Chapter 2
Linear and Nonparametric Quantile Regression

Quantile regression estimates can be presented in tables alongside linear regression estimates. A possible advantage of this approach to presenting quantile regression results is that it is easy to compare the values of the coefficients and standard errors with OLS estimates and across quantiles. As we have seen, quantile estimates actually contain far more information than can be presented in simple tables. The estimates imply a full distribution of values for the dependent variable. It also is easy to show how changes in the explanatory variables affect the distribution of the dependent variable.

The objective of this chapter is to provide some intuition for quantile regression estimates. Some simple Monte Carlo examples help to clarify issues related to interpreting quantile regression. I also provide an introduction to nonparametric estimation of quantile models. Nonparametric estimation turns out to be remarkably easy to implement in a quantile regression framework, and the results can be presented in a quite straightforward in a set of graphs.

2.1 Linear Quantile Regression: Simulated Data

The intuition behind quantile regression is easy to illustrate using a simple simulated data set. The raw data are shown in Fig. 2.1. To make the graphs easier to read, the single explanatory variable, x, is limited to the set of integers from 1 to 10. Each integer occurs 200 times in the simulated data set, leading to 2,000 observations in total. The base regression line is simply $y = 10 - 0.5 * x + u$. To ensure an R^2 of approximately 0.80 for the regression, I set $var(u) = 0.25 * var(x) * \frac{(1-R^2)}{R^2} = var(x)/16 = 0.5159$. After drawing 2,000 values of u from a normal distribution, the raw data look like a classic regression scatter: a clear, downward-sloping function with no systematic tendency toward unusually high or low values around the base regression line. The regression estimates are presented in the first column of

D. P. McMillen, *Quantile Regression for Spatial Data*,
SpringerBriefs in Regional Science, DOI: 10.1007/978-3-642-31815-3_2,

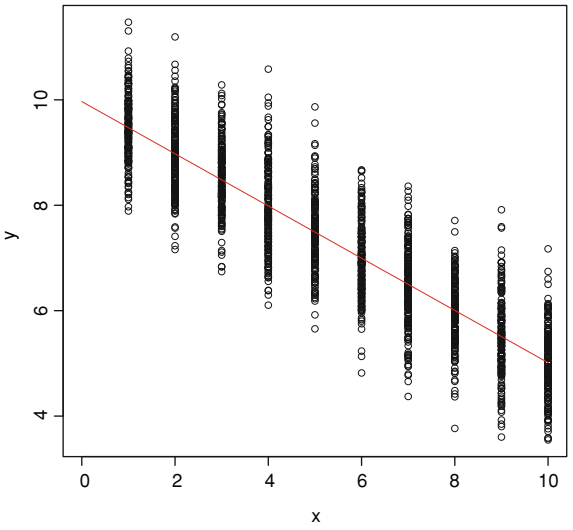

Fig. 2.1 Homoskedastic data

results in Table 2.1. The estimates are very close to the true coefficients, and they are estimated quite accurately, with low standard errors and an R^2 of (as expected) approximately 0.8. The red line in Fig. 2.1 is the estimated regression line.

The estimated quantile regression lines for the 10, 50, and 90 % quantiles are shown in Fig. 2.2, and the coefficient estimates are presented in Table 2.1. To get some intuition for the interpretation of these lines, it actually is easiest to consider a fully nonparametric estimator that takes advantage of the fact that the explanatory variable is limited to 10 integers. Each value of x is associated with 200 values of y. At each x, we can order the values of y from lowest to highest. To estimate the value of y for the 10 % quantile for $x = 1$, the nonparametric estimator would simply pick out the value of y for which 10 % of the values are lower and 90 % are higher, i.e., the 20th value. Similarly, the 50 % quantile would pick the value for which half the values of y at $x = 1$ are lower and half are higher—the 100th of our 200 ordered observations at $x = 1$. Finally, the 90 % would pick the 180th value of y at $x = 1$. We then repeat the procedure for values of y associated with $x = 2$, $x = 3$, and so on. After connecting the dots, the resulting 10, 50, and 90 % nonparametric quantile regression lines would look virtually identical to the lines shown in Fig. 2.2.

The nonparametric procedure cannot be applied so readily to more realistic data sets in which x is continuous. Since it is possible that no two values of x are identical when the variable is continuous, it clearly is not possible to identify the 10th percentile of values of y for given values of x. The intuition carries over to the continuous case, however. A quantile regression line can be thought of as finding the straight line that comes closest to connecting the series of points associated with a given percentile value for y at each value of x.

The fact that the estimated quantile regression lines are parallel in Fig. 2.2 is a direct result of having a constant variance for the errors (and thus for the

Table 2.1 Regression results for homoskedastic data

Variable	OLS	Quantile		
		10 %	50 %	90 %
Constant	9.9673	9.1036	9.9094	10.8698
	(0.0339)	(0.0520)	(0.0413)	(0.0546)
x	−0.4954	−0.4992	−0.4900	−0.4949
	(0.0055)	(0.0095)	(0.0063)	(0.0084)

Notes Standard errors are in parentheses below the estimated coefficients. The R^2 for the OLS regression is 0.8049. The number of observations is 2,000

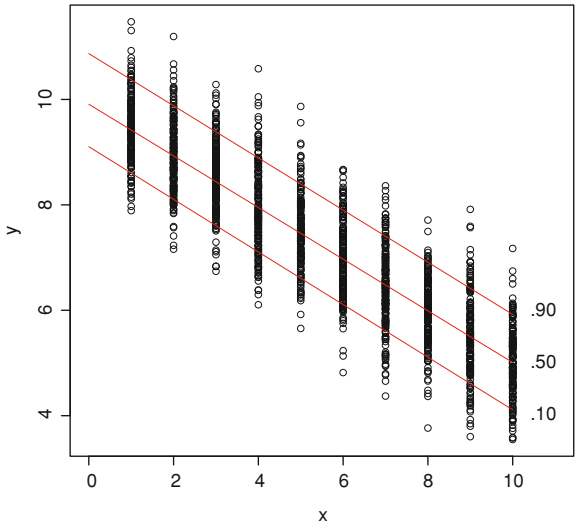

Fig. 2.2 Quantile estimates for homoskedastic data

dependent variable). Since the errors are drawn from a normal distribution and the variance is the same at each value of y, the true quantile lines are parallel, and the estimated lines will be close to parallel except in a case where quite unusual values are drawn for the errors, u.

Quantile regression becomes more interesting when the errors are not homoskedastic. Figure 2.3 shows the raw data for a simulated data set in which the variance is lower at higher values of the explanatory variable. The estimated OLS regression line, which is shown in blue, is nearly identical to the (red) quantile regression line for the median. The coefficient estimates are shown in Table 2.2. Again, we can think of the 10 % quantile regression line as a linear approximation to the set of 10th percentiles for the values of y at each value of x, the 90 % quantile lines as the set of 90th percentiles, and so on.

OLS produces a single set of coefficient estimates. The blue line shows the expected value of y given values for x, i.e., the conditional mean. It is nearly identical to the 50 % quantile regression line because the errors are drawn from a symmetric distribution. The slope is much steeper at the 90 % quantile than at the 10 % quantile, however. The slope of the 90 % line indicates how the value of

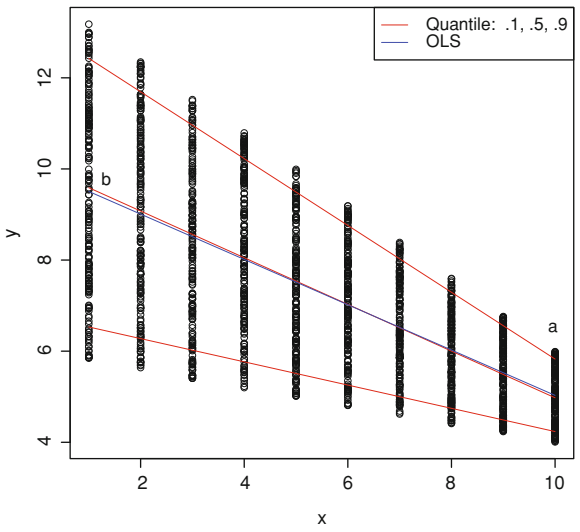

Fig. 2.3 OLS and quantile estimates for heteroskedastic data

Table 2.2 Regression results for heteroskedastic data

Variable	OLS	Quantile		
		10 %	50 %	90 %
Constant	10.0034	6.7859	10.1007	13.1599
	(0.0694)	(0.0934)	(0.1398)	(0.0917)
x	−0.4973	−0.2551	−0.5124	−0.7334
	(0.0112)	(0.0112)	(0.0167)	(0.0107)

Notes Standard errors are in parentheses below the estimated coefficients. The R^2 for the OLS regression is 0.4972. The number of observations is 2,000

y changes with *x* as we move along the 90th percentile of the distribution of values of *y* at each value of *x*. The slope of the 10 % quantile regression line shows how the value of *y* changes with *x* along the 10th percentile of the distribution of values of *y* at each value of *x*. The fact that the slope of the 90 % quantile regression line is much steeper than the 10 % line indicates that the lines are converging as *x* increases. In other words, the distribution of *y* values is less spread out at high values of *x* than at lower values, i.e., the variance of the dependent variable is lower at higher values of the explanatory variables.

It is important to recognize that this interpretation of the quantile regression results is not the same as saying that *x* leads to greater declines in *y* at high values of the dependent variable. This misleading interpretation of quantile results, which is common in the literature, leads to statements such as "the quantile regression results suggest that education adds more to the earnings of high-wage workers," or "greater levels of pollution cause greater declines in the price of high-priced homes." Points *a* and *b* in Fig. 2.3 show why these statements may be misleading. Point *a* is associated with a high value of *x* on the 90 % quantile regression line,

Table 2.3 Tests for differences in coefficients across quantiles

Variable	Constant variance data set		Declining variance data set	
	90–10 %	75–25 %	90–10 %	75–25 %
Constant	1.7663	0.9526	6.3724	3.8544
	(0.0714)	(0.0460)	(0.1234)	(0.1426)
x	0.0044	−0.0027	−0.4781	−0.2811
	(0.0114)	(0.0074)	(0.0146)	(0.0188)

Note Standard errors from 100 bootstrap replications are shown in parentheses

while point *b* represents a low value of *x* on the 50 % quantile regression line. The value of the dependent variable is lower at point *a* than at point *b*. The 90th percentile value of *y* for $x = 10$ is 5.85, the 50 % percentile of values for $x = 1$ is 9.44, and the median value of *y* for the full sample of 2,000 observations is 6.87. Thus, it is *not* the case that an increase in *x* leads to a greater decline in *y* whenever *y* is high. The steeper slope at the 90 % quantile indicates that increases in *x* lead to greater declines in *y* along the 90 % quantile of *y* values than on the 50 % quantile, *conditional on the values of x.*

Sometimes it also is useful to summarize how the spread in the distribution of y changes with x by graphing the difference between quantile regression estimates. In this simple Monte Carlo study, the graphs reveal no new information: the difference between the 10 and 90 % quantile regression estimates do not vary with *x* for the data set with constant variance, but the lines draw closer to one another as *x* increases for the heteroskedastic data set. The "iqreg" command in the statistical software package Stata makes it easy to test whether the differences between quantile regression estimates are different across quantiles. Table 2.3 presents the results for differences between the (a) the 10 and 90 % quantiles and (b) the 25 and 75 % quantiles. For the homoskedastic data set, the coefficients for *x* are not statistically different from one another across either the 10 and 90 % or the 25 and 75 % quantiles. The differences are significantly different for both sets of quantiles for the heteroskedastic data.

2.2 Simulating the Distribution of the Dependent Variable

In general, the conditional quantile function for y given a set of variables X can be written:

$$Q_y(\tau|X) = X\beta(\tau|X) \tag{2.1}$$

where $0 < \tau < 1$. So far, we have limited our attention to a small number of values for the quantile, τ. Focusing on values such as $\tau = 0.10$, 0.50, and 0.90 provides useful information about the distribution of the dependent variable given values of *X*, but it certainly does not provide a complete picture of the full distribution of *y*.

One way to use quantile regression estimates to simulate the distribution of the dependent variable is to draw randomly from possible values of τ and then estimate a separate quantile regression for each value of τ. For example, we might draw 1,000 values of τ from a uniform distribution ranging from 0 to 1, i.e., $\tau \sim U(0,1)$. If we let J represent the number of draws from the U(0,1) distribution, then we have:

$$\widehat{Q}_y(\tau_j|X) = X\widehat{\beta}(\tau_j|X), j = 1 \tag{2.2}$$

With J estimates of the conditional quantile in hand, a standard kernel density function can be applied to $X\widehat{\beta}(\tau_j|X)$ to estimate the density function for the dependent variable.

Since quantile estimates are generally fairly smooth across τ, drawing multiple values of τ from a U(0,1) distribution is a very inefficient way of constructing the density function. Using a limited range of value for τ is more efficient. For example, we might restrict the estimates to $\tau = 0.02, 0.03, \ldots, 0.97, 0.98$, $\tau = 0.02, 0.04, \ldots, 0.96, 0.98$, or a still more limited set of values for τ that provides good coverage of the set of permissible values for τ. Since quantile estimates are likely to have very high variances at extreme values of τ such as 0.01 or 0.99, it generally is a good idea to trim the extreme observations if a grid of values is used for τ.

Figures 2.4 and 2.5 show estimated coefficients for the homoskedastic data set for $\tau = 0.02, 0.03, \ldots, 0.97, 0.98,$. Figures 2.6 and 2.7 are the corresponding graphs for the data set with variances that decline with x. Note the very small range of estimates for the slopes for the homoskedastic data. These estimates imply 97 values for $X\widehat{\beta}(\tau_j|X)$ for each observation for both data sets. Thus, we have 97 × 2,000 = 194,000 implied values for $\widehat{Q}_y(\tau_j|X)$ both data sets. Kernel density estimates for these two large set of estimates leads to the density function estimates shown in Fig. 2.8. Kernel density estimates for the actual values of y are also shown in Fig. 2.8. The quantile estimates are remarkably close to the kernel density estimates for the actual values of the dependent variables.

2.3 The Effect of a Discrete Change in an Explanatory Variable

Unlike standard linear regression, quantile regressions imply interesting effects of a change in the value of an explanatory variable for the full distribution of y. Consider a simple-two variable model, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$. If we want to know the effect of changing the value of x_2 from 1 to 2, then the OLS estimates are simply $y(x_2 = 1) = \beta_0 + \beta_1 x_1 + \beta_2$ and $y(x_2 = 2) = \beta_0 + \beta_1 x_1 + 2\beta_2$. The distribution of y values simply reflects the distribution of x_1, and the distribution shifts to the right by $2\beta_2$if β_2 is positive and to the left by $|2\beta_2|$ if β_2 is negative. The implications of

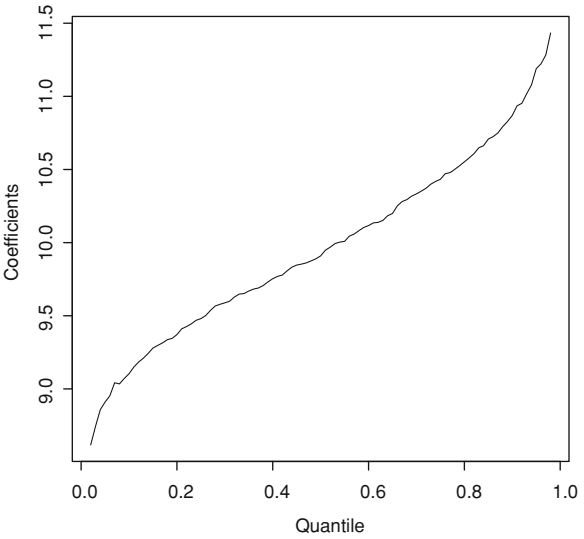

Fig. 2.4 Estimated intercepts for homoskedastic data set

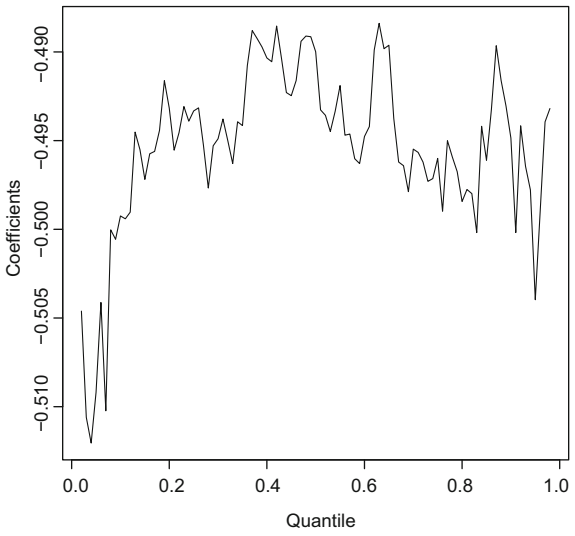

Fig. 2.5 Estimated slopes for homoskedastic data

OLS estimates are even less interesting for a model with a single explanatory variable: the change in the value of the explanatory variable simply identifies another point on the regression line.

Quantile regression estimates can have interesting implications for the distribution of y values even in a model with a single explanatory variable. Consider a model with k explanatory variable in addition to the intercept. After estimating quantile regressions for J quantiles, the predicted values for quantile τ_j are simply:

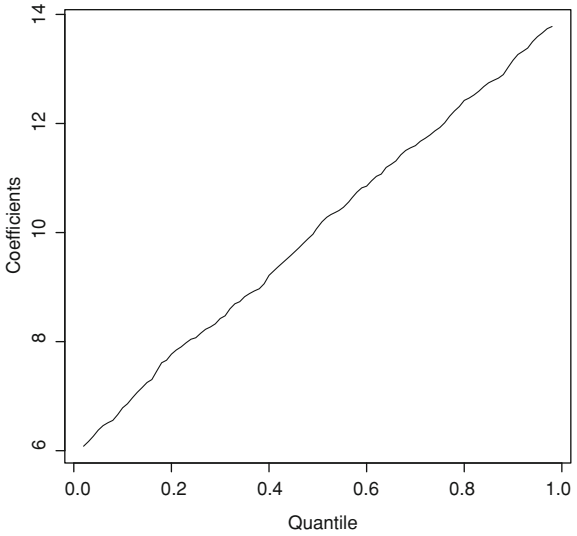

Fig. 2.6 Estimated intercepts for heteroskedastic data

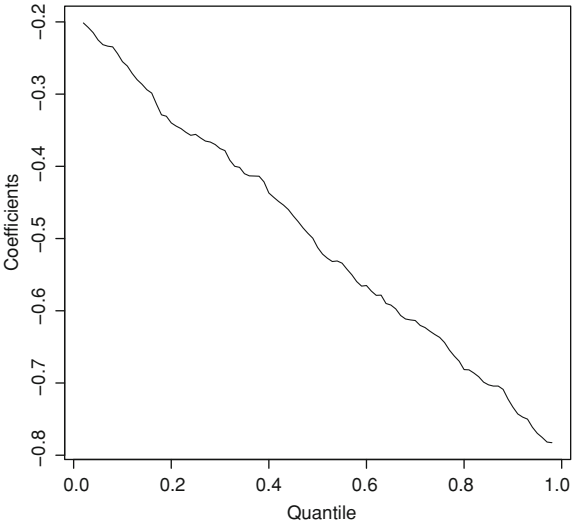

Fig. 2.7 Estimated slopes for heteroskedastic data

$$\widehat{Q}_y(\tau_j|X) = \widehat{\beta}_0(\tau_j) + \widehat{\beta}_1(\tau_j)x_1 + \ldots + \widehat{\beta}_k(\tau_j)x_k,\ j = 1, \ldots, J \tag{2.3}$$

I have simplified the notation by replacing $\beta(\tau_j|X)$ with $\beta(\tau_j)$, but it should be clear that the estimates depend on the observed values of X. Even in the single-explanatory case where k = 1, the implied effect of changing x_1 from δ_0 to δ_1 produces J separate values for

$$\widehat{Q}_y(\tau_j|X, x_1 = \delta_0) = \widehat{\beta}_0(\tau_j) + \widehat{\beta}_1(\tau_j)\delta_0 + \ldots + \widehat{\beta}_k(\tau_j)x_k, j = 1, \ldots, J \tag{2.4}$$

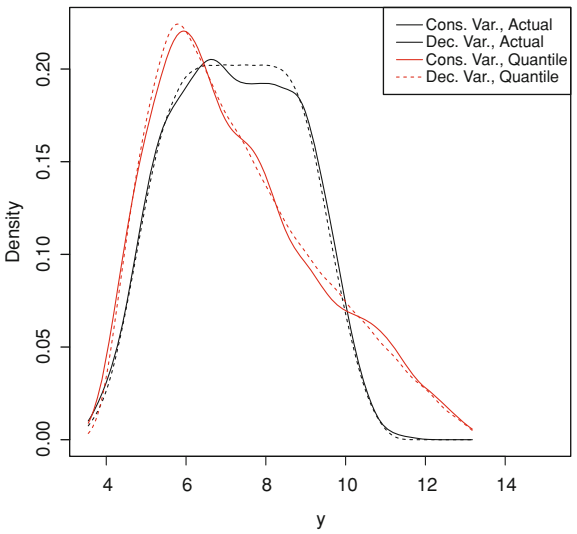

Fig. 2.8 Kernel density estimates

$$\widehat{Q}_y(\tau_j|X, x_1 = \delta_0) = \widehat{\beta}_0(\tau_j) + \widehat{\beta}_1(\tau_j)\delta_1 + \ldots + \widehat{\beta}_k(\tau_j)x_k, j = 1, \ldots, J \qquad (2.5)$$

With J quantiles and n observations, Eqs. (2.4) and (2.5) imply nJ values for the conditional quantile functions. Since $\widehat{\beta}_1(\tau_j)$ is not constant, the conditional quantile functions imply a full distribution of values for y even when x_1 is the only variable in the model.

Consider the effects of changing the single explanatory variable x from 2 to 5 to 8 in our two simulated data sets. After estimating 97 quantile regressions for the assumed values of τ (i.e., for $\tau = 0.02, 0.03, \ldots, 0.97, 0.98$)), we have 97 estimated values of both $\widehat{\beta}_0(\tau)$ and $\widehat{\beta}_1(\tau)$. Thus, we have 97 values for $\widehat{\beta}_0(\tau) + \delta\widehat{\beta}_1(\tau)$, where δ takes on the values of 2, 5, and 8, in turn. We then can calculate kernel density estimates for these three sets of quantile regression predicted values.

The results are shown in Figs. 2.9 and 2.10. For the homoskedastic data, increases in the value of x simply shift the distribution of y parallel to the left. The results appear much different for the heteroskedastic data set. As x increases, the distribution shifts to the left but also becomes much less variable.

Two points are worth emphasizing about these results. First, OLS would simply predict three separate points for each of these cases—one when $x = 2$, one for $x = 5$, and another for $x = 8$. Second, the same results are actually implicit in Figs. 2.4, 2.5, 2.6, 2.7 and, to a lesser extent, in Tables 2.1 and 2.2. The implied effects of changes on in the explanatory variable are much, much more evident when the set of quantile regression estimates is summarized in distribution form, as in Figs. 2.9 and 2.10. A very complex set of results is transformed into very easy-to-read graphs.

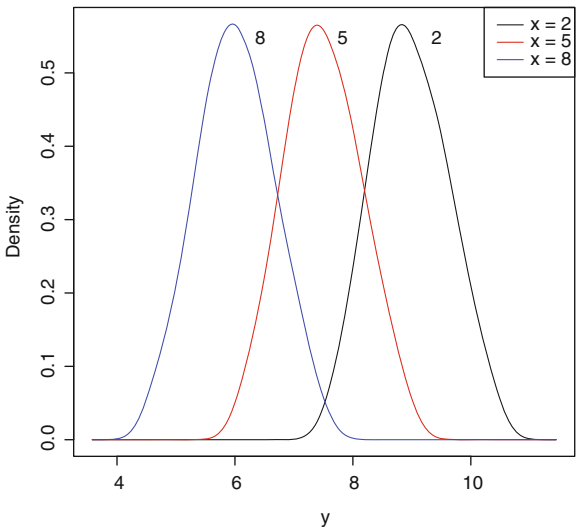

Fig. 2.9 Estimated density for y at 3 values of x, homoskedastic data

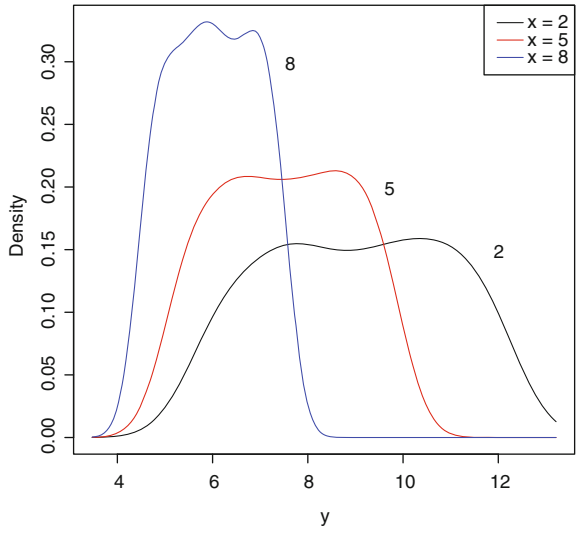

Fig. 2.10 Estimated density for y at 3 values of x, heteroskedastic data

2.4 Nonparametric Quantile Regression

So far we have only considered linear quantile regressions. As I discussed in Chap. 1, nonparametric quantile regressions can sometimes produce much more accurate predictions. Despite their apparent complexity, nonparametric versions of quantile regression are actually quite easy to estimate. The idea is to approximate the results locally with a series of quantile regression that are estimated using a

Table 2.4 Common kernel weight functions

Kernel	Kernel function K(z)
Rectangular	$\frac{1}{2}I(\lvert z\rvert<1)$
Triangular	$(1-\lvert z\rvert)I(\lvert z\rvert<1)$
Epanechnikov	$\frac{3}{4}(1-z^2)I(\lvert z\rvert<1)$
Bi-square	$\frac{15}{16}(1-z^2)^2I(\lvert z\rvert<1)$
Tri-cube	$\frac{70}{81}(1-z^3)^3I(\lvert z\rvert<1)$
Tri-weight	$\frac{35}{32}(1-z^2)^3I(\lvert z\rvert<1)$
Gaussian	$(2\pi)^{-0.5}e^{-z^2/2}$

subset of the observations that are close to a set of target values, with more weight placed on observations that are close to the target points.

In a model with a single explanatory variable, x, the target points are a set of values, x_t, where $t = 1, \ldots, T$. For each target point, define a set of weights that decline with distance, up to some maximum. At larger distances, the weight is set to zero. Any kernel weight function is suitable. Common choices are shown in Table 2.4. For this table, $Z \equiv (x - x_t)/h$, where h is the "bandwidth." In the case of a fixed bandwidth, h is simply a constant such as Silverman's Rule of Thumb bandwidth, $h = 1.06\, var(x)^{-1/5}$. A more common choice when analyzing spatial data set is to use a "window" of observations to set a value of h that varies across target points. For example, a window size of 30 % means that h_t is the 30 % quantile of $|x - x_t|$. In this case, 30 % of the observations receive weight when estimating the quantile regression for target point x_t, and h_t is the maximum distance from the target point of any observation receiving weight.

After defining weights, all that is necessary to estimate a nonparametric quantile regression model is to provide a "weight" option to the *qreg* command in Stata or the *rq* command in the R package quantreg. For example, in Stata, the series for a target value of 2 for x would be:

```
gen dist = abs(x-2)
sum dist, d
scalar h = r(p25)
gen k = (1 - (dist/h)^3)^3
replace k = . if dist>h
qreg y x [aweight=k] if dist<h
```

Comparable commands for R are:

```
library(quantreg)
dist<- abs(x-2)
h=quantile(dist,0.25)
wgt<- (1 - (dist/h)^3)^3
fit<- rq(y ~ x,weights=wgt,subset=(dist<h))
summary(fit,cov=T)
```

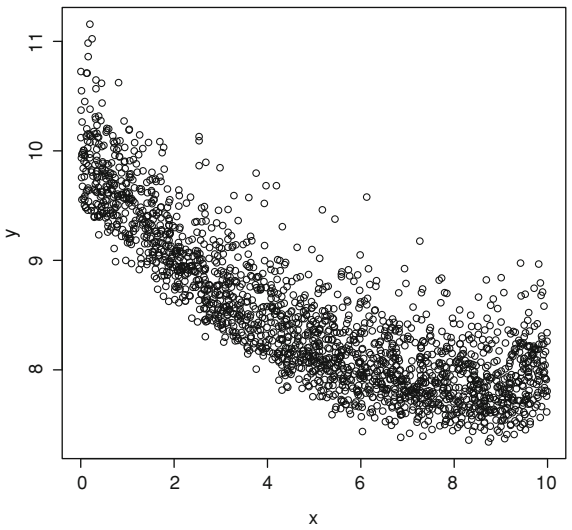

Fig. 2.11 Quadratic function with X^2 errors

The estimates can then be repeated for a series of target points. A brute force method for choosing target points is to use every observation in the data set as a target. This brute force method can be very time consuming even for relatively small data sets. A much quicker method is to take advantage of the estimated function's smoothness by using a set of well-defined points as the target and then interpolating both the coefficients and standard errors to the remaining points in the data set. Loader (1999) discusses methods for choosing target points for nonparametric models. Loader's *locfit* package in R implements these routines.

To illustrate the use of nonparametric quantile regression, consider the following extension of the Monte Carlo study. Instead of a linear relationship between y and x, the base model is $y = 10 - 0.5x + 0.03x^2 + u$, and instead of restricting x to a set of integers, I draw 2,000 values of x from a U(0,10) distribution. To make the quantile regressions different from OLS, I draw u from a χ^2 distribution with 10 of freedom. I then normalize the errors to have a mean of zero and variance of $var(10 - 0.5x + 0.03x^2)(1 - R^2)/R^2$, with $R^2 = 0.8$. Figure 2.11 shows the resulting scatter of values for x and y. In contrast to normally distributed errors, the χ^2 distributions leads to a greater cluster of points at low values of y for any given value for x.

Both OLS and quantile regression estimates will be quite accurate when the model is correctly specified, which in this context means using both x and x^2 as explanatory variables for y. Suppose instead that the estimating equation is misspecified such that only x is included as an explanatory variable. Figure 2.12 shows the results of both correctly specified parametric quantile regression estimates and a nonparametric version of the model that has only x as an explanatory variable. I use a 30 % window and a tri-cube kernel to estimate the model at a set of 14 target points chosen using an adaptive decision tree approach (Loader 1999).

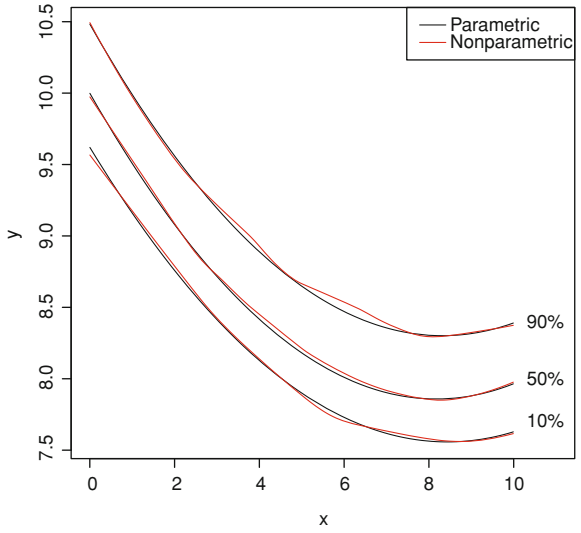

Fig. 2.12 Parametric and nonparametric quantile regression estimates

The 14 target points are 0.00, 1.25, 2.50, 3.12, 3.75, 4.37, 5.00, 5.62, 6.25, 6.87, 7.49, 8.12, 8.74, and 9.99. I then interpolate the results to all 2,000 observations in the data set. Figure 2.12 shows that the nonparametric estimates are remarkably accurate despite being misspecified.

Perhaps surprisingly, it is not much more difficult to simulate the distribution of values for y for nonparametric estimates than is the case for parametric estimates. As before, we can estimate the model for J different values of τ. The estimated coefficients for the constant, x, and x^2 are $\widehat{\beta}_0(\tau_j)$, $\widehat{\beta}_1(\tau_j)$, and $\widehat{\beta}_2(\tau_j)$. Previously, we had one value for each of these coefficients per value of τ. Now, we have n values for the coefficients for each value of τ, so each of these terms is a vector with n entries. After combining the J values for each set of coefficients into an $n \times J$ matrix, the coefficients matrices are $\widehat{\boldsymbol{\beta}}_{\boldsymbol{0}}$, $\widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}$and $\widehat{\boldsymbol{\beta}}_{\boldsymbol{2}}$. Similarly, combine the values of the explanatory variables in $\boldsymbol{x}_o$ (an n-vector of 1's), $\boldsymbol{x}_1$, and $\boldsymbol{x}_2$. Then $\widehat{\boldsymbol{y}} = \boldsymbol{x}_{\boldsymbol{0}}'\widehat{\boldsymbol{\beta}}_{\boldsymbol{0}} + \boldsymbol{x}_{\boldsymbol{1}}'\widehat{\boldsymbol{\beta}}_{\boldsymbol{1}} + \boldsymbol{x}_{\boldsymbol{2}}'\widehat{\boldsymbol{\beta}}_{\boldsymbol{2}}$ is an $n \times J$ matrix of quantile regression predictions of y. Treating this full matrix as a single vector with nJ entries, we use a standard kernel density estimator to display the distribution of predicted values for y.

The density functions for the actual values of y and the matrix of predicted values are shown in Fig. 2.13. I set $\tau = 0.02, 0.03, \ldots, 0.97, 0.98$ for the nonparametric quantile regressions, and used a 30 % window and a tri-cube kernel for each quantile. To put the remarkable similarity of the two density functions into perspective, it should be emphasized that the model is actually misspecified. Whereas the correct set of explanatory variables includes both x and x^2, the nonparametric quantile regressions are estimated without x^2.

The density functions are also easy to calculate for selected values of x. Suppose we want to evaluate the model at $x = 2, 5$, and 8. Let δ represent any of these values. Then the predicted value at $x = \delta$ is simply $\widehat{\boldsymbol{y}} = \widehat{\boldsymbol{\beta}}_{\boldsymbol{0}} + \delta\widehat{\boldsymbol{\beta}}_{\boldsymbol{1}} + \delta^2\widehat{\boldsymbol{\beta}}_{\boldsymbol{2}}$. The results are

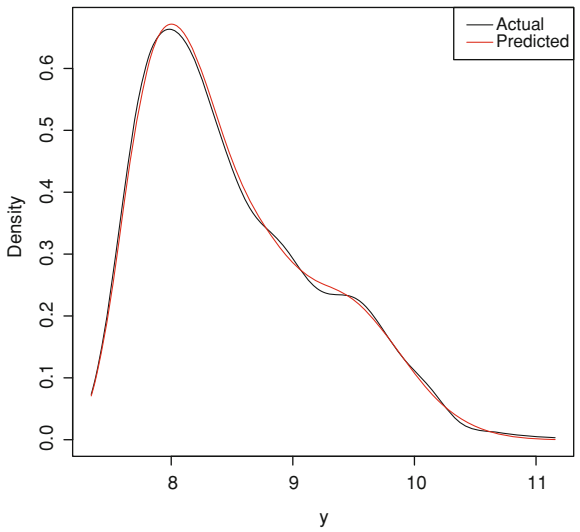

Fig. 2.13 Density functions for actual values and nonparametric quantile predictions

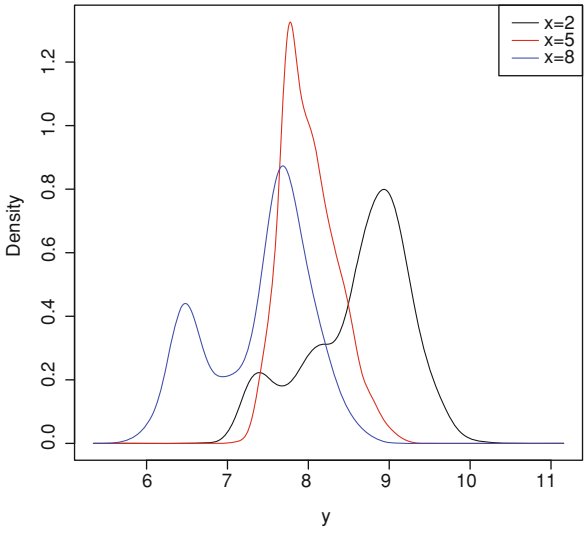

Fig. 2.14 Conditional density estimates for alternative values of x

shown in Fig. 2.14. The distribution of y clearly shifts to the left as x increases. The shapes of the conditional distributions change markedly—a large left tail when $x = 2$, tightly clustered around 8.25 when $x = 3$, and double peaked when $x = 8$.

The nonparametric estimator can potentially be extended directly to models with multiple explanatory variables using appropriate kernel weighting functions. For example, a simple product kernel is often used for the two variable case: $K(\cdot) = K\left(\frac{x_1 - x_{1t}}{h_1}\right) K\left(\frac{x_2 - x_{2t}}{h_2}\right)$, where x_{1t} and x_{2t} are the target values for the two variables. However, nonparametric estimators suffer from a "curse of

dimensionality"—a tendency toward high variance as the number of explanatory variables increases.

The variance can be reduced by imposing some structure on the nonparametric estimates. For example, suppose we are willing to impose that the coefficients are a function of a subset of the variables, so that $y = X\beta(z) + u$. This version of the model is called "conditionally parametric" (CPAR) because the equation simplifies to a standard parametric model given values for z. The CPAR model is used routinely in spatial models, where the coefficients are assumed to vary spatially. In the spatial version of the model, z may represent the geographic coordinates for the observations (e.g., longitude and latitude), or it may simply represent the straight-line distance between each observations and the target location for estimation. A product kernel can be used for the two-dimensional case, while simple univariate kernels can be used for straight-line distances. The CPAR approach is commonly used in spatial regression models, where it is often referred to as "geographically weighted regression," "locally weighted regression", or "local linear regression."

The CPAR is straightforward to apply to quantile estimation. We simply define our kernel weighting function for the target point as $k\left(\frac{z-z_t}{h}\right)$, and then add it to the "weight" options in R and Stata, using $\boldsymbol{X}$ as the set of explanatory variables. Note that the list of explanatory variables, X, can also include the variable (or variables) in z. Thus, we might make the weights a function of longitude and latitude while also directly including these variables in $\boldsymbol{X}$. Alternatively, we define $z - z_t$ to be the straight line distance between an observation and the target point, while also including longitude and latitude as explanatory variables in $\boldsymbol{X}$. The advantage of this approach when there are multiple explanatory variables it that it reduces the variance of the estimates by focusing on the source of the variation in the coefficients—spatial heterogeneity. Within a small geographic area, the model is approximately linear. But we are not requiring that the parametric specification hold globally throughout the sample region.

2.5 Conclusion

Although quantile regression can appear quite complicated, the results turn out to be remarkably easy to summarize with sets of kernel density functions. Comparative statics exercises can be carried out by assuming a few values for one of the explanatory variables while keeping all other variables at their actual values. Kernel density functions for the predicted values of the dependent variables then show how the full distribution of y responds to discrete changes in the explanatory variable. Unlike linear regression, these predictions produce remarkably accurate depictions of the distribution of the dependent variable. Moreover, the approach adapts readily to nonparametric estimation procedures. As we shall see, the CPAR approach is well suited to quantile analysis of spatial data.

http://www.springer.com/978-3-642-31814-6

Quantile Regression for Spatial Data
McMillen, D.P.
2013, IX, 66 p. 47 illus., Softcover
ISBN: 978-3-642-31814-6