1

A visual introduction to quantile regression

Introduction

Quantile regression is a statistical analysis able to detect more effects than conventional procedures: it does not restrict attention to the conditional mean and therefore it permits to approximate the whole conditional distribution of a response variable.

This chapter will offer a visual introduction to quantile regression starting from the simplest model with a dummy predictor, moving then to the simple regression model with a quantitative predictor, through the case of a model with a nominal regressor.

The basic idea behind quantile regression and the essential notation will be discussed in the following sections.

1.1 The essential toolkit

Classical regression focuses on the expectation of a variable Y conditional on the values of a set of variables $\mathbf{X}$, $E(Y|\mathbf{X})$, the so-called regression function (Gujarati 2003; Weisberg 2005). Such a function can be more or less complex, but it restricts exclusively on a specific location of the Y conditional distribution. Quantile regression (QR) extends this approach, allowing one to study the conditional distribution of Y on $\mathbf{X}$ at different locations and thus offering a global view on the interrelations between Y and $\mathbf{X}$. Using an analogy, we can say that for regression problems, QR is to classical regression what quantiles are to mean in terms of describing locations of a distribution.

Quantile Regression: Theory and Applications, First Edition. Cristina Davino, Marilena Furno and Domenico Vistocco.

Companion website: www.wiley.com/go/quantile_regression

QR was introduced by Koenker and Basset (1978) as an extension of classical least squares estimation of conditional mean models to conditional quantile functions. The development of QR, as Koenker (2001) later attests, starts with the idea of formulating the estimation of conditional quantile functions as an optimization problem, an idea that affords QR to use mathematical tools commonly used for the conditional mean function.

Most of the examples presented in this chapter refer to the *Cars93* dataset, which contains information on the sales of cars in the USA in 1993, and it is part of the *MASS* R package (Venables and Ripley 2002). A detailed description of the dataset is provided in Lock (1993).

1.1.1 Unconditional mean, unconditional quantiles and surroundings

In order to set off on the QR journey, a good starting point is the comparison of mean and quantiles, taking into account their objective functions. In fact, QR generalizes univariate quantiles for conditional distribution.

The comparison between mean and median as centers of an univariate distribution is almost standard and is generally used to define skewness. Let Y be a generic random variable: its mean is defined as the center c of the distribution which minimizes the squared sum of deviations; that is as the solution to the following minimization problem:

$$\mu = \underset{c}{\operatorname{argmin}}\, E(Y - c)^2. \tag{1.1}$$

The median, instead, minimizes the absolute sum of deviations. In terms of a minimization problem, the median is thus:

$$Me = \underset{c}{\operatorname{argmin}}\, E|Y - c|. \tag{1.2}$$

Using the sample observations, we can obtain the sample estimators $\hat{\mu}$ and $\hat{Me}$ for such centers.

It is well known that the univariate quantiles are defined as particular locations of the distribution, that is the θ-th quantile is the value y such that $P(Y \leq y) = \theta$. Starting from the cumulative distribution function (CDF):

$$F_Y(y) = F(y) = P(Y \leq y), \tag{1.3}$$

the quantile function is defined as its inverse:

$$Q_Y(\theta) = Q(\theta) = F_Y^{-1}(\theta) = \inf\{y : F(y) > \theta\} \tag{1.4}$$

for $\theta \in [0, 1]$. If $F(.)$ is strictly increasing and continuous, then $F^{-1}(\theta)$ is the unique real number y such that $F(y) = \theta$ (Gilchrist 2000). Figure 1.1 depicts the empirical CDF [Figure 1.1(a)] and its inverse, the empirical quantile function [Figure 1.1(b)],

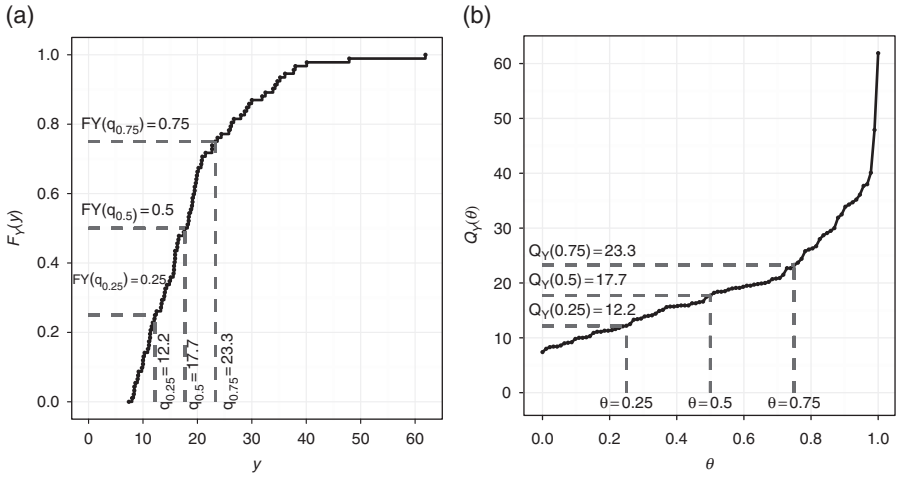

Figure 1.1 Empirical distribution function (a) and its inverse, the empirical quantile function (b), for the Price *variable of the* Cars93 *dataset. The three quartiles of* Price *are represented on the two plots: q_θ corresponds to the abscissa on the* $\mathrm{F_Y(y)}$ *plot, while it corresponds to the ordinate on the* $\mathrm{Q_Y(\theta)}$ *plot; the other input being the value of* θ.

for the *Price* variable of the *Cars93* dataset. The three quartiles, $\theta = \{0.25, 0.5, 0.75\}$, represented on both plots point out the strict link between the two functions.

Less common is the presentation of quantiles as particular centers of the distribution, minimizing the weighted absolute sum of deviations (Hao and Naiman 2007). In such a view the θ-th quantile is thus:

$$q_\theta = \underset{c}{\operatorname{argmin}}\, E[\rho_\theta(Y - c)] \tag{1.5}$$

where $\rho_\theta(.)$ denotes the following loss function:

$$\begin{aligned}\rho_\theta(y) &= [\theta - I(y < 0)]y \\ &= [(1-\theta)I(y \leq 0) + \theta I(y > 0)]|y|.\end{aligned}$$

Such loss function is then an asymmetric absolute loss function; that is a weighted sum of absolute deviations, where a $(1 - \theta)$ weight is assigned to the negative deviations and a θ weight is used for the positive deviations.

In the case of a discrete variable Y with probability distribution $f(y) = P(Y = y)$, the previous minimization problem becomes:

$$\begin{aligned}q_\theta &= \underset{c}{\operatorname{argmin}}\, E[\rho_\theta(Y - c)] \\ &= \underset{c}{\operatorname{argmin}} \left\{ (1-\theta) \sum_{y \leq c} |y - c| f(y) + \theta \sum_{y > c} |y - c| f(y) \right\}.\end{aligned}$$

The same criterion is adopted in the case of a continuous random variable substituting summation with integrals:

$$\begin{aligned} q_\theta &= \underset{c}{\operatorname{argmin}}\, E[\rho_\theta(Y-c)] \\ &= \underset{c}{\operatorname{argmin}} \left\{ (1-\theta) \int_{-\infty}^{c} |y-c| f(y) d(y) + \theta \int_{c}^{+\infty} |y-c| f(y) d(y) \right\} \end{aligned}$$

where $f(y)$ denotes the probability density function of Y. The sample estimator $\hat{q}_\theta$ for $\theta \in [0,1]$ is likewise obtained using the sample information in the previous formula. Finally, it is straightforward to say that for $\theta = 0.5$ we obtain the median solution defined in Equation (1.2).

A graphical representation of these concepts is shown in Figure 1.2, where, for the subset of *small* cars according to the *Type* variable, the mean and the three quartiles for the *Price* variable of the *Cars93* dataset are represented on the x-axis, along with the original data. The different objective function for the mean and the three quartiles are shown on the y-axis. The quadratic shape of the mean objective function is opposed to the V-shaped objective functions for the three quartiles, symmetric for the median case and asymmetric (and opposite) for the case of the two extreme quartiles.

1.1.2 Technical insight: Quantiles as solutions of a minimization problem

In order to show the formulation of univariate quantiles as solutions of the minimization problem (Koenker 2005) specified by Equation (1.5), the presentation of the solution for the median case, Equation (1.2), is a good starting point. Assuming, without loss of generality, that Y is a continuous random variable, the expected value of the absolute sum of deviations from a given center c can be split into the following two terms:

$$\begin{aligned} E|Y-c| &= \int_{y \in \mathcal{R}} |y-c| f(y) dx \\ &= \int_{y<c} |y-c| f(y) dy + \int_{y>c} |y-c| f(y) dy \\ &= \int_{y<c} (c-y) f(y) dy + \int_{y>c} (y-c) f(y) dy. \end{aligned}$$

Since the absolute value is a convex function, differentiating $E|Y-c|$ with respect to c and setting the partial derivatives to zero will lead to the solution for the minimum:

$$\frac{\partial}{\partial c} E|Y-c| = 0.$$

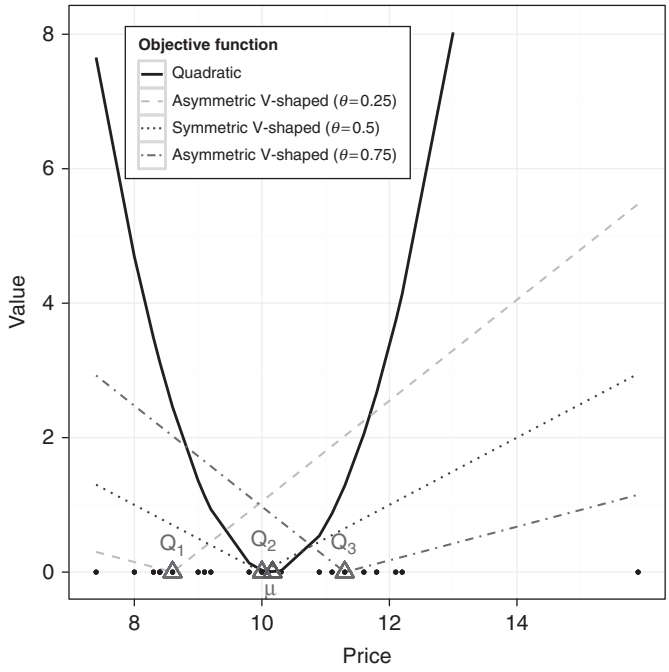

Figure 1.2 Comparison of mean and quartiles as location indexes of a univariate distribution. Data refer to the Price *of* small *cars as defined by the* Type *variable (*Cars93 *dataset). The car prices are represented using dots on the* x*-axis while the positions of the mean and of the three quartiles are depicted using triangles. Objective functions associated with the three measures are shown on the* y*-axis. From this figure, it is evident that the mean objective function has a quadratic shape while the quartile objective functions are V-shaped; moreover it is symmetric for the median case and asymmetric in the case of the two extreme quartiles.*

The solution can then be obtained applying the derivative and integrating per part as follows:

$$\left\{(c-y)f(y)\Big|_{-\infty}^{c} + \int_{y<c} \frac{\partial}{\partial c}(c-y)f(y)dy\right\} +$$

$$\left\{(y-c)f(y)\Big|_{c}^{+\infty} + \int_{y>c} \frac{\partial}{\partial c}(y-c)f(y)dy\right\} = 0$$

Taking into account that:

$$\lim_{x\to-\infty} f(x) = \lim_{x\to+\infty} f(x) = 0$$

for a well-defined probability density function, the integrand restricts in $y = c$[1]:

$$\left\{ \underbrace{(c-y)f(y)\Big|_{y=c}}_{=0 \text{ when } y=c} + \int_{y<c} f(y)dy \right\} + \left\{ \underbrace{(y-c)f(y)\Big|_{y=c}}_{=0 \text{ when } y=c} - \int_{y>c} f(y)dy \right\}.$$

Using then the CDF definition, Equation (1.3), the previous equation reduces to:

$$F(c) - [1 - F(c)] = 0$$

and thus:

$$2F(c) - 1 = 0 \implies F(c) = \frac{1}{2} \implies c = Me.$$

The solution of the minimization problem formulated in Equation (1.2) is thus the median. The above solution does not change by multiplying the two components of $E|Y - c|$ by a constant θ and $(1 - \theta)$, respectively. This allows us to formulate the same problem for the generic quantile θ. Namely, using the same strategy for Equation (1.5), we obtain:

$$\frac{\partial}{\partial c} E[\rho_\theta(Y - c)] = \frac{\partial}{\partial c} \left\{ (1-\theta) \int_{-\infty}^{c} |y - c| f(y)dy + \theta \int_{c}^{+\infty} |y - c| f(y)dy \right\}.$$

Repeating the above argument, we easily obtain:

$$\frac{\partial}{\partial c} E[\rho_\theta(Y - c)] = (1 - \theta)F(c) - \theta(1 - F(c)) = 0$$

and then q_θ as the solution of the minimization problem:

$$F(c) - \theta F(c) - \theta + \theta F(c) = 0 \implies F(c) = \theta \implies c = q_\theta.$$

1.1.3 Conditional mean, conditional quantiles and surroundings

By replacing the sorting with optimization, the above line of reasoning generalizes easily to the regression setting. In fact, interpreting Y as a response variable and $\mathbf{X}$ as a set of predictor variables, the idea of the unconditional mean as the minimizer of Equation (1.1) can be extended to the estimation of the conditional mean function:

$$\hat{\mu}(\mathbf{x}_i, \boldsymbol{\beta}) = \underset{\mu}{\operatorname{argmin}}\, E[Y - \mu(\mathbf{x}_i, \boldsymbol{\beta})]^2,$$

[1] It is worth to recall that our interest is in $y = c$ because it is where $E|Y - c|$ is minimized.

where $\mu(\mathbf{x}_i, \boldsymbol{\beta}) = E[Y|\mathbf{X} = \mathbf{x}_i]$ is the conditional mean function. In the case of a linear mean function, $\mu(\mathbf{x}_i, \boldsymbol{\beta}) = \mathbf{x}_i^\top \boldsymbol{\beta}$, the previous equation becomes:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, E[Y - \mathbf{x}_i^\top \boldsymbol{\beta})]^2$$

yielding the least squares linear regression model. The problem of minimizing the squared error can then be reduced to a problem of numerical linear algebra. Using again the *Cars93* dataset, Figure 1.3(a) shows the geometric interpretation of the least squares criterion in the case of a simple linear regression model where *Price* is the response variable and *Horsepower* is the predictor. The least squares solution provides the line that minimizes the sum of the area of the squares as determined by

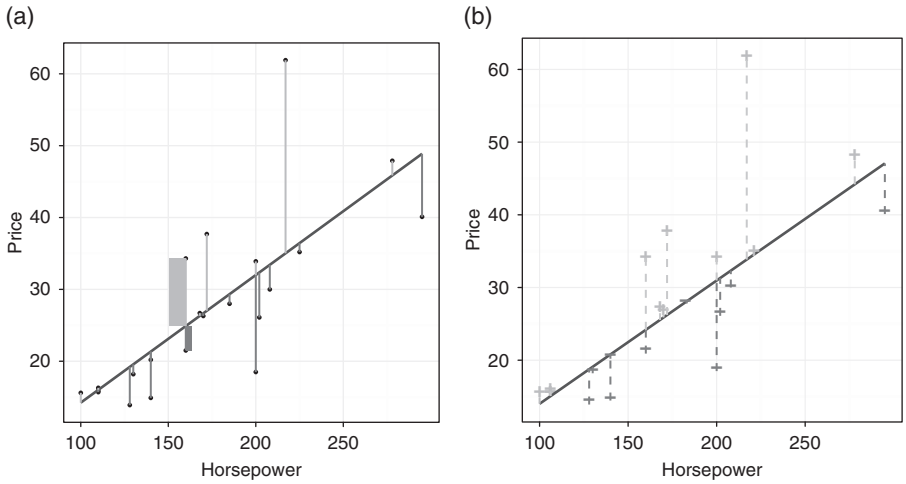

Figure 1.3 A geometric comparison of least squares (a) and least absolute (b) criteria for a simple linear regression problem. Data refer to the relationship between Price *(response variable) and* Horsepower *(predictor variable) for the* Midsize *subset of cars, as determined by the* Type *variable. Minimization of the squared errors, as required by the least squares criterion (a), is geometrically equivalent to minimizing the sum of the squares obtained by projecting the points on the regression line perpendicularly to the* x*-axis. Two of these squares, one corresponding to a negative deviation and one for a positive deviation, are shown on the plot. It is worth noticing that the choice of using non-equal scales on both the axes involves rectangles and not squares in an initial analysis of the chart. (b) shows the least absolute criterion for the median case: the minimization of the sum of the least absolute deviations from the conditional median is equivalent to the overall minimization of the lengths of the segments obtained by the* x*-axis perpendicular projection of the points. Negative* $(-)$ *and positive* $(+)$ *deviations share the same weight in determining the line, involving then a symmetric loss function. As the number of points is even, there is no point that lies on the line.*

the projection of the observed data on the same line using segments perpendicular to the x-axis. Figure 1.3(a), in particular, shows the contribution of two points to this sum of squares, one point lying below the line and one point lying above the line.

The same approach can be used to extend Equation (1.2) to the median, or the more general Equation (1.5) to the generic θ-th quantile. In this latter case, we obtain:

$$\hat{q}_Y(\theta, \mathbf{X}) = \underset{Q_Y(\theta,\, \mathbf{X})}{\operatorname{argmin}} E[\rho_\theta(Y - Q_Y(\theta, \mathbf{X}))],$$

where $Q_Y(\theta, \mathbf{X}) = Q_\theta[Y|\mathbf{X} = \mathbf{x}]$ denotes the generic conditional quantile function. Likewise, for the linear model case the previous equation becomes:

$$\hat{\beta}(\theta) = \underset{\beta}{\operatorname{argmin}} E[\rho_\theta(Y - \mathbf{X}\beta)],$$

where the (θ)-notation denotes that the parameters and the corresponding estimators are for a specific quantile θ. Figure 1.3(b) shows the geometric interpretation of this least absolute deviation criterion in the case of the model:

$$\widehat{Price} = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta) Horsepower,\ \theta = 0.5.$$

The solution to this median case is the line that minimizes the sum of absolute deviations, that is the sum of the lengths of the segments projecting each y_i onto the line perpendicularly to the x-axis.

Figure 1.4 shows the geometric interpretation of the QR solution for the case of $\theta = 0.25$ (first quartile) in Figure 1.4(a), and of $\theta = 0.75$ (third quartile) in Figure 1.4(b), respectively. While in the median case, the deviations $y_i \leq \hat{y}_i$ and the deviations $y_i > \hat{y}_i$ give the same contribution to the criterion, in the first quartile and third quartile cases they bring an asymmetric contribution: for $\theta = 0.25$, the deviations of $y_i \leq \hat{y}_i$ have weight $1 - \theta = 0.75$ with respect to the deviations corresponding to $y_i > \hat{y}_i$, whose weights are $\theta = 0.25$, with m points lying exactly on the line, where m is equal to the number of parameters in the model. This asymmetric weighting system involves an attraction of the line towards the negative deviation. The same happens, inverting the weights and then the direction of attraction, for the case of $\theta = 0.75$. The mathematical formulation of the problem leads to the solution of a linear programming problem. Its basic structure, and the counterpart algorithm solution, will be presented in the next chapter (see Sections 2.1.2 and 2.1.3).

1.2 The simplest QR model: The case of the dummy regressor

In order to introduce quantile regression, it is useful to begin by illustrating the simplest form of linear model; that is a model with a quantitative response variable and a dummy predictor variable. This simple setting aims to study differences in the response variable between the two groups as determined by the dichotomous predictor variable.

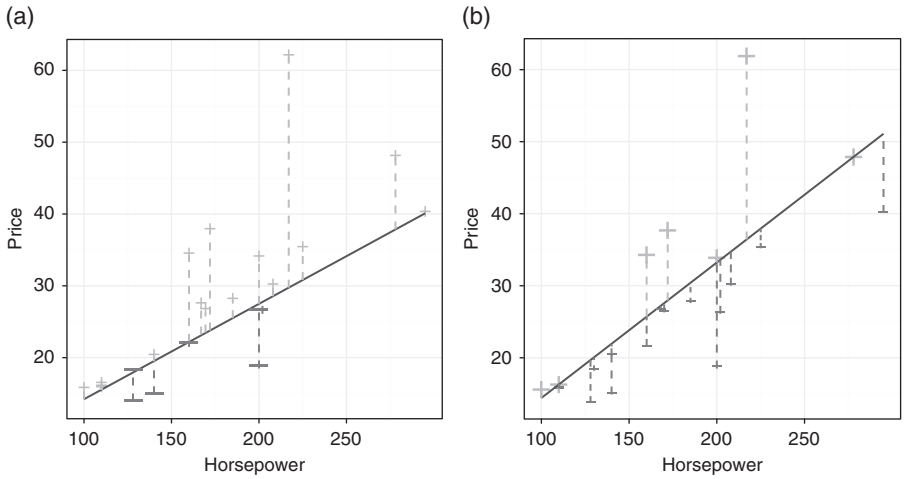

Figure 1.4 The geometric interpretation of the asymmetric least absolute deviation criterion in the case of the conditional first quartile ($\theta = 0.25$*), in (a), and of the conditional third quartile (*$\theta = 0.75$*), in (b). The use of an unbalanced weighting system (weight equal to (1*$-\theta$*) for the sum of negative deviations and weight equal to* θ *for the sum of positive deviations) determines, in the two cases, a different attraction of points lying below or above the regression line. Such different attraction of the two subsets of points is depicted using sizes proportional to the corresponding subset weight. The case of an asymmetric loss function, thus involves a line separating points with almost* θ*% lying below the line and the remainder (1*$-\theta$*)% lying above the line. The explanation of this inaccurate partition is essentially due to the linear programming solution of the problem and is given in detail in Chapter 2. It is also worth noticing that* $m = 2$ *points lies exactly on the line, where* m *is equal to the number of model parameters.*

To illustrate this simple model, we refer again to the *Cars93* dataset in order to compare the distribution of the *Price* of the cars between the two groups of USA and non-USA company manufacturers. In particular, the dummy variable *Origin*, which assumes a unit value for non-USA cars and zero for USA cars, is used as regressor.

Figure 1.5 shows the dotplots for the two groups, USA cars represented on the left-hand side and non-USA cars on the right-hand side. *Price* means for the two distributions are depicted by asterisks while the three quartiles are shown using a box bounded on the first and third quartiles, the box sliced on the median. From the analysis of the two groups' dotplots in Figure 1.5, it is evident that the two samples share a similar mean, but the long right tail for the *non-USA* car distribution gives rise to strong differences in the two extreme quartiles. A different view of the difference between the two distributions is offered by the plot of the *Price* density for the two samples, shown in Figure 1.6(a), which shows a right heavy tail for the non-USA cars distribution. A third graphical representation frequently used to

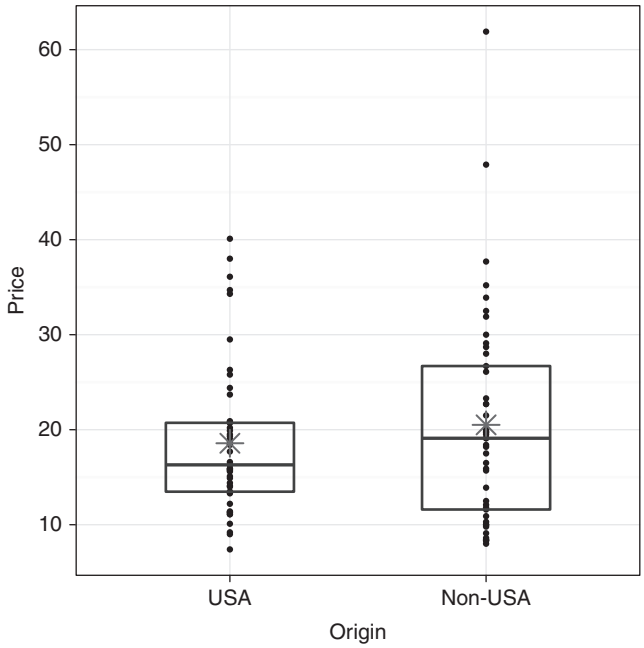

Figure 1.5 Comparison of the distribution of the Price *variable for the USA and non-USA manufactured cars, the two groups being specified by the* Origin *variable (*Cars93 *dataset). On each dotplot, an asterisk for the mean is superimposed. The two boxes are bounded on the first and third quartiles of each distribution and sliced on the corresponding medians. The right boxplot shows a larger right tail for the non-USA subset.*

compare two datasets is the Q–Q plot (Das and Bhattacharjee 2008), which represents the quantiles of the first dataset on the x-axis versus the quantiles of the second dataset on the y-axis, along with a 45° reference line. If the two datasets share a common distribution, the points should fall approximately along the reference line, if the points fall under (above) the line the corresponding set shows a shorter (longer) tail in that part of the distribution. Figure 1.6(b) shows the Q–Q plot for the *Price* of the cars comparing their origin. The representation offers the same information as the density plot, allowing to evaluate shifts in location and in scale, the presence of outliers and differences in tail behavior. Unlike the density plot, which requires one to set the kernel width, the Q–Q plot does not require any tuning parameter. Moreover, it will turn out to be an enlightening representation for the QR output in the case of a simple model consisting of a dummy predictor variable.

It is well known that in this simple case of a model with a unique dummy predictor variable, the classical least square regression:

$$\widehat{Price} = \hat{\beta}_0 + \hat{\beta}_1 Origin$$

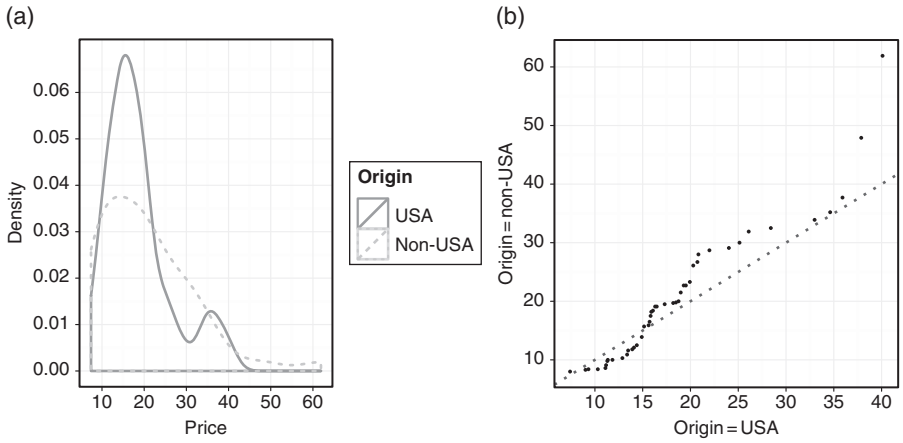

Figure 1.6 Two alternatives to boxplots for comparing distribution: density plot (a) and Q–Q plot (b). Again for the Cars93 *dataset, the two plots refer to the comparison of the* Price *variable for the two subsets as determined by the country manufacturer (USA vs non-USA origin). As well as the dotplots in Figure 1.5, both the density plot and Q–Q plot show a right heavy tail for the non-USA cars. The Q–Q plot in particular, is obtained by depicting quantiles of the USA subset of cars (*x*-axis) and quantiles of the non-USA subset (*y*-axis). It allows us to grasp the QR output in the case of a simple model with a dummy variable.*

is equivalent to the mean comparison between the two groups of USA and non-USA manufactured cars, providing the same results for the classical two samples t-test in the case of inference.

Likewise, the estimation of the QR model:

$$\widehat{Price}_\theta = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta) Origin \tag{1.6}$$

for different values of $\theta \in [0, 1]$ permits us to obtain an estimation of the *Price* quantiles for the two groups of cars. Using the coding USA = 0 and non-USA = 1 for the *Origin* indicator variable in Equation (1.6), it is straigthforward that the estimated price for the USA cars is then:

$$\widehat{Price}_\theta = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta) \times \mathbf{0} = \hat{\beta}_0(\theta),$$

while for the non-USA subset the model becomes:

$$\widehat{Price}_\theta = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta) \times \mathbf{1} = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta).$$

$\hat{\beta}_0(\theta)$ thus provides the estimation of the conditional θ quantile of the *Price* for USA cars, while, for the non-USA cars, the conditional θ quantile is obtained through

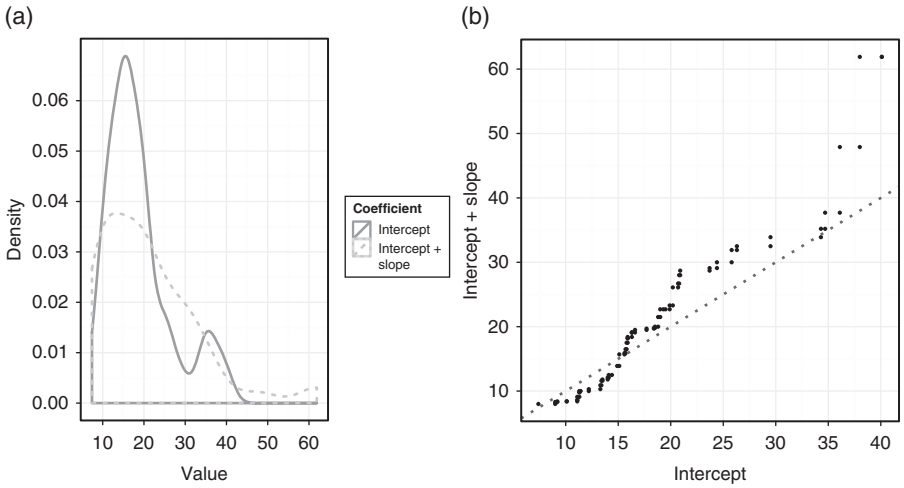

Figure 1.7 Density plot (a) and Q–Q plot (b) for the Price *variable for the two groups (USA vs non-USA) of cars, estimated through the QR model with the use of only the indicator* Origin *variable as predictor. In this simple setting, the set of intercepts for different values of* θ *estimates the* Price *distribution for the USA group while the sum of the intercepts and slopes provides the* Price *estimation for the non-USA group of cars. The QR coefficients correspond to 90 different values of* θ *in [0, 1]. The plots are practically equivalent to the observed ones (see Figure 1.6).*

the combination of the two coefficients; that is as $\hat{\beta}_0(\theta) + \hat{\beta}_1(\theta)$. By varying θ in $[0, 1]$, the set of estimated intercepts offers an estimate of the *Price* distribution for the USA cars. For the non-USA cars, price is obtained through the sum of the sets of intercepts and slopes for the different θ. For this example, using 90 different equally spaced quantiles in the range [0,1], the QR intercepts and slopes are here used to estimate the *Price* distribution for the two groups of cars. The density plot and the Q–Q plot of the estimated distribution reported in Figure 1.7(a) and (b), respectively, are practically equivalent to the observed ones shown in Figure 1.6. From Figure 1.7, it is evident how the different sets of coefficients corresponding to different values of θ are able to fully describe the *Price* distribution conditional on the two levels of the *Origin* variable. In this simplified setting, such conditional distributions represent the estimation of the *Price* distribution for the two different groups of cars computed using the predictor indicator variable.

Numerical results for this example are reported in Table 1.1: the first two rows show the estimated QR coefficients for five selected quantiles, $\theta \in (0.1, 0.25, 0.5, 0.75, 0.9)$, and the third and fourth rows report the estimated quantiles for the two groups through the combination of the two regression coefficients; finally, the fifth and sixth rows show the quantiles computed on the observed data.

Table 1.1 QR coefficients (first two rows) for the simple model $\widehat{Price}_\theta = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta)$ Origin. The combination of the two estimated coefficients for the different θ allows us to estimate the corresponding *Price* quantiles for the two groups of cars (third and fourth row). The last two rows show the unconditional *Price* quantiles.

	$\theta = 0.1$	$\theta = 0.25$	$\theta = 0.5$	$\theta = 0.75$	$\theta = 0.9$
Parameter estimates					
(Intercept)	11.1	13.4	16.3	20.8	34.3
Origin (non-USA)	−2.5	−1.8	2.8	5.9	−0.4
(Intercept)	11.1	13.4	16.3	20.8	34.3
Intercept + Origin (non-USA)	8.6	11.6	19.1	26.7	33.9
Unconditional *Price* quantiles					
Origin (USA)	11.1	13.5	16.3	20.7	30.9
Origin (non-USA)	8.8	11.6	19.1	26.7	33.3

1.3 A slightly more complex QR model: The case of a nominal regressor

The simple model introduced in the previous section can be slightly complicated by replacing the indicator predictor variable by a multiple level categorical variable with *g categories*. The classical regression for this setting is substantially equivalent to a mean comparison among the *g* groups as determined by the levels of the nominal regressors. Similarly, QR allows us to compare the different quantiles among the different *g* groups.

In order to show this setting, again using the *Cars93* dataset, we consider in this section the Airbags variable as a regressor to predict the car price. For such a variable, the three levels – *None*, *Driver only* and *Driver and Passenger* – correspond to the number of *Airbags* installed in the car.

It is widely known that in a regression model, to deal with a *g* level nominal variable requires the introduction of a $g - 1$ dummy variable (Scott Long 1997). Such a coding system then produces the QR model:

$$\widehat{Price}_\theta = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta) I(Driver\ only) + \hat{\beta}_2(\theta) I(Driver\ and\ Passenger),$$

where $I()$ is the indicator function returning 1 if the particular unit assumes the value in parenthesis and 0 otherwise. Moreover, it is well known that the first level, the so-called reference level, is associated with the model intercept.

The previous equation provides the estimation of the conditional θ quantile of the *Price* and the coding system allows us to easily obtain the estimation for the three different levels of the regressor. For the *None* level, associated with the intercept, the model reduces to:

$$\widehat{Price}_\theta = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta) \times \mathbf{0} + \hat{\beta}_2(\theta) \times \mathbf{0} = \hat{\beta}_0(\theta).$$

Table 1.2 QR coefficients (first three rows) for the simple model $\hat{Price}_\theta = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta)I(\textit{Driver only}) + \hat{\beta}_2(\theta)I(\textit{Driver and Passenger})$, where *Price* is the dependent variable and *Airbags* is the regressor. The three estimated coefficients are shown in the first three rows; the combination of the three estimated coefficient for the different θ allows us to estimate the corresponding *Price* quantiles for the two groups of cars (rows four to six); the last three rows show the unconditional *Price* quantiles.

	$\theta = 0.1$	$\theta = 0.25$	$\theta = 0.5$	$\theta = 0.75$	$\theta = 0.9$
(Intercept)	8.4	9.2	12.2	16.3	19.5
Airbags (*Driver only*)	3.4	6.4	7.7	10.0	12.4
Airbags (*Driver and Passenger*)	7.4	8.5	12.2	18.9	20.6
(Intercept)	8.4	9.2	12.2	16.3	19.5
Intercept + Airbags (*Driver only*)	11.8	15.6	19.9	26.3	31.9
Intercept + Airbags (*Driver and Passenger*)	15.8	17.7	24.4	35.2	40.1
Unconditional *Price* quantiles					
Airbags (*None*)	8.4	9.4	11.9	16.2	19.4
Airbags (*Driver only*)	11.9	15.6	19.9	26.2	31.5
Airbags (*Driver and Passenger*)	16.6	18.2	25.5	35.4	38.9

A similar line of reasoning leads to the estimation of the conditional quantiles for the *Drivers only* group:

$$\widehat{Price}_\theta = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta) \times \mathbf{1} + \hat{\beta}_2(\theta) \times \mathbf{0} = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta),$$

and, for the *Driver and Passenger* group:

$$\widehat{Price}_\theta = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta) \times \mathbf{0} + \hat{\beta}_2(\theta) \times \mathbf{1} = \hat{\beta}_0(\theta) + \hat{\beta}_2(\theta).$$

Therefore, in a model with a nominal regressor, for a given quantile θ, the combination of the intercept with the different slopes, allows us to estimate the conditional quantile of the response variable. The estimated effect of the particular group is obtained using the dummy variable associated with the particular slope.

Numerical results for the simple model previously considered are in Table 1.2: the three coefficients are on the first three rows of the table; rows four to six show the estimated conditional quantile for the *Price* variable, while the last three rows show the unconditional *Price* quantiles. Using again 90 quantiles in $(0, 1)$, the estimated conditional densities of the *Price* variable for the three groups of cars are represented

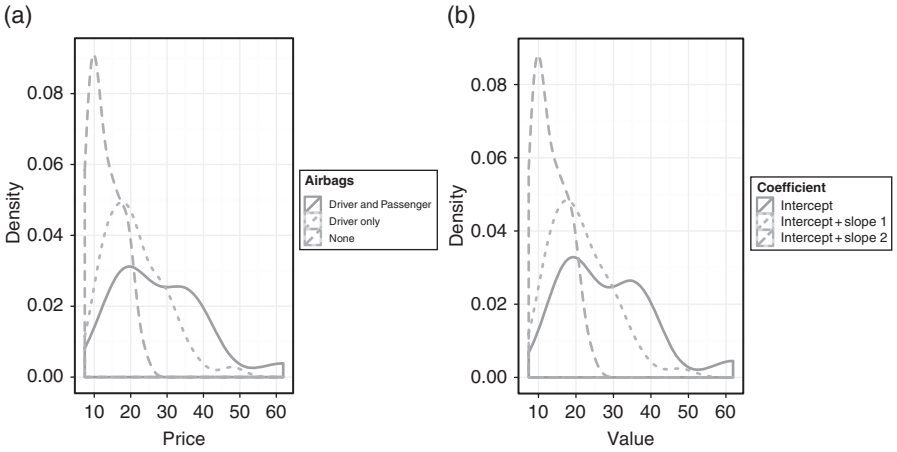

Figure 1.8 Observed density plot (a) for the Price *variable: different lines refer to different groups determined by the nominal variable,* Airbags*. Through the combination of the different coefficients it is possibile to estimate the quantile of the* Price *variable conditional on the* Airbags *variable (b). Comparing (a) and (b), it is evident that the estimated densities are practically equivalent to the observed ones.*

in Figure 1.8(b). They are practically equivalent to three observed distributions, shown in Figure 1.8(a).

1.4 A typical QR model: The case of a quantitative regressor

Once that the general idea behind QR and the two models involving a nominal regressor have been illustrated, we can move to the more common regression setting with a quantitative regressor. A simple example for the *Cars93* dataset is offered by the model:

$$\widehat{Price} = \hat{\beta}_0 + \hat{\beta}_1 Passengers;$$

that is studying the car price starting from the number of passengers they are registered to carry. For the sake of illustration, we restrict our attention to the cars licensed to carry 4, 5, and 6 passengers. Figure 1.9(a) depicts the scatterplot of the *Passengers* variable (x-axis) vs the *Price* variable (y-axis). Given the nature of the *Passengers* variable (discrete variable assuming only three different values), such a scatterplot can be easily interpreted as the combination of three dotplots for the *Price* variable corresponding to the three different numbers of passengers; that is it depicts the three conditional distributions of the *Price* on the *Passengers*. From Figure 1.9(a), differences in location and variability for the three groups are evident. Such differences

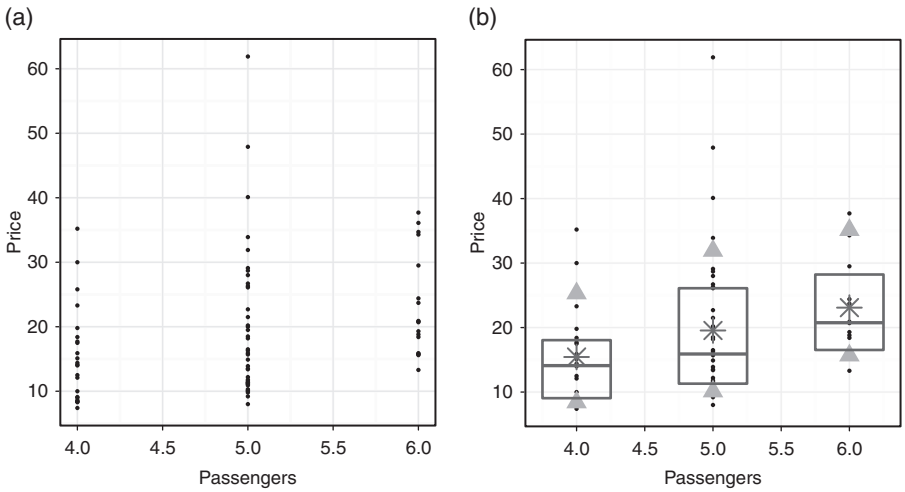

Figure 1.9 Scatterplot of Passengers *vs* Price *for the* Cars93 *dataset (a). The* Passengers *variable assumes only three different values for the considered subset: the plot can then be easily interpreted as the distributions of the* Price *variable conditional on the values of the* Passengers *variable. In (b), the location of the mean (asterisk) and of five quantiles for the three conditional distributions are superimposed on the scatterplot. In particular, the three quartiles are depicted using a box, with boundaries on the two extreme quartiles and sliced on the median, and the two extreme quantiles (*$q_{0.1}$ *and* $q_{0.9}$*) are shown with triangles.*

become clearer by superimposing the information of the mean and of five quantiles of the distributions [Figure 1.9(b)]. The five quantiles used are the common three quartiles (depicted using the box with boundaries on the two extreme quartiles and sliced on the median) and two symmetric extreme quantiles, $q_{0.1}$ and $q_{0.9}$, represented by the triangles in Figure 1.9(b), along with the means (depicted using asterisks).

Using five such values for θ, the QR simple linear model is estimated, and the results are superimposed on Figure 1.10. The ordinary least squares (OLS) line is recognizable by looking at the position of the asterisks. The patterns in Figure 1.10 reveal differences in location, variability, and shape for the different values of the number of passengers for which the car is patented. Such differences in the relationship between the *Price* variable and the *Passengers* variable are more evident by looking at the corresponding three density plots for the *Price* variable shown in Figure 1.11, along with the marginal density plot for the *Price* variable. Each panel of Figure 1.11 refers to a conditional distribution, the latter depicting the marginal distribution: the observed *Price* variable and the QR estimated *Price* variable, conditional on the values of the number of passengers, are represented in each subplot. From this figure it is evident that the estimated QR distribution is able to fully describe the patterns of the *Price*, both for the conditional cases and for the

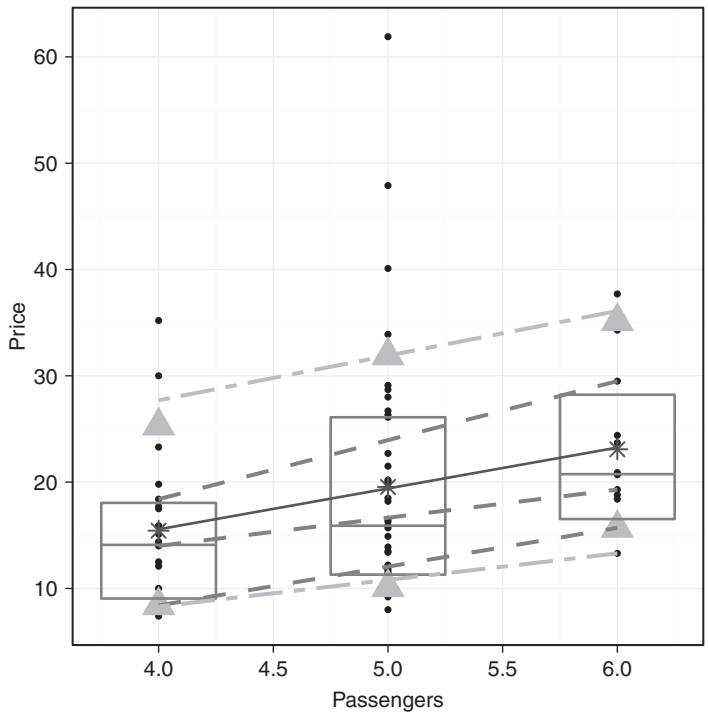

Figure 1.10 QR lines for $\theta = (0.1, 0.25, 0.5, 0.75, 0.9)$ *for the simple linear model* $\widehat{\text{Price}} = \beta_0(\theta) + \beta_1(\theta)$ Passengers. *The OLS line is recognizable by looking at the position of the asterisks, which represent the averages for the three groups.*

marginal case. Density estimation is obtained by simply combining the different estimates for different values of θ and selecting the 'best' model according to the deviation between the observed value and the estimated values. More details on the technicalities behind the density estimation are illustrated in Chapter 4, Section 4.2.

As already mentioned, QR is an extension of the classical estimation of conditional mean models to conditional quantile functions; that is an approach allowing us to estimate the conditional quantiles of the distribution of a response variable Y in function of a set $\mathbf{X}$ of predictor variables.

In the framework of a linear regression, the QR model for a given conditional quantile θ can be formulated as follows:

$$Q_\theta(Y|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}(\theta),$$

where $0 < \theta < 1$ and $Q_\theta(.|.)$ denotes the conditional quantile function for the θ-th quantile.

The parameter estimates in QR linear models have the same interpretation as those of any other linear model, as rates of change. Therefore, in a similar way to the OLS model, the $\beta_i(\theta)$ coefficient of the QR model can be interpreted as the rate of

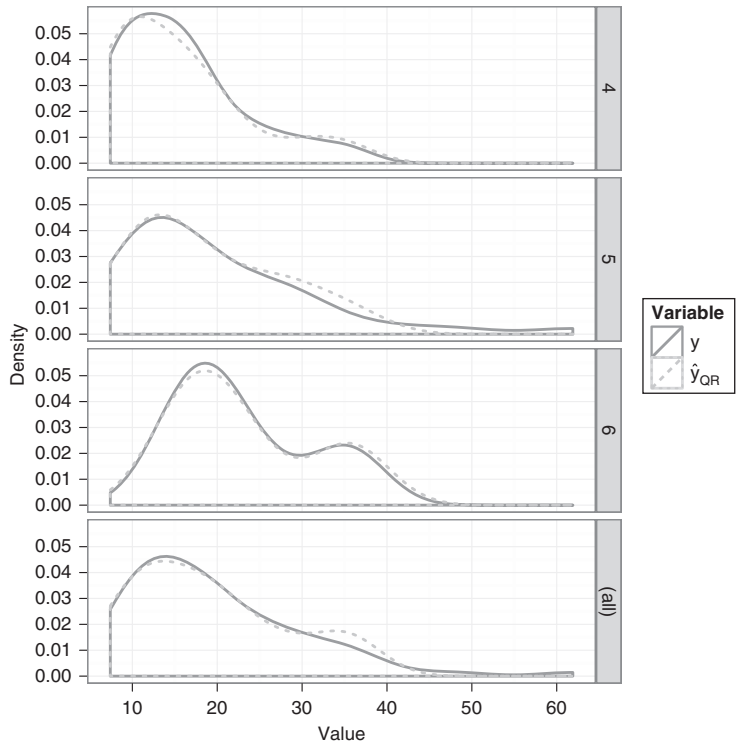

Figure 1.11 The two different lines in each panel represent the densities for the observed Price *variable and the estimated* Price *variable, the latter obtained through QR. Different panels depict the conditional distributions of the three values of the* Passengers *variable, and the bottom panel depicts the marginal distributions.*

change of the θ-th quantile of the dependent variable distribution per unit change in the value of the i-th regressor:

$$\beta_i(\theta) = \frac{\partial Q_\theta(Y|\mathbf{X})}{\partial \mathbf{x}_i}.$$

In Table 1.3, OLS and QR results obtained for the previous simple linear model are limited to the five quantiles $\theta = (0.1, 0.25, 0.5, 0.75, 0.9)$.

Table 1.3 OLS and QR coefficients for the simple linear model predicting car *Price* through the *Passengers* they are licensed to carry.

	OLS	$\theta = 0.1$	$\theta = 0.25$	$\theta = 0.5$	$\theta = 0.75$	$\theta = 0.9$
(Intercept)	0.18	−1.7	−6.20	3.40	−3.80	10.9
Passengers	3.84	2.5	3.65	2.65	5.55	4.2

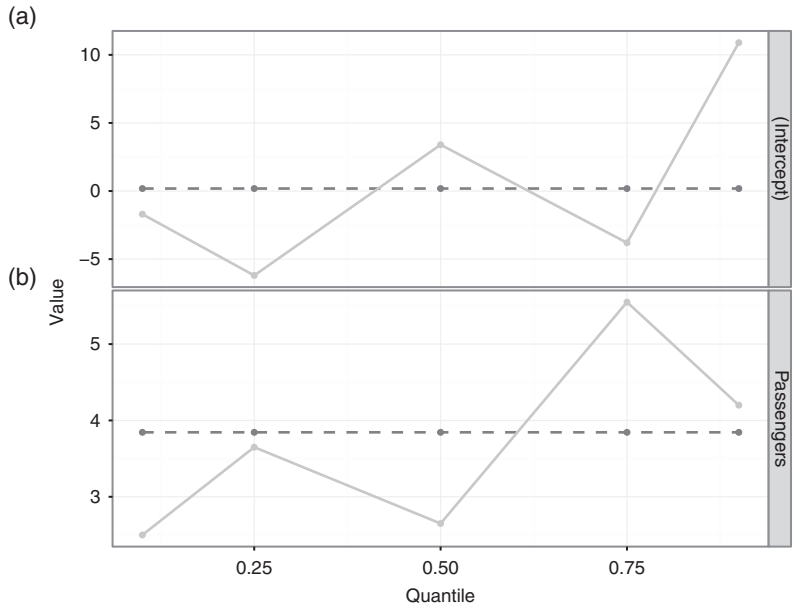

Figure 1.12 Graphical representation of the QR intercept (a) and slope (b) behavior (solid line) for the simple linear model $\widehat{\text{Price}} = \hat{\beta}_0 + \hat{\beta}_1$ Passengers. *The conditional quantiles are represented on the* x*-axis and the coefficient values on the* y*-axis. A dashed line for both coefficients is placed at the corresponding OLS estimates.*

The effect of *Passengers* increases moving from the lower part to the upper part of the *Price* distribution with the exception of the median: this confirms the conjecture of a location-scale model, namely of changes both in the central tendency and in the variability of the response variable. If in the central location (both in mean and in median) a one-unit increase of the *Passengers* variable leads to an increase in the *Price* by 3.84 and 2.65, respectively, such an effect is equal to 3.65 at the 25-th quantile and increases further to 5.55 at the 75-th quantile. In practical terms, QR results suggest that the number of passengers a car is licensed to carry has a greater effect for the more expensive cars, that is for cars with a price greater than or equal to the 75-th quantile ($\hat{\beta}_{\theta=0.75} = 5.55$).

The obtained results can also be graphically inspected. A typical graphical representation of QR coefficients (Koenker 2011) permits us to observe the different behaviors of the coefficients with respect to the different quantiles. Figure 1.12 displays the obtained estimates for intercept and slope: the different conditional quantiles are represented on the *x*-axis while the coefficient values are on the *y*-axis, the solid line with filled symbols represents the point estimates for the five distinct conditional quantiles, while the dashed line is placed at the value of the OLS estimate. With respect to the slope [Figure 1.12(b)], the previous comments on the OLS

and QR results are confirmed: up to the median, QR slope estimates are lower than the OLS counterpart, while the effect of *Passengers* on *Price* seems slightly stronger for the upper quantile. The intercept estimates seem more dependent on the particular quantile.

This chapter restricts itself to simple linear regression models in order to introduce the QR logic. It is worth noticing that the extension to multiple regression follows the same line of reasoning of classical regression (Gujarati 2003; Weisberg 2005). Using the same rules of OLS regression, a categorical explanatory variable can also be included in the model: it is preliminarily transformed into dummy variables, and all of them, except for one, are included in the model, the excluded category being the reference category in the interpretation of the results (Scott Long 1997). More realistic and interesting applications of QR, along with the inference tools, will be presented in the following chapters, starting from Chapter 3.

Finally, for all the examples shown up to now, five conditional quantiles, $\theta = (0.1, 0.25, 0.5, 0.75, 0.9)$, have been used for synthesis purposes. However, although it is possible to extract an infinite number of quantiles, in practice, a finite number is numerically distinct, the so-called *quantile process*; it depends on the number of observations increasing roughly linearly with them (Koenker and D'Orey 1987; Portnoy 1991; Buchinsky 1998). In the case of a particular interest in very high quantiles, larger samples are required (Cade and Noon 2003). For technical details about the quantile process the reader is referred to Chapter 2, Section 2.3.3.

1.5 Summary of key points

- While classical regression gives only information on the conditional expectation, quantile regression extends the viewpoint on the whole conditional distribution of the response variable.
- The mean and the quantiles are particular centers of a distribution minimizing a squared sum of deviations and a weighted absolute sum of deviations, respectively. This idea is easily generalized to the regression setting in order to estimate conditional mean and conditional quantiles.
- A simple linear regression model with a quantitative response variable and a dummy regressor allows us to compare the mean (classical regression) and the quantiles (quantile regression) between the two groups determined by the dummy regressor. Using a nominal regressor, such a comparison is among the g groups corresponding to the g levels of the nominal variable.
- For QR, as well as for classical regression, the parameter estimates in linear models are interpretable as rates of changes: the $\beta_i(\theta)$ coefficient can be interpreted as the rate of change of the θ-th quantile of the dependent variable distribution per unit change in the value of the i-th regressor.
- QR provides location, scale, and shape shift information on the conditional distribution of the response variable.

- QR allows us to approximate the whole distribution of a response variable conditional on the values of a set of regressors.
- We have seen how QR, offering information on the whole conditional distribution of the response variable, allows us to discern effects that would otherwise be judged equivalent using only conditional expectation. Nonetheless, the QR ability to statistically detect more effects can not be considered a panacea for investigating relationships between variables: in fact, the improved ability to detect a multitude of effects forces the investigator to clearly articulate what is important to the process being studied and why.

References

Buchinsky M 1998 Recent advances in quantile regression models: a practical guideline for empirical research. *Journal of Human Resources* **33**, 88–126.

Cade BS and Noon BR 2003 A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment* **1**(8), 412–420.

Das KK and Bhattacharjee D 2008 *101 Graphical Techniques*. Asian Books.

Gilchrist WG 2000 *Statistical Modelling with Quantile Functions*. Chapman & Hall / CRC.

Gujarati DN 2003 *Basic Econometrics*, International Edition. McGraw-Hill.

Hao L and Naiman DQ 2007 *Quantile Regression*. SAGE Publications, Inc.

Koenker R 2001 Linear hypothesis: regression (quantile). In *International Encyclopedia of Social & Behavioral Sciences* (Smelser NJ and Baltes PB eds), 8893–8899. Elsevier.

Koenker R 2005 *Quantile Regression*. Cambridge University Press.

Koenker R 2011 *quantreg: Quantile Regression*. R package version 4.76. http://CRAN.R-project.org/package=quantreg.

Koenker R and Basset G 1978 Regression quantiles. *Econometrica* **46**(1), 33–50.

Koenker RW and D'Orey V 1987 Algorithm AS 229: computing regression quantiles. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **36**(3), 383–393.

Lock RH 1993 1993 New car data. *Journal of Statistics Education* **1**(1).

Portnoy S 1991 Asymptotic behavior of the number of regression quantile breakpoints. *SIAM Journal on Scientific and Statistical Computing Archive* **12**(4), 867–883.

Scott Long J 1997 *Regression Models for Categorical and Limited Dependent Variables*. SAGE Publications.

Venables WN and Ripley BD 2002 *Modern Applied Statistics with S*, 4-th Edition. Springer.

Weisberg S 2005 *Applied Linear Regression*, 3rd Edition. John Wiley & Sons, Ltd.