n(\tau_2 - \tau_1)\big(\bar{Y} - \bar{x}'\hat{\beta}(\tau_2)\big) \\
&\qquad \le \sum_{i=1}^{n} \rho_{\tau_1}\big(Y_i - x_i'\hat{\beta}(\tau_2)\big) + n(\tau_2 - \tau_1)\big(\bar{Y} - \bar{x}'\hat{\beta}(\tau_2)\big) \\
&\qquad = \sum_{i=1}^{n} \rho_{\tau_2}\big(Y_i - x_i'\hat{\beta}(\tau_2)\big) \qquad (3.3.4) \\
&\qquad \le \sum_{i=1}^{n} \rho_{\tau_2}\big(Y_i - x_i'\hat{\beta}(\tau_1)\big) \\
&\qquad = \sum_{i=1}^{n} \rho_{\tau_1}\big(Y_i - x_i'\hat{\beta}(\tau_1)\big) + n(\tau_2 - \tau_1)\big(\bar{Y} - \bar{x}'\hat{\beta}(\tau_1)\big).
\end{aligned}$$

Simplifying, we see that this is equivalent to

$$n(\tau_2 - \tau_1)\big(\bar{x}'\hat{\beta}(\tau_2) - \bar{x}'\hat{\beta}(\tau_1)\big) \ge 0, \tag{3.3.5}$$

from which Theorem 3.2 follows immediately. ■

To illustrate the application of this approach consider the data depicted in Figure 3.1 This is a classical data set in economics and is based on 235 budget surveys of 19th century working class households. Household expenditure on food is measured vertically, and household income is measured horizontally. The data was originally presented by Ernst Engel(1857) to support his hypothesis that food expenditure constitutes a declining share of household income. Following established custom we have transformed both variables to the log-scale, so a slope less than unity is evidence for Engel's Law. We have superimposed 6 estimated linear conditional quantile functions for these data. The fitted lines look roughly parallel, but we might like a formal test of the hypothesis of equality of slopes. Focusing on the inner two lines which represent the fit for the first ($\tau = .25$) and third ($\tau = .75$) quartiles, we find that the difference in the slopes is,

$$\hat{\beta}_2(3/4) - \hat{\beta}_2(1/4) = 0.915 - 0.849 = 0.0661.$$

In Figure 3.2 we illustrate the function $\hat{Q}_Y(\tau|\bar{x})$ for this dataset. The vertical scale is the natural logarithm of food expenditure. The dotted lines forming triangles illustrate the estimation of the sparsity function at the first and third quartiles. The Hall-Sheather bandwidth for both estimates is .097, yielding sparsity estimates of $\hat{s}(1/4) = 0.543$ and $\hat{s}(3/4) = 0.330$. The lower diagonal element of $(X'X)^{-1}$ is 0.022 so the test statistic for the equality of the two slopes is 1.93, which has a p-value of

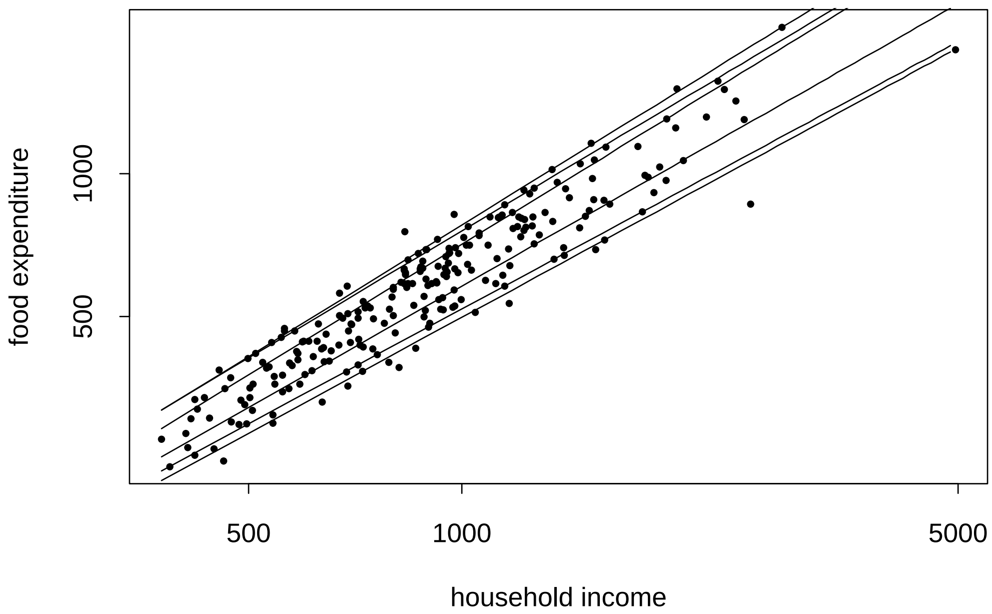

FIGURE 3.1. Engel Curves for Food: This figure plots data taken from Ernst Engel's (1857) study of households' expenditure on food versus annual income. The data consists of 235 observations on European working class households. Superimposed on the plot are six estimated quantile regression lines corresponding to the quantiles $\tau \in \{.05, .1, .25, .75, .9, .95\}$.

.03 for a one-tailed test of the hypothesis of equality of the slopes. This result offers very weak evidence of increasing dispersion in the logarithm of food expenditure with income, a finding which may seem suprising in view of Figure 3.1. In large samples the formal statistical significance of such tests is extremely common. The substantive significance of such heteroscedasticity is, of course, completely application dependent.

There are obviously a number of other possible approaches to the estimation of the sparsity parameter. Welsh(1986) considers a kernel approach which may be interpreted as a weighted average of Siddiqui estimates in which those with narrow bandwidth are given greater weight. Another approach is suggested by simply fitting a local polynomial to $\hat{F}^{-1}(t)$ in the neighborhood of τ, and using the slope of this fitted function at τ as an estimate of sparsity. In Koenker and Bassett (1982) the

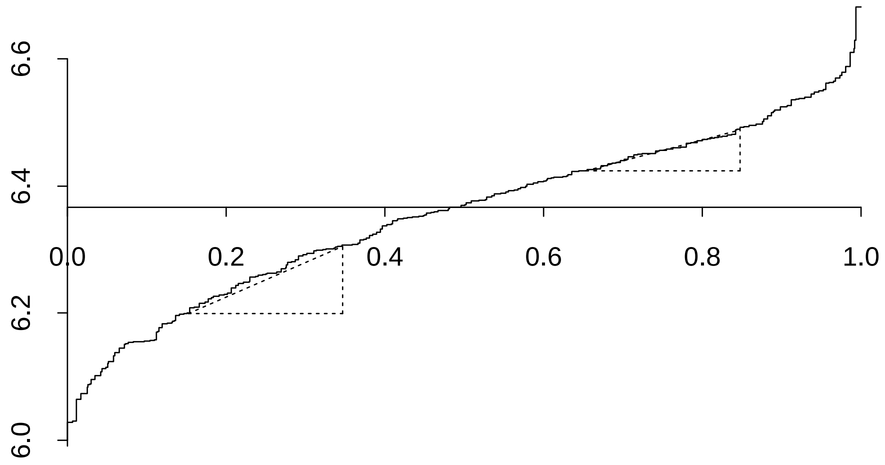

FIGURE 3.2. Sparsity Estimation for the Engel Data: This figure plots $\hat{Q}_Y(\tau|\bar{x}) = \bar{x}'\hat{\beta}(\tau)$ for the Engel data. The vertical scale is logarithmic in expenditure on food. The estimation of the sparsity function by the Siddiqui method is illustrated by the dotted triangles which depict the difference quotient estimator of the sparsity at the first and third quartiles. The estimate of the sparsity is given by the slope of the hypotenuse of the triangles. The Hall-Sheather bandwidth is .097 for this example.

histospline methods of Boneva, Kendall, and Stefanov (1971) are employed to estimate the sparsity function.

4. Inference in non-iid Error Models

The classical iid error linear regression model yields a particularly simple form for the limiting distribution of the quantile regression estimator $\hat{\beta}(\tau)$. However, it might be argued that in the location shift form of the iid error model quantile regression is really superfluous; in this case a reasonable estimator of the conditional central tendency of the response given the covariates is perfectly adequate. In non-iid error settings like the conditional location-scale model introduced in Section 2.5, the asymptotic theory of $\hat\beta(\tau)$ is somewhat more complicated. As we shall show in Section 4.x the limiting covariance matrix of $\sqrt{n}(\hat\beta(\tau) - \beta(\tau))$ takes the form of a Huber sandwich, i.e.,

$$\sqrt{n}(\hat\beta(\tau) - \beta(\tau)) \rightsquigarrow \mathcal{N}(0, H_n^{-1} J_n H_n^{-1})$$

where

$$J_n(\tau) = \tau(1-\tau) n^{-1} \sum_{i=1}^{n} x_i x_i'$$

and

$$H_n(\tau) = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} x_i x_i' f_i(\xi_i(\tau)).$$

The term $f_i(\xi_i(\tau))$ denotes the conditional density of the response, y_i, evaluated at the τ-th conditional quantile. In the iid case these f_i are identical, and the sandwich collapses to the expression we have already considered. We are then faced with the relatively simple problem of estimating a density, or its reciprocal, at a point. However, in the non-iid case we face a more challenging task.

We will describe two approaches to the estimation of the matrix H_n. One is a natural extension of sparsity estimation methods described above, and was suggested by Hendricks and Koenker (1992). The other which is based on kernel density estimation ideas was proposed by Powell (1989).

The Hendricks-Koenker Sandwich

Provided that the τ-th conditional quantile function of $y|x$ is linear, then for $h_n \to 0$ we can consistently estimate the parameters of the $\tau \pm h_n$ conditional quantile functions by $\hat\beta(\tau \pm h_n)$. And the density $f_i(\xi_i)$ can thus be estimated by the difference quotient

$$\hat f_i(\xi_i(\tau)) = 2h_n / x_i'(\hat\beta(\tau + h_n) - \hat\beta(\tau - h_n)),$$

using the same choice of bandwidth discussed above. Substituting this estimate in the expression for H_n above yields an easily implementable estimator for the asymptotic covariance matrix of $\hat\beta(\tau)$ in the non-iid error model. Note that the matrix J_n involves no nuisance parameters and thus is immediately computable.

A potential difficulty with the proposed estimator $\hat f_i(\xi_i(\tau))$ is that there is no guarantee of positivity for every observation in the sample. Indeed, as we have already seen, the quantity

$$d_i = x_i'(\hat\beta(\tau + h_n) - \hat\beta(\tau - h_n))$$

is *necessarily* positive only at $x = \bar x = n^{-1} \sum x_i$. Nevertheless, in practice we find that problems due to "crossing" of the estimated conditional quantile planes occurs only infrequently and in the most extreme regions of the design space. In our implementation of this approach we simply replace $\hat{f}_i$ by its positive part, i.e., we use

$$\hat{f}_i^+ = \max\{0, 2h_n/(d_i - \varepsilon)\}$$

where $\varepsilon > 0$ is a small tolerance parameter intended to avoid dividing by zero in the (rare) cases in which $d_i = 0$ because the ith observation is basic at both $\tau \pm h_n$.

The foregoing approach may be extended easily to the problem of estimating asymptotic covariance matrices for distinct vectors of quantile regression parameters. In these cases we would like to estimate,

$$\text{acov } (\sqrt{n}(\hat{\beta}(\tau_1) - \beta(\tau_1)), \sqrt{n}(\hat{\beta}(\tau_2) - \beta(\tau_2)) = H_n(\tau_1)^{-1} J_n(\tau_1, \tau_2) H_n(\tau_2)^{-1}$$

where now

$$J_n(\tau_1, \tau_2) = [\tau \wedge \tau_2 - \tau_1\tau_2] n^{-1} \sum x_i x_i'.$$

Thus, Wald tests, like the heteroscedasticity tests described above, that involve linear restrictions across several quantile regression parameter vectors can be easily carried out with the same machinery. It should be emphasized that this approach is computationally extremely efficient and thus is particularly attractive for large problems where bootstrapping and the rank test inversion approaches discussed below are impractical.

The Powell Sandwich

Powell (1989) has suggested an alternative, and in some ways even simpler, way to estimate the quantile regression sandwich. Noting that in estimating the matrix $H_n(\tau)$ we are really after a matrix weighted density estimator, he proposes a kernel estimator of the form

$$\hat{H}_n(\tau) = (nh_n)^{-1} \sum K(\hat{u}_i(\tau)/h_n) x_i x_i'$$

where $\hat{u}_i(\tau) = y_i - x_i'\hat{\beta}(\tau)$ and h_n is a bandwidth parameter satisfying $h_n \to 0$ and $\sqrt{n}h_n \to \infty$. He shows that under certain uniform Lipschitz continuity requirements on the f_i, $\hat{H}_n(\tau) \to H_n(\tau)$ in probability. In practice, of course, there remain a number of nettlesome questions about the choice of the kernel function K and the bandwidth parameter h_n we will have more to say about this in Section 3.x where we describe a small Monte-Carlo experiment designed to explore the performance of various inference strategies for quantile regression.

4.1. Other Hypotheses. More general hypotheses are easily accommodated by the Wald approach. For example, as in Koenker and Bassett (1982), we may consider a general linear hypothesis on the vector $\zeta = (\beta(\tau_1)', ..., \beta(\tau_m)')'$ of the form

$$H_0: \; H\zeta = h$$

and test statistic

$$T_n = (H\hat{\zeta} - h)'[H(\Omega \otimes (X'X)^{-1})H']^{-1}(H\hat{\zeta} - h)$$

which is asymptotically χ^2 under H_0. This formulation accommodates a wide variety of testing situations, from simple tests on a single quantile regression coefficient to joint tests involving several covariates and several distinct quantiles. Thus, for example, we might test for the equality of several slope coefficients across several quantiles; such tests provide a robust alternative to conventional least-squares based tests of heteroscedasticity because they can be constructed to be insensitive to outlying response observations. The same formulation can, of course, be adopted to accommodate nonlinear hypotheses on the vector, ζ, by interpreting H_0 above as the Jacobian of the nonlinear hypothesis. Newey and Powell(1987) discuss tests for symmetry employing this approach.

The inherent difficulty of estimating the sparsity, which comes from the form of the the asymptotic covariance matrix Ω, can be avoided by adopting one of several available strategies. One involves replacing hypotheses about a few discrete quantile regression parameters by hypotheses about smoothly weighted functions of quantile regression parameters, another strategy involves turning to the dual formulation of the quantile regression problem and adopting a testing approach which is closely connected to the classical theory of rank tests. We will briefly consider the former approach, before addressing rank tests in the next section.

Gutenbrunner (1994) considers tests based on L-statistics of the form,

$$S_n = \int \hat{\beta}(\tau) d\nu.$$

If ν has a "signed density" J of bounded variation with respect to Lebesgue measure, then the asymptotic variance of S_n under the iid error model takes the form,

$$(3.4.6)\quad \sigma^2(\nu, F) = \int_0^1 \int_0^1 \omega(u,v) d\nu(u) d\nu(v)$$
$$= -\int\int_{u<v} (F^{-1}(u) - F^{-1}(v))^2 d\tilde{J}(u) dJ(v) - (\int_0^1 F^{-1}(u) d\tilde{J}(u))^2,$$

where $\omega(u,v) = (u \wedge v - uv)/f(F^{-1}(u)f(F^{-1}(v)$, and $\tilde{J}(u) = uJ(u)$. The principal advantage of tests based on smooth L-statistics like this is that their asymptotic variance can be estimated at rate $\mathcal{O}_p(n^{-1/2})$ in contrast to tests based on discrete linear combinations of regression quantiles whose asymptotic covariance requires estimation of the sparsity function at a point. Sparsity estimation is circumvented by the use of the smooth weight function, and we can estimate $\sigma^2(\nu, F)$ at rate $\mathcal{O}_p(n^{-1/2})$ by simply plugging in $\hat{Q}_Y(u|\bar{x}) = \bar{x}'\hat{\beta}(u)$ for $F^{-1}(u)$ in (3.4.6).

5. Rank Tests

The classical theory of rank tests as developed in the monograph of Hájek and Šidák (1967) begins with the rankscore functions,

$$\hat{a}_{ni}(t) = \begin{cases} 1 & \text{if } t \leq (R_i - 1)/n \\ R_i - tn & \text{if } (R_i - 1)/n < t \leq R_i/n \\ 0 & \text{if } R_i/n < t \end{cases} \tag{3.5.1}$$

where R_i is the rank of the $i^{\mathbf{th}}$ observation, Y_i, in the sample $\{Y_1, \ldots, Y_n\}$. Integrating $\hat{a}_{ni}(t)$ with respect to various score generating functions φ yields vectors of rank-like statistics which may be used for constructing tests. For example, integrating with respect to Lebesgue measure yields the Wilcoxon scores,

$$b_i = \int_0^1 \hat{a}_{ni}(t)dt = (R_i - 1/2)/n \qquad i = 1, ..., n,$$

while using $\varphi(t) = \text{sgn}(t - 1/2)$ yields the sign scores, $b_i = \hat{a}_{ni}(1/2)$. Invariance of the ranks to monotone transformations means that the R_i's may also be viewed as the ranks of the uniform random sample $\{U_1, ..., U_n\}$ with $U_i = F(Y_i)$, and the rank generating functions $\hat{a}_i(t)$ may be seen as replacing the indicator functions $I(Y_i > F^{-1}(t)) = I(U_i > t)$, by the smoother "trapezoidal" form given by (3.5.1) Thus the rank generating functions behave like an empirical process as the following result shows.

THEOREM 3.3. *(Hájek and Šidák (1967, Thm V.3.5)) Let $(c_{1n}, ..., c_{nn})$ be a triangular array of real numbers satisfying*

$$\max(c_{in} - \bar{c}_n)^2 / \sum_{i=1}^{n} (c_{in} - \bar{c}_n)^2 \to 0 \tag{3.5.2}$$

where $\bar{c}_n = n^{-1} \sum c_{in}$, and assume that $\{Y_1, ..., Y_n\}$ constitute a random sample from an absolutely continuous distribution F. Then, the process

$$Z_n(t) = [\sum_{i=1}^{n} (c_{in} - \bar{c}_n)^2]^{-1/2} \sum_{i=1}^{n} (c_{in} - \bar{c}_n) \hat{a}_i(t)$$

converges weakly to a Brownian Bridge process on $C[0,1]$.

In the two sample problem the c_{in}'s may be taken as simply the indicator of which sample the observations belong to, and the "Lindeberg condition" (3.5.2) is satisfied as long as n_1/n stays bounded away from 0 and 1. A limiting normal theory for a broad class of linear rank statistics of the form

$$S_n = [\sum (c_{in} - \bar{c}_n)^2]^{-1/2} \sum (c_{in} - \bar{c}_n) \hat{b}_i$$

where $\hat{b}_i = -\int \varphi(t) d\hat{a}_i(t)$ is immediate. In particular, for square integrable $\varphi : [0,1] \to \mathbb{R}$ we have the linear representation

$$S_n = [\sum (c_{in} - \bar{c}_n)^2]^{-1/2} \sum (c_{in} - \bar{c}_n)\varphi(U_i) + o_p(1), \tag{3.5.3}$$

and consequently, S_n is asymptotically Gaussian under the null with mean 0 and variance, $A^2(\varphi) = \int (\varphi(t) - \bar{\varphi})^2 dt$, where $\bar{\varphi} = \int \varphi(t) dt$.

Thus, for example, in the two-sample location shift model with local alternatives $H_n : \delta_n = \delta_0 / \sqrt{n}$ we have S_n asymptotically Gaussian with mean $\omega(\varphi, F)(\sum (c_{in} - \bar{c}_n)^2)^{1/2} \delta_0$ and variance $A^2(\varphi)$ where

$$\omega(\varphi, F) = \int f(F^{-1}(t) d\varphi(t).$$

An important virtue of such rank tests is that the test statistic and its limiting behavior under the null hypothesis are independent of the distribution F generating the observations. See Draper(1988) for a detailed discussion of the problems related to the estimation the the nuisance parameter ω in the Wilcoxon case.

How can these ideas be extended to regression when, under the null, a nuisance regression parameter is present? This question was answered by Gutenbrunner and Jurečková(1992) who observed that the Hájek-Šidák rankscores may be viewed as a special case of a more general formulation for the linear model in which the functions $\hat{a}_{ni}(t)$ are defined in terms of the linear program

$$\max\{y'a | X'a = (1-t)X'1,\ a \in [0,1]^n\} \tag{3.5.4}$$

This problem is formally dual to the linear program (1.3.9) defining the regression quantiles. Algorithmic details are given in Koenker and d'Orey (1993). As developed in Gutenbrunner, Jureckova, Koenker and Portnoy (1993), tests of the hypothesis $\beta_2 = 0 \in \mathbb{R}^q$ in the model $y = X_1\beta_1 + X_2\beta_2 + u$ based on the regression rankscore process may be constructed by first computing $\{\hat{a}_{ni}(t)\}$ at the restricted model,

$$y = X_1\beta_1 + u$$

computing the n-vector b with elements $b_i = -\int \varphi(t) d\hat{a}_{ni}(t)$, forming the q-vector,

$$S_n = n^{-1/2} X_2' b,$$

and noting that, under the null $S_n \rightsquigarrow \mathcal{N}(0, A^2(\varphi)Q_0)$ where $A^2(\varphi) = \int_0^1 \varphi^2(t) dt$, $Q_0 = lim_{n\to\infty} Q_n$, $Q_n = (X_2 - \hat{X}_2)'(X_2 - \hat{X}_2)/n$ and $\hat{X}_2 = X_1(X_1'X_1)^{-1}X_1'X_2$. So the test statistic $T_n = S_n' Q^{-1} S_n / A^2(\varphi)$ has an asymptotic χ^2_q null distribution.

In the special case that X_1 is simply a column vector of ones this reduces to the original formulation of Hájek and Šidák. When $\varphi(t) = \text{sgn}(t - 1/2)$ it specializes to the score-test proposed for ℓ_1-regression in Koenker and Bassett (1982), with the important added feature that it resolves the ambiguity about how to treat the (basic) observations with zero residuals.

An important advantage of these rank tests is that they require no estimation of nuisance parameters, since the functional $A(\varphi)$ depends only on the score function and not on the (unknown) distribution of the vector u. This is familiar from the theory of elementary rank tests, but stands in sharp contrast with other methods of testing in the linear model where, typically, some estimation of a scale parameter is needed. For example, in the case of least squares theory, σ^2, is needed. And in the case of quantile regression an estimate the sparsity function is needed for Wald-type tests.

Having found a testing strategy for quantile regression which avoids the estimation of the sparsity function naturally raises the question: could we invert a rank test of this form to provide a method of estimating a confidence interval for quantile regression parameters, thereby circumventing the problem of estimating $s(t)$. Huskova(1994) considers this problem in considerable generality establishing the validity of sequential fixed-width confidence intervals based on rank test inversion for general score functions φ. Unfortunately, for general score functions these intervals are quite difficult to compute. However, in the case of a fixed quantile one particularly natural choice of φ yields extremely tractible computations and we will focus on this case.

These regression rankscore tests may also be viewed as Rao-score tests, or LM tests, since they are based on a restricted estimate of the model under the null hypothesis and involve integrating the gradient of the unrestricted quantile regression problem over the interval $[0, 1]$. This connection is particularly clear in the case of the quantile score function described in the next section.

5.1. Confidence Intervals for $\hat{\beta}(\tau)$ by Inverting Rank Tests. Specializing to the scalar β_2 case and using the τ-quantile score function

$$\varphi_\tau(t) = \tau - I(t < \tau)$$

and proceeding as above, we find that

$$\hat{b}_{ni} = -\int_0^1 \varphi_\tau(t) d\hat{a}_{ni}(t) = \hat{a}_{ni}(\tau) - (1 - \tau) \tag{3.5.5}$$

with

$$\bar{\varphi} = \int_0^1 \varphi_\tau(t) dt = 0$$

and

$$A^2(\varphi_\tau) = \int_0^1 (\varphi_\tau(t) - \bar{\varphi})^2 dt = \tau(1 - \tau).$$

Thus, a test of the hypothesis $H_0 : \beta_2 = \xi$ may be based on $\hat{a}_n$ from solving,

$$\max\{(y - x_2\xi)' a | X_1' a = (1 - \tau) X_1' 1, a \in [0, 1]^n\} \tag{3.5.6}$$

and the fact that

$$S_n(\xi) = n^{-1/2} x_2' \hat{b}_n(\xi) \rightsquigarrow \mathcal{N}(0, A^2(\varphi_\tau) q_n^2) \tag{3.5.7}$$

under H_0; where $q_n^2 = n^{-1} x_2'(I - X_1(X_1'X_1)^{-1}X_1')x_2$. That is, we may compute

$$T_n(\xi) = S_n(\xi)/(A(\varphi_\tau)q_n)$$

and reject H_0 if $|T_n(\xi)| > \Phi^{-1}(1-\alpha/2)$.

0his takes us back to the linear program (3.5.6) which may now be viewed as a one parameter parametric linear programming problem in ξ. In ξ the dual vector $\hat{a}_n(\xi)$ is piecewise constant; ξ may be altered without compromising the optimality of $\hat{a}_n(\xi)$ as long as the sign of the residuals in the primal quantile regression problem do not change. When ξ gets to such a boundary the solution does change, but optimality may be restored by taking one simplex pivot. The process may continue in this way until $T_n(\xi)$ exceeds the specified critical value. Since $T_n(\xi)$ is piecewise constant we can interpolate in ξ to obtain the desired level for the confidence interval. See Beran and Hall (1993) for a detailed analysis of the effect of interpolation like this in the case of confidence intervals for ordinary quantiles. This interval, unlike the Wald type sparsity intervals, is not symmetric; but it *is* centered on the point estimate $\hat{\beta}_2(\tau)$ in the sense that $T_n(\hat{\beta}_2(\tau)) = 0$. This follows immediately from the constraint $X'\hat{a} = (1-\tau)X'1$ in the full problem.

The primary virtue of this approach is that it inherits the scale invariance of the test statistic T_n and therefore circumvents the problem of estimating the sparsity function. Implemented in S, using an adaptation of the algorithm described in Koenker and d'Orey (1993), it has essentially the same computational efficiency as the sparsity methods. A more detailed description of the computational aspects of the parametric programming problem is provided in Section 6.3.

6. Likelihood Ratio Tests

Having now introduced variants of the Wald and score test for quantile regression, it is natural to investigate analogues of the likelihood ratio test as well. In Koenker and Bassett (1982) it was shown that for median regression a test of the hypothesis

$$H_0 : R\beta = r \tag{3.6.1}$$

in the iid-error linear model,

$$y_i = x_i'\beta + u_i \tag{3.6.2}$$

could be based on the statistic

$$T_n = 8(\tilde{V}(\tfrac{1}{2}) - \hat{V}(\tfrac{1}{2}))/s(\tfrac{1}{2}) \tag{3.6.3}$$

where we will denote the value of the objective function at the unrestricted minimizer $\hat{\beta}(\frac{1}{2})$ by,

$$\hat{V}(\tau) = \min_{\{b \in \mathbb{R}^p\}} \sum \rho_\tau(y_i - x_i' b) \tag{3.6.4}$$

and

$$\tilde{V}(\tau) = \min_{\{b \in \mathbb{R}^p | Rb = r\}} \sum \rho_\tau(y_i - x_i' b) \tag{3.6.5}$$

denotes the value under the restricted estimator $\tilde{\beta}(\tau)$. It was shown that under H_0, T_n is asymptotically χ^2_q where $q = \text{rank}(R)$. That this statistic is related to a likelihood ratio is easy to see. Consider the standard Laplace (double exponential) density,

$$f(u) = \tfrac{1}{4}\exp(-\tfrac{1}{2}|u|).$$

Under the assumption that $\{u_i\}$ in (**??**) comes from this density, we have the log-likelihood,

$$\ell(\beta) = -n\log(4) - \tfrac{1}{2}\sum_{i=1}^{n} |y_i - x_i\beta|$$

and the likelihood ratio statistic for testing H_0 would be

$$2(\ell(\hat{\beta}) - \ell(\tilde{\beta})) = 2(\tilde{V}(1/2) - \hat{V}(1/2)),$$

which we would expect to have a χ^2_q asymptotic distribution under H_0. How does this relate to the result (**??**)? Where does the factor 8 come from? Note that in the standard Laplace case $s(1/2) = (f(0))^{-1} = 4$, so the usual theory of the likelihood ratio is vindicated – for this very special case, twice the log likelihood ratio converges in distribution to χ^2_q under H_0. However, when the standard Laplace assumption fails to hold, $s(1/2)$ may be quite different from 4 and the likelihood ratio statistic needs to be modified accordingly.

The simplest example of this, an example which yields another variant of the likelihood ratio, is the Laplace density with free scale parameter, σ. Now,

$$f(u) = \tfrac{1}{4\sigma}\exp\{-\tfrac{1}{2\sigma}|u|\}.$$

so the log-likelihood is

$$\ell(\beta, \sigma) = -n\log(4\sigma) - \tfrac{1}{2\sigma}\sum |y_i - x_i\beta|.$$

It is easily seen that the maximum likelihood estimator of σ is

$$\hat{\sigma} = n^{-1}\hat{V}(\tfrac{1}{2}) = \tfrac{1}{2n}\sum |y_i - x_i'\hat{\beta}|$$

for the unrestricted model and similarly, $\tilde{\sigma} = n^{-1}\tilde{V}(\frac{1}{2})$ is the mle under H_0. Concentrating the likelihood with respect to β we have

$$\ell(\hat{\beta}, \hat{\sigma}) = -n\log\hat{\sigma} - \tfrac{n}{2} - n\log 4$$

and with an analogous definition of $\ell(\tilde{\beta}, \tilde{\sigma})$ the likelihood ratio statistic becomes,

$$2(\ell(\hat{\beta}, \hat{\sigma}) - \ell(\tilde{\beta}, \tilde{\sigma})) = 2n \log(\tilde{\sigma}/\hat{\sigma})$$

Again, we are entitled to expect that the likelihood ratio statistic will have a limiting χ^2_q behavior. This can be easily checked using the earlier result by writing

$$\log(\tilde{\sigma}/\hat{\sigma}) = \log(1 + (\tilde{\sigma} - \hat{\sigma})/\hat{\sigma}) \approx (\tilde{\sigma} - \hat{\sigma})/\hat{\sigma} \tag{3.6.6}$$

an approximation whose validity is easy to establish under the null. Dividing numerator and denominator by $\sigma \equiv E|u|$, and arguing that $\hat{\sigma}/\sigma \to 1$, we have

$$2n \log(\tilde{\sigma}/\hat{\sigma}) = 2n(\tilde{\sigma} - \hat{\sigma})/\sigma + o_p(1) = 2(\tilde{V}(1/2) - \hat{V}(1/2))/\sigma + o_p(1) \tag{3.6.7}$$

Noting that $s(1/2) = 4\sigma$ in this case completes the argument.

While the unknown-scale form of the Laplace model is certainly preferable to the original, fixed-scale form, it is still unsatisfactory in that any departure from the condition $s = 4\sigma$ wreaks havoc with the asymptotic behavior of test statistic under the null. But this is easily rectified by defining the modified likelihood ratio statistics,

$$L_n(1/2) = 8(\tilde{V}(1/2) - \hat{V}(1/2))/s(1/2), \tag{3.6.8}$$

or again using (3.6.6),

$$\Lambda_n(1/2) = 8n\sigma \log(\tilde{\sigma}/\hat{\sigma})/s(1/2), \tag{3.6.9}$$

which may be shown to be asymptotically equivalent with limiting χ^2_q behavior. By directly evaluating the effect of the restriction on the mean absolute deviation from the median regression estimate this result provides a useful complement to alternative Wald and score formulations of tests of H_0. The modified likelihood ratio tests obviously require estimation of the nuisance parameters σ and s, but this is quite straightforward. For σ we may simply use $n^{-1}\hat{V}(1/2)$, while for s we may use the sparsity estimation methods described above.

The same approach we have elaborated for the median may be extended immediately to other quantiles. Define $\hat{V}(\tau)$ and $\tilde{V}(\tau)$ as in Section 1, and let $\hat{\sigma}(\tau) = n^{-1}\hat{V}(\tau)$, $\tilde{\sigma}(\tau) = n^{-1}\tilde{V}(\tau)$, and for the τth quantile consider the test statistics:

$$L_n(\tau) = \frac{2}{\lambda^2(\tau)s(\tau)}[\tilde{V}_n(\tau) - \hat{V}_n(\tau)]$$

and

$$\Lambda_n(\tau) = \frac{2n\hat{\sigma}(\tau)}{\lambda^2(\tau)s(\tau)} \log(\tilde{\sigma}(\tau)/\hat{\sigma}(\tau))$$

where $\lambda^2(\tau) = \tau(1 - \tau)$. Tests of this sort based on the drop in the optimized value of the objective function of an M-estimator when relaxing the restriction imposed by the hypothesis, are termed ρ-tests by **?**. Following this terminology, we will refer below to tests based on $L_n(\tau)$ and $\Lambda_n(\tau)$ as quantile ρ-tests, although the phrase quasi-likelihood ratio tests would also be appropriate.

In some applications it may be useful to formulate joint hypotheses about the relevance of certain groups of covariates with respect to *several* quantiles. For this we require the joint asymptotic distribution of vectors of quantile ρ-test statistics of the form, for example, $(L_n(\tau_1), L_n(\tau_2), ..., L_n(\tau_m))$. Such results may be subsumed in the following general theory for the ρ-test *processes*: $\{L(\tau) : \tau \in [\epsilon, 1-\epsilon]\}$, and $\{\Lambda(\tau) : \tau \in [\epsilon, 1-\epsilon]\}$.

We will adopt the following regularity conditions regarding the model (**??**) and the hypothesis (**??**).

A.1: The error distribution F has continuous Lebesgue density, f, with $f(u) > 0$ on $\{u : 0 < F(u) < 1\}$.

A.2: The sequence of design matrices $\{X_n\} = \{(x_i)_{i=1}^n\}$ satisfy:

- **(i):** $x_{i1} = 1, \quad i = 1, 2, \dots$
- **(ii):** $D_n = n^{-1}X_n'X_n \to D$, a positive definite matrix.
- **(iii):** $n^{-1}\sum \|x_i\|^4 = O(1)$
- **(iv):** $\max_{i=1,\dots,n}\|x_i\| = O(n^{\frac{1}{4}}/\log n)$

A.3: There exists a fixed, continuous function, $\gamma(\tau) : [0,1] \to \Re^q$ such that $R\beta - r = \gamma(\tau)/\sqrt{n}$ for samples of size n.

Remarks. Conditions A.1 and A.2 are by now rather standard in the quantile regression literature. Somewhat weaker conditions on both F and X_n are used in **?** in an effort to extend the theory into the tails, but this doesn't seem critical here so we have reverted to conditions close to those used in **?**. Condition A.3 is a direct analogue of Condition A.3 in **?** allowing us to explore the question of local asymptotic power of tests.

To investigate the asymptotic behavior of $L_n(\tau)$and $\Lambda_n(\tau)$ we require some rather basic theory and notation regarding Bessel processes. Let $W_q(t)$ denote a q-vector of independent Brownian motions and thus, for $t \in [0,1]$,

$$B_q(t) = W_q(t) - tW_q(1)$$

will represent a q-vector of independent Brownian Bridges. Note that for any fixed $t \in (0,1)$,

$$B_q(t) \sim \mathcal{N}(0, t(1-t)I_q). \tag{3.6.10}$$

The normalized Euclidean norm of $B_q(t)$,

$$Q_q(t) = ||B_q(t)||/\sqrt{t(1-t)}$$

is generally referred to as a Bessel process of order q. Critical values for $\sup Q_q^2(t)$ have been tabled by **?** and, more extensively, by **?** using simulation methods. The seminal work on Bessel processes and their connection to K-sample goodness of fit tests seems to be **?**. Again, for any fixed $t \in (0,1)$ we have, from (3.6.10), $Q_q^2(t) \sim \chi_q^2$.

Thus, we may interpret $Q_q^2(t)$ as a natural extension of the familiar univariate χ^2 random variable with q degrees of freedom. Note that in the special case $q = 1$, $\sup Q_1^2(\cdot)$ behaves asymptotically like a squared Kolmogorov-Smirnov statistic.

To characterize the behavior of the test statistic under local alternatives it is helpful to define a non-central version of the squared Bessel process as an extension of the noncentral χ^2 distribution. Let $\mu(t)$ be a fixed, bounded function from $[0,1]$ to $\Re^q$. The standardized squared norm

$$||\mu(t) + B_q(t)||^2/(t(1-t))$$

will be referred to as a non-central Bessel process of order q with noncentrality function $\eta(\tau) = \mu(t)'\mu(t)/(t(1-t))$ and will be denoted by $Q^2_{q,\eta(t)}$. Of course, for any fixed $t \in (0,1)$, $Q^2_{q,\eta(t)} \sim \chi^2_{q,\eta(t)}$, a non-central χ^2_q random variable with q degrees of freedom and non-centrality parameter $\eta(t)$. Finally, the symbol $\Rightarrow$ will denote weak convergence, $\rightsquigarrow$ convergence in distribution, and $\rightarrow$, convergence in probability. The following result appears in **?**.

Theorem 3.4. *Let $\mathcal{T} = [\epsilon, 1-\epsilon]$, for some $\epsilon \in (0, \frac{1}{2})$. Under conditions A.1-3,*

$$L_n(\tau) \Rightarrow Q^2_{q,\eta(\tau)}(\tau) \quad \textit{for } \tau \in \mathcal{T}$$

where $\eta(\tau) = \gamma(\tau)'(RD^{-1}R')^{-1}\gamma(\tau)/\omega^2(\tau)$, and $w(t) = \lambda(\tau)s(\tau)$. And, under the null hypothesis (**??**)

$$\sup_{\tau\in\mathcal{T}} L_n(\tau) \rightsquigarrow \sup_{\tau\in\mathcal{T}} Q_q^2(\tau)$$

The alternative form of the quantile ρ-process based on the location-scale form of the ρ-test has the same asymptotic behavior.

Corollary 3.1. *Under conditions A1-3, $\Lambda_n(\tau) = L_n(\tau) + o_p(1)$, uniformly on $\mathcal{T}$.*

The foregoing results enable the investigator to test a broad range of hypotheses regarding the joint effects of covariates while permitting one to restrict attention to specified ranges of the family of conditional quantile functions. Thus, for example, we may focus attention on only one tail of the conditional density, or on just a neighborhood of the median, without any prejudgment that effects should be constant over the whole conditional density as in the conventional location shift version of the regression model.

7. The Regression Rankscore Process Revisited

A disadvantage of the quantile ρ-tests proposed above is the required estimation of the nuisance sparsity function, $s(\tau)$. Clearly, analogous Wald type tests would share this disadvantage and consequently we will not pursue them here. However, an interesting connection can be drawn at this point with the theory of rank based tests for linear models proposed in ?, which we will refer to hereafter as GJKP.

These tests, which may be regarded as considerably refined forms of the simple score tests for median regression considered in ?, are based on the regression rank score process, introduced by ?. To simplify the exposition we will, following GJKP, consider the special "exclusion form" of the hypothesis (??), i.e.

$$\begin{bmatrix} 0 \vdots I_q \end{bmatrix} \begin{bmatrix} \beta_1(\tau) \\ \beta_{2n}(\tau) \end{bmatrix} = 0. \tag{3.7.11}$$

with the understanding that, as noted in the proof of Theorem 2.1, the model may always be reparameterized so that the hypothesis can be expressed in this form. The regression rank score process for the restricted version of the model under H_0 is,

$$\hat{a}_n(\tau) = \operatorname{argmax}\{y'a | X_1'a = (1-\tau)X_1'e,\ a \in [0,1]^n\} \tag{3.7.12}$$

where e denotes an n-vector of ones, and X has been partitioned as $[X_1 \vdots X_2]$ to conform with the partitioning of the hypothesis. We adopt the following standard notation for partitioning the matrix D defined in Condition A.2.ii above: $D_{ij}\ \ i,j = 1,2$ will denote the ij^{th} block of D, and D^{ij} will denote the ij block of D^{-1}. To illustrate, recall $D^{22} = (D_{22} - D_{21}D_{11}^{-1}D_{12})^{-1}$. The problem posed in (3.7.12) is the formal dual problem corresponding to the (primal) quantile regression linear program. Theorem 5.1 of GJKP considers tests of the exclusion form of the null hypothesis (3.7.11) based on the test statistic,

$$T_n = S_n' D_n^{22} S_n / A^2(\varphi)$$

where

$$\begin{aligned} S_n &= n^{-1/2}(X_2 - \hat{X}_2)'\hat{b}_n, \\ \hat{X}_2 &= X_1(X_1'X_1)^{-1}X_1'X_2, \\ D_n^{22} &= ((X_2 - \hat{X}_2)'(X_2 - \hat{X}_2)/n)^{-1} \\ \hat{b}_n &= (-\int \varphi(t)\hat{a}_{in}(t)dt)_{i=1}^n, \end{aligned}$$

φ is a score generating function of bounded variation and $A^2(\varphi) = \int_0^1 (\varphi(t) - \bar{\varphi})^2 dt$ with $\bar{\varphi} = \int_0^1 \varphi(t)dt$. Here we restrict attention to the special quantile score function,

$$\varphi_\tau(t) = \tau - I(t < \tau).$$

so $\bar{\varphi} = 0$, and $A^2(\varphi) = \lambda^2(\tau) = \tau(1-\tau)$. The familiar special case is sign scores with $\tau = 1/2$. The quantile score function which yields rankscores,

$$\hat{b}_n = \hat{a}_n(\tau) - (1-\tau),$$

allows us to define the rank test *process*

$$T_n(\tau) = S_n(\tau)' D_n^{22} S_n(\tau)/\lambda^2(\tau).$$

For fixed $\tau \in \mathcal{T}$, Theorem 5.1 of GJKP implies that under the conditions A1-3 above, $T_n(\tau) \rightsquigarrow \chi^2_{q,\eta(\tau)}$, where $\eta(\tau) = \gamma(\tau)'(D^{22})^{-1}\gamma(\tau)/\omega^2(\tau)$.Note that $(s(\tau))^{-1} = -\int_0^1 \varphi_\tau(t) df(F^{-1}(t))$ in (5.5) of GJKP in this case. This result may be extended to the whole process for $\tau \in \mathcal{T}$ providing a rank test link to the ρ-test processes already introduced. We will refer to the process $T_n(\tau)$ form as the quantile score process in the sequel. We summarize the foregoing discussion in the following result.

THEOREM 3.5. *Under the conditions A1-3, $T_n(\tau) = L_n(\tau) + o_p(1)$, uniformly on $\mathcal{T}$.*

A crucial feature of this form of the test which distinguishes it from the corresponding ρ-test processes is that, since the rank score process $a_n(\tau)$ is scale invariant, T_n does not require any estimate of the nuisance parameters $s(\tau)$ or $\sigma(\tau)$. This is a very substantial advantage over both the ρ-test and Wald approaches to testing in quantile regression, as was already stressed in **?**. Indeed, this observation leads to an important extension of the theory to cover a much broader class of nulls than those possible under the iid error assumption of model (1.1). In effect, the iid error assumption requires that in the null model the effect of the covariates is a pure location shift. However, the the theory for T_n can be extended to models in which under the null the conditional quantile functions are linear, but are no longer required to be parallel. This includes, but is by no means limited to, models with linear conditional scale effects.

8. Resampling Methods

There has been considerable recent interest in resampling methods for estimating confidence intervals for quantile-type estimators. However, despite the fact that confidence intervals for quantiles was one of the earliest success stories for the bootstrap (in contrast to the delete-1 jackknife which fails in this case) recent results have been considerably more guarded in their enthusiasm. Hall and Martin (1989) conclude:

> It emerges from these results that the standard bootstrap techniques perform poorly in constructing confidence intervals for quantiles... The percentile method does no more than reproduce a much older method with poor coverage accuracy at a fixed level: bias corrections fail for the same reason; bootstrap iteration fails to improve the order of coverage accuracy; and percentile-t is hardly an efficacious alternative because of non-availability of suitable variance estimates.

Nevertheless, there has been considerable recent interest, particularly among econometricians, in using the bootstrap to compute standard errors in quantile regression applications. See Buchinsky (1994), Hahn (1993), Fitzenberger (1996), and Horowitz (1996) for recent examples.

There are several possible implementations of the bootstrap for quantile regression applications. As in other regression applications we have a choice between the residual bootstrap and the xy-pairs bootstrap. The former resamples with replacement from the residual vector and adds this to the fitted vector $X\hat{\beta}_n(\tau)$ and reestimates, in so doing it assumes that the error process is iid. The latter resamples xy pairs, and therefore is able to accommodate some forms of heteroskedasticity. As in the sparsity estimation approaches we may consider replacing the residual EQF by the EQF obtained directly from the the regression quantile process, but this maintains the iid error assumption. More interesting is the possibility of resampling directly from the full regression quantile process which we will call the Heqf bootstrap. By this we mean for each bootstrap realization of n observations we draw n p-vectors from the estimated process $\hat{\beta}_n(t)$. There are, say, J distinct such realizations

$$\hat{\beta}_n(t) = \hat{\beta}_n(t_j) \quad \text{for} \quad t_j \leq t < t_{j+1}$$

$j = 1, ..., J$ and each is drawn with probability $\pi_j = t_{j+1} - t_j$. For each design row x_i we associate the bootstrapped y observation which is the inner product of that design row and the corresponding ith draw from the regression quantile process. This procedure has the virtue that it is again capable of accommodating certain forms of heteroskedastic regression models, in particular those with linear conditional quantile functions.

Finally, we will describe a new resampling method due to Parzen, Wei and Ying (1993) which is quite distinct from the bootstrap. It arises from the observation that the function

$$S(b) = n^{-1/2} \sum_{i=1}^{n} x_i(\tau - I(y_i \leq x_i'b)) \tag{3.8.1}$$

which is the estimating equation for the τth regression quantile is a pivotal quantity for the true τth quantile regression parameter $b = \beta_\tau$. That is, its distribution may be generated exactly, *irrespective of the true F generating the observations*, by generating a random vector U which is a weighted sum of independent, recentered Bernoulli variables which play the role of the indicator function. They show further that for large n the distribution of $\hat{\beta}_n(\tau) - \beta_\tau$ can be approximated by the conditional distribution of $\hat{\beta}_U - \hat{\beta}_n(\tau)$, where $\hat{\beta}_U$ solves an augmented quantile regression problem with $n+1$ observations and $x_{n+1} = -n^{1/2}u/\tau$ and y_{n+1} is sufficiently large that $I(y_{n+1} \leq z_{n+1}'b)$ is always zero. This is essentially the same as solving $S(b) = -u$ for given realization of u. This approach, by exploiting the asymptotically pivotal role of the quantile regression "gradient condition", also achieves a robustness to certain heteroskedastic quantile regression models. In practice, one might be able to exploit the fact that the solution to the augmented problems is close to the original one, since they differ by only one observation, but we have not tried to do this in our simulation experiments which are reported in the next section.

9. Monte-Carlo Comparison of Methods

In this final subsection we report on a small Monte-Carlo experiment designed to compare the performance of the methods described above. We focus primarily on the computationally less demanding sparsity estimation and inverted rankscore methods, but some results are reported for three of the resampling methods. Preliminary results indicated that the Hall and Sheather bandwidths performed considerably better than the Bofinger choice so we have restricted our reported results mainly to this form of sparsity estimation. We also focus exclusively on the problem of confidence intervals for the median regression parameters, partly because this is the most common practical problem, and also because it restricts the amount of computation and reporting required. In subsequent work, it is hoped to provide a much more exhaustive Monte-Carlo experiment.

We considered first an iid error model in which both x's and y's were generated from the Student t distribution. The degrees of freedom parameter varies over the set $\{1, 3, 8\}$ for both x's and y's. The first column of the design matrix is ones, all other entries are iid draws from the specified t distribution. For each cell of the experiment the design matrix is drawn once, and 1000 replications of the response vector, y, are associated with this fixed design matrix. Throughout, we have studied only the sample size $n = 50$.

All of the computations were carried out in the 'S' language of Becker, Chambers, and Wilks(1988) on a Sun workstation.

In Table 1 we report observed Monte-Carlo coverage frequencies for nine situations and three non-resampling methods. Confidence intervals are computed for all three slope coefficients for each situation so in each cell we report the number of times the interval covers the true parameter (zero in all cases) in 3000 trials. Throughout the experiment the nominal size is .10. In these iid error situations we see that the size of the rank inversion method is quite accurate throughout, as is the Hall-Sheather sparsity estimate. However the Bofinger results are considerably less satisfactory. Generally, the rank-inversion intervals are shorter than the sparsity intervals except for the anomalous cases of Cauchy design.

To compare the performance of the resampling methods we report in Table 2 results for 3 iid error situations and 5 methods. Since the resampling methods are quite slow, 500 resamples are done for each of them, we restrict attention to only

TABLE 3.1. Confidence Interval Performance – IID Errors

	coverage			length		
	dfy= 1	dfy= 3	dfy= 8	dfy= 1	dfy= 3	dfy= 8
dfx= 1						
rank-inverse	0.892	0.923	0.922	0.320	0.392	0.359
sparsity-HS	0.893	0.907	0.909	0.240	0.142	0.079
sparsity-BS	0.932	0.927	0.931	0.288	0.153	0.083
dfx= 3						
rank-inverse	0.875	0.904	0.890	0.625	0.504	0.501
sparsity-HS	0.923	0.911	0.923	0.614	0.505	0.544
sparsity-BS	0.954	0.932	0.937	0.736	0.544	0.577
dfx= 8						
rank-inverse	0.887	0.885	0.884	0.791	0.617	0.585
sparsity-HS	0.941	0.920	0.919	0.921	0.683	0.640
sparsity-BS	0.968	0.948	0.935	1.107	0.737	0.680

the diagonal cases of the previous table with the degrees of freedom parameter for x's and y's equal. We are primarily interested in resampling as a means of achieving consistent confidence intervals in heteroskedastic situations so we restrict attention to the Parzen-Wei-Ying (PWY) approach, the heteroskedastic empirical quantile function bootstrap (Heqf), and the xy-pairs bootstrap. It can be seen from the table that again the rank-inversion method is quite reliable in terms of size, and also performs well with respect to length. The PWY resampling method has empirical size less than half the nominal 10 percent, while the xy-bootstrap is also undersized. The Heqf-bootstrap is accurately sized except for the Cauchy situation. It is obviously difficult to compare the lengths achieved by various methods, given the discrepancies in size, however the rank inversion approach seems to perform reasonably well in this respect.

TABLE 3.2. Confidence Interval Performance – IID Errors

	coverage			length		
	df= 1	df= 3	df= 8	df= 1	df= 3	df= 8
rank-inverse	0.900	0.893	0.879	0.335	0.427	0.558
sparsity-HS	0.872	0.922	0.915	0.217	0.455	0.613
PWY	0.961	0.957	0.957	0.411	0.520	0.680
Heqf-BS	0.802	0.881	0.895	0.220	0.380	0.512
XY-BS	0.929	0.948	0.945	0.331	0.486	0.640

A more challenging problem for estimation of confidence intervals for quantile regression problems involves heteroskedastic situations. We consider a simple case which bears a close resemblance to the previous iid error situations. Again, we generate 3 columns of the design matrix X as iid draws, this time from the lognormal distribution. The response vectors are then drawn from a Student t distribution with location 0 and scale given by $\sigma_i = \sum_{i=1}^{4} x_i/5$. For all i, $x_{1i} = 1$. Again the design is fixed for a given configuration, and hence scale is fixed. In this model all the conditional quantile functions are linear, so the Heqf-bootstrap is applicable, however the simple sparsity estimation approach is obviously not consistent under these conditions.

TABLE 3.3. Confidence Interval Performance – Heteroskastic Errors

	coverage			**length**		
	df= 1	df= 3	df= 8	df= 1	df= 3	df= 8
rank-inverse	0.887	0.902	0.878	1.196	0.793	0.621
sparsity-HS	0.763	0.717	0.656	0.702	0.552	0.357
PWY	0.953	0.950	0.946	1.557	0.907	0.715
Heqf-BS	0.754	0.813	0.804	0.971	0.655	0.486
XY-BS	0.907	0.911	0.897	1.332	0.799	0.612

Again the rank-inversion approach seems to perform well. As expected the sparsity approach fails miserably. The Parzen-Wei-Ying resampler is again substantially undersized – a rather puzzling result. The xy bootstrap also performs very well, but the Heqf version of the bootstrap has very poor coverage frequencies suggesting that this approach is probably not reliable. Since the rank-inversion method is on the order of 10 times faster than any of the bootstrap methods even for moderate sized problems it appears to have a substantial advantage.

10. Problems

1. *Suppose* $X_1, ...X_n$ *are iid from df* F *with* $f(x) = F'(x) > 0$ *on the real line. Generalize* (3.2.1) *for a single quantile to obtain the result* (3.2.2) *on the joint asymptotic distribution of several sample quantiles.*

2. *Let* $X_1, \ldots, X_n$ *be iid from* F, *and denote the order statistics* $X_{(1)}, \ldots, X_{(n)}$. *Show that the probability that random interval* $I_n(r) = [X_{(r)}, X_{(s)}]$ *for* $1 \le r \le s = n - r + 1$ *covers the* τ*th quantile* $\xi_\tau = F^{-1}(\tau)$ *is,*

$$\begin{aligned} P(\xi_\tau \in I_n(r)) &\equiv P(X_{(r)} \le \xi_\tau \le X_{(s)}) \\ &= P(X_{(r)} \le \xi_\tau) - P(X_{(s)} < \xi_\tau) \\ &\ge C_n(r) \end{aligned}$$

where $C_n(r) = \sum_{i=r}^{n-r} \binom{n}{i} \tau^i (1-\tau)^{n-1}$. *Equality holds in the last line iff* F *is continuous at* ξ_τ.

3. *An interesting extension of Problem 2 is provided by Guilbaud (1979) who shows that for intervals of the form*

$$I_n(r,t) = \left[\frac{X_{(r)} + X_{(r+t)}}{2}, \frac{X_{(s)} + X_{(s+t)}}{2}\right]$$

where $1 \leq r \leq s = n - r + 1$, *and* $0 \leq t < s - r$. *we have* for the median,

$$P(\xi_{1/2} \in I_n(r,t)) \geq \frac{1}{2}C_n(r) + \frac{1}{2}C_n(r+t)$$

for general F, *and for continuous, strictly increasing* F,

$$P(\xi_{1/2} \in I_n(r,t)) \leq C_{n-t}(r)$$

with equality in the latter expression iff F *is symmetric.*

4. *Show that the interval in Problem 1 may be interpreted as an inversion of the following sign test of the hypothesis* $H_0 : \xi_\tau = \xi$: *let* Q_n *be the number of observations* $\{X_1, \dots, X_n\}$ *less than* ξ, *since* Q_n *is Bin* (n, τ), *we can choose* $\gamma \in (0,1)$ *and* c_α *such that under* H_0,

$$P(Q_n < c_\alpha)) + \gamma P(Q_n = c_\alpha) = \alpha$$

and reject H_0 *when* $Q_n < c_\alpha$ *or with probability* γ *when* $Q_n = c_\alpha$, *for a one-sided alternative. For a two-sided alternative we choose a corresponding upper critical value as well. Note that for* n *moderate to large we can approximate the binominial probabilities with their normal approximation. See e.g., Lehmann (1959, problem 3.28).*

5. *To explore the role of the condition* $f(0) > 0$, *reconsider Problem 1 assuming instead that* $F(0) = 1/2$, $\quad f(0) = 0$, *and* $f'(0\pm) = \pm 1$. *How does this alter a.) the rate of convergence of* $\hat{\mu}$, *and b.) the form of the limiting distribution?*

6. *To explore a simple version of the asymptotics for quantile regression consider the scalar parameter (regression through the origin model)*

$$y_i = x_i \beta + u_i$$

where the u_i*'s are iid and satisfy the* F *conditions of Problem 1. Formulate conditions on the design sequence* $\{x_i\}$ *which ensure that*

$$\hat{\beta}_n = argmin_{b \in \mathbb{R}} \sum_{i=1}^{n} |y_i - x_i b|$$

satisfies

$$\sqrt{n}(\hat{\beta}_n - \beta) \rightsquigarrow N(0, \omega^2 Q^{-1})$$

and $Q = \lim n^{-1} \sum x_i^2$.

CHAPTER 4

Asymptotic Theory of Quantile Regression

While the finite sample distributio theory of regression quantiles can be represented explicitly as we have illustrated in Theorem 3.1 above, the practical application of this theory would entail a host of hazardous assumptions and an exhausing computational effort. It is generally conceded throughout statistics that approximation methods involving local linearization and the central limit theorem play an indispensible role in the analysis of the performance of statistical procedures and in rendering such procedures practical tools of statistical inference. The zealous pursuit of these objectives is inevitably met with accusations that we live in a cloud-cuckoo land of "asymptopia", but life is full of necessary compromises and approximations. And it is fair to say that those who try to live in the world of "exact results" in finite sample sample statistical distribution theory are exiled to an even more exotic territory.

Fortunately, there are many tools available to help us evaluate the adequacy of our asymptotic approximations. As we have already seen, Monte-Carlo simulation can be an extremely valuable tool. Higher ordre expansions, although particularly challenging in the present context, may offer useful assessments of the accuracy of simpler aproximations and possible refinement strategies. And the rapid development of resampling methods for statistical inference offer many new options for inference.

The fundamental task of asymptotic theory is to impose some discipline and rigor on the process of developing statistical procedures. The natural enthusiasm that arises from the first few "successful" applications of a new technique can be effectively tempered by some precisely cast questions of the form: suppose data arose according the conditions A, does the procedure produce a result that converges in some appropriate sense to object B? Under what precise conditions does the procedure "work"? And, if possible, how well does it "work" relative to other competing procedures.

As important virtue of of quantile regression, one that we have stressed thoughout, is the natural interpretability of the conditional quantile functions, as an objective for data analysis. Unlike the many obscure objects of desire introduced *en masse* by the robustness literature, for example, the conditional quantile functions offer an easily interpretable target for statistical analysis. In this Chapter we will survey the existing literature on the asymptotic theory of quantile regression and describe how these results are used.

1. Consistency

2. Bahadur Representation

3. Weak Convergence

4. Applications to Inference

CHAPTER 5

L-Statistics and Weighted Quantile Regression

1. L-Statistics for the Linear Model

Linear combinations of functions of a few sample quantiles often provide a surprisingly efficient, yet extremely simple, means of estimating salient characteristics of a random variable. Weighted averages of the quartiles as an estimator of location, interquantile ranges for scale, Hill's (1975) estimator of the tail exponent of a distribution function are all of this form. In extreme cases such estimators can even be optimal: the median as a estimator of location for the Laplace, or double exponential, distribution, or the midrange for the location of the uniform distribution for example. But generally, L-statistics with smoother weight functions are preferred for reasons of both efficiency and robustness.

The leading example of the "smooth" L-estimator is undoubtedly the trimmed mean. Denoting the τth quantile by

$$\hat{\xi}(\tau) = \text{argmin}_{\xi} \sum \rho_\tau(y_i - \xi)$$

we may express the α trimmed mean as

$$\hat{\mu}_\alpha = \int_0^1 \varphi_\alpha(\tau)\hat{\xi}(\tau)d\tau$$

where $\varphi_\alpha(\tau) = I(\alpha \leq \tau \leq 1-\alpha)/(1-2\alpha)$, which is simply the sample average of the central $[(1-2\alpha)n]$ order statistics. This estimator has a long history of applications in astronomy and other fields, and has been singled out by several prominent authors as the quintessential robust estimator of location. See, e.g., Bickel and Lehmann (1975) and Stigler (1977).

L-statistics were initially regarded as "quick and dirty" substitutes when maximum likelihood estimation was either infeasible, because we were unsure about the parametric form of the model, or intractable due to computational difficulties. But it was quickly recognized that asymptotically fully efficient L-statistics could be constructed to compete with maximum likelihood estimator on a equal footing. An interesting example of the contrast between M- and L-estimators is the well-known "least-favorable" density of Huber (1964) which takes the form,

$$f_\varepsilon(x) = \begin{cases} c\ \exp(-x^2/2) & \text{if } |x| \leq k \\ c\ \exp(k^2/2 - k|x|) & \text{otherwise} \end{cases}$$

where $c = (1-\varepsilon)/\sqrt{2\pi}$ and k satisfies $2\phi(k)/k - 2\Phi(k) = \varepsilon(1-\varepsilon)$ with ϕ, and Φ denoting the standard normal density and distribution function, respectively. This density, which is Gaussian in the center with exponential tails, has minimal Fisher information in the class

$$\mathcal{F} = \{F = (1-\varepsilon)\Phi + \varepsilon H | H \in \mathcal{H}\}$$

where $\mathcal{H}$ denotes the set of distribution functions symmetric about zero, and ε is a fixed number between zero and one. Given ε, an M-estimator of location for this "minimax" family of densities may be constructed by solving,

$$\hat{\mu}_\varepsilon = \text{ argmin } \sum \rho_{H_k}(y_i - \xi)$$

where $\rho_{H_k}(x) = [x^2 I(|x| \leq k) + k^2 I(|x| > k)]$. Since scale is generally unknown, and $\hat{\mu}_\varepsilon$ as formulated above is not scale equivariant, in practice we need to replace it by something of the form

$$\hat{\mu}_\varepsilon = \text{ argmin }_\xi \sum \rho_{H_k}((y_i - \xi)/\hat{\sigma})$$

for some scale equivariant estimator $\hat{\sigma}$, or jointly minimize with respect to location and scale parameters. A dramatically simpler, yet asymptotically fully efficient, alternative to such M-estimators for this least favorable model, is the $\alpha = \varepsilon/2$ trimmed mean. The latter has the advantage of being automatically scale equivariant as well as avoiding the necessity of computing the Huber constant $k(\varepsilon)$. The form of the α-trimmed mean weight function $\varphi_\alpha(\cdot)$ reflects clearly the form of the Huber least favorable density with the central Gaussian, portion of the sample receiving constant weight $(1-2\alpha)^{-1}$, and the exponential tail portion receiving weight zero. Since the density was designed to minimize the information provided by the tails it seems hardly surprising that he tail observations are uninformative about location.

Tukey (1965) posed the crucial question, "which part of the sample contains the information?" For the Huber density the question is clearly answered at least in an asymptotic sense by "the central $[(1-2\alpha)n]$ order statistics." The tail observations contribute nothing asymptotically to the design of an efficient estimator of location for this density. Other models are even more extreme in this respect. We have already mentioned the Laplace density for which the sample median contains, again we should stress – asymptotically, all the available sample information about location. In general, the optimal L-estimator weight function provides a concise, asymptotic answer to Tukey's question. For location, this weight function takes the form

$$\varphi_0(t) = -(\log f)''(F^{-1}(t)).$$

For scale, it takes the form,

$$\varphi_1(t) = [(\log f)' + x(\log f)''](F^{-1}(t)).$$

In Figure 5.1 we illustrate these weight functions for a variety of familiar densities. In addition, for purpose of comparison, we include a sketch of the influence function of these estimators which can also be interpreted as the ψ-function which serves to determine the optimal M-estimator in each case.

We note first that the Gaussian density is unique in its treatment of each observation as equally informative about location. For the logistic, Student-t and Huber densities the location weights fall off sharply in the tails. The Student densities illustrate a curious phenomenon. When, as in the Student cases, the tail behavior of the density is heavier than exponential, then the optimal weight function can be negative in the tails. This has the apparently paradoxical effect that, given a sample, if we move observations, say in the right tail further to the right, we may actually move our estimate to the left. Accentuating the outliers in the Student-t model increases our willingness to discount them. This paradox is only a thinly disguised variant of the familiar tendency for an enthusiastic movie review by a dubious reviewer to send us in the opposite direction looking for another film.

Not surprisingly, virtually all the sample information *about location* in our asymmetric densities is contained in the first few order statistics. However, note that for the last Weibull example all of the order statistics receive positive weight. The L-estimators of scale exhibit rather different behavior than their corresponding location estimators. For symmetric densities the weight functions for the optimal scale. L-estimators are odd functions, thus assuring location invariance. In the Gaussian case we have

$$\varphi_1(t) = \Phi^{-1}(t)$$

so our estimator is

$$\hat{\sigma} = \int_0^1 \Phi^{-1}(t)\hat{\xi}(t)dt$$

$\hat{\xi}(t) \to \mu + \sigma\Phi^{-1}(t)$, so we have, changing variables,

$$\hat{\sigma} \to \sigma \int_0^1 (\Phi^{-1}(t))^2 dt = \sigma \int_{-\infty}^{\infty} x^2 d\Phi(t) = \sigma.$$

As in the case of location estimation, heavy tails in the model density induces a downweighting of tail observations for scale L-estimators too. The χ^2 densities yield a constant weighting for scale estimation, while for the Weibull model the corresponding weight functions are sharply increasing in the tails reflecting their "light" tails.

Most aspects of the theory of L-estimation in the univariate case can be carried forward to the linear model directly via quantile regression. For example, consider the linear location-scale model

$$y_i = x_i'\beta + (x_i'\gamma)u_i \tag{5.1.1}$$

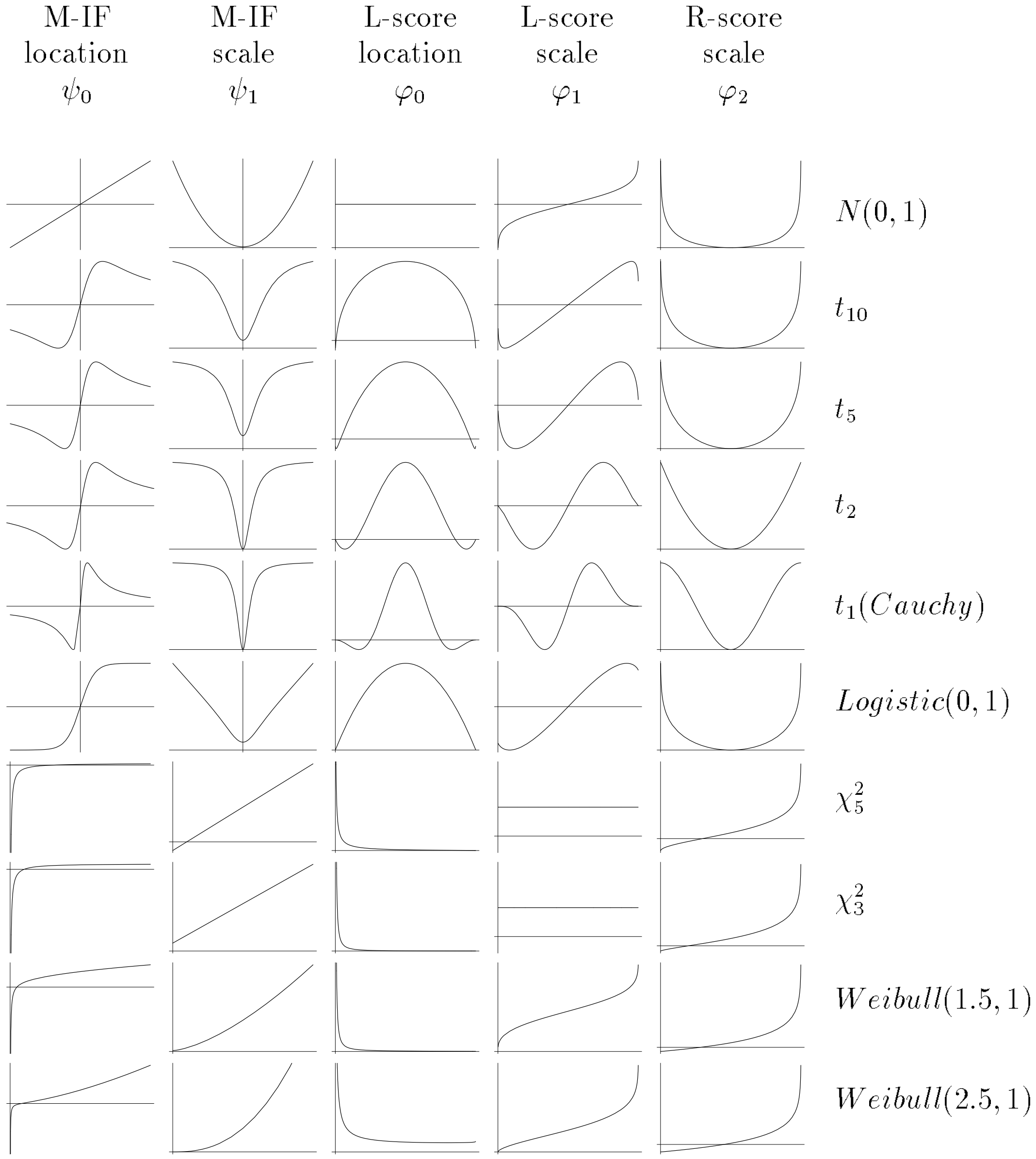

FIGURE 5.1. Comparative Anatomy of Score Functions. This figure illustrates the shapes of the optimal score functions for L-estimators of location and scale for some representative distributions. In addition, we provide the shapes of the optimal ψ-functions for optimal M-estimators for these distributions, which can also be interpreted as the influence functions for the corresponding L-estimators. Finally, we include the optimal score functions for the rank test of scale for each distribution. Table 1 of Hájek and Šidák (1967) provides analytic expressions for many of these functions.

with u_i iid from F. Asymptotically efficient estimators of β and γ are available as

$$\begin{aligned}\hat{\beta} &= \int_0^1 \varphi_o(t)\hat{\beta}(t)dt \\ \hat{\gamma} &= \int_0^1 \varphi_1(t)\hat{\beta}(t)dt\end{aligned}$$

where φ_0 and φ_1 are as defined above, and $\hat{\beta}(t)$ denotes the p-dimensional quantile regression process. Obviously, the integrals should be interpreted as computed coordinatewise in this case.

In some cases, such "efficient" L-statistics are a bit surprising in their departure from more familiar M-estimators. For example, in the Gaussian case the estimator

$$\hat{\beta} = \int_0^1 \hat{\beta}(t)dt$$

which simply averages each coordinate of the quantile regression process appears quite different from the familiar least-squares estimator. However, under Gaussian conditions, it performs quite well in Monte-Carlo comparison with least-squares, see Portnoy and Koenker (1989). In the case of the Gaussian scale estimator, we now have, for F Gaussian,

$$\hat{\beta}(t) \to \beta + \gamma\Phi^{-1}(t)$$

so

$$\hat{\gamma} = \int_0^1 \Phi^{-1}(t)(\hat{\beta})dt \to \gamma.$$

Indeed, for general F we obtain

$$\hat{\gamma} = \int_0^1 \Phi^{-1}(t)\hat{\beta}(t)dt \to \gamma \int_0^1 \Phi^{-1}(t)F^{-1}(t)dt$$

so we estimate γ "consistently up to scale," i.e., ratios of the coordinates of $\hat{\gamma}$ are consistent for corresponding ratios of the vector γ.

It is a significant advantage of the L-statistic approach to the estimation of linear models that in cases in which there is both heteroscedasticity and asymmetric innovations it is possible to distinguish the location and scale effects of the covariates, while with conventional least squares based methods this proves to be much more difficult. Consider, for example, the simple version of (5.1.1),

$$y_i = \alpha + \beta x_i + \gamma x_i^2 u_i,$$

with u_i iid from F, and F asymmetric so $Eu_1 = \mu \neq 0$. Least squares estimation of the quadratic model would estimate the parameters of the conditional mean function

$$E(y_i|x_i) = \alpha + \beta x_i + \mu\gamma x_i^2.$$

Now, if we proceed conventionally, we would regress absolute residuals on the available covariates, assuming,

$$E|y_i - \hat{y}_i| = c_0 + c_1 x_i + c_2 x_i^2$$

it is easy to see that least squares estimation of this model would yield, $\hat{c}_0 \to 0, \hat{c}_1 \to 0$ and $\hat{c}_2 \to \gamma E|u_i - \mu|$. However, if in the first step we estimated the simpler linear model and then tried to introduce the quadratic effect in only the second step, we introduce a bias resulting from the approximation of a quadratic function by something linear and this misspecification is then transmitted to the next step as well. In contrast, the quantile regression estimation of the quadratic model yields, via the L-statistic approach, a family of estimators of γ which are all "consistent up to scale." In addition, as we shall explore in more detail in Chapter 3, rank based tests for the heteroscedasticity parameter, γ, may be based on preliminary quantile regression estimation of the simpler linear model.

L-estimation also provides a convenient approach to adaptive estimation of the linear model. By *estimating* the optimal score functions φ_0 and φ_1 we can achieve full optimality for a broad class of location-scale models. This approach is developed for the pure location version of the linear model in Koenker and Portnoy (1990).

Finally, we should add that the integrals which appear in the foregoing L-statistics are usually quite simple to compute due to the piecewise constant form of the quantile regression process. We can illustrate this for the case of scale estimation with the Gaussian weight function, $\Phi^{-1}(t)$,

$$\begin{aligned} \hat{\gamma} &= \int_0^1 \Phi^{-1}(t)\hat{\beta}(t)dt \\ &= \sum_{j=1}^{J} \hat{\beta}(t_{j+1}) \int_{t_j}^{t_{j+1}} \Phi^{-1}(t)dt \\ &= \sum_{j=1}^{J} \beta(t_{j+1})[\phi(\Phi^{-1}(t_{j+1})) - \phi(\Phi^{-1}(t_j))]. \end{aligned}$$

Note that by the left-continuous convention for quantile functions, $\hat{\beta}(t) = \hat{\beta}(t_{j+1})$ for $t \in (t_j, t_{j+1}]$. A robustified version of this Gaussian scale estimator is easily adapted by restricting the domain of the integration to avoid the tails, and possibly the center of the unit interval. Welsh and Morrison (1990) discuss such estimators in considerable detail.

2. Kernel Smoothing for Quantile Regression

A number of recent papers have suggested that kernel smoothing may be used to improve the performance of univariate quantile estimators. The idea is quite simple and may be extended in a straightforward manner to regression using the L-estimation approach of the previous section.

Let $k(x)$ be a kernel function of some conventional form, for purposes of illustration we may take the Epanechnikov kernel,

$$k(t) = \frac{3}{4}(1 - \frac{1}{5}t^2)I(|t| < \sqrt{5})/\sqrt{5}.$$

We wish to consider replacing $\hat{\beta}_n(\tau)$ by the smoothed quantile regression process

$$\tilde{\beta}_n(\tau) = \int_0^1 \hat{\beta}_n(\tau)k((\tau - t)/h_n(\tau))/h_n(\tau)dt \tag{5.2.1}$$

The degree of smoothing is determined by the bandwidth function $h_n(\tau)$. Since the function $\hat{\beta}_n(\tau)$ is piecewise constant we can simplify the computation by an integration by parts,

$$\begin{aligned} \tilde{\beta}_n(\tau) &= \int_0^1 K((\tau - t)/h_n(\tau))d\hat{\beta}(t) - \hat{\beta}(t)K((\tau - t)/h_n) \mid_0^1 \\ &= \sum_{i=1}^{J} K((\tau - t_i)/h_n(\tau))\Delta\hat{\beta}(t_i) + \hat{\beta}(0) \end{aligned}$$

where $K(t) = \int_{-\infty}^t k(s)ds$, and the summation is over the jump points of the $\hat{\beta}_n(\tau)$ process. Since most of the popular kernel functions take the form of familiar densities, direct methods to evaluate the corresponding distribution functions are easily available, and the computation can be carried out quite efficiently.

In Table 5.1 we report the results of a small Monte Carlo experiment in which we have generated data from several iid linear models and compared the performance of several smoothed quantile regression estimators. Data for the linear model

$$y_i = \beta_1 + \beta_2\Phi^{-1}(i/(n+1)) + \beta_3|\Phi^{-1}(i/(n+1))| + u_i \qquad i = 1, \ldots, n, \tag{5.2.2}$$

was generated with and y_i iid F, with F chosen as Gaussian. This Gaussian iid error setting is particularly favorable for kernel smoothing of the slope parameters and we see considerable improvement due to smoothing for these parameters. For the intercept parameter there is an obvious bias effect due to smoothing, but still the smoothing has generally favorable consequences for experimental mean square errors. This bias would extend to the estimation of slope parameters as well as the intercept parameter in heteroscedastic models. We have observed earlier that the estimated quantile regression surfaces may intersect near the boundary of the observed design space. Some smoothing of the estimates offers a simple approach to ameliorating this

quantile	bandwidth	β_1	β_2	β_3
0.75	0.200	0.02674	0.00494	0.01498
0.75	0.100	0.02454	0.00743	0.02253
0.75	0.050	0.02515	0.00833	0.02477
0.75	0.020	0.02705	0.00908	0.02667
0.75	0.000	0.02910	0.00964	0.02879
0.90	0.080	0.04925	0.00813	0.02651
0.90	0.050	0.04057	0.01159	0.03742
0.90	0.020	0.04162	0.01335	0.04136
0.90	0.010	0.04336	0.01419	0.04311
0.90	0.000	0.04586	0.01515	0.04596
0.95	0.040	0.07829	0.01257	0.04114
0.95	0.020	0.06149	0.01826	0.05900
0.95	0.010	0.06286	0.01945	0.06149
0.95	0.005	0.06549	0.02017	0.06429
0.95	0.000	0.06975	0.02185	0.06950

TABLE 5.1. Monte-Carlo Mean Squared Errors for Kernel Smoothed Quantile Regression Estimators. The table reports Monte-Carlo mean squared errors for several kernel smoothed quantile regression estimators. The data for the experiment was generated from the model (5.2.2) with $\{u_i\}$ iid standard normal, $n = 200$ and 1000 replications per cell of the experiment. In each cell we compute 5 different degrees of smoothing, represented by the bandwidth parameter, h_n. It can be seen that more smoothing is advantageous for the slope parameters, since the kernel is averaging estimates of a common quantity. However, for the intercept, kernel smoothing introduces a bias, which is reflected in the poor performance of the largest bandwidth in each panel of the table.

effect. This lunch, of course, isn't free. Even in the iid error model there is inevitably some bias introduced in the intercept estimation due to the smoothing, and in non-iid models there may be bias in slope parameter estimation as well. Replacing the "locally constant" model of smoothing implicit in the usual kernel approach by a locally polynominal model as described for example in Hastie and Loader (1994) may offer further improvements.

3. Weighted Quantile Regression

The location-scale model of regression takes the form

$$Y_i = \mu(x_i) + \sigma(x_i)u_i \tag{5.3.1}$$

with $\{u_i\}$ independent and identically distributed (iid) as F. If the location and scale functions could be parameterized by $\theta \in \Theta \subset \mathbb{R}^p$ then the conventional maximum likelihood, M-estimation approach would suggest solving,

$$\min_{\theta\in\Theta} \sum_{i=1}^{n} [\rho((Y_i - \mu(x_i,\theta))/\sigma(x_i,\theta)) + \log \sigma(x_i,\theta)], \tag{5.3.2}$$

for some appropriate choice of ρ. Ideally, we would like to choose $\rho = -\log f$ when f is the density of the u_i's, but robustness considerations might suggest other choices if f were unknown. Ruppert and Carroll (1988) offer an excellent treatment of the state of the art based on M-estimation in this parametric setting.

If we are only willing to assume some smoothness in $\mu(x), \sigma(x)$, not an explicit parametric form, the situation is less clear. Various kernel and nearest neighbor approaches have been suggested, but penalized likelihood seems very attractive. We might begin by ignoring the scale heterogeneity and minimize

$$\sum \rho(Y_i - \mu(x_i)) + \lambda_\mu \int (\mu''(x))^2 dx$$

over μ in the say, the Sobolev space of functions with absolutely continuous first derivative and square integrable second derivative. Having estimated $\mu(x)$ in this manner, we could proceed to estimate the scale function by minimizing,

$$\sum \rho(\hat{u}_i/\sigma(x_i)) + \log \sigma(x_i) + \lambda_\sigma \int (\sigma''(x))^2 dx$$

where, $\hat{u}_i = Y_i - \hat{\mu}(x_i)$ and again ideally, $\rho = -\log f$. Iteration of this scheme would yield a penalized maximum likelihood estimator for the location-scale regression model.

An alternative approach based on L-estimation seems somewhat more flexible. Since the conditional quantile functions of Y_i in the location-scale regression model are simply

$$Q_Y(\tau|x) = \mu(x) + \sigma(x) Q_u(\tau)$$

we may estimate these functions by minimizing

$$\sum \rho_\tau(Y_i - \xi(x_i)) + \lambda_\tau \int (\xi''(x))^2 dx$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$, as usual, and λ_τ denotes a penalty parameter that governs the smoothness of the resulting estimate. However, now, as we will emphasize in chapter X, the $\mathcal{L}_2$ roughness penalty may be more conveniently chosen to be the L_1 or L_∞ norm of ξ''. Denoting the minimizer as $\hat{Q}_Y(\tau|x)$, standard L-estimation ideas may be employed to average over τ thereby obtaining estimates of μ and σ. Again, optimality at a known f would suggest

$$\hat{\mu}(x) = \int_0^1 \varphi_0(t)\hat{Q}_Y(t|x)dt$$

$$\hat{\sigma}(x) = \int_0^1 \varphi_1(t)\hat{Q}_Y(t|x)dt$$

where $\varphi_0(t) \equiv (\log f)''(Q(t)) = (f'/f)'(Q(t))$, and $\varphi_1(t) \equiv (xf'/f)'(Q(t))$

A particularly simple example of the foregoing approach is offered by the case where the functions $\mu(x)$ and $\sigma(x)$ are assumed to be *linear in parameters*, so we may rewrite (5.3.1) as

$$Y_i = x_i'\beta + (x_i'\gamma)u_i. \tag{5.3.3}$$

We should emphasize that linearity in the covariates is not essential, so this formulation includes various "series-expansion" models in which the x_i's may be interpreted as basis functions evaluated at the observed values of the original covariates.

In model (5.3.3) the conditional quantile functions are simply

$$Q_Y(\tau|x) = x'(\beta + \gamma Q_u(\tau))$$

and the linear quantile regression estimator

$$\hat{\beta}(\tau) = \operatorname{argmin}_{b\in\mathbb{R}^p} \sum \rho_\tau(Y_i - x'b)$$

converges under rather mild conditions to $\beta(\tau) = \beta + \gamma F^{-1}(\tau)$. Nevertheless, there is something inherently unsatisfactory about $\beta(\tau)$ in this context. This is reflected clearly in the asymptotic covariance of $\hat{\beta}(\tau)$.

As in the more familiar context of least squares estimation, the presence of heteroscedasticity resulting from the dependence of $x'\gamma$ on x in (5.3.3) introduces no asymptotic bias in $\hat{\beta}(\tau)$, but it does introduce a source of asymptotic inefficiency. We shall see that the asymptotic covariance matrix of $\sqrt{n}(\hat{\beta}(\tau) - \beta(\tau))$ is $\omega(\tau, F)D_{(1)}^{-1}D_{(0)}D_{(1)}^{-1}$, an expression analogous to the famous Eicker-White sandwich formula, where $D_{(r)} = \lim X'\Gamma^{-r}X$ and Γ is the diagonal matrix with typical element $x'\gamma$ and $\omega(\tau, F) = \tau(1-\tau)/f^2(F^{-1}(\tau))$. Reweighting the quantile regression minimization problem we may define a weighted quantile regression estimator as,

$$\hat{\beta}(\tau, \gamma) = \operatorname{argmin}_{b\in\mathbb{R}^p} \sum_{i=1}^n \rho_\tau(Y_i - x_i'b)/(x_i'\gamma)$$

and show that $\sqrt{n}(\hat{\beta}(\tau,\gamma) - \beta(\tau))$ is asymptotically Gaussian with mean 0 and covariance matrix $\tau(1-\tau)D_{(2)}^{-1}$. It is straightforward to show that $D_{(1)}^{-1}D_{(0)}D_{(1)}^{-1} - D_{(2)}^{-1}$ is nonnegative definite. Again, as in the least-squares case we may estimate γ, and $\hat{\beta}(\tau, \hat{\gamma})$ will have the same asymptotic behavior as $\hat{\beta}(\tau, \gamma)$ for any $\sqrt{n}$ consistent estimator

$\hat{\gamma}$. Simple $\sqrt{n}$ consistent estimators of γ may be easily constructed as interquantile ranges, i.e.

$$\hat{\gamma}_n = \hat{\beta}_n(\tau_1) - \hat{\beta}_n(\tau_0).$$

It is evident that such estimators need only be consistent "up to scale" that is we require only that

$$\hat{\gamma}_n = \kappa\gamma + O_p(n^{-1/2})$$

for some scalar κ, since κ plays no role in the minimization problem defining $\hat{\beta}(\tau, \gamma)$. For the interquantile range estimator we would have, for example,

$$\kappa^{-1} = Q_u(\tau_1) - Q_u(\tau_0).$$

Improved estimators of γ may be constructed as L-estimators with smooth weight functions along the lines described in Section 1.X. In practice there may be cases in which $x_i'\hat{\gamma}_n$ is negative for some indices i, and it may be reasonable to take absolute values in these cases. Since we must assume that $\sigma(x) = x'\gamma$ is strictly positive over the entire design space, this must occur with probability tending to zero.

Since $\rho_\tau(\cdot)$ is piecewise linear, and $\hat{\sigma}(x) = x'\hat{\gamma}_n > 0$, at least eventually, we may rewrite the weighted quantile regression model as

$$\tilde{Y}_i = \tilde{x}_i'\beta + u_i$$

where $\tilde{Y}_i = Y_i/\hat{\sigma}(x_i)$, and $\tilde{x}_i = x_i/\hat{\sigma}(x_i)$. In these transformed variables we have

$$Q_{\tilde{Y}_i}(\tau|x_i) = \hat{\sigma}_i^{-1}(x_i)x_i(\beta + \gamma Q^{-1}(\tau)) = \tilde{x}_i'\beta + \hat{\sigma}^{-1}(x_i)x_i'\gamma Q^{-1}(\tau),$$

and since $\hat{\sigma}(x_i) \to x_i'\gamma$, in probability it follows that the weighted quantile regression estimator $\hat{\beta}(\tau, \hat{\gamma})$ converges to $\beta - \gamma Q^{-1}(\tau)$ like its unweighted counterpart, but because the weighted model now has iid errors it achieves full efficiency. This is the basic message of Koenker and Zhao (1995) which provides a detailed analysis of this approach. Newey and Powell (1990) study a related weighted quantile regression estimator in the censored case and show that the estimator attains a semiparametric efficiency bound.

Within the general class of linear quantile regression models, that is the class of models with conditional quantile functions which are linear in parameters, the location scale models (5.3.3) are quite special. Before plunging ahead with reweighting as we have just described one may wish to test whether the location-scale hypothesis is plausible. A simple test of this form could be based on the p-vector of ratios,

$$T_n = (T_{ni}) = \left(\frac{\hat{\beta}_{ni}(\tau_1) - \hat{\beta}_{ni}(\tau_0)}{\hat{\beta}_{ni}(\tau_1') - \hat{\beta}_{ni}(\tau_0')} \right).$$

Under the location scale hypothesis the components of T_n would all converge to

$$\frac{Q_u(\tau_1) - Q_u(\tau_0)}{Q_u(\tau_1') - Q_u(\tau_0')}$$

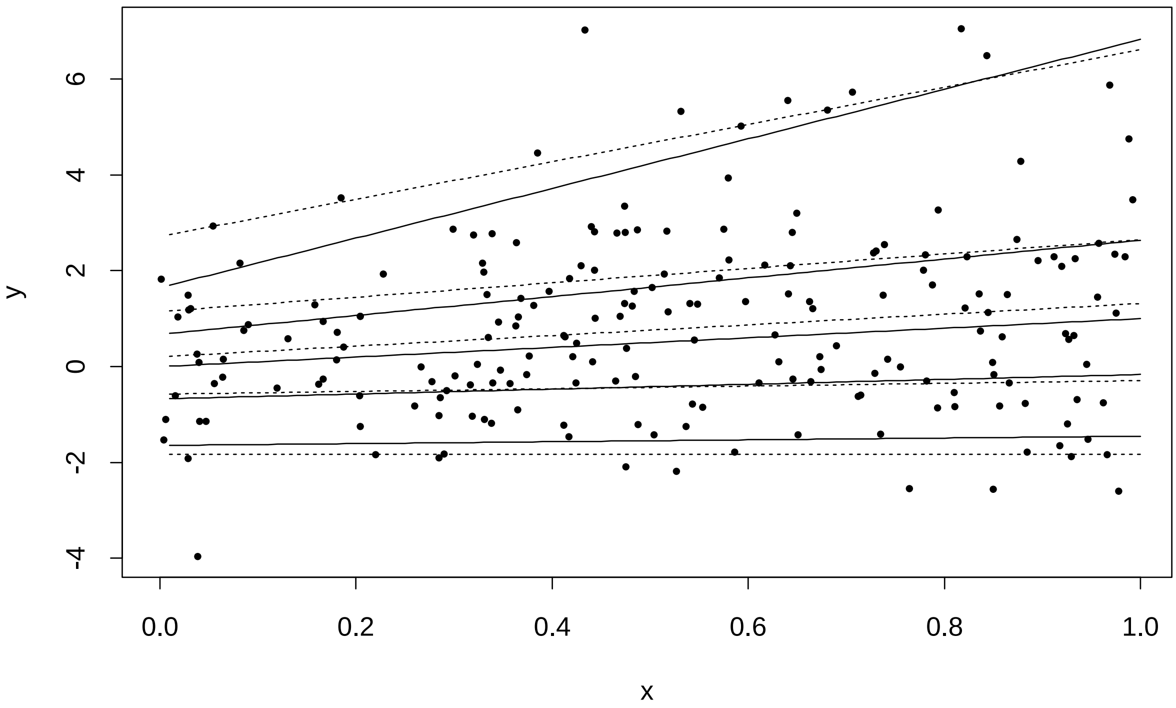

FIGURE 5.2. Non-location-scale linear quantile regression: The 200 points plotted points are generated from a model with linear conditional quantile functions illustrated by the solid lines in the figure. Corresponding unweighted quantile regression estimates appear as dotted lines. Note that although all the conditional quantile functions are linear, the model is not of the location scale form: the the conditional distribution of the response is symmetric at $x = 0$ but quite asymmetric at $x = 1$.

and consequently one could base a test on some measure of the maximal discrepancy between the observed components. We will describe more sophisticated strategies for carrying out tests of this type in Section X.x.

A simple example of a non-location-scale linear conditional quantile function model is given in Figure 3. We have generated 200 observations from a model with conditional quantile functions

$$Q_Y(\tau|x) = \Phi^{-1}(\tau) + exp(\Phi^{-1}(\tau))x$$

with the x's generated as uniform on the interval $[0, 10]$. It is immediately apparent that the *shape* of the conditional density of Y is very different at $x = 0$ than it is at $x = 10$. At 0, Y is conditionally standard normal and the solid lines which indicate the true conditional quantile functions are symmetric, while at $x = 10$ the conditional density is quite asymmetric, reflecting the effect of the lognormal component proportional to x. The corresponding fitted quantile regression lines appear as dotted lines in the figure. It is easy to show that the coefficients of these unweighted estimates are consistent for their corresponding population parameters, but as in the location scale model the question arises: Can we improve upon the unweighted estimators?

This question has a straightforward answer. The appropriate weights for the τth quantile regression are the vector of conditional densities evaluated at the τth quantile, $w_i = f_Y(Q_Y(\tau|x_i)) = (x'\dot{\beta}(\tau))^{-1}$ where $\dot{\beta}(\tau) = d\beta(\tau)/d\tau$ Estimating these weights is quite straightforward given estimates of the of the entire unweighted quantile regression process, using any of the sparsity estimation methods discussed in Section S.x. In the location-scale case the situation is somewhat simplified because $(x'\dot{\beta}(\tau))^{-1} = (x'\gamma)\dot{Q}_u(\tau)$ for some p-vector γ and consequently we can ignore the sparsity estimation problem because it would contribute the same factor to each weight. In general, this isn't the case and we are required to estimate $\dot{\beta}$.

CHAPTER 6

Computational Aspects of Quantile Regression

While early advocates of absolute error methods like Boscovitch, Laplace, and Edgeworth all suggested ingenious methods for minimizing sums of absolute errors for bivariate regression problems, it was not until the introduction of the simplex algorithm in the late 1940's, and the formulation of the ℓ_1 regression problem as a linear program somewhat later, that a practical, general method for computing absolute error regression estimates was made available.

We have already seen that the linear programming formulation of quantile regression is an indispensable tool for understanding its statistical behavior. Like the Euclidean geometry of the least squares estimator, the polyhedral geometry of minimizing weighted sums of absolute errors plays a crucial role in understanding these methods. This chapter begins with a brief account of the classical theory of linear programming stressing its geometrical aspects and introducing the simplex method. The simplex approach to computing quantile regression estimates will then be described and we will emphasize the special role of simplex-based methods for "sensitivity analysis" or parametric programming in a variety of quantile regression contexts in Section 2.

In Section 3 we will describe some recent developments in computation which rely on "interior point" methods for solving linear programs. These new techniques are especially valuable in large quantile regression applications where the simplex approach becomes impractical. Interior point methods are also highly relevant for nonlinear quantile regression problems, a topic addressed in Section 4. Some specialized topics dealing with applications to non-parametric quantile regression are treated in the final section of the chapter.

1. Introduction to Linear Programming

Problems which seek to optimize a linear function subject to linear constraints are called *linear programs*. Such problems play an important role throughout applied mathematics. One of the earliest explicit examples is the so-called diet problem: an individual has a choice of n foods which he may consume in quantities $x = (x_1, ..., x_n)$.

The foods provide nutrients in varying degrees, and we may represent the requirements for such nutrients by the m linear constraints,

$$\begin{array}{cccc} a_{11}x_1 & +\ldots+ & a_{1n}x_n & \geq & b_1 \\ \vdots & & \vdots & & \vdots \\ a_{m1}x_1 & +\ldots+ & a_{mn}x_n & \geq & b_m \end{array} \tag{6.1.4}$$

where a_{ij} denotes the amount of nutrient i provided by food j, and $b_1, ..., b_m$ denote the annual requirements of each of the m nutrients. The cost of the diet x may be represented as

$$c(x) = c_1x_1 + ... + c_nx_n$$

so we may concisely formulate the problem of finding the least expensive diet achieving our nutritional requirements as

how do we know that all these constraints will be active at a solution? Formally, this seems only necessary when we insist that $Ax = b$.

$$\min\{c'x | Ax \geq b,\ x \geq 0\}. \tag{6.1.5}$$

The first formulation of this problem to be solved by formal linear programming methods was that of Stigler (1945). The simplex method applied to Stigler's problem produced a rather appalling diet consisting of flour, corn meal, evaporated milk, peanut butter, lard, beef liver, cabbage, potatoes and spinach and achieved the staggering annual cost of \$39.47, a saving of almost 50 cents per year over a diet found earlier by Stigler by somewhat less systematic methods. According to Dantzig(1951) computing this simplex solution by hand in 1947 required 120 man hours.

Why were so few foods represented in the "optimal diet", when the original problem offered an enticing menu of 77 different foods? The answer to this question is fundamental to the basic understanding of linear programming so it is worth considering in some detail. Stigler's formulation of the diet problem involved nine different nutritional requirements. It is no coincidence that the optimal diet consisted of exactly nine distinct foods.

1.1. Vertices. To see this we must delve a bit more deeply into the geometry of the constraint set $S = \{x | Ax \geq b, x \geq 0\}$. Since S is polyhedral and convex, being the intersection of a finite number of halfspaces, vertices of S have a special status. The vertices of S are extreme points, or "corners", isolated points which do not lie on a line connecting distinct points is S. To further characterize such vertices, consider the augmented, $n + m$ by n system of linear inequalities,

$$\begin{pmatrix} A \\ I_n \end{pmatrix} x \geq \begin{pmatrix} b \\ 0 \end{pmatrix}$$

Associated with any point $x \in S$, the *active constraints* will refer to the constraint rows which hold with equality. Nonbinding constraints, which hold with strict inequality, will be called *inactive*. A vertex of S is a point $x \in S$ whose submatrix of active constraints contains at least one subset of n linearly independent rows. It is this linear independence which prohibits vertices from being expressed as proper linear combinations of two or more distinct points in S. The crucial role of vertices in the theory of linear programming is revealed by the following proposition.

PROPOSITION 6.1. *If the linear program* (6.1.5) *has a bounded solution, then it has a vertex solution.*

This proposition has a rather self-evident geometric interpretation. We may regard the constraint set S as an n-dimensional object like a cut diamond with flat facets, and linear edges connecting a finite number of distinct vertices. Level surfaces of the "cost", or objective, function $c(x)$ are simply a family of parallel hyperplanes so the solution may be envisioned as gradually "lowering" the cost planes until they just touch the set S. This "tangency" may occur at a vertex, in which case the solution is unique, or it may occur along an entire edge, or facet in higher dimensions, in which case the solution consists of an entire convex set delimited by vertices. In either case, the crucial role of vertex solutions is apparent. If the objective function can be reduced without bound while maintaining feasibility, the notional solution "at infinity" does not occur at a vertex. A formal proof of this proposition, which is somewhat more arduous than our geometric intuition, may be found for example in Gill, Murray and Wright(1991, see Theorem 7.7.4)

Presuming that the m constraint rows defined by the matrix A are linearly independent, we can form vertices of S by replacing m rows of I_n by the rows of A, and setting the remaining $n-m$ coordinates of x to zero. Let h denote the indices of the active constraints thus selected, and partition the constraints, writing them as

$$\begin{pmatrix} A(h) & A(\bar{h}) \\ 0 & I_{n-m} \end{pmatrix} \begin{pmatrix} x(h) \\ x(\bar{h}) \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix},$$

where $\bar{h} = \{1, \ldots, n\} \backslash h$, and $A(h)$ denotes a submatrix of A consisting of the columns corresponding to the active indices h. Solving, we have, presuming $A(h)^{-1}$ exists,

$$x(h) = A(h)^{-1}b \tag{6.1.6}$$

$$x(\bar{h}) = 0. \tag{6.1.7}$$

Provided $x(h) \geq 0$, this point is a vertex of S. Whether this vertex is optimal remains to be seen, but the proposition assures us that we need only check the finite number of such vertices and pick the one that achieves the lowest cost. This comment may not have come as a great solace to early pioneers of linear programming computation, since there are $\binom{n}{m}$ such vertices, each requiring the solution of an $m \times m$ linear system of equations. Even for Stigler's problem $\binom{77}{9} \geq 1.6 \times 10^{11}$ appears prohibitively large. Fortunately, we need not visit every vertex. We need only to find an intelligent way of passing from one vertex to the next; the convexity of the constraint set and the linearity of the objective function assures us that starting at any vertex there is a path through adjacent vertices along which the objective function decreases on each edge. To explore such strategies we need: a criterion for deciding that a vertex is optimal, and a rule for choosing an adjacent vertex if the optimality condition is not satisfied.

1.2. Directions of Descent. Let x_0 be an initial, not necessarily feasible point and consider the the feasibility of a step of length σ in the direction p with respect to the ith constraint, $a_i'x \geq b_i$. Since

$$a_i'(x_0 + \sigma p) = a_i'x_0 + \sigma a_i'p$$

we may write the step length to constraint i from x_0 in direction p as,

$$\sigma_i = \frac{a_i'x_0 - b}{-a_i'p} \text{ if } a_i'p \neq 0$$

If $a_i'p = 0$, the step moves x_0 parallel to the constraint and we regard the step length to the constraint as infinite with sign determined by the sign of the initial "residual" associated with the constraint.

Given that x_0 were feasible we need to know what directions are feasible in the sense that the point $x_0 + \sigma p$ remains feasible for some sufficiently small σ. Suppose first that constraint i is inactive at x_0, so that $a_i'x_0 > b_i$. In this case, any direction p is feasible: if $a_i'p \geq 0$, $x_0 + \sigma p$ remains feasible for any $\sigma > 0$. If $a_i'p < 0$, both numerator and denominator of the step length to the ith constraint are positive and there is a strictly positive σ_i at which the constraint becomes active. We conclude that inactive constraints do not restrict the set of feasible directions, only the length of the feasible step is constrained.

For active constraints the situation is reversed. If $a_i'x_0 = b_i$ and $a_i'p < 0$, even the smallest step $\sigma > 0$ violates the feasibility condition. On the other hand, if $a_i'p \geq 0$, feasibility is maintained for all $\sigma > 0$. Thus, for active constraints, the feasible directions are constrained, but once a direction is determined to be feasible active constraints impose no further restriction on step length.

Feasible directions with respect to several constraints may be derived immediately from this analysis of a single constraint. If p is feasible at x_0 for the system of constraints $Ax \geq b$, then for some $\sigma > 0$, $A(x_0 + \sigma p) \geq b$ and this requires, in turn, that for the active constraints, i.e., the rows, h, such that $A(h)x_0 = b$ we have $A(h)p \geq 0$.

Note that since equality constraints of the form $Ax = b$ must be active, and since they require both $Ax \geq b$ and $Ax \leq b$, so $-Ax \geq -b$, the preceding argument implies that any feasible direction p must satisfy both $Ap \geq 0$ and $-Ap \geq 0$, so for equality constraints the only feasible directions are those that lie in the null space of A, i.e., such that $Ap = 0$.

This brings us to the existence of feasible directions of *descent*. Since

$$c'(x_0 + \sigma p) = c'x_0 + \sigma c'p$$

a direction p is a direction of descent at x_0 iff $c'p < 0$. Given an initial point, x_0, how do we determine whether we have a feasible direction of descent? Let $\mathcal{S}(A')$ denote the space spanned by the column vectors of A',

$$\mathcal{S}(A') = \{y | y = A'v \text{ for some } v\}$$

then if $y \in \mathcal{S}(A')$ and $Ap = 0$, there exists a v such that $y'p = v'Ap = 0$, so p is not a descent direction. On the other hand, if $y \notin \mathcal{S}(A')$, then there exists a $p \in \Re^n$ such that $Ap = 0$ and $y'p < 0$, and p is a direction of descent at x_0.

1.3. Conditions for Optimality. In the special case of linear programs with only equality constraints the conditions for optimality of a solution are quite simple. Since this simple case illustrates certain aspects of the more general problem we may begin by considering the problem

$$\min\{c'x | Ax = b\}$$

There are three possible situations:

1. The constraints $Ax = b$ are inconsistent, so the feasible set is empty and no optimal point exists.
2. The constraints $Ax = b$ are consistent, so the feasible set is nonempty, and either
   (a) $c \in \mathcal{S}(A')$, so there exists v such $c = A'v$, and for any feasible point x,
   $$c(x) = c'x = v'Ax = y'b$$
   so any feasible point achieves the same value of the objective function. Thus, all feasible points are optimal, or
   (b) $c \notin \mathcal{S}(A')$, so there exists a direction p such that $Ap = 0$, and $c'p < 0$. This direction is feasible and thus starting from any feasible point, x, the objective function can be driven to $-\infty$ by taking a step $x + \sigma p$ for σ arbitrarily large.

None of these options seem particularly attractive and together they illustrate the crucial role of inequality constraints in determining vertex solutions in linear programming.

The situation for problems with inequality constraints is somewhat more challenging. In this case we must carefully distinguish between active and inactive constraints in determining feasible direction, and obviously sets of active and inactive constraints may depend on the initial feasible point. Let h denote the index set of active constraints and $\bar{h}$ the index set of inactive constraints at an initial (feasible) point x_0. A feasible direction of descent, p, exists if

$$A(h)p \geq 0 \text{ and } c'p < 0$$

The point x_0 is optimal iff no direction of descent exists.

For equality constrained linear programs it was possible to provide a simple test for the optimality of an initial feasible point based on whether c was contained in the span of A'. For inequality constrained LP's the situation is somewhat more complicated and relies on the following classical result.

LEMMA 6.2. *(Farkas (1902)) Let B be a p by n matrix and c be a vector in $\Re^n$, then*

$$c'p \geq 0 \textit{ for all } p \textit{ such that } Bp \geq 0$$

iff

$$c = B'v \textit{ for some } v \geq 0$$

Applying the lemma to obtain optimality conditions for the purely inequality constrained linear program we have the following theorem.

THEOREM 6.3. *(GMW 1992, Theorem 7.7.2) For the problem,*

$$\min\{c'x | Ax \geq b\}$$

either:

1. *The constraints $Ax \geq b$ are inconsistent and therefore the problem has no solution, or*
2. *There is a feasible point $\hat{x}$, and a vector $\hat{v}$ such that*

$$c = A(h)'\hat{v} \textit{ with } \hat{v} \geq 0$$

   where $A(h)$ denotes the submatrix of A corresponding to the active constraints at $\hat{x}$. In this case $c(\hat{x})$ is the unique minimum value of $c'x$ for $\{x|Ax \geq b\}$ and $\hat{x}$ is an optimizer, or,
3. *The constraints $Ax \geq b$ are consistent, but the conditions $(*)$ are not satisfied for any feasible point x. In this case, the solution is unbounded from below.*

1.4. Complementary Slackness. The optimality conditions (2) require that a weighted sum of the m possible columns of the matrix A' equals the cost vector c. In the Theorem we have not specified the dimension of the vector $\hat{v}$, but we have seen that if $\hat{x}$ is a vertex, then $A(h)'$ must consist of at least n linearly independent columns from the full set of m columns of A'. It is perhaps more convenient, and obviously equivalent to take $u \in \Re^m$ and let $v = u(h)$, and set $u(\bar{h}) = 0$. We may then express (2) as

$$c = A'u \text{ and } u \geq 0$$

with the added "complementary slackness" condition that

$$r_i u_i = 0 \qquad i = 1, \ldots, m$$

where $r_i = a_i' x_0 - b$. For $i \in h$ we are assured that $r_i u_i = 0$, because $r_i = 0$, by definition, for active constraints, while for $i \in \bar{h}, u_i = 0$. The vector $u \in \Re^m$ may be regarded as a vector of Lagrange multipliers corresponding to the constraints $Ax \geq b$. For binding constraints we expect these multipliers, which may be interpreted as the "marginal costs" of their respective constraints, to be positive. For nonbinding (inactive) constraints tightening the constraint infinitesimally imposes no cost so the multiplier is zero.

mention KKT c
tions here?

Combining the conclusions for equality and inequality constraints we have the following result.

THEOREM 6.4. *Consider the linear program*

$$\min\{c'x \mid Ax = b, \quad x \geq 0\}$$

and suppose $\hat{x}$ *is a feasible point. The point* $\hat{x}$ *is optimal iff there exists* $\hat{u} \in \Re^n$ *and* $\hat{v} \in \Re^m$, *such that,*

$$c = A'\hat{v} + \hat{u} \quad u \geq 0$$

with

$$\hat{x}_i \hat{u}_i = 0 \qquad i = 1, \ldots, n$$

The canonical form of LP given in Theorem 6.4 appears somewhat restrictive, but apparently more general forms may be transformed into this canonical form by the introduction of new variables. For example, consider the problem

$$\min\{c'x \mid Ax = b, Dx \geq d\}.$$

Rewriting this as

$$\min\{\tilde{c}'z \mid \tilde{A}z = \tilde{b}, \quad z \geq 0\}$$

where $\tilde{c} = (c, 0)', \tilde{b} = (b', d')', z = (x', y')'$ and

$$\tilde{A} = \begin{pmatrix} A & 0 \\ D & -I \end{pmatrix}$$

we are back to canonical form. The new variables y are usually called "slack" variables connoting that they take the value zero when the original constraints are active, and "take up the slack" associated with inactive constraints.

1.5. Duality. The Lagrange multiplier interpretation of the variables u in Theorem Theorem 6.4 is our first hint of the elegant theory of duality rising in linear programming and related contexts. Corresponding to any primal linear program we may formulate a *dual* linear program, a reflection of the original problem "through the looking-glass" in which minimizing with respect to the original variables turns into maximizing with respect to the Lagrange multipliers of the dual formulation.

In the dual formulation of the diet problem, for example, we have seen Theorem 6.4 that at a solution $\hat{x}$, there exists a vector of Lagrange multipliers, say, $\hat{v} = (\hat{y}', \hat{z}')'$, such that

$$c = A'\hat{y} = \hat{z}$$

and such that $a_i\hat{x}_i - b_i)\hat{y}_i = 0,\ i = 1, ..., m$ and $\hat{x}_i\hat{z}_i = 0,\ i = 1, ..., n$. At such a point, note that

$$c'\hat{x} = \hat{y}'A\hat{x} + \hat{z}'\hat{x} = \hat{y}'b.$$

This suggests that we might consider reformulating the original problem of finding the optimal diet, $\hat{x}$, as a problem of finding the vector of Lagrange multipliers, v, solving the linear program,

$$\max\{b'y | A'y + z = c, (y, z) \geq 0\}.$$

Since the slack vector, z, is simply a "residual" immediately obtained from the constraint

$$A'y + z = c$$

once we have specified y, we can view this dual problem as asking us to find a vector of "shadow prices", y, corresponding to the m nutritional requirements. These prices may be interpreted at a solution as the prices which consumers would be willing to pay for dietary supplements corresponding to each of the m nutritional requirements. One way to interpret the dual of the diet problem is as a search for a revenue maximizing vector of these "diet supplement" shadow prices which would be sustainable given the current prices of the available foods of the primal problem. Obviously, these prices are constrained by the current levels of food prices and the quantities of each nutritional requirement provided by the foods. Finding the optimal shadow prices is equivalent to finding the optimal diet for the original problem, because in effect it identifies a subset of active constraints, ones for which the dual constraints are binding, these active constraints define a basis h which in turm can be used to find the explicit primal solution as in (6.1.6). Note also that the requirement that at a solution all the nutrients have positive shadow prices ensures that all of the rows of A appear in the active set of the primal solution and this is the essential requirement which ensures that there will be precisely m foods in the optimal diet.

To examine this duality more generally consider the canonical (primal) LP,

$$\min\{c'x | Ax = b, \quad x \geq 0\} \tag{6.1.8}$$

for which we have seen that a feasible point, $\hat{x}$, is optimal iff there exist vectors $\hat{u}$ and $\hat{v}$ such that

$$c = A'\hat{v} + \hat{u}, \quad \hat{u} \geq 0 \quad \text{with} \quad \hat{x}_i\hat{u}_i = 0 \qquad i = 1, \ldots, n.$$

These optimality conditions suggest the following dual formulation of the problem:

$$\max\{b'v | A'v \leq c\}.$$

Note that at a solution $(\hat{x}, \hat{u}, \hat{v})$ we have

$$c'\hat{x} = \hat{v}'A\hat{x} + \hat{u}'\hat{x} = \hat{v}'b$$

so at the optimum the primal and dual values of the objective function are equal. At any other feasible point (x, u, v) we have

$$c'x - v'b = u'x \geq 0.$$

The left hand side is usually called the "duality gap", the discrepancy between the primal and dual objective functions, and since it is equal to the inner product of two nonnegative vectors, it can be taken as a direct measure of the departure from the complementary slackness condition $u'x = 0$ which indicates optimality. This relationship plays a central role in the theory of interior point algorithms for solving linear programs.

An important observation regarding the dual problem is that the primal vector $\hat{x}$ may be regarded as the Lagrange multiplier vector corresponding to the dual constraints $A'v \leq c$, at the optimum. Thus, finding the dual of the dual formulation returns us to the primal formulation of the problem. The apparent symmetry between dual and primal formulations may be further accentuated by rewriting the dual in the equivalent form

$$\max\{b'v | A'v + u = 0, \quad u \geq 0\}.$$

In the next Section we explore the application of the foregoing ideas in the specific context of the quantile regression optimization problem.

2. Quantile Regression

The primal formulation of the basic linear quantile regression problem,

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

may be written as,

$$\min\{\tau e_n'u + (1-\tau)e_n'v | y - Xb = u + v, \ b \in \mathbb{R}^p, \ (u, v) \in \mathbb{R}_+^{2n}\},$$

which may be seen to be (almost) in canonical form (6.1.8) by making the following identifications: $c = (0_p', \tau e_n', (1-\tau)e_n')'$, $x = (b', u', v')'$, $A = [X \vdots I \vdots -I]$ and $b = y$. Full adherence to the canonical form would require the further decomposition of the vector b into its positive and negative parts. As we have already noted, the polyhedral nature of the constraint set and the linear objective function imply that we may focus attention on the vertices of the constraint set. These vertices, as we observed in Chapter 2, may be indexed by the $\binom{n}{p}$ elements $h \in \mathcal{H}$ and take the form,

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

$$u(h) = v(h) = 0,$$

$$u(\bar{h}) = (y - Xb(h))^{+},$$

$$v(\bar{h}) = (y - Xb(h))^{-}.$$

Clearly, at any such vertex, the complementary slackness conditions, $u_i v_i = 0$ hold, and there are at least p indices, $i \in h$, with $u_i = v_i = 0$. Such points provide an exact fit to the p observations in the subset h, and set the corresponding u and v vectors of the solution equal to the positive and negative parts of the resulting residual vector.

The primal quantile regression problem has corresponding dual problem,

$$\max\{y'd | X'd = 0,\ d \in [\tau - 1, \tau]^n\}$$

Equivalently, we may reparametrize the dual problem to solve for

$$a = d + (1 - \tau)e_n$$

yielding the new problem,

$$\max\{y'a | X'a = (1 - \tau)X'e_n,\ a \in [0, 1]^n\}.$$

In dual form, the problem may be seen to be one of optimizing with respect to a vector that lies in a unit cube. Such "bounded variables" problems have received considerable special attention in the linear programming literature. The dual also provides a critical link to the theory of linear rank statistics, generalizing the rank score functions of Hájek , as described in Gutenbrunner and Jurečková (1992).

Since, at any solution, $\{(\hat{b}, \hat{u}, \hat{v}), \hat{d}\}$, we must have

$$\tau e_n' \hat{u} + (1 - \tau) e_n' \hat{v} = y'\hat{d}$$

we see that,

$$\hat{d}_i = \begin{cases} \tau & \text{if} \quad \hat{u}_i > 0 \\ (\tau - 1) & \text{if} \quad \hat{v}_i > 0 \end{cases}$$

while for observations $i \in h$ with $\hat{u}_i = \hat{v}_i = 0$, $\hat{d}(h)$ is determined from the equality constraint, $X'd = 0$, as,

$$\hat{d}(h) = -[X(h)']^{-1} X(\bar{h})' \hat{d}(\bar{h}).$$

The dual vector $\hat{d}(h)$ at a solution corresponding to the basic observations h is thus the same as the vector ξ_h of Theorem 2.1. The fact that at a solution $\hat{d}(h) \in [\tau - 1, \tau]^p$ is precisely the optimality requirement that appears in that result.

It is conventional in describing implementations of simplex-type algorithms for solving linear programs to speak of a phase I of the algorithm in which an initial feasible vertex of the problem is found, and then a phase II in which we proceed from one such vertex to another until optimality is acheived. In some linear programs just determining whether a feasible vertex *exists* can be quite difficult, however in the quantile regression problem we can choose any subset, h, and define a basic feasible point to the problem provided that the matrix $X(h)$ is of full rank, thereby completing phase I. Thus, we will focus attention below on Phase II of the computation, describing an approach implemented in the path-breaking algorithm of Barrodale and Roberts (1973).

Let $h_0 \in \mathcal{H}$ denote the index set corresponding to an initial feasible vertex of the constraint set. Consider the directional derivitive of the objective function,

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

evaluated at $b(h_0) = X(h_0)^{-1}y(h_0)$ in direction δ, which we may write as in Section 2.2,

$$\nabla R(b(h_0), \delta) = -\sum_{i=1}^{n} \psi_\tau^*(y_i - x_i'b(h_0), -x_i'\delta)x_i'\delta,$$

where $\psi_\tau^*(u, w) = \tau - I(u < 0)$ if $u \neq 0$ and $\psi_\tau^*(u, w) = \tau - I(w < 0)$ if $u = 0$. Since we are already at a vertex, and the constraint set is convex, we can restrict attention to certain extreme directions, δ, which correspond to moving away from the vertex along the edges of the constraint set which intersect at the current vertex. These edges may be represented algebraically as

$$d(\alpha, h, \delta_j) = b(h) + \alpha\delta_j \tag{6.2.9}$$

where δ_j is the jth column of the matrix $X(h)^{-1}$ and α is a scalar which controls where we are along the edge. This representation of the edge is just the usual parametric representation of a line through the point $b(h)$ in direction δ_j. Here, δ_j, is the vector orthogonal to the constraints formed by the remaining basic observations with the jth removed. Note that α can be either positive or negative in (6.2.9) and the directional derivitive will be different for δ_j and $-\delta_j$.

As in Theorem 2.1, let

$$\xi(h) = -\sum_{i \in \bar{h}} \psi_\tau^*(y_i - x_i'b(h))x_i'X(h){-}1$$

and note that if δ_j is the jth column of $X(h)^{-1}$, then for $i \in h$, $x_i'\delta_j = 0$ for $j \neq i$ and equals one for $i = j$, so for such δ_j,

$$\nabla R(b(h_0), \delta_j) = \xi_j(h) + 1 - \tau$$

and

$$\nabla R(b(h_0), -\delta_j) = -\xi_j(h) + \tau.$$

If these directional derivitives are all non-negative for $j = 1, ..., p$, we have the optimality condition of Theorem 2.1,

$$\xi(h) \in [\tau - 1, \tau]^p,$$

otherwise there is an edge which is a direction of descent leading away from the point $b(h)$, and we can reduce $R(b)$ by moving in this direction way from $b(h)$.

Which edge should be chosen? The conventional choice, adopted by BR (**?**) is the one of "steepest descent". This is the one for which the directional derivitive $\nabla R(b(h), \pm\delta_j)$ is most negative. Having selected a direction $\delta^* = \sigma\delta_j$ for some j and $\sigma \in \{-1, 1\}$ we face the question: how far should we travel in this direction? BR answered this question quite innovatively.

Rather than simply adopting the usual simplex strategy of traveling only as far as the next vertex, that is only as far as needed to drive one of the non-basic observation's residuals to zero, they proposed to continue in the original direction as long as doing so continued to reduce the value of the objective function. Thus, as we travel in the direction, δ^*, we encounter points at which observations not in the original basis have residuals which are eventually driven to zero. Conventional simplex strategies, when they encountered such a situation, would introduce the new observation into to the basis, recompute a descent direction, and continue. Instead, BR elect to continue in the originally determined direction as long as it remains a viable direction of descent. In this way, they dramatically reduce the number of simplex pivot operations required when the basis changes. At the intermediate points all that is required is a change of sign in the contribution to the gradient of the observation whose residual is passing through zero. In effect, this strategy constitutes what is sometimes called a Cauchy algorithm in general nonlinear optimization theory. see e.g. Nazareth (1991). We select a direction of steepest descent at each iteration and then do a one dimensional line search, a minimization along a ray in the chosen direction.

The resulting line search has a simple interpretation which takes us back to the discussion of the regression through the origin example of Chapter 1. In effect, at each step, we are solving a problem of the form,

$$min_{\alpha \in \mathbb{R}} \sum_{i=1}^{n} \rho_\tau(y_i - x_i' b(h) - \alpha x_i' \delta^*)$$

which is just the regression through the origin quantile regression problem in α with the response $y_i - x_i' b(h)$ and a design consisting of the single variable $x_i' \delta^*$. This subproblem is easily solved by finding a weighted quantile, we simply need to generalize slightly the weighted median characterization already described in Chapter 1. This is done in Appendix A of this Chapter.

The algorithm continues in this manner until there is no longer a direction of descent at which point optimality has been acheived, and the algorithm terminates. We will see that the BR algorithm provides an extremely efficient approach to quantile regression computations for problems of modest size. For problems with n up to several hundred observations the modified BR algorithm described in KO is comparable in speed to least squares estimation of the conditional mean regression as implemented in current software packages like Splus and Stata. In very large problems, however, the simplex approach of BR, or perhaps more precisely, the exterior point approach of BR – the path along the exterior of the constraint set – becomes painfully slow, relative to least squares. In Section 4 we will describe some recent developments which significantly improve upon the performance of BR in large problems. But before introducing these new methods we will describe several applications of parametric programming ideas, which offer extremely effective exterior point computational strategies for important *families* of quantile regression problems.

3. Parametric Programming for Quantile Regression Problems

The aspect of quantile regression computation which seems most puzzling to newcomers to the subject is the idea that we can solve

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

efficiently for *all* $\tau \in [0,1]$. One solution to a linear programming problem for fixed τ seems bad enough, how is it possible to find a solution for a continuum of τ's? The answer to this question again lies in the elementary geometry of linear programming.

Consider the primal version of the quantile regression LP and imagine that we are at a vertex solution for some initial $\tau = \tau_0$. What happens when we decide that we would also like to know the solution for some $\tau_1 > \tau_0$? Geometrically, changing τ tilts the orientation of the (hyper-)planes representing the level surfaces of the objective function, but has no effect on the constraint set of the primal problem. Thus if we are at a unique vertex solution corresponding to τ_0, there is a neighborhood of τ_0 within which the solution remains unperturbed. Eventually, of course, the plane tilts enough so that not only the original solution at τ_0, but an entire edge of the constraint set solves the perturbed problem. There will be a whole line segment of solutions in $\mathbb{R}^p$ corresponding to the “tangency” of the plane representing the minimal attainable level of the objective function on an edge of the constraint set.

Tilting the plane a bit beyond this τ_1 restores the uniqueness of the solution at the opposite end of the line segment defining the edge. What has happened algebraically is quite simple. The edge may be represented as in the previous subsection as

$$d(\alpha, h, \delta_j) = b(h) + \alpha\delta_j$$

where $b(h)$ is the initial vertex solution at τ_0 and δ_j is the jth column of $X(h)^{-1}$. As in our description of the BR algorithm we travel along this edge until we come to the next adjacent vertex. At this new vertex we have driven a new residual, say the kth to zero, and at this value τ_1 any point between $b(h)$ and $b(h')$ with $h' = k \cup h \setminus j$ solves the problem. For $\tau > \tau_1$ the plane representing the objective function again tilts away from the edge and we have a unique solution at $b(h')$.

Proceeding in this manner we identify breakpoints $\tau_j \in \{0 = \tau_0, \tau_1, ..., \tau_J = 1\}$ at which the primal solution flips from one basis to another. (At these points we have an interval of solutions, elsewhere the solution is unique.) Of course, in the simplest one sample setting where $x_i \equiv 1$ we have exactly n of these breakpoints at $\tau_j = j/n$. However, in general the number and location of the τ_j's depends in a complicated way on the design configuration as well as the observed response. Under mild regularity conditions which can be expected to hold in most quantile regression applications, Portnoy (1989) has shown that the number of breakpoints J is $\mathcal{O}_p(n \log n)$. There is an extensive LP literature on related problems of this sort which are usually called parametric programming problems or sensitivity analysis. See Gal () for a detailed treatment of this literature.

This isn't quite right we might need to mention the pathology of intervals of MOS....

Do we need some empirical evidence on this???

An explicit formulation of the computations described above returns us to the dual form of the problem. At the initial solution $b(h)$, at $\tau = \tau_0$ we have the dual constraint,

$$X'\hat{a}(\tau) = (1-\tau)X'e$$

for some $\hat{a}(\tau) \in [0,1]^n$, and for the non-basic observations, $\hat{a}_i(\tau) = I(u_i > 0)$ for $i \in \bar{h}$. Define

$$\mu = (X(h)')^{-1}[X'e - \sum_{i \in \bar{h}} x_i \hat{a}_i(\tau)]$$

and

$$\lambda = (X(h)')^{-1}X'e$$

so for τ sufficiently close to τ_0,

$$\hat{a}_i(\tau) = \mu_i - \lambda_i \tau \quad i \in h.$$

To find the *next* τ we need to find among all the τ's which solve either,

$$\mu_i - \lambda_i \tau = 0,$$

or,

$$\mu_i - \lambda_i \tau = 1,$$

the one that changes least. To accomplish this we simple examine the set of $2p$ numbers,

$$\mathcal{T} = \{\mu_i/\lambda_i, (\mu_i - 1)/\lambda_i,\ i \in h\}.$$

Presuming that we are looking for the next *largest* τ, we select,

$$\tau_1 = min\{\tau \in \mathcal{T} | \tau > \tau_0\}.$$

This selection determines not only the length of the interval for which the point $b(h)$ remains optimal, but also identfies which edge is optimal at the new breakpoint τ_1. The direction of movement along this edge is given by

$$\sigma = \begin{cases} 1 & \text{if } \tau_1 = \mu_i/\lambda_i \; i \in h \\ -1 & \text{if } \tau_1 = (\mu_i - 1)/\lambda_i \; i \in h \end{cases}$$

Let the index of the minimizing edge be $i_0 \in h$, then the parametric representation of the edge is $b(h) + \sigma\delta_{i_0}$ where δ_{i_0} is the i_0 column of $X(h)^{-1}$. Finally, to determine how far we can go in this direction we define the ratios,

$$\mathcal{S} = \{s_j = r_j/(\sigma x_j' \delta_{i_0}), \; j \in \bar{h}\},$$

The smallest positive element of $\mathcal{S}$ identifies the distance we may travel along our selected edge before forcing one of the non-basic residuals to become zero. The j so selected now replaces the deleted basic observation i_0 in h and we proceed as from the beginning. In this way we can find the entire quantile regression process $\{\hat{\beta}(t) : t \in [0,1]\}$ and the corresponding dual process, $\{\hat{a}(t) : \; t \in [0,1]\}$, in roughly $n \log n$ simplex pivots. For modest n this is extremely quick; for large n we suggest some alternative computational strategies which can significantly reduce the computational effort without much sacrifice in the informational content of the estimated processes.

3.1. Parametric Programming for Regression Rank Tests. Another important parametric family of quantile regression problems arises in the computation of the inverted rank test confidence intervals described in Chapter 3. In this case, we begin with the solution to a $p+1$ dimensional quantile regression problem. And we would like to construct a confidence interval for the jth parameter of the model by inverting a particular form of the GJKP () rank test of the hypothesis,

$$H_0 : \beta_j = \xi$$

that is, we would like to find an interval $(\hat{\beta}_j^L, \hat{\beta}_j^U)$ with asymptotic coverage $1-\alpha$. Statistical aspects of this confidence interval are described in Chapter 3, here we will focus on describing the computational details of the procedure.

Let $\tilde{X}$ denote the full $(p+1)$-column design matrix of the problem and X, the reduced design matrix with the jth column deleted. Let $\hat{h}$ denote the index set of the basic observations corresponding to the solution of the full problem, and let u denote the jth row of the matrix $\tilde{X}(\hat{h})^{-1}$. Our first task is to reduce the basis $\hat{h}$ by one element, in effect finding a new basis, $\tilde{h}$, for the dual problem,

$$\max\{y - X_j\xi)'a | X'a = (1-\tau)X'e, a \in [0,1]^n\}$$

for ξ near $\hat{\beta}_j$. Here we denote the jth column of $\tilde{X}$ by X_j. Clearly, this reduced solution has $\tilde{a}_i = \hat{a}_i$ for $i \notin h$, but we must identify the observation to be removed from the basic set. Consider the ratios, for $i \in \hat{h}$,

$$s_i = \begin{cases} (\hat{a}_i - 1)/u_i & \text{if} u_i < 0 \\ \hat{a}_i/u_i & \text{if} u_i \geq 0 \end{cases}$$

and let $k \in \hat{h}$ denote the index of the minimal s_i, i.e.

$$w^* = s_k = min_{i \in \hat{h}}\{s_i : i \in \hat{h}\}.$$

The new dual solution is thus,

$$\tilde{a}_i = \hat{a}_i - w^* u_i \; i \in \tilde{h} = \hat{h} \setminus k.$$

Note that this modification of the dual solution doesn't alter the primal solution, we have simply reduced the rank of the porblem by one and incorporated the effect of the jth covariate into the response variable.

Now we are ready to begin the parametric programming exercise. But in this case we must focus attention on the dual form of the quantile regression problem. As we change ξ in the dual problem we again may view this as tilting the plane of the objective function, this time in the dual formulation, while the dual constraint set remains fixed. And again we are looking for a sequence of steps around the exterior of the constraint set; this time the path terminates at the value of ξ for which we first reject H_0.

Each step begins by identifying the nonbasic observation which will next enter the basis. This is the observation whose residual is first driven to zero by the process of increasing ξ. This is fundementally different than the primal parametric problem over τ. In that case, the primal solution $\hat{\beta}(\tau)(\tau)$, was piecewise constant in τ and the dual solution $\hat{a}(\tau)$ was piecewise linear in τ. Now the situation is reversed with the dual solution, viewed as a function of ξ, $\hat{a}(\xi)$, piecewise constant and the primal solution $\hat{\beta}(\xi)$, piecewise linear in ξ. If $b(h)$ is the unique vertex solution to the reduced problem at $\xi = \hat{\beta}_j$, then there is a neighborhood around this value for which the basic observations, h, remain optimal and we can write the residual vector of the perturbed problem as,

$$\begin{aligned} r(\xi) &= y - X_j\xi - XX(h)^{-1}(y(h) - X_j(h)\xi) \\ &= y - XX(h)^{-1}y(h) - (X_j - XX(h)^{-1}X_j(h))\xi. \end{aligned}$$

The incoming observation i is the minimal element corresponding to

$$\delta^* = \min_i\{y_i - x_i'X(h)^{-1}y(h)/(x_{ij} - x_i'X(h)^{-1}X_j(h)) > 0\},$$

presuming, of course that we are interested in *increasing* ξ. Finally, we must find the observation leaving the basis. Let

$$v = X(h)^{-1}x_i(h).$$

If the incoming residual

$$r_{i*} = y_{i*} - x_{i*}'X(h)^{-1}(y(h) - X_j(h)\hat{\beta}_j)$$

is greater than zero set,

$$g_i = \begin{cases} -\hat{a}_i/v_i & \text{if } v_i < 0 \\ (1-\hat{a}_i)/v_i & \text{otherwise} \end{cases}$$

or if $r_i* < 0$, set

$$g_i = \begin{cases} \hat{a}_i/v_i & \text{if } v_i < 0 \\ (\hat{a}_i - 1)/v_i & \text{otherwise.} \end{cases}$$

The outgoing observation is the one corresponding to the minimal value of the g_i's. We can now update the basis and continue the process, until we we reject H_0. The process may be repeated to find the other endpoint of the confidence interval and continued to determine the confidence intervals for each parameter appearing in the model. As can be seen, for example, in Table 2.2, these confidence intervals are asymmetric so they cannot be represented in the form of the usual "estimate $\pm$ k_α standard deviation", so in this way they resemble the confidence intervals one might obtain from the bootstrap percentile method.

When sample sizes are large there are typically a large number vertices which must be traversed to find solutions using the simplex approach we have described above. Indeed, there are infamous examples, notably that of Klee and Minty (1972) which have shown that in problems of dimension, n, simplex methods can take as many as 2^n pivots, each requiring $\mathcal{O}(n)$ effort. Such worst case examples are admittedly pathological, and one of the great research challenges of recent decades in numerical analysis has been to explain why simplex is so quick in more typical problems, see Shamir (1993) for an excellent survey of this literature. Nonetheless, we will see that for quantile regression problems with p fixed and $n \to \infty$, the modified algorithm of Barrodale and Roberts exhibits $\mathcal{O}(n)$ behavior in the number of pivots and therefore has $\mathcal{O}(n^2)$ growth in cpu-time. In the next section we introduce interior point methods for solving linear programs which have been shown to dramatically improve upon the computational efficiency of simplex large quantile regression problems.

4. Interior Point Methods for Canonical LP's

Although prior work in the Soviet literature offered theoretical support for the idea that linear programs could be solved in polynomial time, thus avoiding the pathological exponential growth of the worst-case Klee-Minty examples, the paper of

Karmarker (1984) constituted a watershed in the numerical analysis of linear programming. It offered not only a cogent argument for the polynomiality of interior point methods of solving LP's, but also provided for the first time direct evidence that interior point methods were demonstrably faster than simplex in specific, large, practical problems.

The close connection between the interior point approach of Karmarkar (1984) and earlier work on barrier methods for constrained optimization, notably Fiacco and McCormick (1968) was observed by Gill, et al. (1986) and others, and has led to what may be called without much fear of exaggeration a paradigm shift in the theory and practice of linear and nonlinear programming. Remarkably, some of the fundamental ideas required for this shift appeared already in the 1950's in a sequence of Oslo working papers by the economist Ragnar Frisch. This work is summarized in Frisch (1956). We will sketch the main outlines of the approach, with the understanding that further details may be found in the excellent expository papers of Wright (1992), Lustig, Marsden and Shanno (1994), and the references cited there.

Consider the canonical linear program

$$\min \ \{c'x \mid Ax = b, \ \ x \geq 0\}, \tag{6.4.10}$$

and associate with this problem the following logarithmic barrier (potential-function) reformulation,

$$\min \ \{B(x, \mu) \mid Ax = b\} \tag{6.4.11}$$

where

$$B(x, \mu) = c'x - \mu \sum \log x_k.$$

In effect, (6.4.11) replaces the inequality constraints in (6.4.10) by the penalty term of the log barrier. Solving (6.4.11) with a sequence of parameters μ such that $\mu \to 0$ we obtain in the limit a solution to the original problem (6.4.10). This approach was elaborated in Fiacco and McCormick (1968) for general constrained optimization, but was revived as a linear programming tool only after its close connection to the approach of Karmarkar (1984) was pointed out by Gill, et al. (1986). The use of the logarithmic potential function seems to have been introduced by Frisch (1956), who described it in the following vivid terms,

> My method is altogether different than simplex. In this method we work systematically from the interior of the admissible region and employ a logarithmic potential as a guide – a sort of radar – in order to avoid crossing the boundary.

Suppose that we have an initial feasible point, x_0, for (6.4.10), and consider solving (6.4.11) by the classical Newton method. Writing the gradient and Hessian of B with respect to x as

$$\begin{aligned} \nabla B &= c - \mu X^{-1} e \\ \nabla^2 B &= \mu X^{-2} \end{aligned}$$

where $X = \operatorname{diag}(x)$ and e denotes an n-vector of ones, we have at each step the Newton problem

$$\min_p \{c'p - \mu p' X^{-1} e + \tfrac{1}{2} \mu p' X^{-2} p \mid Ap = 0\}. \tag{6.4.12}$$

Solving this problem, and moving from x_0 in the resulting direction p toward the boundary of the constraint set maintains feasibility and is easily seen to improve the objective function. The first order conditions for this problem may be written as

$$\mu X^{-2} p + c - \mu X^{-1} e = A'y \tag{6.4.13}$$

$$Ap = 0 \tag{6.4.14}$$

where y denotes an m-vector of Lagrange multipliers. Solving for y explicitly, by multiplying through in the first equation by AX^2 and using the constraint to eliminate p, we have

$$AX^2 A'y = AX^2 c - \mu AXe. \tag{6.4.15}$$

These normal equations may be recognized as generated from the linear least squares problem

$$\min_y \| XA'y - Xc - \mu e \|_2^2 \tag{6.4.16}$$

Solving for y, computing the Newton direction p from (6.4.13), taking a step in the Newton direction toward the boundary constitute the essential features of the primal log barrier method. A special case of this approach is the affine scaling algorithm in which we take $\mu = 0$ at each step in (6.4.15), an approach anticipated by Dikin (1967) and studied by Vanderbei, Meketon and Freedman (1986) and numerous subsequent authors.

Recognizing that similar methods may be applied to the primal and dual formulations simultaneously, recent theory and implementation of interior point methods for linear programming have focused on attacking both formulations. The dual problem corresponding to (6.4.10) may be written as

$$\max\{b'y \mid A'y + z = c, \;\; z \geq 0\}. \tag{6.4.17}$$

Optimality in the primal implies

$$c - \mu X^{-1} e = A'y \tag{6.4.18}$$

so setting $z = \mu X^{-1}e$ we have system

$$
\begin{aligned}
Ax &= b \quad x > 0 \\
A'y + z &= c \quad z > 0 \\
Xz &= \mu e.
\end{aligned} \tag{6.4.19}
$$

Solutions $(x(\mu), y(\mu), z(\mu))$ of these equations constitute the central path of solutions to the logarithmic barrier problem, which approach the classical complementary slackness condition $x'z = 0$, as $\mu \to 0$, while maintaining primal and dual feasibility along the path.

If we now apply Newton's method to this system of equations, we obtain

$$
\begin{pmatrix} Z & 0 & X \\ A & 0 & 0 \\ O & A' & I \end{pmatrix} \begin{pmatrix} p_x \\ p_y \\ p_z \end{pmatrix} = \begin{pmatrix} \mu e - Xz \\ b - Ax \\ c - A'y - z \end{pmatrix} \tag{6.4.20}
$$

which can be solved explicitly as,

$$
\begin{aligned}
p_y &= (AZ^{-1}XA')^{-1}[AZ^{-1}X(c - \mu X^{-1}e - A'y) + b - Ax] \\
p_x &= XZ^{-1}[A'p_y + \mu X^{-1}e - c + A'y] \\
p_z &= -A'p_y + c - A'y - z.
\end{aligned} \tag{6.4.21}
$$

Like the primal method, the real computational effort of computing this step is the Choleski factorization of the diagonally weighted matrix $AZ^{-1}XA'$. Note that the consequence of moving from a purely primal view of the problem to one that encompasses both the primal and dual is that $AX^{-2}A'$ has been replaced by $AZ^{-1}XA'$ and the right hand of the equation for the y-Newton step has altered somewhat. But the computational effort is essentially identical. To complete the description of the primal-dual algorithm we would need to specify how far to go in the Newton direction p, how to adjust μ as the iterations proceed, and how to stop.

In fact, the most prominent examples of implementations of the primal-dual log barrier approach now employ a variant due to Mehrotra (1992), which resolves all of these issues. We will briefly describe this variant in the next section in the context of a slightly more general class of linear programs But before doing so, we will illustrate the ideas by describing a simple example of the foregoing theory.

4.1. Newton to the Max: An Elementary Example. To illustrate the shortcomings of the simplex method, or indeed of any strategy for solving linear programs which relies on an iterative path along the *exterior* of the constraint set, consider the problem depicted in Figure 6.1. We have a random polygon whose vertices lie on the unit circle and our objective is to find a point in the polygon that maximizes the sum of its coordinates, that is, the point furthest north-east in the figure.

Since any point in the polygon can be represented as a convex weighting of the extreme points, the problem may be formulated as

$$\max\{e'u | X'd = u, \;\; e'd = 1, \;\; d \in \Re^n_+\}, \tag{6.4.22}$$

where e denotes a (conformable) vector of ones, X is an $n \times 2$ matrix with rows representing the n vertices of the polygon and d is the vector of convex weights to be determined. Eliminating u we may rewrite (6.4.22) somewhat more simply as

$$\max\{s'd | e'd = 1, \;\; d \in \Re^n_+\}, \tag{6.4.23}$$

where $s = Xe$. This is an extremely simple linear program which serves as a convenient geometric laboratory animal for studying various approaches to solving such problems. Simplex is particularly simple in this context, because the constraint set *is* literally a simplex. If we begin at a random vertex, and move around the polygon until optimality is achieved, we pass through $O(n)$ vertices in the process. Of course, a random initial vertex is rather naive, and one could do much better with an intelligent "Phase 1" approach that found a *good* initial vertex. In effect we can think of the "interior point" approach we will now describe as a class of methods to accomplish this, rendering unnecessary further travel around the outside of the polygon.

The log barrier formulation of Frisch is,

$$\max\{s'd + \mu \sum_{i=1}^{n} \log d_i | e'd = 1\} \tag{6.4.24}$$

where the barrier term $\mu \sum \log d_i$ serves as a penalty which keeps us away from the boundary of the positive orthant in the space of the dual vector d. By judicious choice of a sequence $\mu \to 0$ we might hope to converge to a solution of the original problem.

Restricting attention, for the moment, to the primal log-barrier formulation (6.4.24) and defining,

$$B(d, u) = s'd + \mu \sum \log d_i \tag{6.4.25}$$

we have $\nabla B = s + \mu D^{-1}e$ and $\nabla^2 B = -\mu D^{-2}$ where $D = \text{diag}(d)$. Thus, at any initial feasible, d, we have the associated Newton subproblem

$$\max_p\{(s + \mu D^{-1}e)'p - \frac{\mu}{2}p'D^{-2}p | e'p = 0\}.$$

This problem has first order conditions

$$\begin{aligned} s + \mu D^{-1}e - \mu D^{-2}p \;\; &= ae \\ e'p \;\; &= 0 \end{aligned}$$

and multiplying through by $e'D^2$, and using the constraint, we have,

$$e'D^2 s + \mu e'De = ae'D^2 e.$$

Thus solving for the Lagrange multiplier $\hat{a}$ we obtain the Newton direction

$$p = \mu^{-1}D^2s + De - \hat{a}e \tag{6.4.26}$$

where $\hat{a} = (e'D^2e)^{-1}(e'D^2s + \mu e'De)$. Pursuing the iteration $d \leftarrow d + \lambda p$, thus defined, *with μ fixed*, yields the central path $d(\mu)$ which describes a to the solution d^* of the original problem (6.4.22). We must be careful to keep the step lengths λ small enough to maintain the interior feasibility of d. Note that the initial feasible point $d = e/n$ represents $d(\infty)$.

As emphasized by Gonzaga (1992) and others, this central path is a crucial construct for the interior point approach. Algorithms may be usefully evaluated on the basis of how well they are able to follow this path. Clearly, there is some tradeoff between staying close to the path and moving along the path, thus trying to reduce μ, iteration by iteration. Improving upon existing techniques for balancing these objectives is the subject of a vast outpouring of current research. Excellent introductions to the subject are provided in the survey paper of Margaret Wright (1992) and the recent monograph of Stephen Wright (1996).

Thus far, we have considered only the primal version of our simple polygonal problem, but it is also advantageous to consider the primal and dual forms together. The dual of (6.4.22) is very simple:

$$\min\{a | ea - z = s, \quad z \geq 0\}. \tag{6.4.27}$$

The scalar, a, is the Lagrange multiplier on the equality constraint of the primal introduced above, while z is a vector of "residuals," or slack variables in the terminology of linear programming. This formulation of the dual exposes the real triviality of the problem – we are simply looking for the maximal element of the vector $s = Xe$. This is a very special case of the linear programming formulation of finding any ordinary quantile. But the latter would require us to split z into its positive and negative parts, and would also introduce upper bounds on the variables, d, in the primal problem.

Another way to express the central path, one that nicely illuminates the symmetric roles of the primal and dual formulations of the original problem, is to solve the equations,

$$\begin{aligned} e'd &= 1 \\ ea - z &= s \\ Dz &= \mu e. \end{aligned} \tag{6.4.28}$$

That solving these equations is equivalent to solving (6.4.24) may be immediately seen by writing the first order conditions for (6.4.24) as

$$\begin{aligned} e'd &= 1 \\ ea - \mu D^{-1}e &= s, \end{aligned}$$

and then appending the definition $z = \mu D^{-1}e$. The equivalence then follows from the negative definiteness of the Hessian $\nabla^2 B$. This formulation is also useful in highlighting a crucial interpretation of the log-barrier penalty parameter, μ. For any feasible pair (z, d) we have

$$s'd = a - z'd,$$

so $z'd$ is equal to the duality gap, the discrepancy between the primal and dual objective functions at the point (z, d). At a solution, we have the complementary slackness condition $z'd = 0$, thus implying a duality gap of zero. Multiplying through by e' in the last equation of (6.4.28), we may take $\mu = z'd/n$ as a direct measure of progress toward a solution.

Applying Newton's method to these equations yields

$$\begin{pmatrix} Z & 0 & D \\ e' & 0 & 0 \\ 0 & e & -I \end{pmatrix} \begin{pmatrix} p_d \\ p_a \\ p_z \end{pmatrix} = \begin{pmatrix} \mu e - Dz \\ 0 \\ 0 \end{pmatrix}, \tag{6.4.29}$$

where we have again presumed initial, feasible choices of d and z. Solving for p_a we have

$$\hat{p}_a = (e'Z^{-1}De)^{-1}e'Z^{-1}(Dz - \mu e)$$

which yields the primal-dual Newton direction:

$$p_d = Z^{-1}(\mu e - Dz - Dep_a) \tag{6.4.30}$$

$$p_z = ep_a. \tag{6.4.31}$$

It is of obvious interest to compare this primal-dual direction with the purely primal step derived above. In order to do so, however, we need to specify an adjustment mechanism for μ.

To this end we will now describe an approach suggested by Mehrotra(1992) that has been widely implemented by developers of interior point algorithms, including the interior point algorithm for quantile regression described in Portnoy and Koenker(1997). Given an initial feasible triple (d, a, z), consider the affine-scaling Newton direction obtained by evaluating the first equation of (6.4.29) at $\mu = 0$. Now compute the step lengths for the primal and dual variables respectively using

$$\lambda_d = argmax\{\lambda \in [0, 1] | d + \lambda p_d \geq 0\}$$

$$\lambda_z = argmax\{\lambda \in [0, 1] | z + \lambda p_z \geq 0\}.$$

But rather than precipitously taking this step, Mehrotra suggests adapting the direction somewhat to account for both the "recentering effect" introduced by the μe term in (6.4.29) and also for the nonlinearity introduced by the last of the first order conditions.

Consider first the recentering effect. If we contemplate taking a full step in the affine scaling direction we would have,

$$\hat{\mu} = (d + \lambda_d p_d)'(z + \lambda_z p_z)/n,$$

while at the current point we have,

$$\mu = d'z/n.$$

Now, if $\hat{\mu}$ is considerably smaller than μ, it means that the affine scaling direction has brought us considerably closer to the optimality condition of complementary slackness: $z'd = 0$. This suggests that the affine scaling direction is favorable, that we should reduce μ, in effect downplaying the contribution of the recentering term in the gradient. If, on the other hand, $\hat{\mu}$ isn't much different than μ, it suggests that the affine-scaling direction is unfavorable and that we should leave μ alone, taking a step which attempts to bring us back closer to the central path. Repeated Newton steps with μ fixed put us exactly on this path. These heuristics are embodied in Mehrotra's proposal to update μ by

$$\mu \leftarrow \mu(\hat{\mu}/\mu)^3.$$

To deal with the nonlinearity, Mehrotra (1992) proposed the following "predictor-corrector" approach. A full affine scaling step would entail

$$(d + p_d)'(z + p_z) = d'z + d'p_z + p_d'z + p_d'p_z.$$

The linearization implicit in the Newton step ignores the last term, in effect predicting that since it is of $\mathcal{O}(\mu^2)$ it can be neglected. But since we have already computed a preliminary direction, we might as well reintroduce this term to correct for the nonlinearity as well to accomplish the recentering. Thus, we compute the modified direction by solving

$$\begin{pmatrix} Z & 0 & D \\ e' & 0 & 0 \\ 0 & e & I \end{pmatrix} \begin{pmatrix} \delta_d \\ \delta_a \\ \delta_z \end{pmatrix} = \begin{pmatrix} \mu e - Dz - P_d p_z \\ 0 \\ 0 \end{pmatrix},$$

where $P_d = diag(p_d)$. This modified Newton direction is then subjected to the same step-length computation and a step is finally taken. It is important in more realistic problem settings that the linear algebra required to compute the solution to the modified step has already been done for the affine scaling step. This usually entails a Cholesky factorization of a matrix which happens to be scalar here, so the modified step can be computed by simply backsolving the same system of linear equations already factored to compute the affine scaling step.

In Figure 6.1 we provide an example intended to illustrate the advantage of the Mehrotra modified step. The solid line indicates the central path. Starting from the same initial point $d = e/n$, the dotted line represents the first affine scaling step. It is successful in the limited sense that it stays very close to the central path, but it only

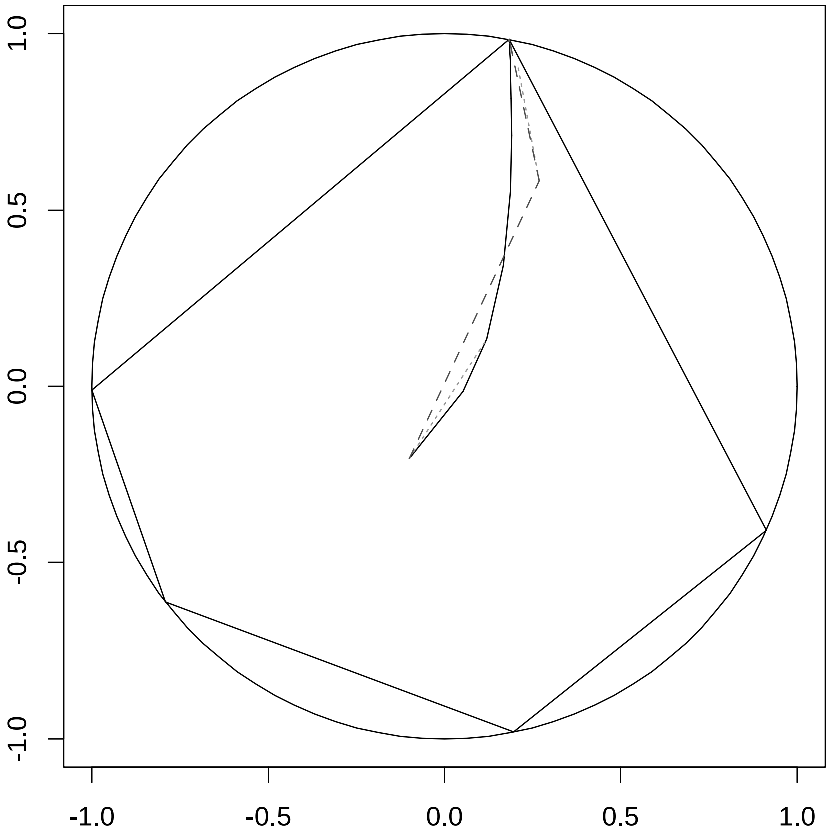

FIGURE 6.1. A Simple Example of Interior Point Methods for Linear Programming: The figure illustrates a random pentagon of which we would like to find the most northeast vertex. The central path beginning with an equal weighting of the 5 extreme points of the polygon is shown as the solid curved line. The dotted line emanating from the this center is the first affine scaling step. The dashed line is the modified Newton direction computed according to the proposal of Mehrotra. Subsequent iterations are unfortunately obscured by the scale of the figure.

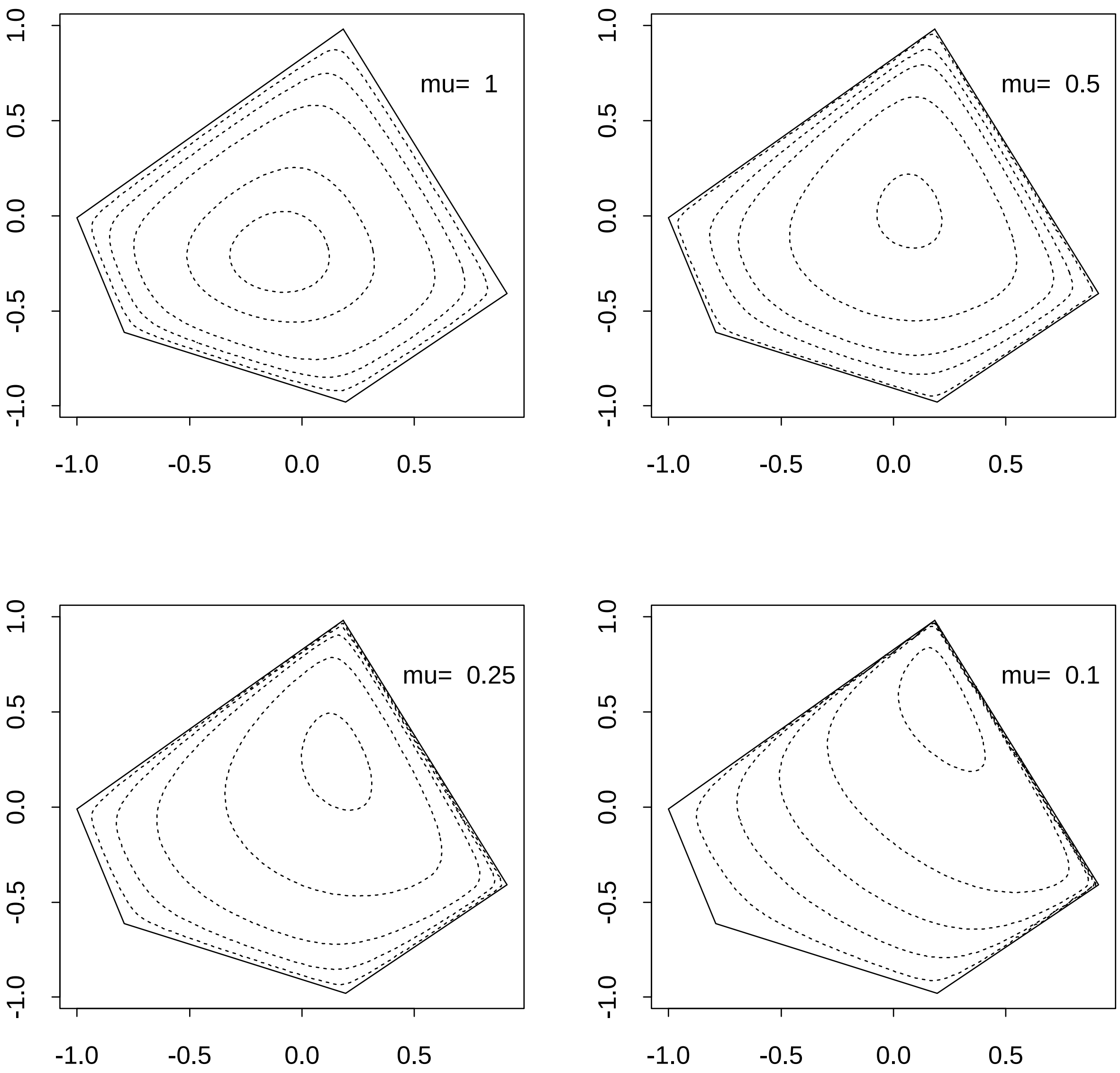

FIGURE 6.2. Contours of the Log Barrier Objective Function for the Simple Polygonal Linear Program: The figure illustrates four different contour plots of the log barrier objective function (6.4.24) corresponding to four different choices of μ. In the first panel, $\mu = 1$ and the contours are centered in the polygon. As μ is reduced the penalized objective function is less influenced by the penalty term and more strongly influenced by the linear component of the original LP formulation of the problem. Thus, for $\mu = .1$ we find that the unconstrained maximum of the log barrier function occurs quite close to the optimal vertex of the original LP. The locus of solutions to the log barrier problems for various μ's is called the central path, and is illustrated in 6.1 by the solid curved line.

takes a short step toward our final destination. In contrast, the first modified step, indicated by the dashed line, takes us much further. By anticipating the curvature of the central path, it takes a step more than twice the length of the unmodified, affine-scaling step. On the second step the initial affine-scaling step is almost on target, but again somewhat short of the mark. The modified step is more accurately pointed at the desired vertex and is thus, again, able to take a longer step.

In Figure 6.2 we try to illustrate the log barrier approach by plotting 4 versions of the contours corresponding to the penalized objective function for four distinct values of the penalty parameter μ. In the first panel, with $\mu = 1$ we are strongly repelled from the boundary of the constraint set and the unconstrained maximum of the barrier function occurs near the center of the polygon. In the next panel, with μ reduced to $\frac{1}{2}$ the barrier penalty exerts a somewhat weaker effect and the contours indicate that the unconstrained maximum occurs somewhat closer to the upper vertex of the polygon. This effect is further accentuated in the $\mu = \frac{1}{4}$ figure, and in the last figure with $\mu = \frac{1}{10}$ we find that the maximum occurs quite close to the vertex. The path connecting the maximum of the family of fixed-μ problems is generally called the central path. As emphasized by Gonzaga (1992) and others, this central path is a crucial construct for the interior point approach. Competing algorithms may be usefully evaluated on the basis of how well they are able to follow this path. Clearly, there is some tradeoff between staying close to the path and moving along the path, thus trying to reduce μ, iteration by iteration. Improving upon existing techniques for balancing these objectives is the subject of a vast outpouring of current research. Excellent introductions to the subject are provided in the survey paper of Margaret Wright (1992) and the recent monograph of Stephen Wright (1996).

It is difficult in a single example like this to convey a sense of the overall performance of these methods. After viewing a large number of realizations of these examples, one comes away convinced that the Mehrotra modified step consistently improves upon the affine scaling step, a finding that is completely consistent with the theory.

5. Interior Point Methods for Quantile Regression

We have seen that the problem of solving

$$\min_{b \in \Re^p} \sum_{i=1}^{n} \rho_\tau(y_i - x_i' b) \tag{6.5.32}$$

where $\rho_\tau(r) = r(\tau - I(r < 0))$ for $\tau \in (0,1)$. may be formulated as the linear program,

$$\min\{\tau e'u + (1-\tau)e'v \mid y = Xb + u - v, (u,v) \in \Re_+^{2n}\}, \tag{6.5.33}$$

and has dual formulation,

$$\max\{y'd \mid X'd = 0, \ \ d \in [\tau - 1, \tau]^n\} \tag{6.5.34}$$

or, setting $a = d + 1 - \tau$,

$$\max\{y'a \mid X'a = (1-\tau)X'e, \;\; a \in [0,1]^n\}. \tag{6.5.35}$$

The dual formulation of the quantile regression problem fits nicely into the standard formulations of interior point methods for linear programs with bounded variables. The function $a(\tau)$ that maps $[0,1]$ to $[0,1]^n$ plays a crucial role in connecting the statistical theory of quantile regression to the classical theory of rank tests as described in Gutenbrunner and Jurečková (1992) and Gutenbrunner, Jurečková , Koenker and Portnoy (1993). See Koenker and d'Orey (1987,1993) for a detailed description of modifications of the Barrodale and Roberts(1974) simplex algorithm for this problem.

Adding slack variables s, and the constraint $a + s = e$, we obtain the barrier function

$$B(a, s, \mu) = y'a + \mu \sum_{i=1}^{n} (\log a_i + \log s_i), \tag{6.5.36}$$

which should be maximized subject to the constraints, $X'a = (1-\tau)X'e$ and $a+s = e$. The Newton step, δ_a, solving

$$\max y'\delta_a + \mu\delta_a'(A^{-1} - S^{-1})e - \tfrac{1}{2}\mu\delta_a'(A^{-2} + S^{-2})\delta_a \tag{6.5.37}$$

subject to $X'\delta_a = 0$, satisfies

$$y + \mu(A^{-1} - S^{-1})e - \mu(A^{-2} + S^{-2})\delta_a = Xb \tag{6.5.38}$$

for some $b \in \Re^p$, and δ_a such that $X'\delta_a = 0$. As before, multiplying through by $X'(A^{-2} + S^{-2})^{-1}$ and using the constraint, we can solve explicitly for the vector b,

$$b = (X'WX)^{-1}X'W(y + \mu(A^{-1} - S^{-1})e) \tag{6.5.39}$$

where $W = (A^{-2}+S^{-2})^{-1}$. This is a form of the primal log barrier algorithm described above. Setting $\mu = 0$ in each step yields an affine scaling variant of the algorithm. We should stress again that the basic linear algebra of each iteration is essentially unchanged, only the form of the diagonal weighting matrix W has changed. We should also emphasize that there is nothing especially sacred about the explicit form of the barrier function used in (6.5.36). Indeed, one of the earliest proposed modifications of Karmarkar's original work was the affine scaling algorithm of Vanderbei, Meketon and Freedman (1986), which used, implicitly, $\mu \sum_{i=1}^{n} \log(\min(a_i, s_i))$ in lieu of the additive specification.

Again, it is natural to ask if a primal-dual form of the algorithm could improve performance. In the bounded variables formulation we have the Lagrangian,

$$L(a, s, b, u, \mu) = B(a, s, \mu) - b'(X'a - (1-\tau)X'e) - u'(a + s - e), \tag{6.5.40}$$

and setting $v = \mu A^{-1}$ we have the first order conditions, describing the central path, see Gonzaga(1992),

$$\begin{aligned} X'a &= (1-\tau)X'e \\ a + s &= e \\ Xb + u - v &= y \\ USe &= \mu e \\ AVe &= \mu e, \end{aligned} \tag{6.5.41}$$

yielding the Newton step,

$$\begin{aligned} \delta_b &= (X'WX)^{-1}((1-\tau)X'e - X'a + X'W\xi(\mu)) \\ \delta_a &= W(X\delta_b + \xi(\mu)) \\ \delta_s &= -\delta_a \\ \delta_u &= \mu A^{-1}e - Ue - A^{-1}U\delta_a \\ \delta_v &= \mu S^{-1}e - Ve + S^{-1}V\delta_s \end{aligned} \tag{6.5.42}$$

where $\xi(\mu) = y - Xb + \mu(S^{-1} - A^{-1})e$. The most successful implementations of this approach to date employ the predictor-corrector step of Mehrotra (1992) which is described in the context of bounded variables problems in Lustig, Marsden and Shanno (1992). A related earlier approach is described in Zhang(1992). In Mehrotra's approach we proceed somewhat differently. Rather than solving for the Newton step (6.5.42) directly, we substitute the step directly into (6.5.41), to obtain,

$$\begin{aligned} X'(a+\delta_a) &= (1-\tau)X'e \\ (a+\delta_a) + (s+\delta_s) &= e \\ X(b+\delta_b) + (u+\delta_u) - (v+\delta_v) &= y \\ (U+\Delta_u)(S+\Delta_s) &= \mu e \\ (A+\Delta_a)(V+\Delta_v) &= \mu e, \end{aligned} \tag{6.5.43}$$

where $\Delta_a, \Delta_v, \Delta_u, \Delta_s$ denote the diagonal matrices with diagonals, $\delta_a, \delta_v, \delta_u, \delta_s$ respectively. As noted by Lustig, Marsden and Shanno, the primary difference between solving this system and the prior Newton step is the presence of the nonlinear terms $\Delta_u\Delta_s$, $\Delta_a\Delta_v$ in the last two equations. To approximate a solution to these equations, Mehrotra (1992) suggests first solving for an affine primal-dual direction by setting $\mu = 0$ in (6.5.42). Given this preliminary direction, we may then compute the step length using the following ratio test,

$$\hat{\gamma}_P = \sigma \min\{\min_j\{-a_j/\delta_{a_j}, \delta_{a_j}\}, \min_j\{-s_j/\delta_{s_j}, \delta_{s_j}\}\} \tag{6.5.44}$$

$$\hat{\gamma}_D = \sigma \min\{\min_j\{-u_j/\delta_{u_j}, \delta_{u_j}\}, \min_j\{-v_j/\delta_{v_j}, \delta_{v_j}\}\}. \tag{6.5.45}$$

using scaling factor $\sigma = .99995$, as in Lustig, Marsden, and Shanno. Then defining the function,

$$\hat{g}(\hat{\gamma}_P, \hat{\gamma}_D) = (s + \hat{\gamma}_P \delta_s)'(u + \hat{\gamma}_D \delta_u) + (a + \hat{\gamma}_P \delta_a)'(v + \hat{\gamma}_D \delta_v) \tag{6.5.46}$$

the new μ is taken as,

$$\mu = \left(\frac{\hat{g}(\hat{\gamma}_P, \hat{\gamma}_D)}{\hat{g}(0,0)} \right)^3 \frac{\hat{g}(0,0)}{2n}. \tag{6.5.47}$$

To interpret (6.5.46) we may use the first three equations of (6.5.41) to write, for any primal-dual feasible point (u, v, s, a)

$$\tau e'u + (1 - \tau)e'v - (a - (1 - \tau)e)'y = u's + a'v. \tag{6.5.48}$$

So the quantity $u's + a'v$ is equal to the duality gap, the difference between the primal and dual objective function values at (u, v, s, a), and $\hat{g}(\hat{\gamma}_P, \hat{\gamma}_D)$ is the duality gap after the tentative affine scaling step. Note that the quantity $a - (1 - \tau)e$ is simply the vector d appearing in the dual formulation (6.5.34). At a solution, classical duality theory implies that the duality gap vanishes, that is the values of the primal and dual objective functions are equal and the complementary slackness condition, $u's + a'v = 0$ holds. If, in addition to feasibility, (u, v, s, a) happened to lie on the central path, the last two equations of (6.5.41) would imply that,

$$u's + a'v = 2\mu n.$$

Thus, the function $\hat{g}$ in (6.5.46) may be seen as an attempt to adapt μ to the current iterate in such a way that for any given value of the duality gap, μ is chosen to correspond to the point on the central path with that gap. By definition, $\hat{g}(\hat{\gamma}_P, \hat{\gamma}_D)/\hat{g}(0,0)$ is the ratio of the duality gap after the tentative affine-scaling step to the gap at the current iterate. If this ratio is small the proposed step is favorable and we should reduce μ further, anticipating that the recentering and nonlinearity adjustment of the modified step will yield further progress. If, on the other hand, $\hat{g}(\hat{\gamma}_P, \hat{\gamma}_D)$ isn't much different from $\hat{g}(0,0)$, the affine scaling direction is unfavorable, and further reduction in μ is ill-advised. Since leaving μ fixed in the iteration brings us back to the central path, such unfavorable steps are intended to enable better progress in subsequent steps iy bringing the current iterate back to the vicinity of the central path. The rationale for the cubic adjustment in (6.5.47) which implements these heuristics, is based on the fact that the recentering of the Newton direction embodied in the terms $\mu A^{-1}e$ and $\mu S^{-1}e$ of (6.5.42) and (6.5.49) accomodates the $\mathcal{O}(\mu)$ term in the expansion of the duality gap function $\hat{g}$ while the nonlinearity adjustment described below accomodates the $\mathcal{O}(\mu^2)$ effect of the $\delta_s \delta_u$ and $\delta_a \delta_v$ terms.

We compute the following approximation to the solution of the system (6.5.43) with this μ and the nonlinear terms $\Delta_s \Delta_u$ and $\Delta_a \Delta_v$ taken from the preliminary primal-dual affine direction:

$$
\begin{aligned}
\delta_b &= (X'WX)^{-1}((1-\tau)X'e - X'a + X'W\xi(\mu))\\
\delta_a &= W(X\delta_b + \xi(\mu))\\
\delta_s &= -\delta_a \qquad\qquad (6.5.49)\\
\delta_u &= \mu A^{-1}e - Ue - A^{-1}U\delta_a + A^{-1}\Delta_s\Delta_u e\\
\delta_v &= \mu S^{-1}e - Ve + S^{-1}V\delta_s + S^{-1}\Delta_a\Delta_v e.
\end{aligned}
$$

The iteration proceeds until the algorithm terminates when the duality gap $y'a-(1-\tau)e'Xb+e'v$ becomes smaller than a specified ϵ. Recall that the duality gap is zero at a solution, and thus, this criterion offers a more direct indication of convergence than is usually available in iterative algorithms.

6. Interior vs. Exterior: Some Computational Experience

Our expectations about satisfactory computational speed of regression estimators are inevitably strongly conditioned by our experience with least squares. In Figure 4.1 we illustrate the results of a small experiment to compare the computational speed of 3 ℓ_1 algorithms: the Barrodale and Roberts (1973) simplex algorithm which is employed in many contemporary statistical packages, Meketon's affine scaling algorithm, and our implementation of Mehrotra's (1992) predictor-corrector version of the primal-dual log barrier algorithm. The former is indicated in the figure as `mek` and the latter as `rqfn` for regression quantiles *via* Frisch-Newton. The two interior point algorithms were coded in Fortran employing Lapack, (Anderson, *et al* (1995)), subroutines for the requisite linear algebra. They were then incorporated as functions into Splus and timings are based on the Splus function `unix-time()`. The Barrodale and Roberts timings are based on the Splus implementation `l1fit(x,y)`. For comparison purposes we also illustrate timings for least squares estimation based on Splus function `lm(y ~ x)`.

Such comparisons are inevitably fraught with qualifications about programming style, system overhead, *etc.*. We have chosen to address the comparison within the Splus environment because a.) it is the computing environment in which we feel most comfortable, a view widely shared by the statistical research community, and b.) it offers a convenient means of incorporating new functions in lower level languages, like Fortran and C, providing a reasonably transparent and efficient interface with the rest of the language. We have considerable experience with the Barrodale and Roberts (1974) Fortran code as implemented in Splus for `l1fit`. This code also underlies the quantile regression routines described in Koenker and d'Orey (1987, 1993) and still represents the state-of-the-art after more than 20 years. The Splus function `l1fit` incurs a modest overhead getting problems into and out of BR's Fortran, but this overhead is quickly dwarfed by the time spent in the Fortran in large problems. Similarly, we have tried to write the interior point code to minimize the Splus overhead, although some improvements are still possible in this respect.

Least-squares timings are also potentially controversial. The Splus function `lm` as described by Chambers (1992) offers three method options: QR decomposition, Cholesky, and singular value decomposition. All of our comparisons are based on the default choice of the QR method. Again there is a modest overhead involved in getting the problem descriptions into and the solutions out of the lower level Lapack routines which underlie `lm`. We have run some very limited timing comparisons outside Splus directly in Fortran to evaluate these overhead effects and our conclusion from this is that any distortions in relative performance due to overhead effects are slight.

We would stress that the code underlying the least squares computations we report is the product of decades of refinement, while our interior point routines are still in their infancy. There is still considerable scope for improvement in the latter.

Several features of the figures are immediately striking. For small problems all the ℓ_1 algorithms perform impressively. They are all faster than the QR implementation of least squares which is generally employed in `lm`. For small problems the simplex implementation of Barrodale and Roberts is the clear winner, but its roughly quadratic (in sample size) growth over the illustrated range quickly dissipates its initial advantage. The interior point algorithms do considerably better than simplex at larger sample sizes, exhibiting roughly linear growth, as does least-squares. Meketon's affine scaling algorithm performs slightly better than the primal- dual algorithm, which is somewhat surprising, but for larger p the difference is hardly noticeable.

Beyond the range of problem sizes illustrated here, the advantage of the interior point method over simplex grows exorbitant, fully justifying the initial enthusiasm with which Karmarkar (1984) was received. Nevertheless, there is still a significant gap between ℓ_1 and ℓ_2 performance in large samples. We explore this gap from the probabilistic viewpoint of computational complexity in the next section.

7. Computational Complexity

In this section we investigate the computational complexity of the interior point algorithms for quantile regression described above. We should stress at the outset, however, that the probabilistic approach to complexity analysis adopted here is rather different than that employed in the rest of the interior-point literature where the focus on worst case analysis has led to striking discrepancies between theoretical rates and observed computational experience. The probabilistic approach has the virtue that the derived rates are much sharper and consequently more consonant with observed performance. A similar gap between worst-case theory and average practice can be seen in the analysis of parametric linear programming via the simplex algorithm, where it is known that in certain problems with an n by p constraint matrix there can be as many as n^p distinct solutions. However, exploiting some special aspects of the

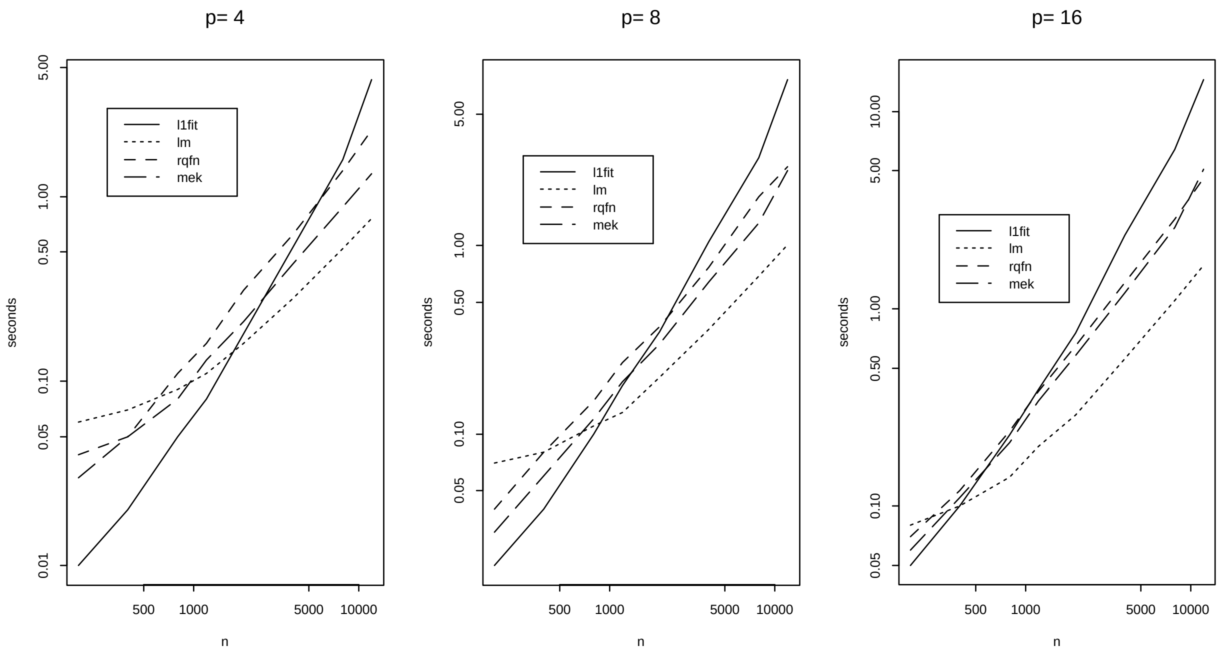

FIGURE 6.3. Timing comparison of three ℓ_1-algorithms for median regression: Times are in seconds for the median of five replications for iid Gaussian data. The parametric dimension of the models is $p+1$ with p indicated above each plot, p columns are generated randomly and an intercept parameter is appended to the resulting design. Timings were made at 8 design points in n: 200, 400, 800, 1200, 2000, 4000, 8000, 12000. The solid line represents the results for the simplex-based Barrodale and Roberts algorithm implemented in Splus as `l1fit`, the `rqfn` dashed line represents a primal-dual interior point algorithm, `mek` uses an affine scaling form of the interior point approach and the dotted line represents least squares timings based on `lm(y ~ x)` as a benchmark

quantile regression problem and employing a probabilistic approach, Portnoy (1991) was able to show that the number of distinct vertex solutions (in τ) is $\mathcal{O}_p(n \log n)$, a rate which provides excellent agreement with empirical experience.

For interior point methods the crux of the complexity argument rests on showing that at each iteration the algorithm reduces the duality gap by a proportion, say $\theta_n < 1$. Thus after K iterations, an initial duality gap of Δ_0 has been reduced to $\theta_n^K \Delta_0$. Once the gap is sufficiently small, say, less than ε, there is only one vertex of the constraint set at which the duality gap can be smaller. This follows obviously from the fact that the vertices are discrete. Thus, the vertex with the smaller duality gap must be the optimal one, and this vertex may be identified by taking p simplex-type steps. This process, called purification in Gonzaga (1992, Lemma 4.7), requires in our notation p steps involving $\mathcal{O}(np^2)$ operations, or $\mathcal{O}(np^3)$ operations. Hence, the number of iterations, K, required to make, $\theta_n^K \Delta_0 < \epsilon$ is,

$$K < \log(\Delta_0/\epsilon)/(-\log\theta_n).$$

In the worst-case analysis of the interior point literature, ε is taken to be 2^{-L} where L is the total number of binary bits required to encode the entire data of the problem. Thus, in our notation ε would be $\mathcal{O}(np)$. Further, the conventional worst-case analysis employs the bound $\theta_n < (1 - cn^{-1/2})$, and takes Δ_0 independent of n so the number of required iterations is $\mathcal{O}(\sqrt{n}L)$. Since each iteration requires a weighted least-squares solutions of $\mathcal{O}(np^2)$ operations, the complexity of the algorithm as a whole would be $\mathcal{O}(n^{5/2}p^3)$, apparently hopelessly disadvantageous relative to least squares. Fortunately, however, in the random problems for which quantile regression methods are designed, the ε bound on the duality gap at the second best vertex can be shown to be considerably larger, at least with probability tending to 1, than this worst case value of 2^{-L}. Lemma A.1 of the appendix to Portnoy and Koenker (1997) provides the bound $\log\epsilon = \mathcal{O}_p(p\log n)$ under mild conditions on the underlying regression model. This leads to a considerably more optimistic view of these methods for large problems.

Renegar (1988) and numerous subsequent authors have established the existence of a large class of interior point algorithms for solving linear programs which, starting from an initially feasible primal-dual point with duality gap Δ_0, can achieve convergence to a prescribed accuracy ϵ in $\mathcal{O}(\sqrt{n}\log(\Delta_0/\epsilon))$ iterations *in the worst case.* More recently, Sonnevend, Stoer, and Zhao (1991) have shown under somewhat stronger nondegeneracy conditions that this rate can be improved to $\mathcal{O}(n^a\log(\Delta_0/\epsilon))$ with $a < 1/2$. We will call an algorithm which achieves this rate an n^a-algorithm. They give explicit conditions, which hold with probablity one if the the y's have a continuous density, for the case $a = 1/4$. The following result then follows immediately from previously cited Lemma.

THEOREM 6.5. *In the linear model $Y_i = x_i'\beta + u_i \quad i = 1, \ldots, n$, assume:*
(i) $\{(x_i, Y_i), i = 1, \ldots, n\}$ are iid with a bounded continuous density in $\Re^{p+1}$.
(ii) $E|x_{ij}|^p < \infty$ and $E|Y_i|^a < \infty$, for some $a > 0$.
An n^a-algorithm for median regression converges in $\mathcal{O}_p(n^a p\log n)$ iterations. And with $\mathcal{O}(np^2)$ operations required per iteration and $\mathcal{O}(np^3)$ operations required for the final "purification" process such an algorithm has complexity, $\mathcal{O}_p(n^{1+a}p^3\log n)$.

Mizuno, Todd and Ye (1993) provide an alternative probabilistic approach to the existence of an n^a-algorithm, with $a < 1/2$ and provide a heuristic argument for $a = 1/4$. They also conjecture that n^a might be improvable to $\log n$, by a more refined probabilistic approach. This would improve the overall complexity in the above Theorem to $\mathcal{O}_p(np^3 \log^2 n)$ and seems quite plausible in light of the empirical evidence reported below, and elsewhere in the interior point literature. In either case we are still faced with a theoretical gap between ℓ_1 and ℓ_2 performance that substantiates the empirical experience reported in the previous section. We now introduce a new form of preprocessing for ℓ_1 problems that has been successful in further narrowing this gap.

8. Preprocessing for Quantile Regression

Many modern linear programming algorithms include an initial phase of preprocessing which seeks to reduce problem dimensions by identifying redundant variables and dominated constraints. See, for example, the discussion in Section 8.2 of Lustig, Marsden, and Shanno(1993) and the remarks of the discussants. Bixby, in this discussion, reports reductions of 20-30% in the row and column dimensions of a sample of standard commercial test problems due to "aggressive implementation" of preprocessing. Standard preprocessing strategies for LP's are not, however, particularly well-suited to the statistical applications which underlie quantile regression. In this section we describe some new preprocessing ideas designed explicitly for quantile regression, which can be used to reduce dramatically the effective sample sizes for these problems.

The basic idea underlying our preprocessing step rests on the following elementary observation. Consider the directional derivative of the median regression, ℓ_1, problem

$$\min_b \sum_{i=1}^n |y_i - x_i' b|$$

which may be written in direction w as

$$g(b, w) = \sum_{i=1}^n x_i' w \ \mathrm{sgn}^*(y_i - x_i' b, x_i' w),$$

where

$$\mathrm{sgn}^*(u, v) = \begin{cases} \mathrm{sgn}(u) & \text{if } u \neq 0, \\ \mathrm{sgn}(v) & \text{if } u = 0. \end{cases}$$

Optimality may be characterized as a b^* such that $g(b^*, w) \geq 0$ for all $w \in \Re^p$. Suppose for the moment that we "knew" that a certain subset J_H of the observations $N = \{1, \ldots n\}$ would fall above the optimal median plane and another subset J_L would fall below. Then consider the revised problem

$$\min_{b \in \Re^p} \sum_{i \in N \setminus J_L \cup J_H} |y_i - x_i' b| + |y_L - x_L' b| + |y_H - x_H' b|$$

where $x_K = \sum_{i \in J_K} x_i$, for $K \in \{H, L\}$ and y_L, y_H can be chosen as arbitrarily small and large enough, respectively, to ensure that the corresponding residuals remain negative and positive. We will refer in what follows to these combined pseudo-observations as "globs". The new problem, under our provisional hypothesis, has exactly the same gradient condition as the original one, and therefore the same solutions, but the revision has reduced effective sample size by $\#\{J_L, J_H\} - 2$, i.e. by the number of observations in the globs.

How might we know J_L, J_H? Consider computing a preliminary estimate $\hat{\beta}$ based on a subsample of m observations. Compute a simultaneous confidence band for $x_i'\beta$ based on this estimate for each $i \in N$. Under plausible sampling assumptions we will see that the length of each interval is proportional to $p/\sqrt{m}$, so if M denotes the number of y_i falling inside the band, $M = \mathcal{O}_p(np/\sqrt{m})$. Take J_L, J_H to be composed of the indices of the observations falling outside the band. So we may now create the "globbed" observations $(y_K, x_K), K \in \{L, H\}$ and reestimate based on $M + 2$ observations. Finally, we must check to verify that all the observations in J_H, J_L have the anticipated residual signs; if so, we are done, if not, we must repeat the process. If the coverage probability of the bands is P, presumably near 1, then the expected number of repetitions of this process is the expectation of a geometric random variable, Z, with expectation P^{-1}. We will call each repetition a cycle.

8.1. Implementation. In this subsection we will sketch some further details of the preprocessing strategy. We should emphasize that there are many aspects of the approach that deserve further research and refinement. In an effort to encourage others to contribute to this process we have made all of the code described below available at the website `http://www.econ.uiuc.edu/research/rqn/rqn.html`. We will refer in what follows to the Frisch-Newton quantile regression algorithm *with preprocessing* as `prqfn`.

The basic structure of the current `prqfn` algorithm looks like this:

```
k ← 0
l ← 0
m ← [2n^{2/3}]
while(k is small){
    k = k + 1
    solve for initial rq using first m observations
    compute confidence interval for this solution
    reorder globbed sample as first M observations
    while(l is small){
        l = l + 1
        solve for new rq using the globbed sample
        check residual signs of globbed observations
        if no bad signs: return optimal solution
```

```
        if only few bad: adjust globs, reorder sample, update M, continue
        if too many bad: increase m and break to outer loop
        }
    }
```

The algorithm presumes that the data has undergone some initial randomization so the first m observations may be considered representative of the sample as a whole. In all of the experiments reported below we use the Mehrotra-Lustig-Marsden-Shanno primal-dual algorithm to compute the subsample solutions. For some "intermediately large" problems it would be preferable to use the simplex approach, but we postpone this refinement. Although the affine scaling algorithm of Meketon(1986) exhibited excellent performance on certain subsets of our early test problems, like those represented in Figure 4.1, we found its performance inconsistent in other tests. It was consequently abandoned in favor of the more reliable primal-dual formulation. This choice is quite consistent with the general development of the broader literature on interior point methods for linear programming, but probably also deserves further exploration.

8.2. Confidence Bands. The confidence bands used in our reported computational experiments are of the standard Scheffé type. Under iid error assumptions the covariance matrix of the initial solution is given by

$$V = \omega^2 (X'X)^{-1}$$

where $\omega^2 = \tau(1-\tau)/f^2(F^{-1}(\tau))$; the reciprocal of the error density at the τth quantile is estimated using the Hall-Sheather (1986) bandwidth for Siddiqui's(1960) estimator. Quantiles of the residuals from the initial fit are computed using the Floyd-Rivest(1975) algorithm. We then pass through the entire sample computing the intervals,

$$B_i = (x_i'\hat{\beta} - \zeta||\hat{V}^{1/2}x_i||, x_i'\hat{\beta} + \zeta||\hat{V}^{1/2}x_i||).$$

The parameter ζ is currently set naively, at 2, but could, more generally, be set as $\zeta = (\Phi^{-1}(1-\alpha) + \sqrt{2p-1})/\sqrt{2} = \mathcal{O}(\sqrt{p})$ to achieve $(1-\alpha)$ coverage for the band, and thus assures that the number of cycles is geometric. Since, under the moment condition of the previous Theorem, if $p \to \infty$, the quantity $||\hat{V}^{1/2}x_i||$ also behaves like the square-root of a χ^2 random variable, the width of the confidence band is of $\mathcal{O}_p(p/\sqrt{m})$.

Unfortunately, using the Scheffé bands requires $\mathcal{O}(np^2)$ operations, a computation of the same order as that required by least-squares estimation of the model. It seems reasonable, therefore, to consider alternatives. One possibility, suggested by the Studentized range, is to base intervals on the inequality,

$$|x_i'\hat{\beta}| \leq \max_j \left\{ \left|\hat{\beta}_j\right| / s_j \right\} \times \sum_{j=1}^{p} |x_{ij}| \, s_j \,, \tag{6.8.50}$$

where s_j is $\hat{\omega}$ times the jth diagonal element of the $(X'X)^{-1}$ matrix, and $\hat{\omega}$ is computed as for the Scheffé intervals. This approach provides conservative (though not "exact") confidence bands with width $c_q \sum_{j=1}^{p} |x_j| \, s_j$. Note that this requires only $\mathcal{O}(np)$ operations, thus providing an improved rate. Choice of the constant, c_q, is somewhat problematic, but some experimentation with simulated data showed that c_q could be taken conservatively to be approximately one, and that the algorithm was remarkably independent of the precise value of c_q. For these bands the width is again $\mathcal{O}_p(p/\sqrt{m})$, as for the Scheffé bands. Although these $\mathcal{O}(np)$ confidence bands worked well in simulation experiments, and thus merit further study, the computational experience reported here is based entirely on the more traditional Scheffé bands.

After creating the globbed sample, we again solve the quantile regression problem, this time with the M observations of the globbed sample. Finally we check the signs of the globbed observations. If they all agree with the signs predicted by the confidence band we may declare victory and return the optimal solution. If there are only a few incorrect signs we have found it expedient to adjust the globs, reintroduce these observations into the new globbed sample and resolve. If there are too many incorrect signs, we return to the initial phase, increasing the initial sample size somewhat, and repeat the process. One or two repetitions of the inner (fixup) loop are not unusual; more than two cycles of the outer loop is highly unusual given current settings of the confidence band parameters.

8.3. Choosing m. The choice of the initial subsample size, m, and its implications for the complexity of an interior point algorithm for quantile regression *with preprocessing* is resolved by the next lemma.

THEOREM 6.6. *Under the conditions of Theorem 6.5, for any nonrecursive quantile regression algorithm with complexity, $\mathcal{O}_p(n^\alpha p^\beta \log n)$, for problems with dimension (n, p), there exists a confidence band construction based on an initial subsample of size m with expected width, $\mathcal{O}_p(p/\sqrt{m})$, and consequently, the optimal initial subsample size is $m^* = \mathcal{O}((np)^{2/3})$. With this choice of m^*, M is also $\mathcal{O}((np)^{2/3})$. Then, with $\alpha = 1 + a$, and $\beta = 3$, from Theorem 6.5, the overall complexity of the algorithm with preprocessing is, for any n^a underlying interior point algorithm,*

$$\mathcal{O}_p((np)^{2(1+a)/3} p^3 \log n) + \mathcal{O}_p(np).$$

For $a < 1/2$, n sufficiently large, and p fixed, this complexity is dominated by the complexity of the confidence band computation, and is strictly smaller than the $\mathcal{O}(np^2)$ complexity of least-squares.

Proof: Formally, we treat only the case of p fixed, but we have tried to indicate the role of p in the determination of the constants, where possible. Thus, for example, for $p \to \infty$, we have suggested above that the width of both the Scheffé bands and the Studentized range bands are $\mathcal{O}_p(p/\sqrt{m})$. For p fixed this condition is trivially satisfied. By independence we may conclude that the number of observations inside such a confidence band will be,

$$M = \mathcal{O}_p(np/\sqrt{m}),$$

and minimizing, for any constant c,

$$m^\alpha p^\beta \log m + (cnp/\sqrt{m})^\alpha p^\beta \log(cnp/\sqrt{m}) \tag{6.8.51}$$

yields,

$$m^* = \mathcal{O}((np)^{2/3}).$$

Substituting this m^* back into (6.8.51), Theorem 6.5 implies that we have complexity,

$$\mathcal{O}((np)^{2(1+a)/3} p^3 \log n),$$

for each cycle of the preprocessing. The number of cycles required is bounded in probability since it is a realization of a geometrically distributed random variable with a finite expectation. The complexity computation for the algorithm as a whole is completed by observing that the required residual checking is $\mathcal{O}(np)$ for each cycle, and employing the Studentized range confidence bands also requires $\mathcal{O}(np)$ operations per cycle. Thus the contribution of the confidence band construction and residual checking is precisely $\mathcal{O}_p(np)$, and for any $a < 1/2$ the complexity of the ℓ_1 algorithm is therefore dominated by this term for any fixed p and n sufficiently large. ∎

Remarks. (1.) Clearly these results above apply not only to median regression, but to quantile regression in general. (2.) If the explicit rates in p of the Theorem hold for $p \to \infty$, and if the Mizuno-Todd-Ye conjecture that n^a can be improved to $\log n$ holds, then the complexity of the algorithm becomes,

$$\mathcal{O}(n^{2/3} p^3 \log^2 n) + \mathcal{O}_p(np).$$

The contribution of the first term in this expression would then assure an improvement over least squares for n sufficiently large, provided $p = o(n^{1/5})$, a rate approaching the domain of nonparametric regression applications. (3.) It is tempting to consider the recursive application of the preprocessing approach described above, and this can be effective in reducing the complexity of the solution of the initial subsample m problem, but it does not appear possible to make it effective in dealing with the globbed sample. This accounts for the qualifier "nonrecursive" in the statement of the theorem.

9. More Computational Experience

In this section we provide some further evidence on the performance of our implementation of the algorithm on both simulated and real data. In Figure 6.4 we compare the performance of `l1fit` with the new `prqfn`, which combines the primal-dual algorithm with preprocessing. With the range of sample sizes 20,000 - 120,000, the clear superiority of `prqfn` is very striking. At $n = 20,000$ `prqfn` is faster than `l1fit` by a factor of about 10, and it is faster by a factor of 100 at $n = 120,000$. The quadratic growth in the `l1fit` timings is also quite apparent in this figure.

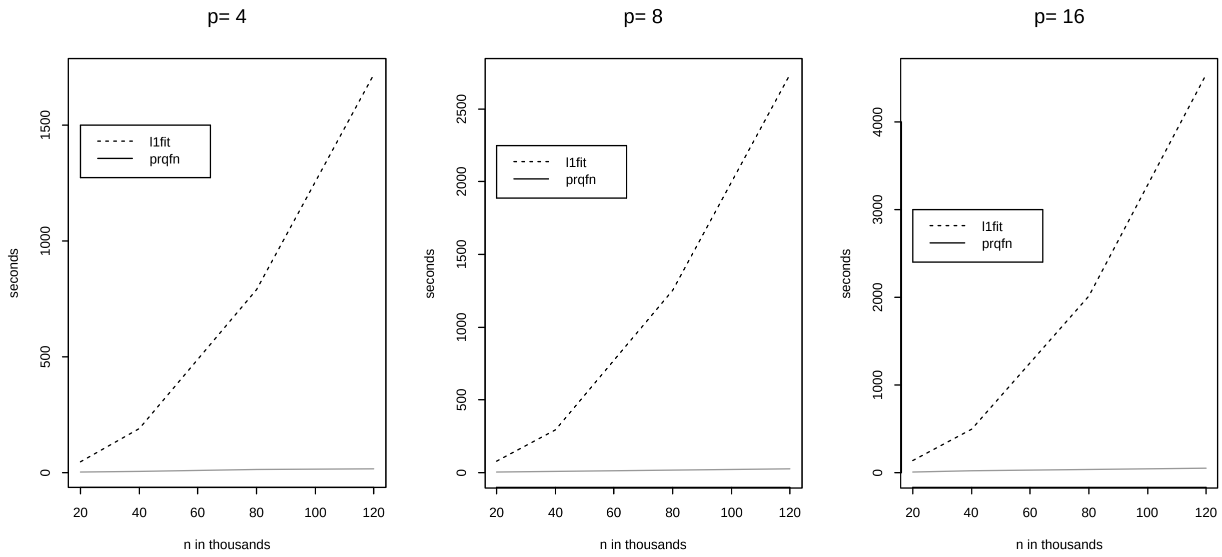

FIGURE 6.4. Timing comparison of two ℓ_1-algorithms for median regression: Times are in seconds for the mean of five replications for iid Gaussian data. The parametric dimension of the models is $p + 1$ with p indicated above each plot, p columns are generated randomly and an intercept parameter is appended to the resulting design. Timings were made at 4 design points in n: 20000, 40000, 80000, 120000. The dotted line represents the results for the simplex-based Barrodale and Roberts algorithm `l1fit`, which increases roughly quadratically in n. The solid line represents `prqfn`, the timings of the Frisch-Newton interior point algorithm, with preprocessing,

In Figure 6.5 we illustrate another small experiment to compare `rqfn` and `prqfn` with `lm` for n up to 180,000. Patience, or more accurately the lack thereof, however, doesn't permit us to include further comparisons with `l1fit`. Figure 6.5 displays the improvement provided by preprocessing, and shows that `prqfn` is actually slightly faster than `lm` at for $p = 4$ and quite close to least squares speed for $p = 8$ for this range of sample sizes. It may be noted that internal Fortran timings of of `prqfn` have shown that most of the time is spent in the primal-dual routine `rqfn` for $n < 200,000$. The results of Sections 5-6 suggest that the greatest value of preprocessing appears when n is large enough that the time needed to create the globs and check residuals is comparable to that spent in `rqfn`.

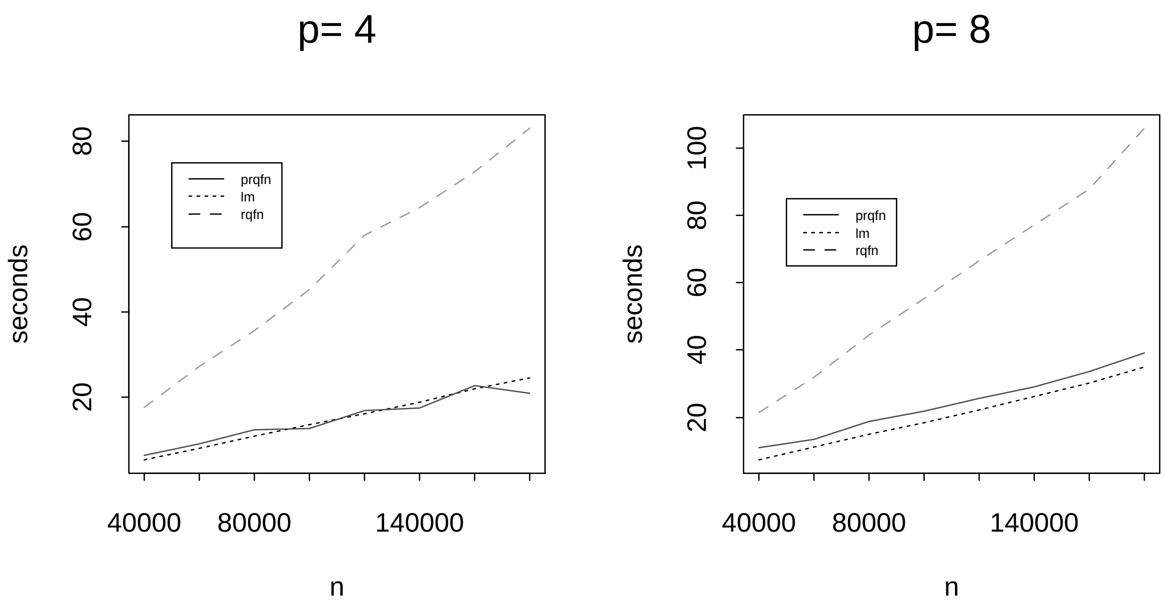

FIGURE 6.5. Timing comparison of two ℓ_1-algorithms for median regression: Times are in seconds for the mean of ten replications for iid Gaussian data. The parametric dimension of the models is $p+1$ with p indicated above each plot, p columns are generated randomly and an intercept parameter is appended to the resulting design. Timings were made at 8 design points in n: 40,000, 60,000, 80,000, 100,000, 120,000, 140,000, 160,000, 180,000. The `rqfn` dashed line represents a primal-dual interior point algorithm, `prqfn` is `rqfn` with preprocessing, and the dotted line represents least squares timings based on `lm(y ~ x)` as a benchmark.

Finally, we report some experience with a moderately large econometric application. This is a fairly typical wage equation as employed in the labor economics literature. See Buchinsky (1994,1995) for a much more extensive discussion of related results. The data are from the five percent sample of the 1990 U.S. Census, and consists of annual salary and related characteristics on 113,547 men from the state of Illinois who responded that they worked 40 or more weeks in the previous year and who worked on average 35 or more hours per week.

covariate	τ =0.05	τ =0.25	τ =0.5	τ =0.75	τ =0.95	ols
intercept	7.60598 (0.028468)	7.95888 (0.012609)	8.27162 (0.009886)	8.52930 (0.010909)	8.54327 (0.025368)	8.21327 (0.010672)
exp	0.04596 (0.001502)	0.04839 (0.000665)	0.04676 (0.000522)	0.04461 (0.000576)	0.05062 (0.001339)	0.04582 (0.000563)
exp2	−0.00080 (0.000031)	−0.00075 (0.000014)	−0.00069 (0.000011)	−0.00062 (0.000012)	−0.00056 (0.000028)	−0.00067 (0.000012)
education	0.07034 (0.001770)	0.08423 (0.000784)	0.08780 (0.000615)	0.09269 (0.000678)	0.11953 (0.001577)	0.09007 (0.000664)
white	0.14202 (0.014001)	0.17084 (0.006201)	0.15655 (0.004862)	0.13930 (0.005365)	0.10262 (0.012476)	0.14694 (0.005249)
married	0.28577 (0.011013)	0.24069 (0.004878)	0.20120 (0.003824)	0.18083 (0.004220)	0.20773 (0.009814)	0.21624 (0.004129)

TABLE 6.1. Quantile Regression Results for a U.S. Wage Equation

We seek to investigate the determinants of the logarithm of individuals' reported wage or salary income in 1989 based on their attained educational level, a quadratic labor market experience effect, and other characteristics. Results are reported Table 6.1 for five distinct quantiles. Least squares results for the same model appear in the final column of the table. The standard errors reported in parentheses were computed by the sparsity method described in Koenker (1994) using the Hall-Sheather bandwidth. There are a number of interesting findings. The experience profile of salaries is quite consistent across quantiles, with salary increasing with experience at a decreasing rate. There is a very moderate tendency toward more deceleration in salary growth with experience at the lower quantiles. The white-nonwhite salary gap is highest at the first quartile, with whites receiving a 17 percent premium over non-whites with similar characteristics, but this appears to decline both in the lower tail and for higher quantiles. Marriage appears to entail an enormous premium at the lower quantiles, nearly a 30 percent premium at the fifth percentile for example, but this premium declines somewhat as salary rises. The least squares results are quite consistent with the median regession results, but we should emphasize that the pattern of estimated quantile regression coefficients in the table as a whole is quite inconsistent with the classical iid-error linear model, or indeed, any of the conventional models accomodating some form of parametric heteroscedasticity.

method	τ =0.05	τ =0.25	τ =0.5	τ =0.75	τ =0.95
prqfn	9.92	9.78	19.91	7.68	8.64
rqfn	41.07	42.34	28.33	40.87	59.69
rq	565.97	2545.42	3907.42	3704.50	3410.49

TABLE 6.2. Timing Comparisons for 3 Methods in Wage Equation Example: Results are given in seconds for three different quantile regression algorithms described in the text.

In Table 6.2 we report the time (in seconds) required to produce the estimates in the previous table, using three alternative quantile regression algorithms. The time required for the least squares estimates reported in the last column of Table 6.1 was 7.8 seconds, roughly comparable to the `prqfn` times. Again, the interior-point approach with preprocessing as incorporated in `prqfn`, is considerably quicker than the interior point algorithm applied to the full data set in `rqfn`. The simplex approach to computing quantile regression estimates is represented here by the modification of the Barrodale and Roberts(1974) algorithm described in Koenker and d'Orey (1987), and denoted by `rq` in the table. There is obviously a very substantial gain in moving away from the simplex approach to computation in large problems of this type.

10. Conclusion

In 1887, six years after publishing his path breaking work in economics *Mathematical Psychics*, F.Y. Edgeworth began a series of papers "On a new method of reducing observations relating to several quantities." Edgeworth's new method, which he called the "plural median" was intended to revive the Boscovich/Laplace *methode de situation* as a direct competitor to the least squares approach championed by Galton and others. Edgeworth (1888) proposed dropping the zero-mean constraint on residuals, employed by his predecessors, arguing that it conflicted with the median intent of the absolute error approach. And appealing to the univariate results of Laplace, he conjectured that the plural median should be more accurate than the least squares estimator when the observations were more "discordant" than those from the Gaussian probability law. Finally, he proposed a rather arcane geometric algorithm for computing the plural median and remarked rather cryptically:

> ...the probable error is increased by about 20 percent when we substitute the Median for the Mean. On the other hand, the labour of extracting the former is rather less: especially, I should think in the case of many unknown variables. At the same time, that labour is more "skilled". There may be needed the attention of a mathematician; and, in the case of many unknowns, some power of hypergeometrical conception.

The "20 percent" is a bit optimistic. At the normal model, for example, we know that the median would have confidence intervals which are about 25 percent wider than those based on the mean. But many of the details of Edgeworth's conjectures concerning the improvements achieveable by the plural median over comparable least squares methods of inference in discordant situations have been filled in over the last 20 years. And we are now on the verge of fully vindicating Edgeworth's other claim that the "plural median" is less laborious, as well as more robust, than its least squares competitor. In the metaphor of Portnoy and Koenker (1997), the common presumption that Laplace's old ℓ_1 tortoise was forever doomed to lag behind the quicker Gaussian hare representing least squares, may finally be overturned.

Appendix A

1. Weighted Univariate Quantiles

Consider the through-the-origin quantile regression problem,

$$\min_{b in \mathbb{R}} \sum \rho_\tau(y_i - x_i b)$$

Basic solutions, $b(h)$, are of the simple form $b_i = y_i/x_i$, Directional derivitives take the form,

$$\nabla R(b,\delta) = -\sum \psi_\tau^*(y_i - x_i b, -x_i\delta)x_i\delta$$

and at a solution must be nonnegative for $\delta \in \{-1,1\}$. Adopting the temporary convention that $sgn(0) = 1$ we may write

$$\begin{aligned}\sum(\tau - I(y_i < x_i b)x_i &= \sum(\tau - \tfrac{1}{2} - \tfrac{1}{2}sgn(y_i - x_i b)x_i \\ &= \sum[\tau - \tfrac{1}{2} - \tfrac{1}{2}sgn(y_i/x_i - b)sgn(x_i)]x_i \\ &= (\tau - \tfrac{1}{2})\sum x_i - \tfrac{1}{2}\sum |x_i| + I(y_i/x_i < b)|x_i|\end{aligned}$$

Thus, as with the somewhat simpler case of the median which we discussed in Chapter 1, we may order the candidate slopes $b_i = y_i/x_i$ as $b_{(i)}$ and look for the smallest index j such that the corresponding sum of $|x_i|$'s exceeds the quantity

$$-(\tau - \tfrac{1}{2})\sum x_i + \tfrac{1}{2}\sum |x_i|.$$

This idea is, perhaps, best illustrated by the following simple S function which returns the weighted quantile estimate $\hat{b}$ and the index of the optimal basic observation pair.

```
> wquantile <- function(x, y, t = 0.5)
{
#weighted univariate quantile
#
ord <- order(y/x)
b <- (y/x)[ord]
wabs <- abs(x[ord])
k <- sum(cumsum(wabs) < ((t - 0.5) * sum(x) + 0.5 * sum(wabs)))
return(b = b[k + 1], k = ord[k + 1])
}
```

APPENDIX A

Non-parametric Quantile Regression

1. Kernel Methods

Start with Stone and nearest neighbor ideas and then develop the Chaudhuri, Welsh, et al work on locally polynomial QR.

2. Regression Splines

Again Stone plus Ng, He, etc.

3. Smoothing Splines

4. Multivariate Extensions

APPENDIX B

Frontiers (Grab-Bag??!!)of Quantile Regression

1. Time Series
2. Endogoneity and Sample Selection Problems
3. Discrete Response Models
4. Extreme Regression Quantiles
5. Multivariate Quantile Regression

REFERENCES

ANDREWS, D.F., P.J. BICKEL, F.R. HAMPEL, P.J. HUBER, W.H. ROGERS, AND J.W. TUKEY, (1972) *Robust Estimates of Location: Survey and Advances*, Princeton U. Press: Princeton.

BARRO, R. J. AND X. SALA-I-MARTIN, (1995). *Economic Growth*, McGraw-Hill: New York.

BARRODALE, I. AND R.D.K. ROBERTS, (1974). Solution of an overdetermined system of equations in the ℓ_1 norm, *Communications ACM*, **17**, 319-320.

BARTELS, R. AND A. CONN, (1980) Linearly constrained discrete ℓ_1 problems, *Transactions of the ACM on Mathematical Software*, **6**, 594-608.

BASSETT, G.W. AND R. KOENKER, (1982). An empirical quantile function for linear models with iid errors. *Journal of the American Statistical Association.* **77**, 407-415.

BECKER, R.A., J.M. CHAMBERS, AND A.R. WILKS, (1988) *The New S Language,* Wadsworth: Pacific Grove, CA.

BERAN, R. AND P. HALL, (1993). Interpolated nonparametric prediction intervals and confidence intervals, J. Royal Stat. Soc. (B), **55**, 643-652.

BICKEL, P. J. AND E.L. LEHMANN,, (1975) Descriptive statistics for nonparametric models. I: Introduction, and II: Location *The Annals of Statistics*, **3**, 1038-1069.

BLOOMFIELD, P. AND W.L. STEIGER, (1983). *Least Absolute Deviations: Theory, Applications, and Algorithms*, Birkhauser: Boston.

BOFINGER, E. (1975). Estimation of a density function using order statistics, *Australian J. of Statistics*, **17**, 1-7.

BONEVA, L.I., D. KENDALL, AND I. STEFANOV, (1971), Spline transformations: three new diagnostic aids for the statistical Data-analyst,

BROWN, G.W., AND A.M. MOOD, (1951) On median tests for linear hypotheses, *Proceedings of the 2nd Berkeley Symposium*, J. Neyman, (ed.) U. of California Press: Berkeley.

BUCHINSKY, M. (1994). Changes in US Wage Structure 1963-87: An Application of Quantile Regression, *Econometrica*, **62**, 405-458.

BUCHINSKY, M. (1995). Quantile Regression, the Box-Cox Transformation Model and U.S. Wage Structure 1963-1987, *J. of Econometrics*, **65**, 109-154.

CHARNES, A., W.W. COOPER AND R.O. FERGUSON, (1955). Optimal estimation of executive compensation by linear programming, *Management Science*, **1**, 138-151.

CLARKE, F.H., (1990), *Optimization and Nonsmooth Analysis*, Siam: Philadelphia.

DRAPER, D. (1988). Rank-based robust analysis of linear models, *Statistical Science*, **3**, 239-271.

FALK, M. (1986), On the estimation of the quantile density function, *Statistics & Probability Letters*, **4**, 69-73.

FERGUSON, T.S. (1967) *Mathematical Statistics*, Academic Press: New York.

FITZENBERGER, B. (1996). A Guide to Censored Quantile Regressions, forthcoming in C.R. Rao and G.S. Maddala (eds.), *Handbook of Statistics,* **15**, North-Holland: New York.

GOLDBERGER, A.S. (1983), Abnormal Selection Bias, *Studies in Econometrics, Time-Series, and Multivariate Statistics*, S. Karlin, T. Amemiya and L. Goodman, (eds.) Academic Press: New York.

GUTENBRUNNER, C. (1994) Tests for heteroscedasticity based on regression quantiles and regression rank scores, *Asymptotic Statistics: Proceedings of the Fifth Prague Symposium*, Mandl, P. and Hušková , (eds.), Physica-Verlag: Heidelberg.

GUTENBRUNNER, C. AND J. JUREČKOVÁ , (1992) Regression quantile and regression rank score process in the linear model and derived statistics, *Ann. Statist.* **20**, 305-330.

GUTENBRUNNER, C., J.JUREČKOVÁ , R. KOENKER, AND S. PORTNOY, (1993). Tests of linear hypotheses based on regression rank scores, *J. of Nonparametric Statistics*, **2**, 307-33.

HAHN, J. (1995) Bootstrapping quantile regression estimators, *Econometric Theory*, **11**, 105-121.

HÁJEK , J. AND ŠIDÁK , Z. (1967). *Theory of Rank Tests*, Academia, Prague.

HALL, P. AND M.A. MARTIN, (1989), A note on the accuracy of bootstrap percentile method confidence intervals for a quantile *Statistics & Probability Letters*, **8**, 197-200.

HALL, P. AND S. SHEATHER, (1988) On the distribution of a studentized quantile, *JRSS-B*, **50**, 381-391.

HAMPEL, F. (1974), The influence curve and its role in robust estimation, *J. of the Amerian Statistical Association*, **69**, 383-393.

HARTER, H. L. (1974, 1975) The method of least squares and some alternatives, *International Statistical Review*, **42**, 147-174, 235-264, **43**, 1-44, 125-190, 269-272.

HASTIE, T. AND LOADER, C. (1993) Local regression: Automatic kernel carpentry, (with discusssion), *Statistical Science* **8**, 120-129

HECKMAN, J. J. (1979), Sample selection bias as a specification error, *Econometrica*, **47**, 153-162

HILL, B. M. (1975), A simple general approach to inference about the tail of a distribution, *Annals of Statistics*, **3**, 1163-1174.

HOGG, R. V. (1975), Estimates of percentile regression lines using salary data, *Journal of the American Statistical Association*, **70**, 56-59.

HOROWITZ, J. (1996) Bootstrap methods for median regression models, preprint.

HUBER, P. J. (1964), Robust estimation of a location parameter *The Annals of Mathematical Statistics*, **35**, 73-101

HUŠKOVÁ , M. (1994) Some sequential procedures based on regression rank scores, *J. of Nonparametric Statistics*, **3**, 285-298.

JUREČKOVÁ , J. (1991) Tests of Kolmogorov-Smirnov type based on regression rank scores, *Transactions of the 11th Prague Conference on Information Theory and Statistical Decision Functions*, Visek, J.A. (ed), Academia: Prague.

KOENKER, R. AND G. BASSETT, (1978) Regression Quantiles, *Econometrica*, **46**, 33-50.

KOENKER, R. AND G. BASSETT, , (1982). Tests of linear hypotheses and ℓ_1 estimation, *Econometrica*, **50**, 1577-1584.

KOENKER, R. AND G. BASSETT, (1982). Robust tests for heteroscedasticity based on regression quantiles, *Econometrica,* **50**, 43-61.

KOENKER, R. AND V. D'OREY, (1987) Computing Regression Quantiles, *Applied Statistics*, **36**, 383-393.

KOENKER, R. AND V. D'OREY,(1993) A Remark on Computing Regression Quantiles, *Applied Statistics*, **43**, 410-414.

KOENKER, R.W. AND S. PORTNOY, (1987). L-Estimation for the linear model., *Journal of the American Statistical Association,* **82**, 851-857.

KOENKER, R.W. AND Q. ZHAO, (1994) L-Estimation for linear heteroscedastic models, *J. of Nonparametric Statistics*, **3**, 223-235.

KOLMOGOROV, A. N. (1931), The method of the median in the theory of errors, *Mat. Sb.* reprinted in *Selected Works of A.N. Kolmogorov*, vol II, A.N. Shiryayev, (ed), Kluwer: Dordrecht.

MOSTELLER, F. AND J.W. TUKEY, (1977). *Data Analysis and Regression: A Second Course in Statistics*, Addison-Wesley: Reading, MA.

NEWEY, W.K., AND J. L. POWELL (1987) Asymmetric least squares estimation and testing, *Econometrica*, **55**, 819-847.

NEWEY, W.K., AND J. L. POWELL (1990) Estimation of Type I censored regression models under conditional quantile restrictions, *Econometric Theory*, **6**, 295-317.

PARZEN, E. (1979), Nonparametric statistical data modeling, *Journal of the American Statistical Association*, **74**, 105-121.

PARZEN, M. I., L.J. WEI, AND Z. YING, (1994) A resampling method based on pivotal estimating functions, *Biometrika*, **81**, 341-350.

PORTNOY, S. (1984) Tightness of the sequence of empiric cdf processes defined from regression fractiles. *Robust and Nonlinear Time Series Analysis*, Franke, J., Hardle, W. and Martin, D. (eds) Springer-Verlag: New York.

PORTNOY, S. (1989). Asymptotic behavior of the number of regression quantile breakpoints, *SIAM J. Scientific & Statistical Computing,* **12**, 867-883.

PORTNOY, S. AND R. KOENKER, (1989). Adaptive L-estimation of linear models, *Ann. Statist.,* **17**, 362-381.

POWELL, J.L. (1986). Censored regression quantiles, *J. Econometrics,* **32**, 143-155.

POWELL, J.L. (1989). Estimation of monotonic regression models under quantile restrictions, *Nonparametric and Semiparametric Methods in Econometrics*, Barnett, W.A., Powell, J.L. and Tauchen, G. (eds), Cambridge U. Press: Cambridge.

POWELL, J.L. (1994). Estimation of semiparametric models, *Handbook of Econometrics*, Engle, R.F. and McFadden, D.L. (eds). North-Holland: New York.

ROCKAFELLAR, R.T. (1970), *Convex Analysis*, Princeton U Press: Princeton.

ROUSSEEUW, P. AND A. LEROY, (1987), *Robust Regression and Outlier Detection* John Wiley & Sons: New York.

RUPPERT, D. AND R.J. CARROLL,. (1980). Trimmed least squares estimation in the linear model, *Journal of the American Statistical Association.* **75**, 828-838.

SHEATHER, S.J. AND J.S. MARITZ (1983) An estimate of the asymptotic standard error of the sample median, *Australian J. of Statistics*, **25**, 109-122.

SHEATHER, S. J. AND J.S. MARRON,, (1990) Kernel quantile estimators *Journal of the American Statistical Association*, **85**, 410-416.

SHEYNIN, O.B., (1973), R.J. Boscovich's work on probability, *Archive for History of Exact Sciences*, **9**, 306-324.

SIDDIQUI, M., (1960) Distribution of Quantiles from a Bivariate Population, *Journal of Research of the National Bureau of Standards,* 64B, 145-150.

SMITH, R., (1994) Nonregular regression, *Biometrika*, **81**, 173-183.

STIGLER, S. M. (1977) Do robust estimators work with real data? *Annals of Statistics*, **5**, 1055-1077.

STIGLER, S.M. (1986) *The History of Statistics.* Cambridge: Harvard Press.

THEIL, H. (1950) A rank-invariant method of linear and polynomial regression analysis, *Ned. Akad. Wetensch. Proc. Ser. A*, **53**, 386-392.

TUKEY, J. (1965), Which part of the sample contains the information? *Proceedings of the National Academy*, **53**, 127-134.

WAGNER, H.M., (1959). Linear programming techniques for regression analysis, *Journal of American Statistical Association*, **54**, 206-212.

WELSH, A.H. (1988). Asymptotically efficient estimation of the sparsity function at a point, *Stat. and Prob. Letters*, **6**, 427-432.

WELSH, A. H., AND H.L. MORRISON, (1990), Robust L estimation of scale with an application in astronomy, *Journal of the American Statistical Association*, **85**, 729-743.

YING, Z, S.H. JUNG AND L.J. WEI (1992) Survival Analysis with Median regression Model, preprint.

Zhou, K.Q. and S.L. Portnoy, (1996) Direct use of regression quantiles to construct confidence sets in linear models, *Annals of Statistics*, **24**, 287-306.