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 begin with a discussion of quantile treatment effects in the two-sample treatment-control model. In this context, the difference between the empirical quantile functions of the treatment and control observations provides a natural generalization of conventional mean measures of the treatment effect. This "quantile treatment effect" is precisely what is delivered by the the quantile regression estimator of a model with a single binary indicator variable. We describe some basic characteristics of the quantile regression estimator, its equivariance properties, and robustness. The interpretation of estimated quantile regression parameters is described in the context of several applications. Some issues of misspecification are raised, and the chapter concludes with an interpretation of the quantile regression model as a random coefficient model.

2.1 QUANTILE TREATMENT EFFECTS

The simplest formulation of regression is the classical two-sample treatment-control model. We begin by reconsidering a general model of two-sample treatment response introduced by Lehmann and Doksum in the 1970s. 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

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

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

then the distribution of treatment responses, G, is stochastically larger than the distribution of control responses, F. Thus, in the context of survival analysis for clinical trials, for example, we could say that the treatment was unambiguously beneficial. However, if we encounter a crossing of the survival functions, the benefit of the treatment must be regarded as ambiguous.

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

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

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

$$\Delta(x) = G^{-1}(F(x)) - x. \tag{2.1}$$

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

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

Note that we can recover the mean treatment effect by simply integrating the quantile treatment effect over τ; that is,

$$\bar{\delta} = \int_0^1 \delta(\tau)d\tau = \int G^{-1}(\tau)d\tau - \int F^{-1}(\tau)d\tau = \mu(G) - \mu(F),$$

where $\mu(F)$ is the mean of the distribution F.

Doksum provides a thorough axiomatic analysis of this formulation of treatment response. Figure 2.1 illustrates the basic idea. At the median there is a positive treatment effect that becomes larger as we move upward into the right tail of the distribution. However, in the left tail, the treatment is actually disadvantageous. Figure 2.2 illustrates several variants of the quantile treatment effect for location, scale, and location-scale shifts of the normal distribution. In the upper panels of the distribution functions, the control distribution appears as black and the treatment as gray. In the middle panels we have the corresponding density functions, and in the lower panels we show the quantile treatment effects. In the location model, the quantile treatment effect is obviously constant, but in the scale and location-scale models the quantile treatment effect is an affine transformation of the control distribution and actually crosses the zero axis, indicating that the treatment is not always advantageous. A less-conventional example is illustrated in Figure 2.3, where the treatment alters the skewness of the distribution from highly left-skewed to highly right-skewed; this results in a ∪-shaped quantile treatment effect.

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

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central,
http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

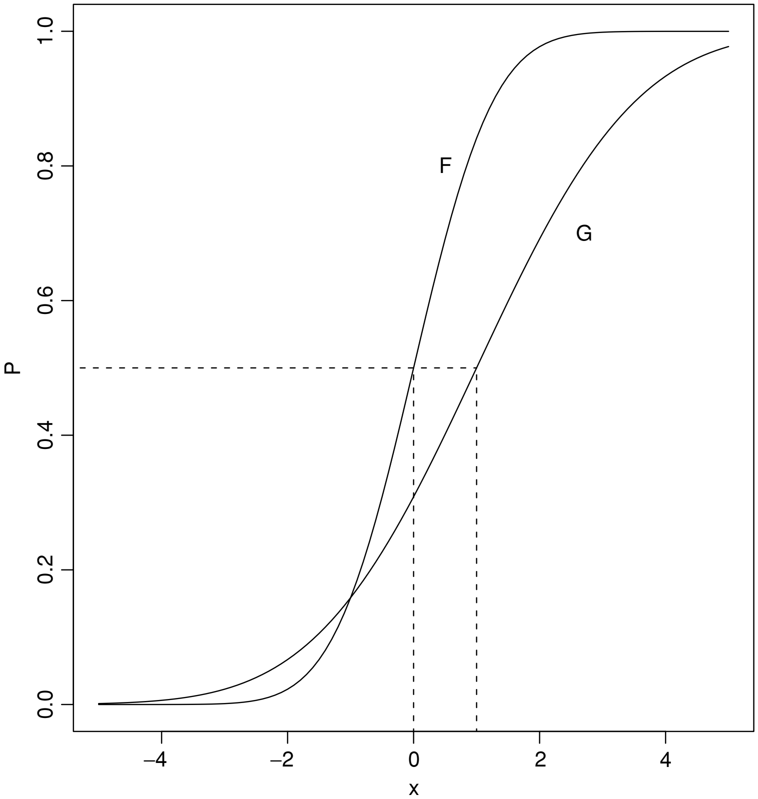

Figure 2.1. Lehmann quantile treatment effect. Horizontal distance between the treatment and control (marginal) distribution functions.

formulate the quantile regression model for the binary treatment problem as

$$Q_{Y_i}(\tau|D_i) = \alpha(\tau)(1 - D_i) + \beta(\tau)D_i, \tag{2.2}$$

where D_i denotes the treatment indicator, with $D_i = 1$ indicating treatment and $D_i = 0$ the control, we obtain the estimates $\hat{\alpha}(\tau) = \hat{F}_m^{-1}(\tau)$ and $\hat{\beta}(\tau) = \hat{G}_n^{-1}(\tau)$. However, if we consider

$$Q_{Y_i}(\tau|D_i) = \alpha(\tau) + \delta(\tau)D_i, \tag{2.3}$$

then we may estimate the quantile treatment effect directly.

To illustrate, Doksum reconsiders a study by Bjerkedal (1960) 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 treatment 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.

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central,
http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

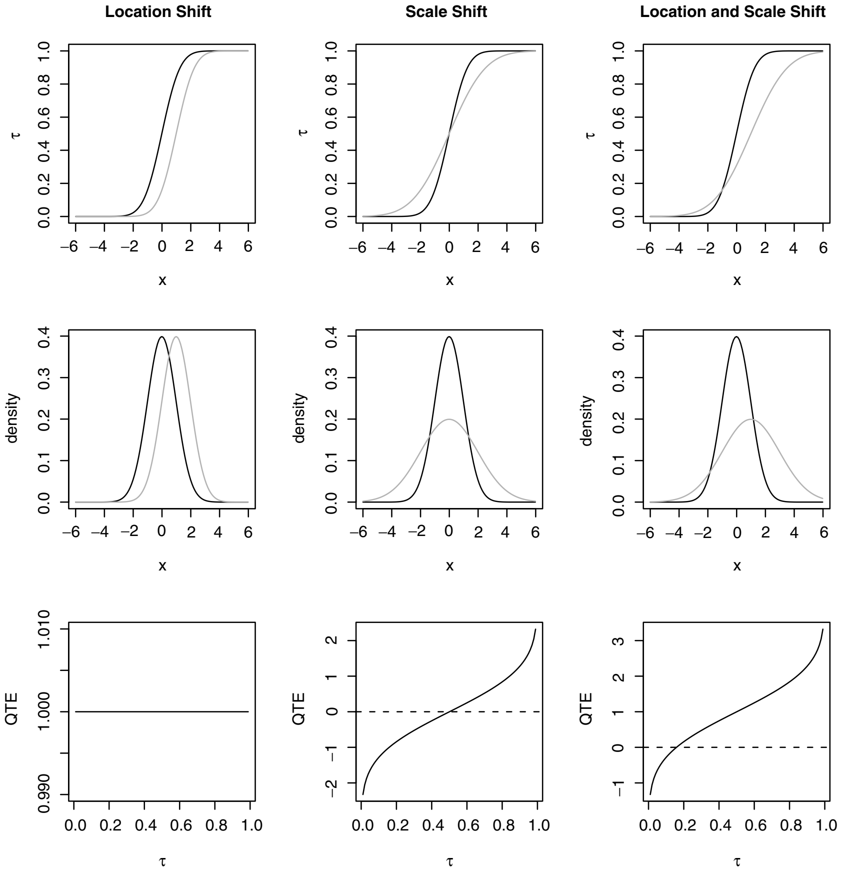

Figure 2.2. Lehmann quantile treatment effect for three examples. Location shift, scale shift, and location-scale shift.

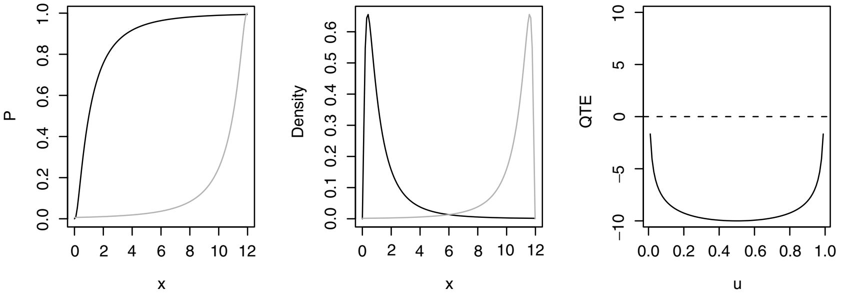

Figure 2.3. Lehmann quantile treatment effect for an asymmetric example. The treatment reverses the skewness of the distribution function.

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central,
http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

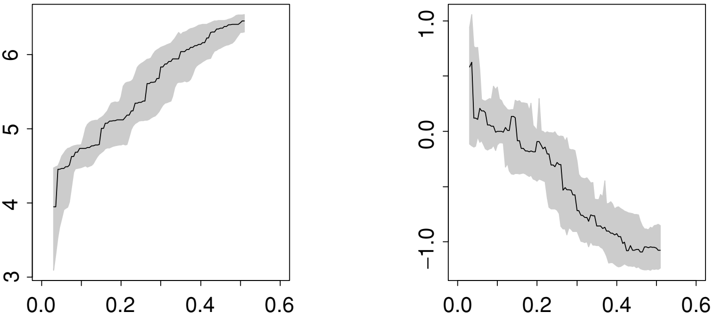

Figure 2.4. Quantile regression results for the guinea pig example analyzed by Doksum and taken from Bjerkedal. Fitted model (2.3) for log survival times is based on a sample of 107 controls and 60 treatment subjects injected with the tubercle bacilli. In the left-hand panel we plot the function $\hat{\alpha}(\tau)$, representing the empirical quantiles of the log survival time distribution for the control sample. In the right-hand 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 distribution functions 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 shaded region illustrates a 90% confidence band for the estimated effects.

In Figure 2.4 we plot the estimated functions $\hat{\alpha}(\tau)$ and $\hat{\delta}(\tau)$. The plots are "censored" beyond $\tau = 0.6$ 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 the right-hand panel, is evidently neither a location shift, which would appear as a horizontal line, nor a scale shift, which would appear as a proportional dilation of the "control effect" depicted in the left-hand (intercept) panel. Here, animals receiving the treatment injection of bacilli appear to benefit from the treatment in the lower tail of the distribution, whereas the treatment shows a strongly significant adverse effect on survival in the upper tail. 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 to 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; that is, $(Y_i | D_i = 0) = \alpha(\tau)$. 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,

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central,
http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

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 described by $\delta(\tau)$. In fact, the treatment may miraculously make weak subjects especially robust and turn the strong into jello. All we can observe from experimental evidence, however, is the difference in the two marginal survival distributions, and it is natural to associate the treatment effect with the difference in the corresponding quantiles of these 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 to 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”; thus, these terms play the same role as “propensity 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-Escudero (2001) in a study of the earnings of identical twins.

Finally, it may be noted that quantile treatment effect (2.1) is intimately tied to the traditional two-sample QQ-plot, which has a long history as a graphical diagnostic device. Following Parzen (1979), there is an extensive related literature on the issue of the “comparison density function”

$$\frac{d}{du} G(F^{-1}(u)) = \frac{g(F^{-1}(u))}{f(F^{-1}(u))},$$

an approach that is closely related to the PP-plot. 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. This connection between the Lehmann–Doksum treatment effect and the

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central,
http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

QQ-plot is explored by Doksum and Sievers (1976) and Nair (1982) for the p-sample problem. Quantile regression may be viewed as a method 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.

2.2 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 inflating a beach ball centered at y until it touches the subspace spanned by X. The quantile-regression ρ_τ "distance,"

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

has diamond-shaped polyhedral contours. Replacing Euclidean beach balls with polyhedral diamonds raises some new problems, but many nice features and insights persist. We do not obtain an elegant "closed-form" solution such as

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

but the algorithm that leads to the quantile regression estimates is really no more esoteric than, say, the sequence of Householder transformations, which are employed to produce the standard QR decomposition of X and lead eventually to the closed-form least-squares estimate.

To minimize

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

we differentiate to obtain the "normal equations"

$$\nabla_\beta \| y - \hat{y}(\beta) \|^2 = X^\top (y - X\beta) = 0$$

and solve for $\hat{\beta}$. These normal equations yield a unique solution provided that the design matrix X has full column rank.

In quantile regression we proceed similarly, but we need to exercise some caution about the differentiation step. The objective function,

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

is piecewise linear and continuous. It is differentiable except at the points at which one or more residuals, $y_i - x_i^\top \beta$, are zero. At such points, $R(\beta)$ has *directional* derivatives in all directions, depending, however, on the direction

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central,
http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

of evaluation. The directional derivative of R in direction w is given by

$$\begin{aligned}\nabla R(\beta, w) &\equiv \frac{d}{dt} R(\beta + tw) \mid_{t=0} \\ &= \frac{d}{dt} \sum_{i=1}^{n} (y_i - x_i^\top \beta - x_i^\top tw)[\tau - I(y_i - x_i^\top \beta - x_i^\top + tw) < 0)] \mid_{t=0} \\ &= -\sum \psi_\tau^*(y_i - x_i^\top \beta, -x_i^\top w) x_i^\top w,\end{aligned}$$

where

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

If, at a point $\hat{\beta}$, the directional derivatives are all nonnegative (i.e., $\nabla R(\hat{\beta}, w) \geq 0$ 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. It simply requires that the function is increasing as we move away from the point $\hat{\beta}$ regardless of the direction in which we decide to move. We can visualize the nature of such solutions by imagining a polyhedral constraint set with flat faces connected by straight edges meeting at vertices like a cut-glass bowl. Minimizing a linear function subject to such a constraint set typically yields a solution at a vertex; to verify the optimality of such a vertex solution, one needs only to verify that the objective function is nondecreasing along all edges emanating from the vertex.

These vertex solutions, as we will show in more detail in Chapter 6, correspond to points in parameter space at which p observations are interpolated when p parameters are being estimated. This "exact-fit" property of the solution is sometimes regarded with suspicion: "Aren't all the other observations being 'ignored'?" Of course, all observations participate equally in determining which ones are eventually interpolated. Just as the median identifies one middle observation, the median regression estimator identifies a group of p observations that define a hyperplane that best represents the conditional median function. This property was already recognized by Gauss (1809):

> Laplace made use of another principle for the solution of linear equations the number of which is greater than the number of the unknown quantities, which had been previously proposed by Boscovich, namely, that the sum of the errors themselves taken positively, be made a minimum. It can be easily shown, that a system of values of unknown quantities derived from this principle alone, must necessarily (except in the special cases in which the problem remains, to some extent, indeterminate) exactly satisfy as many equations out of the number proposed, as there are unknown quantities so that the remaining quantities come under consideration only so far as they help to determine the choice.

Copyright © 2005. Cambridge University Press. All rights reserved.

2.2.1 Regression Quantiles Interpolate *p* Observations

In the terminology of linear programming, these p-element subsets are called basic solutions. They may be seen as extreme points of the polyhedral constraint

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central,
http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

set: vertices of the polyhedron that 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 an entire facet of the constraint set. We will have more to say about nonuniqueness later; for now, it will suffice to observe that, even when multiple solutions occur, the basic solutions play a fundamental role because 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 consideration of these p-element subsets of observations, we require 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 let $X(h)$ 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 that passes through the points $\{(x_i, y_i), i \in h\}$ as

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

presuming, of course, that the matrix $X(h)$ is nonsingular. There are obviously too many of these basic solutions, $\binom{n}{p} = O(n^p)$, to simply search through them as one searches through 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.

Copyright © 2005. Cambridge University Press. All rights reserved.

2.2.2 The Subgradient Condition

We are now ready to introduce the basic optimality condition for the quantile regression problem:

$$\min_{b \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^\top b). \tag{2.4}$$

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 need not 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

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central,
http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

nonnegative in all directions. To check this at $b(h)$, we must consider

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

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

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

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

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

where

$$\xi^\top = \sum_{i \in \bar{h}} \psi_\tau^*(y_i - x_i^\top b(h), -x_i^\top X(h)^{-1} v) x_i^\top X(h)^{-1}.$$

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

$$0 \leq -(\tau - 1) - \xi_i(h) \qquad i = 1, \ldots, p,$$

whereas for $v = -e_i$ we have

$$0 \leq \tau + \xi_i(h) \qquad i = 1, \ldots, p.$$

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

$$-\tau 1_p \leq \xi(h) \leq (1 - \tau) 1_p.$$

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

Definition 2.1. *We 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 $h \in \mathcal{H}$,*

$$y_i - x_i^\top b(h) \neq 0 \qquad \textit{for any } i \notin h.$$

Copyright © 2005. Cambridge University Press. All rights reserved.

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 quantile-regression problem (2.4) of the form $b(h) = X(h)^{-1}y(h)$ if and only if, for some $h \in \mathcal{H}$,*

$$-\tau 1_p \leq \xi(h) \leq (1-\tau)1_p, \tag{2.5}$$

where $\xi^\top(h) = \sum_{i \in \bar{h}} \psi_\tau(y_i - x_i^\top b(h)) x_i^\top X(h)^{-1}$ 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 2.1. 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; therefore, we have more than p zero residuals – either at a solution or, more generally, in exterior point algorithms like a 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 inequalities (2.5) are satisfied only weakly. This occurs, typically, when the xs are discrete, so that sums of the x_is, weighted by τ or $(\tau - 1)$, sum exactly to τ or $\tau - 1$. If the xs 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 arises from discrete xs and multiple optimal solutions from discrete ys.

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 1.4, 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 optimality condition (2.5) is the following result that shows, provided the design matrix "contains an intercept," there will be roughly $n\tau$ negative residuals and $n(1 - \tau)$ positive ones.

Theorem 2.2. *Let P, N, and Z denote the proportion of positive, negative, and zero elements of the residual vector $y - X\hat{\beta}(\tau)$. If X contains an intercept, that is, if there exists $\alpha \in \mathbb{R}^p$ such that $X\alpha = 1_n$, then for any $\hat{\beta}(\tau)$, solving (1.19), we have*

$$N \leq n\tau \leq N + Z$$

and

$$P \leq n(1-\tau) \leq P + Z.$$

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central,
http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

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

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

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

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

which yields

$$\tau P - (1-\tau)N - (1-\tau)Z \leq 0.$$

Similarly, for $w = -\alpha$, we obtain

$$-\tau P + (1-\tau)N - \tau Z \leq 0.$$

Combining these inequalities and using the fact that $n = N + P + Z$ completes the proof. ■

Corollary 2.1. *As a consequence, if $Z = p$, 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 $(1-\tau)$,

$$\frac{P}{n} \leq 1 - \tau \leq \frac{P+p}{n}.$$

Remark 2.2. 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 is not 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 $y = (y_1^\top, y_2^\top)^\top$ 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

$$\hat{\beta}(\tau) = (\hat{\beta}_1(\tau), \hat{\beta}_2(\tau))^\top;$$

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central,
http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

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

$$-\sum_{i=1}^{n} \psi_\tau^*(y_i - b, -x_i^\top w)x_i^\top w \geq 0$$

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

$$-\sum_{i=1}^{n_j} \psi_\tau^*(y_{ij} - b_j, -w_j)w_j \geq 0 \quad j = 1, 2,$$

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

Remark 2.3. Our formulation of the optimality conditions for quantile regression in this section is fully equivalent to the approach based on the subgradient described by Rockafellar (1970). To make this connection more explicit, recall that the subgradient of a function $f\colon X \to \mathbb{R}$, at x, denoted $\partial f(x)$, is the subset of the dual space X^* given by

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

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

2.2.3 Equivariance

Several important features of the least-squares regression estimator are sometimes taken for granted in elementary treatments of regression, but they 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 reparameterize 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 are 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.* Several such properties can be grouped together under the heading of *equivariance* and treated quite explicitly because they are often an important aid in careful interpretation of statistical results. To facilitate this treatment, we will explicitly denote a τth 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 $\tau \in [0, 1]$,*

(i) $\hat{\beta}(\tau; ay, X) = a\hat{\beta}(\tau; y, X)$

(ii) $\hat{\beta}(\tau; -ay, X) = -a\hat{\beta}(1 - \tau; y, X)$

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

(iii) $\hat{\beta}(\tau; y + X\gamma, X) = \hat{\beta}(\tau; y, X) + \gamma$

(iv) $\hat{\beta}(\tau; y, XA) = A^{-1}\hat{\beta}(\tau; y, X)$.

Remark 2.4. Properties (i) and (ii) imply a form of scale equivariance, property (iii) is usually called shift or regression equivariance, and property (iv) is called equivariance to reparameterization of design.

Presuming that X "contains an intercept" (i.e., there exists $\gamma \in \mathbb{R}^p$ such that $X\gamma = 1_n$), the effect of our temperature scale change is simply to shift $\hat{\beta}(\tau; y, X)$ to $5/9(\hat{\beta}(\tau; y, X) - 32\gamma)$. Typically, in this example, γ would be the first unit basis vector e_1 and 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 5/9. In the second example, the situation is even simpler. The result of reparameterizing the xs 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,

$$Q_{h(Y)}(\tau) = h(Q_Y(\tau)); \tag{2.6}$$

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 considered earlier, or other exceptional circumstances. Condition 2.6 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

$$h(y_i, \lambda) = x_i^{\top}\beta + u_i,$$

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

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

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central,
http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

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 it were constructed for $E(y|x)$. One often sees $h^{-1}(x^{\top}\hat{\beta})$ used in place of $E(y|x)$ in such circumstances – $\exp(x^{\top}\hat{\beta})$ when the model has been specified as $\log(y) = x^{\top}\beta$, for example – but this is difficult to justify formally. Of course, the linearity of the expectation operator is also a particularly convenient property and is not shared by the quantiles except under very special circumstances (see Section 2.6).

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, $x^{\top}\hat{\beta}$, for the conditional median of $h(y)$ given x, we are perfectly justified in interpreting $h^{-1}(x^{\top}\hat{\beta})$ 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.

2.2.4 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

$$y_i^* = x_i^{\top}\beta + u_i \quad i = 1, \ldots, n, \tag{2.7}$$

where $\{u_i\}$ is independently and identically distributed (iid) from a distribution function F with density 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:

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

where Δ_i denotes the censoring indicator, $\Delta_i = 1$ if the ith observation is censored, and $\Delta_i = 0$ otherwise. For F Gaussian, this leads to an estimate of the

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

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 a discussion of this bias in some typical cases.

Powell (1986) observed that the equivariance of the quantiles to monotone transformations implies that in this model the conditional quantile functions of the response depend only on the censoring point but are independent of F. Formally, we may express the τth conditional quantile function of the observed response, y_i, in model (2.7) as

$$Q_{y_i}(\tau | x_i) = \max\{0, x_i^\top \beta + F_u^{-1}(\tau)\}. \tag{2.8}$$

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

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

by

$$\min_b \sum_{i=1}^n \rho_\tau(y_i - \max\{0, x_i^\top b\}), \tag{2.9}$$

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 model (2.8) slightly to accommodate a linear scale (heteroscedasticity) effect,

$$y_i^* = x_i^\top \beta + (x_i^\top \gamma) u_i \quad i = 1, \ldots, n \tag{2.10}$$

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

$$Q_{y_i}(\tau | x_i) = \max\{0, x_i^\top \beta + x_i^\top \gamma F_u^{-1}(\tau)\} \tag{2.11}$$

can also be estimated by solving (2.9). Because 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 that we observe

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

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central,
http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

then, by the same argument that led to (2.9) (see Fitzenberger, 1996), we would now have

$$\min_b \sum_{i=1}^{n} \rho_\tau(y_i - \max\{\underline{y}_i, \min\{\bar{y}_i, x_i^\top b\}\}). \tag{2.12}$$

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 (2.9) and (2.12) until Chapter 6. For computational purposes, the nonlinear “kinks” in the response function created by the censoring require careful attention because 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 (1995) 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.

2.3 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 density of 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).

2.3.1 The Influence Function

The modern view of robustness of statistical methods, strongly influenced by Tukey (see, e.g., Andrews, Bickel, Hampel, Huber, Rogers, and Tukey, 1974), 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 analog 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 we may consider contaminating F by replacing a small amount of mass ε from F by an equivalent

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

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 distribution function that 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,

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

and so

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

whereas for the median (see Problem 2.5),

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

$$IF_{\tilde{\theta}}(y, F) = \text{ sgn } (y - \tilde{\theta}(F))/f(F^{-1}(1/2)), \tag{2.13}$$

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, following Tukey, we will call the "sparsity" at the median, because 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 has already been 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 (2.13) by τ. The boundedness of the quantile influence function is obviously maintained, provided that the sparsity at τ is finite. Extending the *IF* to median regression is straightforward, but we now need F to represent the joint distribution of the pairs (x, y). Writing dF in the conditional form

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

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

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

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

where

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

Again, we see that the estimator has bounded influence in y because y only appears 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 farther and farther 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 well known. We will consider several proposals to robustify the behavior of quantile regression to influential design observations in Section 8.5, where we deal with the breakdown point of quantile regression estimators.

The robustness of the quantile regression estimator to outlying ys 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 farther 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 $(y_i - x_i^\top \hat{\beta}(\tau))$ 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 the 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 on 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 $i = 1, \ldots, n$; then*

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

where $\hat{u} = y - X\hat{\beta}(\tau; y, X)$.

As long as we do not alter the sign of the residuals, *any* of the y observations may be altered without altering the initial solution. Although this may, at first thought, appear astonishing, even bizarre, a second thought assures us that

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

without it we could not 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.

2.3.2 The Breakdown Point

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, and yet 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 nonprobabilistic formulation poses certain inherent difficulties. He, Jurečková, Koenker, and Portnoy (1990) showed that the breakdown point of regression estimators is closely tied to a measure of tail performance introduced by Jurečková (1981) for location estimators.

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

$$B(a, T_n) = \frac{-\log P_\theta(|T_n - \theta_0| > a)}{-\log(1 - F(a))},$$

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

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

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

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

for some $c > 0$ and $r > 0$. In this case, $\bar{X}_n$ attains optimal tail performance, with log of the probability of a large error tending to zero n times faster than the

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

log of the probability that a single observation exceeds the bound a. Whereas, on the contrary, for F with algebraic tails, so

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

for some $m > 0$, and $B(a, T_n)$ 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, $T_n(X_1, \dots, X_n)$, that are monotone in each argument, it can be shown (Theorem 2.1 of (He, Jurečková, Koenker, and Portnoy, 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 $\lim_{a\to\infty} \log(1 - F(a + c))/\log(1 - F(a)) = 1$ for any $c > 0$,

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

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 $h_{ii} = x_i^\top(X^\top X)^{-1}x_i$, for iid Gaussian errors, but again $\lim B(a, T_n) = 1$ for Fs with algebraic tails. For quantile regression estimators, a trivial upper bound on tail performance and breakdown is given by $\lim B(a, \hat{\beta}(\tau)) \le [\min\{\tau, 1-\tau\}n] + 1$, 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

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

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

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

Then $\lim B(a, \hat{\beta}(1/2)) \ge m_* + 1$ for algebraic tailed F, and the breakdown point m^* of $\hat{\beta}(1/2)$ satisfies $m_* + 1 \le m^* \le m_* + 2$. Although it is somewhat difficult to compute precisely the value of m_* for general designs in higher

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

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 $1 - 1/\sqrt{2} \approx 0.29$, a quite respectable breakdown point. For regression quantiles other than the median, breakdown is determined by similar considerations. Section 8.5 describes some proposals for high-breakdown quantile regression methods.

2.4 INTERPRETING QUANTILE REGRESSION MODELS

In the classical linear regression model, where

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

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 on both β_1 and β_2 and perhaps the effect depends more significantly on the value of z, at which we choose to evaluate the derivative. This is illustrated in the birth-weight analysis of Chapter 1.

In the transformation model,

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

there is a strong temptation to write

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

This is a common practice in logarithmic models, that is, 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.

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central,
http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

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

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

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

$$Q_{h(Y)}(\tau|X = x) = x^\top \beta(\tau),$$

then

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

So, for example, if we specify

$$Q_{\log(Y)}(\tau|X = x) = x^\top \beta(\tau),$$

then it follows that

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

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

The interpretation of the partial derivative itself, $\partial Q_Y(\tau|X = x)/\partial x_j$, often requires considerable care. We 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 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, 1996), we are typically unable to identify features of the joint distribution of control and treatment responses because we do not 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.

2.4.1 Some Examples

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

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central,
http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

Table 2.1. *The union wage premium. Quantile regression estimates of the union wage premium in the U.S. as estimated by Chamberlain (1994) based on 5358 observations from the 1987 CPS data on workers with 20–29 years of experience*

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

The Union Wage Premium

Chamberlain (1994) considers the union wage premium – that is, the percentage wage premium that union workers receive over comparable nonunion employees. Based on 1987 data from the U.S. Current Population Survey (CPS), 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.

The results for the union wage effect are summarized in Table 2.1 for manufacturing and nonmanufacturing 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 nonmanufacturing 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

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central,
http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

covariates similar to those used by Chamberlain was considerably smaller for union workers than for nonunion 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 much econometric literature on this aspect of the interpretation, which stresses the endogoneity of union status. Individuals are obviously not randomly assigned to union or nonunion 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 8.8.

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 Health 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% of the respondents reported zero consumption, and so for quantiles with $\tau < 0.4$, we have no demand response to either price or income. The income elasticity is fairly constant at about $\hat{\beta} \approx 0.25$, with some evidence of a somewhat less elastic response near $\tau = .4$ and $\tau = 1$. More interesting is the pattern of the price elasticity, $\beta_1(\tau)$, which is most elastic at moderate consumption levels with $\tau \approx 0.7$ and becomes very inelastic (unresponsive to price changes) for individuals with either very low levels of consumption ($\tau = 0.4$), 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 demand, and this is what the quantile regression analysis provides.

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

Daily Melbourne Temperatures

The third example reconsiders a semiparametric AR(1) model for daily temperature in Melbourne, Australia. Hyndman, Bashtannyk, and Grunwald (1996) recently analyzed these data using a modal regression approach. The quantile regression approach is strongly complementary and offers a somewhat more complete view of the data. Figure 2.5 provides an AR(1) scatterplot 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 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-hand side of the plot, where the central tendency of the scatter follows the 45-degree line quite closely. On the right-hand side, however, corresponding to summer conditions, the pattern

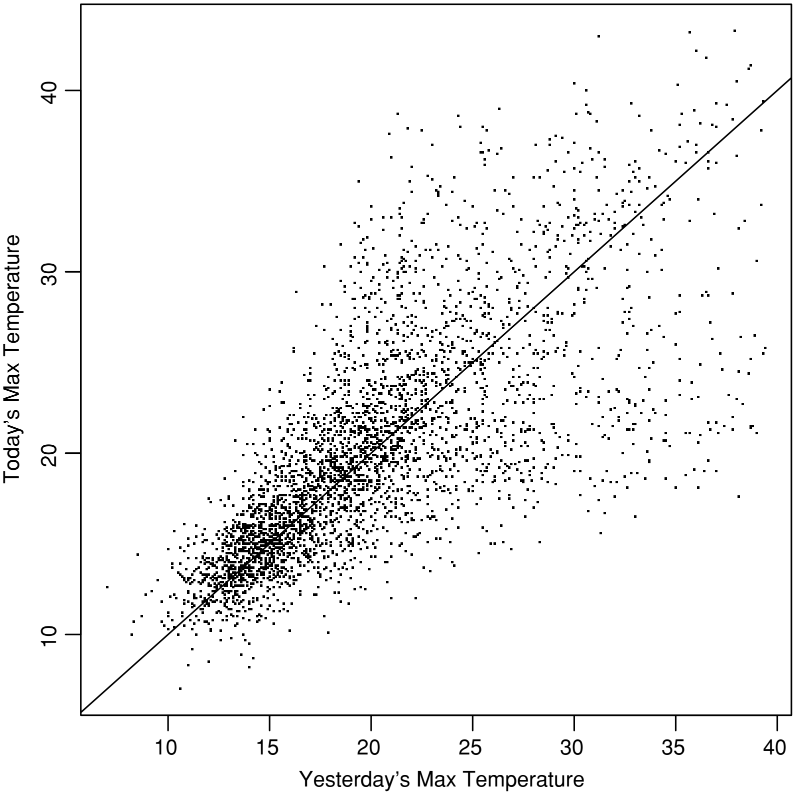

Figure 2.5. Melbourne maximum daily temperature. The plot illustrates 10 years of daily maximum (Centigrade) temperature data for Melbourne, Australia as an AR(1) scatterplot. Note that, conditional on hot weather on the prior day, the distribution of maximum temperature on the following day appears to be bimodal.

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

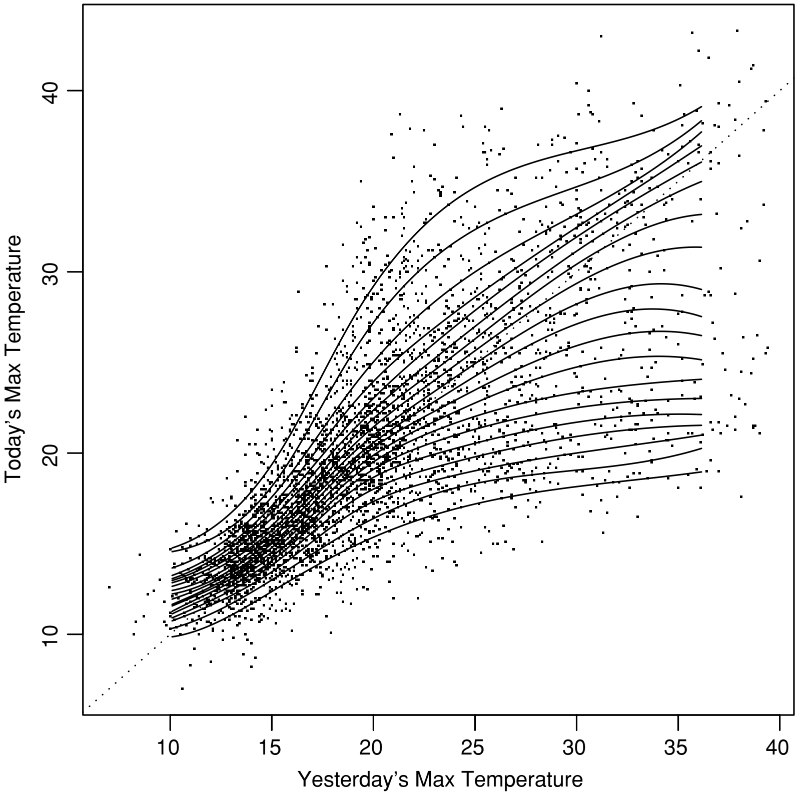

Figure 2.6. Melbourne daily maximum temperature. Superimposed on the AR(1) scatterplot of daily maximum temperatures are 12 estimated conditional quantile functions. These functions support the view that the conditional density of maximum temperature conditional on prior warm weather is bimodal.

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°C.

Figure 2.6 superimposes 19 estimated quantile regression curves on the scatterplot. 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 $\{\phi_i(\cdot) : i = 1, \ldots, p\}$ denote the basis functions of the spline. Once the knot positions of the spline have been selected, such models are linear in

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

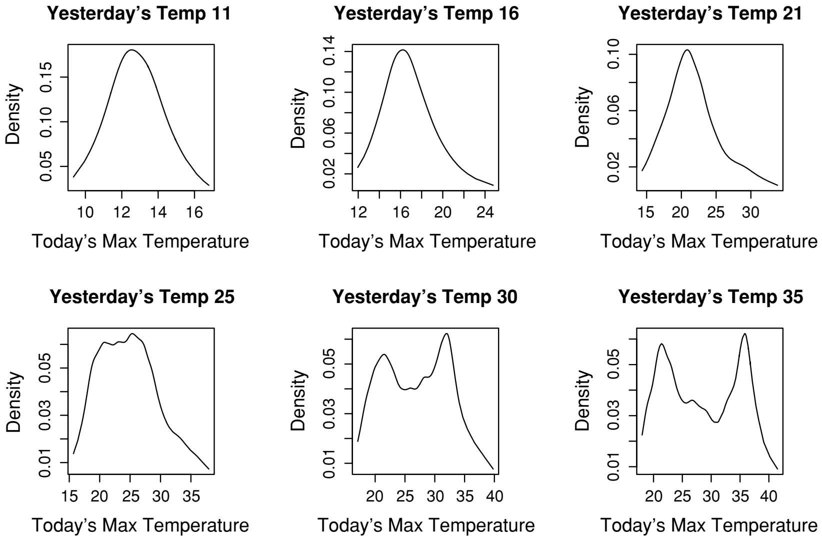

Figure 2.7. Melbourne daily maximum temperature. Conditional density estimates of today's maximum temperature for several values of yesterday's maximum temperature, based on the Melbourne data. Note that today's temperature is bimodal when yesterday was hot. When yesterday was cold, the density of today's temperature is unimodal and concentrated at somewhat warmer temperatures.

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. Figure 2.7 illustrates this approach with several density estimates based on the Melbourne data. In the last panel of this figure, we clearly see the bimodal form of the conditional density for the summer days 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: 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 are inadequate, and the recent wave of models for conditional scale, variance, and so on 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.

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central,
http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

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 2×2 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 2.8. As can be seen in these figures, the relationship is very weak until we reach the upper tail. Only the 0.95 and 0.99 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 occasionally 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 exerts 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 cited earlier.

Copyright © 2005. Cambridge University Press. All rights reserved.

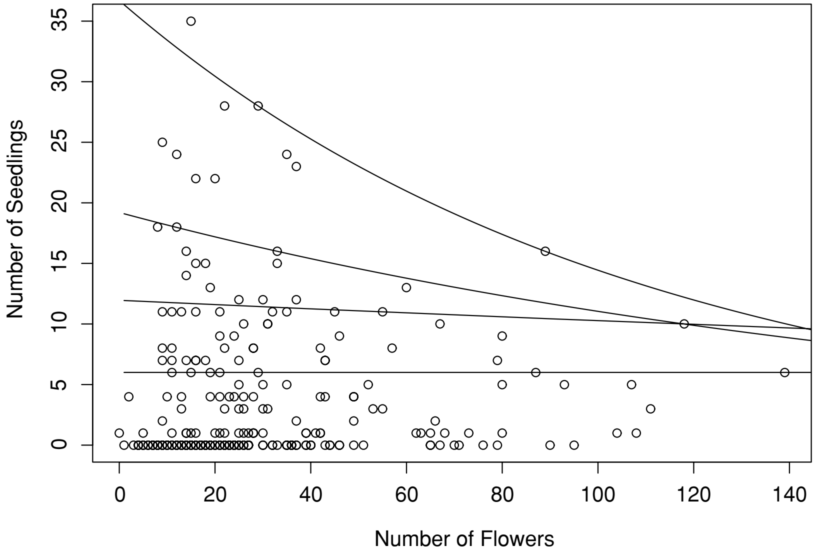

Figure 2.8. Glaciar lily seedling counts. Plotted observations on flower and seedling counts for 256 contiguous 2 × 2 m quadrats of subalpine meadow in western Colorado. As in Cade *et al.*'s (1999) study, one outlying count of 72 seedlings in a region with 16 flowers was omitted from the plot but included in the fitting. The four plotted curves are estimates of the $\tau \in \{0.75, 0.9, 0.95, 0.99\}$ 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 suggested in 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. Section 4.7 offers a more extensive review of the literature on extreme quantile regression estimation.

A somewhat unsettling feature of Figure 2.8 is the crossing of the fitted conditional quantile functions at the extreme right-hand side of the figure. Given the sparsity of data in this region, this might be viewed as inevitable: the fit of a globally quadratic model determined by the data to the left of the crossing cannot anticipate problems in regions without any observations. Some further consideration of this problem is provided in the next section.

Copyright © 2005. Cambridge University Press. All rights reserved.

2.5 CAUTION: QUANTILE CROSSING

An attractive feature of quantile regression that has been repeatedly emphasized is that it enables us to look at slices of the conditional distribution without any

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central,
http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

reliance on global distributional assumptions. Only local information near the specified quantile is employed. This can be seen in Theorem 2.4 in the insensitivity of the fit to perturbations of the observations above and below the fit. It is also apparent in the asymptotic behavior described in the next chapter. But even the most appealing mathematical features can occasionally assume a more malevolent aspect. The virtues of independently estimating a family of conditional quantile functions can sometimes be a source of serious embarrassment when we find that estimated quantile functions cross, thus violating the basic principle that distribution functions and their associated inverse functions should be monotone increasing.

It is of some comfort to recognize that such crossing is typically confined to outlying regions of the design space. The following result shows that at the centroid of the design, $\bar{x} = n^{-1} \sum x_i$, the estimated conditional quantile function

$$\hat{Q}_Y(\tau|\bar{x}) = \bar{x}^\top \hat{\beta}(\tau)$$

is monotone in τ.

Theorem 2.5. *The sample paths of $\hat{Q}_Y(\tau|\bar{x})$ are nondecreasing in τ on* [0, 1].

Proof. We will show that $\tau_1 < \tau_2$ implies $\bar{x}^\top \hat{\beta}(\tau_1) \leq \bar{x}^\top \hat{\beta}(\tau_2)$. For any $b \in \mathbb{R}^p$,

$$\sum_{i=1}^{n} \left[\rho_{\tau_2}\big(Y_i - x_i^\top b\big) - \rho_{\tau_1}\big(Y_i - x_i^\top b\big) \right] = n(\tau_2 - \tau_1)(\bar{Y} - \bar{x}^\top b). \tag{2.14}$$

This equation follows directly from the definition of ρ_τ:

$$\begin{aligned} &\rho_{\tau_2}\big(Y_i - x_i^\top t\big) - \rho_{\tau_1}\big(Y_i - x_i^\top t\big) \\ &\quad = (\tau_2 - \tau_1)\big(Y_i - x_i^\top t\big)^+ + \big[(1-\tau_2) - (1-\tau_1)\big]\big(Y_i - x_i^\top t\big)^- \\ &\quad = \big(\tau_2 - \tau_1\big)\Big[\big(Y_i - x_i^\top t\big)^+ - \big(Y_i - x_i^\top t\big)^-\Big] \qquad (2.15) \\ &\quad = \big(\tau_2 - \tau_1\big)\big(Y_i - x_i^\top t\big). \end{aligned}$$

Now, using the definition of $\hat{\beta}(\tau)$ as a minimizer of ρ_τ, and applying (2.14) with $b = \hat{\beta}(\tau_k)$ for $k = 1, 2$, we have

$$\begin{aligned} &\sum_{i=1}^{n} \rho_{\tau_1}\big(Y_i - x_i^\top \hat{\beta}(\tau_1)\big) + n\big(\tau_2 - \tau_1\big)\big(\bar{Y} - \bar{x}^\top \hat{\beta}(\tau_2)\big) \\ &\qquad \leq \sum_{i=1}^{n} \rho_{\tau_1}\big(Y_i - x_i^\top \hat{\beta}(\tau_2)\big) + n\big(\tau_2 - \tau_1\big)\big(\bar{Y} - \bar{x}^\top \hat{\beta}(\tau_2)\big) \\ &\qquad = \sum_{i=1}^{n} \rho_{\tau_2}\big(Y_i - x_i^\top \hat{\beta}(\tau_2)\big) \qquad (2.16) \end{aligned}$$

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

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

Simplifying, we see that this is equivalent to

$$
n\big(\tau_2 - \tau_1\big)\big(\bar{x}^\top \hat{\beta}(\tau_2) - \bar{x}^\top \hat{\beta}(\tau_1)\big) \geq 0, \tag{2.17}
$$

from which the result follows immediately. ■

Of course monotonicity at $x = \bar{x}$ is not a guarantee that $\hat{Q}_Y(\tau|x)$ will be monotone in τ at other values of x. Indeed it is obvious that if $\hat{Q}_Y$ is linear in variables then there must be crossing sufficiently far away from $\bar{x}$. It may be that such crossing occurs outside the convex hull of the x observations, in which case the estimated model may be viewed as an adequate approximation within this region. But it is not unusual to find that crossing has occurred inside this region as well. It is easy to check whether $\hat{Q}_Y(\tau|x)$ is monotone at particular x points. If there is a significant number of observed points at which this condition is violated, then this can be taken as evidence of misspecification of the covariate effects.

To illustrate the consequences of such misspecification, consider the simple location-scale shift model:

$$
y_i = \beta_0 + x_i\beta_1 + (\gamma_0 + \gamma_1 x_i)\nu_i. \tag{2.18}
$$

Suppose that the ν_i are iid with distribution function F. When the scale parameter $\gamma_1 = 0$, then we have a pure location shift model and the family of conditional quantile functions are parallel. When $\gamma_0 = 0$ and $\gamma_1 > 0$, we have a family of conditional quantile functions that all pass through the point $(0, \beta_0)$. This is fine as long as we contemplate using the model only in the region of positive xs. If, however, venturing into more dangerous territory, we let x_is take both positive and negative values, then the quantile functions are no longer linear. If ν has quantile function $F^{-1}(\tau)$ then $-\nu$ has quantile function $-F^{-1}(1-\tau)$, and so the quantile quantile functions for our location-scale model may be written in piecewise linear form:

$$
Q_Y(\tau|x_i) = \begin{cases} \beta_0 + x_i\beta_1 + (\gamma_0 + \gamma_1 x_i)F^{-1}(\tau) & \text{if } \gamma_0 + \gamma_1 x_i \geq 0 \\ \beta_0 + x_i\beta_1 + (\gamma_0 + \gamma_1 x_i)F^{-1}(1-\tau) & \text{otherwise.} \end{cases}
$$

If we persist in fitting linear models in the face of this nonlinearity of the true model, we can be badly misled. In Figure 2.9 we illustrate two cases: in the left panel the quantile functions are kinked at the point $x = 0$ and we observe xs uniformly distributed on the interval $[-0.1, 5]$. In this case there is very little bias introduced by the kinks in the conditional quantile functions. The gray lines represent the fitted linear approximations to the piecewise linear conditional quantile functions, and there is little distortion except at the 0.01

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

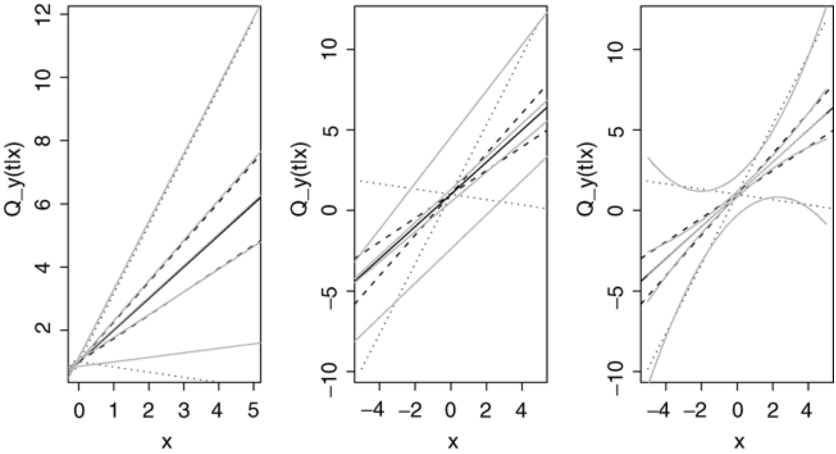

Figure 2.9. Quantile regression under misspecification. In model (2.18), with $\beta_0 = 1, \beta_1 = 1, \gamma_0 = 0, \gamma_1 = 1$, and ν_is iid $\mathcal{N}(0, 0.5^2)$, the conditional quantile functions depicted as dashed and dotted lines appear to cross but are actually kinked at the point $x = 0$. When linear conditional quantile functions are estimated, the degree of distortion, or bias, depends on the location of the kink relative to the support and density of the x observations. In the left-most panel, x is uniform on the interval $[-0.1, 5]$. The gray lines show the fitted conditional quantile lines for $\tau \in \{0.01, 0.3, 0.5, 0.7, 0.99\}$. There is little distortion except for the $\tau = 0.01$ fit. In the middle panel, we repeat the exercise with $x \sim [-5, 5]$, and so the kink occurs in the middle of the domain. Now all the fitted lines are distorted except the median, and the fitted quantile functions mimic an iid error model with parallel quantile functions even though the true conditional quantile functions are highly nonlinear. In the right-most panel, the fitted conditional quantile functions are respecified as quadratic in x and now quite accurately mimic the piecewise linear form of the true functions, except in the most extreme x regions.

Copyright © 2005. Cambridge University Press. All rights reserved.

quantile. In the middle panel, the xs are now uniform on $[-5, 5]$, and we see that the attempt to fit the piecewise linear truth with a strictly linear model completely misrepresents reality, except at the median where reality is still really linear. The true $\vee$-shaped and $\wedge$-shaped quantile functions appear, due to the symmetry of the xs, to produce parallel quantile functions as if there were iid error. In the rightmost panel of the figure, the fitted curves are respecified as quadratic, and it is apparent that they now provide a somewhat more accurate representation of the model.

In some applications we may wish to impose monotonicity in some stronger form across the quantile functions. One strategy for accomplishing this has been proposed by He (1997), who adopts the location-scale shift model,

$$y_i = x_i^\top \beta + (x_i^\top \gamma) u_i, \tag{2.19}$$

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

with u_i iid. He suggests estimating this model in three steps:

1. a median regression of y_i on x_i to obtain $\hat{\beta}$ and associated residuals, $\hat{u}_i = y_i - x_i^\top \hat{\beta}$,
2. a median regression of $|\hat{u}_i|$ on x_i to obtain $\hat{\gamma}$ and associated fitted values, $s_i = x_i^\top \hat{\gamma}$; and
3. a bivariate quantile regression of $\hat{u}_i$ on s_i constrained through the origin to determine scalar coefficients $\hat{\alpha}(\tau)$.

Provided that the scale estimates s_i are nonnegative, we may take

$$\hat{Q}_Y(\tau|x) = x^\top(\hat{\beta} + \hat{\alpha}(\tau)\hat{\gamma})$$

as an estimate of the conditional quantile functions. It is guaranteed to be monotone in τ at all x since $\hat{\alpha}$ must be monotone by Theorem 2.5. We cannot be sure that the s_is will be nonnegative, and so it might be useful to consider constraining step 2 to produce nonnegative fitted values. This could be relatively easily accomplished using the approach described by Koenker and Ng (2004). However, He suggests quite sensibly that the unconstrained approach is preferable – if it produces negative fitted values, this may be taken as diagnostic evidence of misspecification of the model.

Location-scale model (2.19) imposes quite stringent restrictions on the family of conditional quantile functions. In Section 3.8 we will consider testing the plausibility of these restrictions.

2.6 A RANDOM COEFFICIENT INTERPRETATION

When we write the linear quantile regression model as

$$Q_Y(\tau|x) = x^\top \beta(\tau), \tag{2.20}$$

and claim that the model holds for all $\tau \in (0, 1)$, we have specified a complete stochastic mechanism for generating the response variable Y. Recall that a random variable, Y with distribution function F, can be simulated by generating a standard uniform random variable, $U \sim U[0, 1]$, and then setting $Y = F^{-1}(U)$. Thus, in model (2.20), Y conditional on $X = x$ can be simulated by setting $Y = x^\top \beta(U)$ for $U \sim U[0, 1]$. Given a mechanism for generating a sequence of design vectors $(x_1, \ldots, x_n)$, we can draw independent random uniforms $(u_1, \ldots, u_n)$ and generate sample responses as

$$y_i = x_i^\top \beta(u_i), \quad i = 1, \ldots, n. \tag{2.21}$$

Note that this procedure allows us to generate xs as a dependent sequence, perhaps recursively, depending on the σ-field generated by the lagged y_is.

This strategy for simulating observations from the linear quantile regression model also suggests a new random coefficient interpretation of the model. The vector $\beta(U)$ has the special feature that all its coordinates are determined by a single draw of the uniform variable U. Thus, in contrast to most of the

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central,
http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

literature on random coefficient models – typically those employing multivariate Gaussian structure – marginal distributions of the coefficients $\beta_j(U)$ from (2.20) can take a quite arbitrary form but are tied together by a strong form of dependence. Implicit also in the formulation of model (2.20) is the requirement that $Q_Y(\tau|x)$ is monotone increasing in τ for all x. In some circumstances, this necessitates restricting the domain of x; in other cases, when the coordinates of x are themselves functionally dependent, monotonicity may hold globally.

To illustrate, consider the quadratic model

$$Q_Y(\tau|x(z)) = \beta_0(\tau) + \beta_1(\tau)z + \beta_2(\tau)z^2. \tag{2.22}$$

If $\beta_2(\tau) = 0$, and $\beta_0(\tau)$ and $\beta_1(\tau)$ are monotone increasing, then $Q_Y(\tau|x(z))$ is monotone in τ for any $z \geq 0$. However, under these conditions, unless $\beta_1(\tau)$ is a constant function, monotonicity will fail for some sufficiently negative values of z. This need not be considered a fatal flaw of the model, particularly in circumstances under which z is expected to be positive. The linear specification may provide a perfectly adequate approximation over the relevant domain. When $\beta_2(\tau)$ is nonzero, it is easy to find open neighborhoods of the parameter space for which monotonicity of $Q_Y(\tau|x(z))$ holds globally for all $z \in \mathbb{R}$. The distinction between specifications that are linear in parameters and those that are linear in variables needs to be kept in mind.

Evaluation of the monotonicity of a fitted function

$$\hat{Q}_Y(\tau|x) = x^\top \hat{\beta}(\tau) \tag{2.23}$$

at the observed design points $\{x_i : i = 1, \ldots, n\}$ is relatively straightforward and provides a useful check on model adequacy. At one central design point, monotonicity is assured by Theorem 2.5.

Of course if the coordinates of x are functionally related, as in the quadratic example, then $\bar{x}$ itself needs not be a valid design point. But the fact that $Q_Y(\tau|x)$ is monotone in τ at $\bar{x}$ means that it will also be monotone in τ *near* $\bar{x}$, and consequently we can reparameterize the model so that the random coefficients are comonotone. Comonotonicity was introduced by Schmeidler (1986) and plays an important role in variants of expected utility theory based on the Choquet integral discussed in Section 3.9.

Definition 2.2. *Two random variables $X, Y : \Omega \to \mathbb{R}$ are said to be comonotonic if there exists a third random variable $Z : \Omega \to \mathbb{R}$ and increasing functions f and g such that $X = f(Z)$ and $Y = g(Z)$.*

A crucial property of comonotonic random variables is the behavior of the quantile functions of their sums. For comonotonic random variables X, Y, we have

$$F_{X+Y}^{-1}(u) = F_X^{-1}(u) + F_Y^{-1}(u).$$

This is because, by comonotonicity, we have $U \sim U[0, 1]$ such that $g(U) = F_X^{-1}(U) + F_Y^{-1}(U)$, where g is left-continuous and increasing, and so by

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central,
http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

monotone invariance of the quantile function we have $F_{g(U)}^{-1} = g \circ F_U^{-1} = F_X^{-1} + F_Y^{-1}$. In the language of classical rank correlation, comonotonic random variables are perfectly concordant (i.e., have rank correlation 1). The classical Fréchet bounds for the joint distribution function H of two random variables, X and Y, with univariate marginals F and G is given by

$$\max\{0, F(x) + G(y) - 1\} \leq H(x, y) \leq \min\{F(x), G(y)\}.$$

Comonotonic X and Y achieve the upper bound. The extremal nature of comonotonicity is further clarified by the following result.

Theorem 2.6 (Major, 1978). *Let X, Y be random variables with marginal distribution functions F and G, respectively, and finite first absolute moments. Let $\rho(x)$ be a convex function on the real line; then*

$$\inf E\rho(X - Y) = \int_0^1 \rho(F^{-1}(t) - G^{-1}(t))dt,$$

where the inf *is over all joint distributions, H, for (X, Y) having marginals F and G.*

It is easy to see that the bound is achieved when X and Y are comonotonic because, in this case, for $U \sim U[0, 1]$, we have

$$E\rho(X - Y) = E\rho(F^{-1}(U) - G^{-1}(U)) = \int_0^1 \rho(F^{-1}(u) - G^{-1}(u))du.$$

Mallows (1972) formulates Major's result for the special case of $\rho(u) = u^2$ and notes that it implies, among other things, that the maximal Pearson correlation of X and Y occurs at the Fréchet bound, where

$$\max \int xydF(x, y) = \int_0^1 F^{-1}(t)G^{-1}(t)dt.$$

It is dangerous to leap immediately to the conclusion that the random coefficient vector in quantile regression model (2.20), $\beta(U)$, should be comonotonic. In fact, in most parameterizations of the model there is no reason to expect that the functions $\beta_i(\tau)$ will be monotone. What is crucial is that there exists a reparameterization that does exhibit comonotonicity. Recall from Theorem 2.3 that we can always reparameterize (2.20) as

$$Q_Y(\tau|x) = x^\top AA^{-1}\beta(\tau) = z^\top \gamma(\tau). \tag{2.24}$$

Suppose that we now choose $p = \dim(\beta)$ design points $\{x^k : k = 1, \ldots, p\}$ where model (2.20) holds. This is always possible for x^ks sufficiently close to $\bar{x}$. The matrix A can be chosen so that $Ax^k = e^k$, the kth unit basis vector. Then for any x^k we have that, conditional on $X = x^k$,

$$Y = (e^k)^\top \gamma(U) = \gamma_k(U). \tag{2.25}$$

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central,
http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

Inside the convex hull of the x^k points, that is, conditioning on a point $x = \sum w_k x^k$ for $0 \leq w_k \leq 1$ with $\sum w_k = 1$, we have

$$Y = \sum w_k \gamma_k(U) \tag{2.26}$$

and we have a comonotonic random coefficient representation of the model. In effect, we have done nothing more than reparameterize the model so that the coordinates

$$\gamma_k(\tau) = F_{Y|X}^{-1}(\tau|x^k) \quad k = 1, \ldots, p$$

are the conditional quantile functions of Y at the points x^k. The fact that quantile functions of weighted sums of comonotonic random variables with nonnegative weights are weighted sums of the marginal quantile functions allows us to interpolate linearly between the chosen x^k. Of course, linear extrapolation is also possible, but, as usual, with extrapolation we should be cautious.

The simplest example of the foregoing theory is the two-sample treatment-control model. If we write the model to estimate the Lehmann treatment effect as in (2.3), then there is no particular reason to expect that the parameter $\delta(\tau) = G^{-1}(\tau) - F^{-1}(\tau)$ will be monotone increasing. However, if we reparameterize the model as in (2.2), then we clearly have monotonicity in both coordinates of the population parameter vector.

2.7 INEQUALITY MEASURES AND THEIR DECOMPOSITION

The extensive literature on the measurement of inequality has devoted considerable attention to the problem of decomposing, or attributing, changes in inequality to various causal factors. An important class of inequality measures for this purpose is based on Gini's mean difference,

$$\gamma = (2\mu)^{-1} E|X - Y|.$$

Gini's γ measures inequality in the distribution of wealth by computing the expected disparity in two random draws, X and Y, from the prevailing distribution of wealth, normalized by mean wealth μ. This might be interpreted as a measure of "expected envy" for two randomly selected members of the society. A more convenient definition for present purposes for the Gini coefficient is formulated by considering the Lorenz curve.

Let Y be a positive random variable with distribution function F_Y and mean $\mu < \infty$; then the Lorenz curve may be written in terms of the quantile function of Y, $Q_Y(\tau) = F_Y^{-1}(\tau)$, as

$$\lambda(t) = \mu^{-1} \int_0^t Q_Y(\tau) d\tau.$$

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central,
http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

The Lorenz curve describes the proportion of total wealth owned by the poorest proportion t of the population. Gini's mean difference may be expressed as

$$\gamma = 1 - 2\int_0^1 \lambda(t)dt,$$

that is, as twice the area between the 45° line and the Lorenz curve.

A family of such inequality measures allowing the investigator to accentuate the influence of certain ranges of wealth can be formulated by considering monotone transformations of Y. Let $\mu_h = Eh(Y)$, and write

$$\lambda_h(t) = \mu_h^{-1}\int_0^t Q_{h(Y)}(\tau)d\tau = \mu_h^{-1}\int_0^t h(Q_Y)(\tau)d\tau.$$

Thus, for example, by considering log wealth rather than wealth itself, we can downplay the influence of the upper tail and focus more attention on the lower tail of the distribution.

The advantages of the Gini/Lorenz formulation of inequality measurement for exploring decompositions based on quantile regression is apparent if we consider the model

$$Q_{h(Y)}(\tau|x) = x^\top \beta(\tau)$$

for the conditional quantile functions of wealth. As a consequence, we have an immediate additive decomposition of the Lorenz curve,

$$\lambda_h(t|x) = \mu_h^{-1}\int_0^t Q_{h(Y)}(\tau|x)d\tau = \mu_h^{-1}\sum_{j=1}^{p} x_j \int_0^t \beta_j(\tau)d\tau,$$

and this yields an additive decomposition of the Gini coefficient as well. Such decompositions allow the investigator to explore the evolution of aggregate changes in Gini's coefficient over time: a portion of the change may be attributed to change in the distribution of the covariates, and another portion to changes in the "structure of wealth" as represented by the evolution over time of $\beta(\tau)$.

Doksum and Aaberge (2002) consider the decomposition of the Gini coefficient in somewhat more detail, and Machado and Mata (2001) have investigated related decompositions in the context of applications to wage inequality in labor economics. The asymptotic behavior of the Lorenz curve and related statistics has been extensively studied by Goldie (1977) and subsequent authors. A multivariate extension of the Lorenz/Gini approach is explored by Mosler (2002).

2.8 EXPECTILES AND OTHER VARIATIONS

We have seen that minimizing sums of asymmetrically weighted absolute errors yields the sample quantiles. What if we try to minimize asymmetrically weighted sums of squared errors, or use some other asymmetrically weighted loss function? This question has been explored by several authors, including

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central,
http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

Aigner, Amemiya, and Poirier (1976), Newey and Powell (1987), Efron (1992), and Jones (1994). Minimizing the asymmetrically weighted least-squares criterion,

$$R(\xi) = \sum_{i=1}^{n} \tau(y_i - \xi)_+^2 + (1-\tau)(y_i - \xi)_-^2, \tag{2.27}$$

where u_+ and u_- denote the positive and negative parts of u; this yields what Newey and Powell (1987) call the expectiles. The central case, $\tau = 1/2$, gives the sample mean. This centering around the mean can be viewed as a virtue, as Efron has argued in the context of count data. But it also raises some concern about the robustness of estimation procedures designed for the expectiles and their interpretation. In contrast to the quantiles, which depend only on local features of the distribution, expectiles have a more global dependence on the form of the distribution. Shifting mass in the upper tail of a distribution has no impact on the quantiles of the lower tail, but it does have an impact on *all* the expectiles.

In location-scale settings, linear conditional quantile functions imply linear conditional expectile functions, and so there is a convenient rescaling of the expectiles to obtain the quantiles. But in more complicated settings, the relationship between the two families is more opaque. Figure 2.10 illustrates a

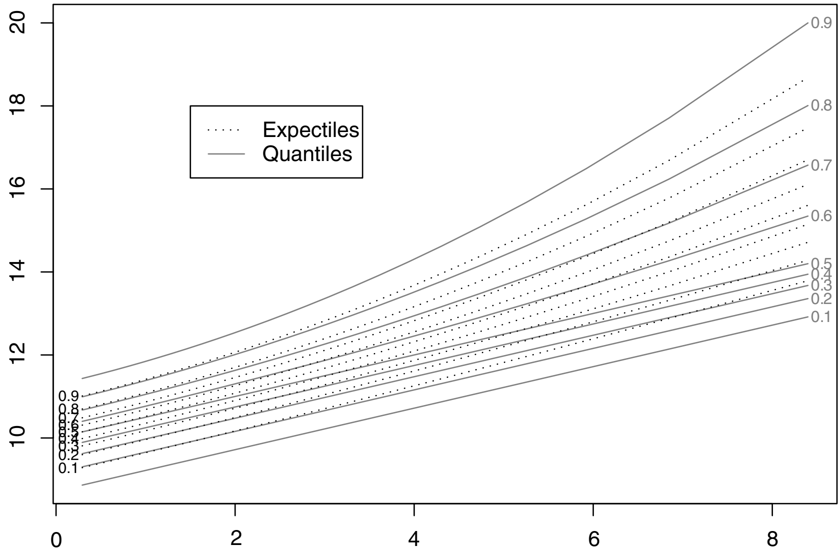

Figure 2.10. Quantiles and expectiles. Families of conditional quantile functions and their associated conditional expectile functions are illustrated. The model has linear quantile functions in the lower tail (i.e., with $\tau < 1/2$) and quadratic conditional quantile functions for $\tau > 1/2$. Note that all of the (dotted) expectile functions exhibit a nonlinearity that is induced by the nonlinearity of the (gray) quantiles in the upper tail.

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

family of conditional quantile functions that are linear in the lower tail, that is, for $\tau < 1/2$, but the quantile functions are quadratic functions of the covariate in the upper tail. The conditional expectile functions for this example, shown as dotted lines, exhibit nonlinearity throughout the entire range. See Problems 2.5 and 2.6.

2.9 INTERPRETING MISSPECIFIED QUANTILE REGRESSIONS

In the classical least-squares setting, if we consider the model

$$y_i = \theta(z_i) + u_i \tag{2.28}$$

with the $\{u_i\}$ iid from the distribution F, and independent of the z_i, but, mistakenly, we estimate the linear model,

$$y_i = \sum_{j=1}^{p} x_j(z_i)\beta_j + v_i, \tag{2.29}$$

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

$$\hat{\beta} \to E_Z(x(Z)x(Z)^\top)^{-1} x(Z)\theta(Z)$$

Thus, we can view the population version of the misspecified least-squares projection as occurring in two steps: in the first, the response y is projected to obtain the correct conditional mean response θ. 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 the distribution of design points, because it minimizes the quantity $(\theta(Z) - x(Z)^\top b)^2$.

In quantile regression, the analysis of the consequences of misspecification is somewhat more complicated, but there is still a sense in which we have an $\mathcal{L}_2$ approximation of the true conditional quantile function by a function linear in parameters. The following result is due to Angrist, Chernozhukov, and Fernandez (2003) who provide a somewhat different proof.

Theorem 2.7. *Suppose Y has τth conditional quantile function $Q_Y(\tau|Z)$ and conditional density function $f_Y(\cdot|Z)$; then*

$$\begin{aligned}\beta(\tau) &\equiv \text{argmin}_{b\in\mathbb{R}^p} E\rho_\tau(Y - x(Z)^\top b)\\ &= \text{argmin}_{b\in\mathbb{R}^p} E_Z w(Z,b)\Delta^2(Z,b),\end{aligned}$$

where $\Delta(Z,b) = x(Z)^\top b - Q_Y(\tau|Z)$ and

$$w(Z,b) = \int_0^1 (1-u) f_Y(Q_Y(\tau|Z) + u\Delta(Z,b)|Z)du.$$

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central,
http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

Proof. Let $U = Y - Q_Y(\tau|Z)$ and write

$$\beta(\tau) = \operatorname{argmin}_{b \in \mathbb{R}^p} E(\rho_\tau(U - \Delta) - \rho_\tau(U)).$$

Taking expectations of Y conditional on Z and then with respect to Z, and using Knight's (1998) identity (4.3), yields

$$\begin{aligned} E(\rho_\tau(U-\Delta)-\rho_\tau(U)) &= E\left[\int_0^\Delta (I(U \le s) - I(U \le 0))ds - \Delta\psi_\tau(U)\right] \\ &= E_Z\left[\int_0^\Delta (F_Y(Q_Y(\tau|Z) + s|Z) - \tau)ds\right] \\ &= E_Z[R_Y(Q_Y(\tau|Z) + \Delta|Z) - R_Y(Q_Y(\tau|Z)|Z) - \tau\Delta)], \end{aligned}$$

where $\psi_\tau(u) = \tau - I(u < 0)$ and $R_Y(s|Z) = \int_{-\infty}^s F_Y(t|Z)dt$. Now recall that for twice-differentiable functions g we have

$$g(b) = g(a) + (b - a)g'(a) + \int_a^b (b - x)g''(x)dx;$$

therefore, setting $b = a + \Delta$ and transforming $x \to a + u\Delta$, we have

$$g(a + \Delta) - g(a) = \Delta g'(a) + \Delta^2 \int_0^1 (1 - u)g''(a + u\Delta)du.$$

Setting $a = Q_Y(\tau|Z)$ and $g(s) = R_Y(s|Z)$ completes the proof. ■

Therefore, in misspecified situations the usual quantile regression estimator *does* minimize a quadratic measure of discrepancy between the true conditional quantile function and the best linear predictor thereof, but this measure of discrepancy must be weighted by a factor that depends on the conditional density of the response. When the discrepancy is small so that the conditional density is locally almost constant in the interval between $x(Z)^\top \beta(\tau)$ and $Q_Y(\tau|Z)$, then the weighting is simply proportional to $f_Y(Q_Y(\tau|Z)|Z)$, thus assigning more weight to regions of design space where the density is large, and estimation is consequently more accurate. It is interesting to contrast this weighting with the use of unweighted least squares estimation criteria in repeated measurement applications by, for example, Chamberlain (1994) and Bassett, Tam, and Knight (2002).

2.10 PROBLEMS

Problem 2.1. The quantile treatment effect described in Section 2.1 assumes that subjects at the τth quantile of the control distribution will also fall at the τth quantile of the treatment distribution. Suggest some alternative models of the joint distribution of the control and treatment outcomes, *given fixed*

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central,
http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.

marginals, and explain how knowledge of the joint distribution might be relevant to treatment assignment.

Problem 2.2. In many randomized treatment settings, *the copula is unidentified*. Explain the slogan in light of the preceeding problem and suggest ways that might be used to circumvent it by enriching the experimental setup.

Problem 2.3. Extend Corollary 2.2 to the p-sample problem with design matrix

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

Problem 2.4. 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.

Problem 2.5. Show that, if the real-valued random variable Y has distribution function F with finite expectation μ, then the τth expectile of Y is the unique root of the equation

$$G(y) \equiv \frac{P(y) - yF(y)}{2(P(y) - yF(y)) + y - \mu} = \tau,$$

where $P(y) = \int_{-\infty}^{x} y dF(y)$. (Jones, 1994)

Problem 2.6. Because the expectiles and the quantiles in Figure 2.10 are quite different, one may ask: Are there distributions for which the quantiles and expectiles coincide? Show that the distribution function

$$F(y) = 1/2(1 + \operatorname{sgn}(y)\sqrt{1 + 4/(4 + y^2)})$$

satisfies this requirement. Would minimizing $R(\xi)$ in (2.27) yield a consistent estimator of the expectiles for this distribution?

Copyright © 2005. Cambridge University Press. All rights reserved.

Koenker, Roger. Quantile Regression, Cambridge University Press, 2005. ProQuest Ebook Central,
http://ebookcentral.proquest.com/lib/uchicago/detail.action?docID=239194.
Created from uchicago on 2019-01-26 11:22:44.