Quantile Regression

Relevant Texts

Roger Koenker , R., K. F. Hallock (2001), Quantile Regression. *The Journal of Economic Perspectives*, 15(4): 143-156.

Roger Koenker (2005), ***Quantile Regression***. CUP

Lingxin Hao & Daniel Q. Naiman (2006), ***Quantile Regression***. Sage Publications.

Chen, Jau-er (2010), ***Investigations on Quantile regression: Theories and Applications for Time Series Models***.

Davino, C., M. Furno, and D. Vistocco (2014). ***Quantile Regression: Theory and Application***. Wiley.

Marilena Furno, and Domenico Vistocco (2018). ***Quantile Regression: Estimation and Simulation***. Wiley.

Koenker, Chernozhukov, He and Peng (2018). ***Handbook of Quantile Regression***. CRC Press.

Joshua Angrist, Jörn-Steffen Pischke (2009), ***Mostly Harmless Econometrics***. Princeton University Press.

A. Colin **Cameron** & Pravin K. **Trivedi** (2005), ***Microeconometrics: Methods and Application***. CUP

1. Quantiles and Quantile Function

1. Unconditional mean, unconditional quantiles and surroundings

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:

$\mu = \text{argmin}_c \, E(Y-c)^2$. (here, c indicates mean). (1.1)

The median of a finite list of numbers can be found by arranging all the numbers from smallest to greatest.

The median, instead, **minimizes the absolute sum of deviations**.

In terms of a minimization problem, the median is thus:

$\text{Median} = \text{argmin}_c E|Y-c|$. (here, c indicates median). (1.2)

Using the sample observations, we can obtain the sample estimators $\hat{\mu}$ and $\widehat{Median}$ 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):

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

The quantile function is defined as its inverse:

$Q_Y(\theta) = Q(\theta) = F_Y^{-1}(\theta) = \inf\{y: F(y) > \theta\}$, for $\theta \in [0, 1]$. (1.4)

If $F(.)$ is strictly increasing and continuous, then $F^{-1}(\theta)$ is the unique real number y such that $F(y) = \theta$.

The following figure depicts the empirical CDF and its inverse, the empirical quantile function for the Price variable of the Cars93 dataset.

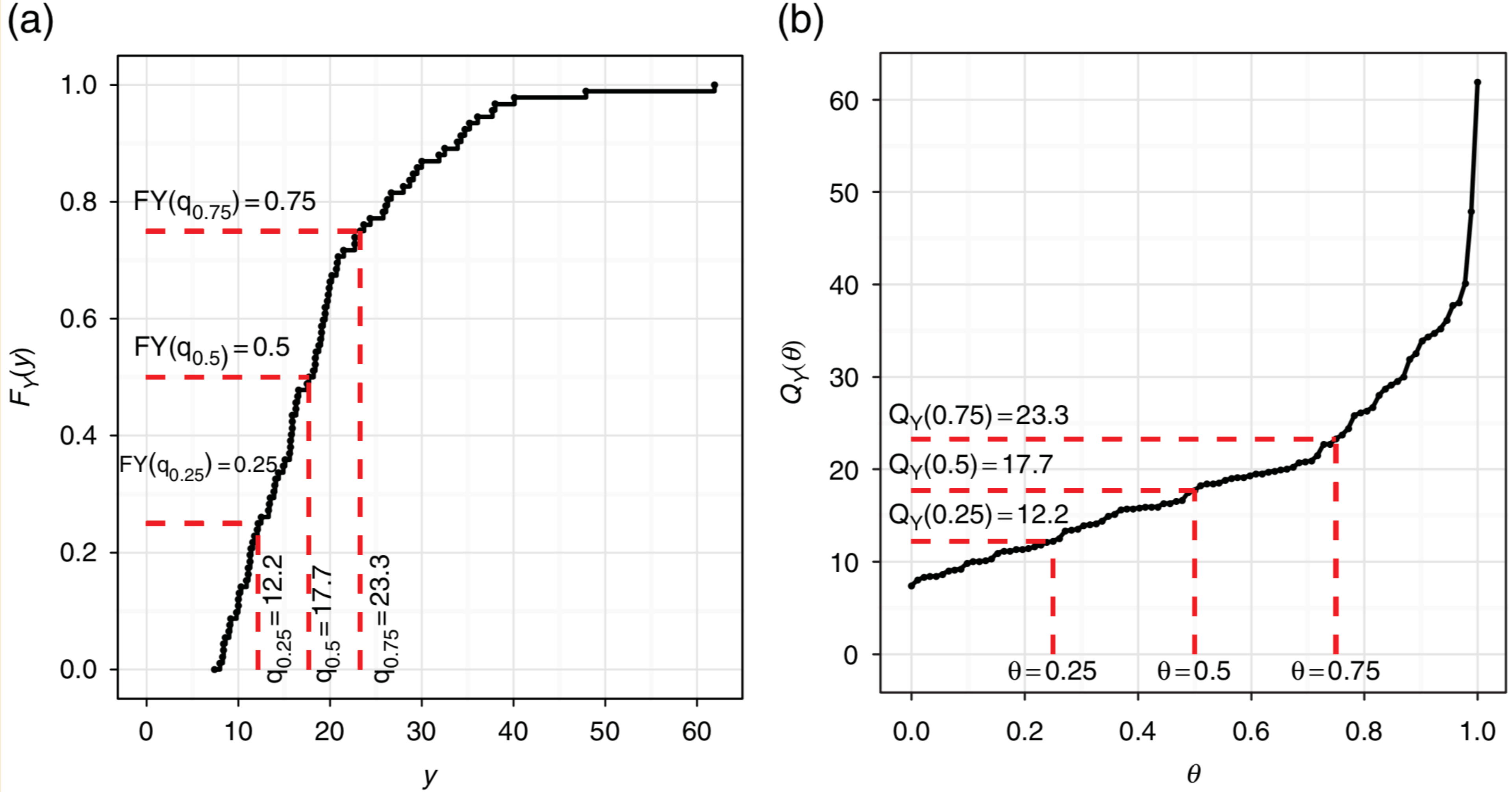
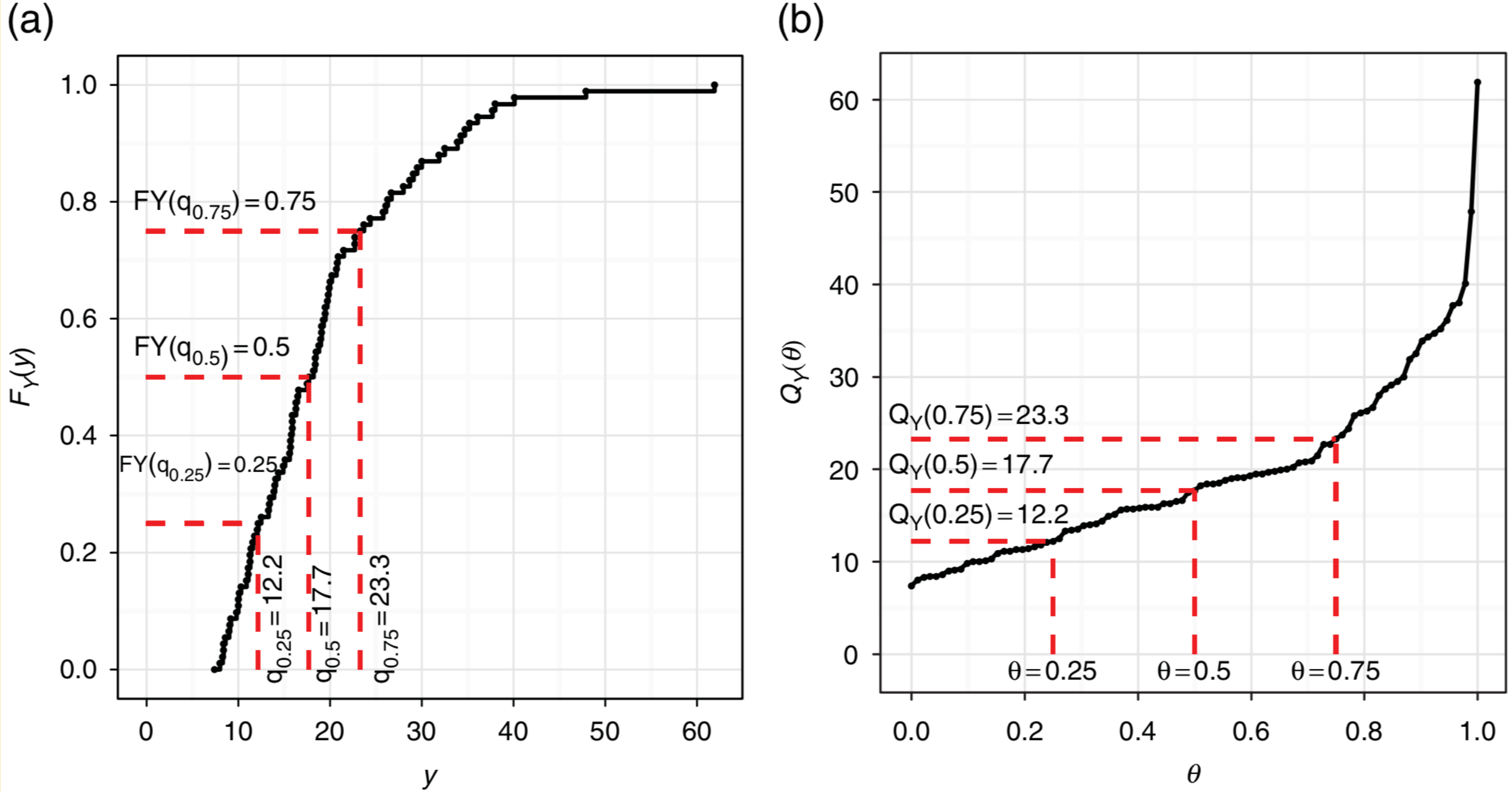
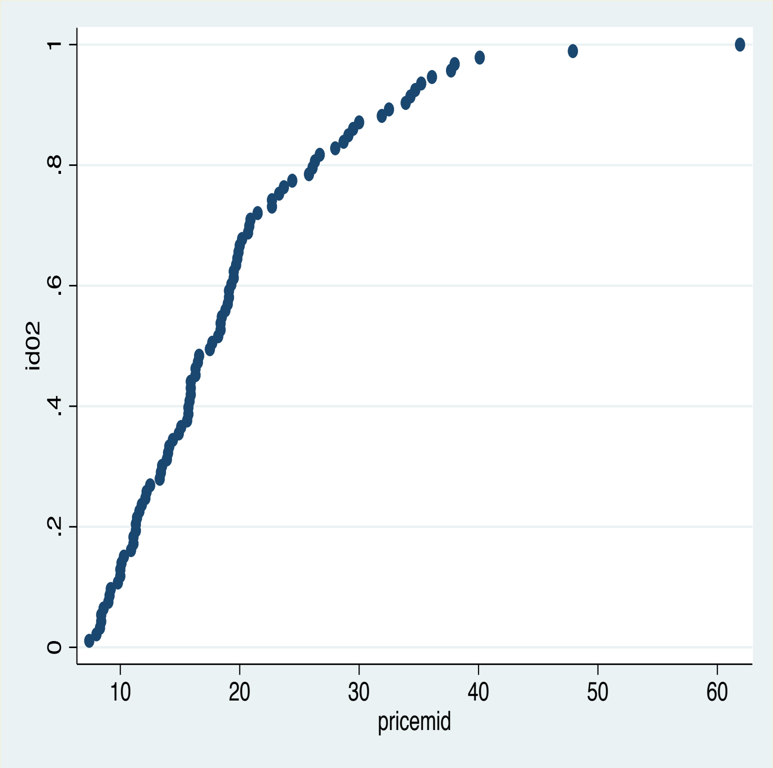


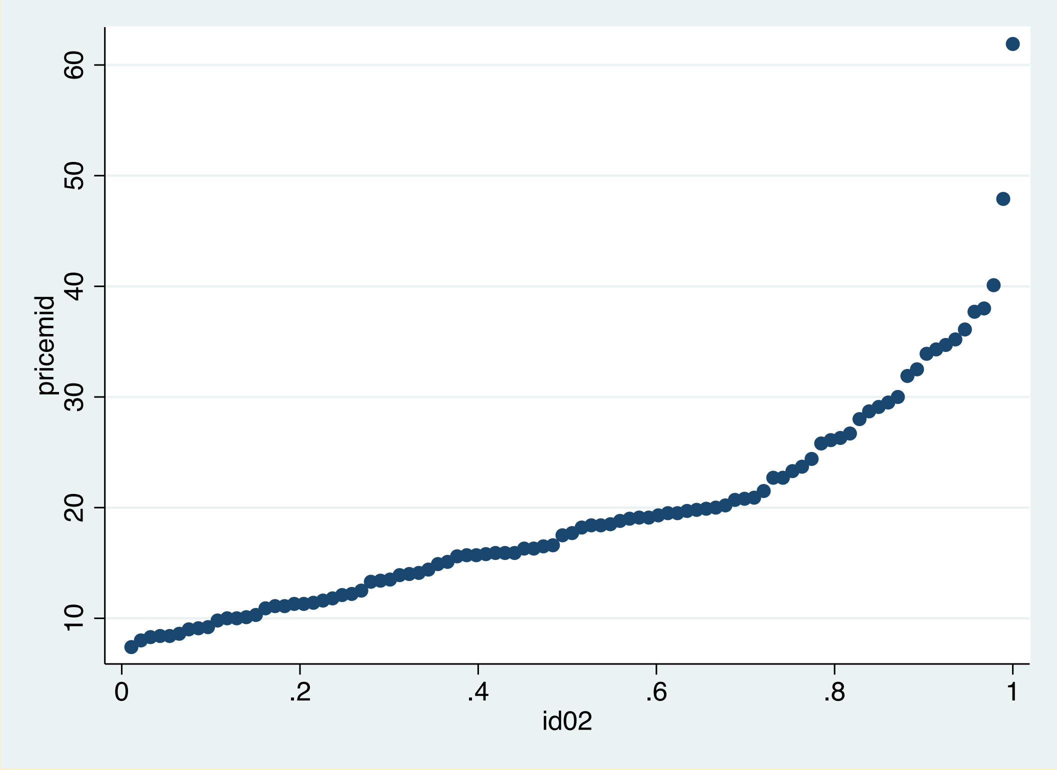
Figure 1.1 Empirical distribution function (a) and its inverse, the empirical quantile function (b), for the Price *variable of the* Cars93 *dataset. The three quartiles of* Price *are represented on the two plots: qθ corresponds to the abscissa on the* $F_Y(y)$ *plot, while it corresponds to the ordinate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



The three quantiles, θ={0.25, 0.5, 0.75}, represented on both plots point out the strict link between the two functions.

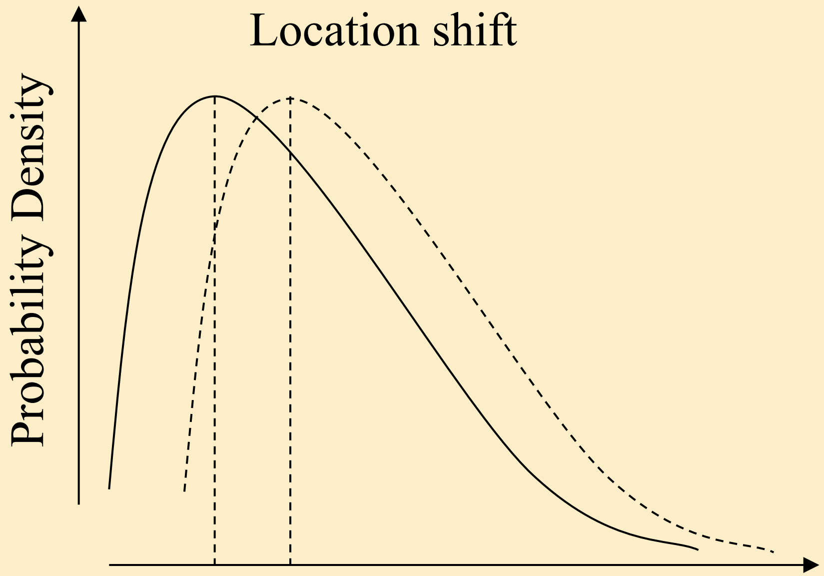


0
.2
.4
.6
.8
1
id02
10
20
30
40
50
60
pricemid



10
20
30
40
50
60
pricemid
0
.2
.4
.6
.8
1
id02

The ***location*** and ***spread*** of a distribution are inadequate for fully describing a distribution.



Suppose we are told that the mean household income for whites (W) exceeds that of blacks (B) by $20,500.

This could be described simply by a ***shift*** in the ***location*** of the distribution while retaining the shape.

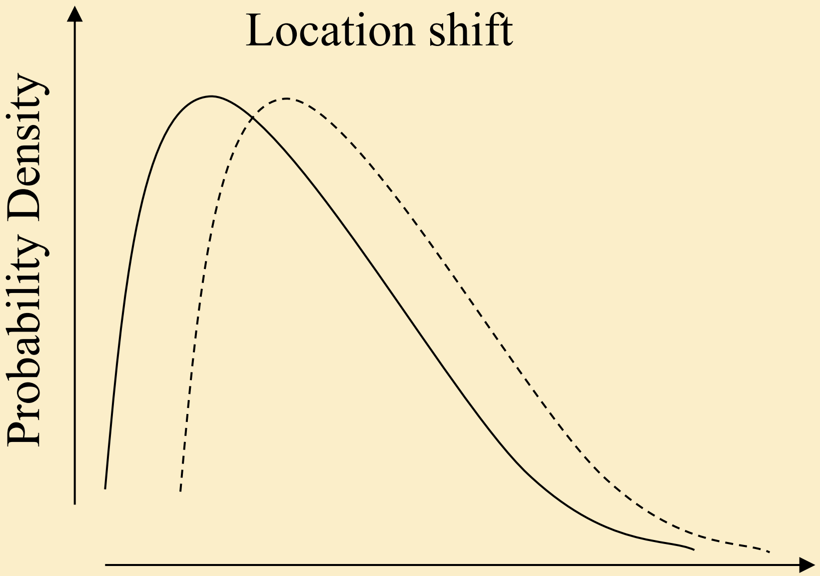
The relationship between the distributions is expressed by

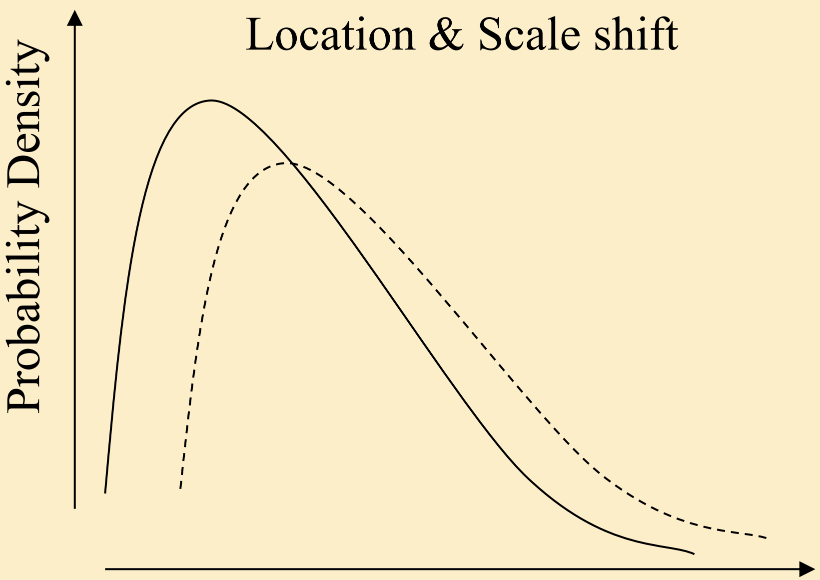
$F^B(y) = F^W(y - 20{,}500)$.

The difference in distributions may consist of a change in both ***location*** and ***scale***,

so that the relationship between the distributions takes the general form $F^B(y) = F^W(ay - c)$ for constants a and c $(a > 0)$.

This is the case when both **the mean** and **the variance of y** differ between populations W and B.





Knowledge of measures of ***location*** and ***scale***, for example, the mean and standard deviation, or alternatively the ***median*** and ***interquartile range***, enables us to compare the attribute Y between the two distributions.

As distributions become less symmetrical, more complex summary measures are needed.

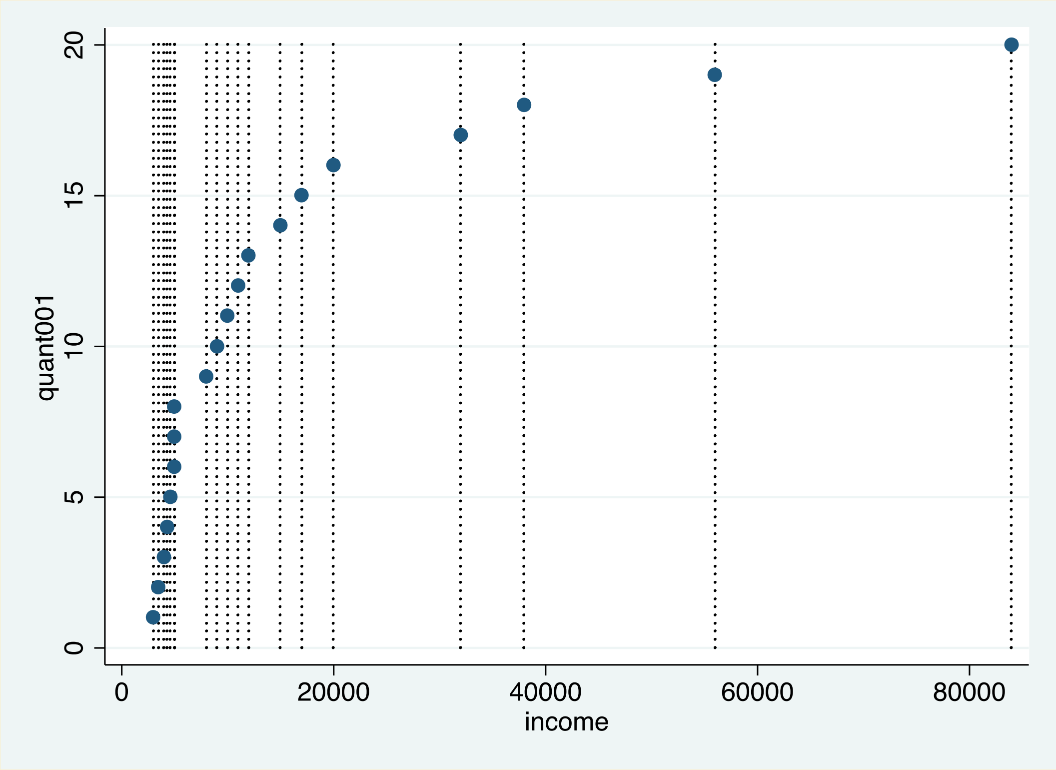
Consideration of **quantiles** and **quantile functions** leads to a rich collection of summary measures.

Continuing the discussion of a CDF, F, for some population attribute, **the p^{th} quantile of this distribution**,

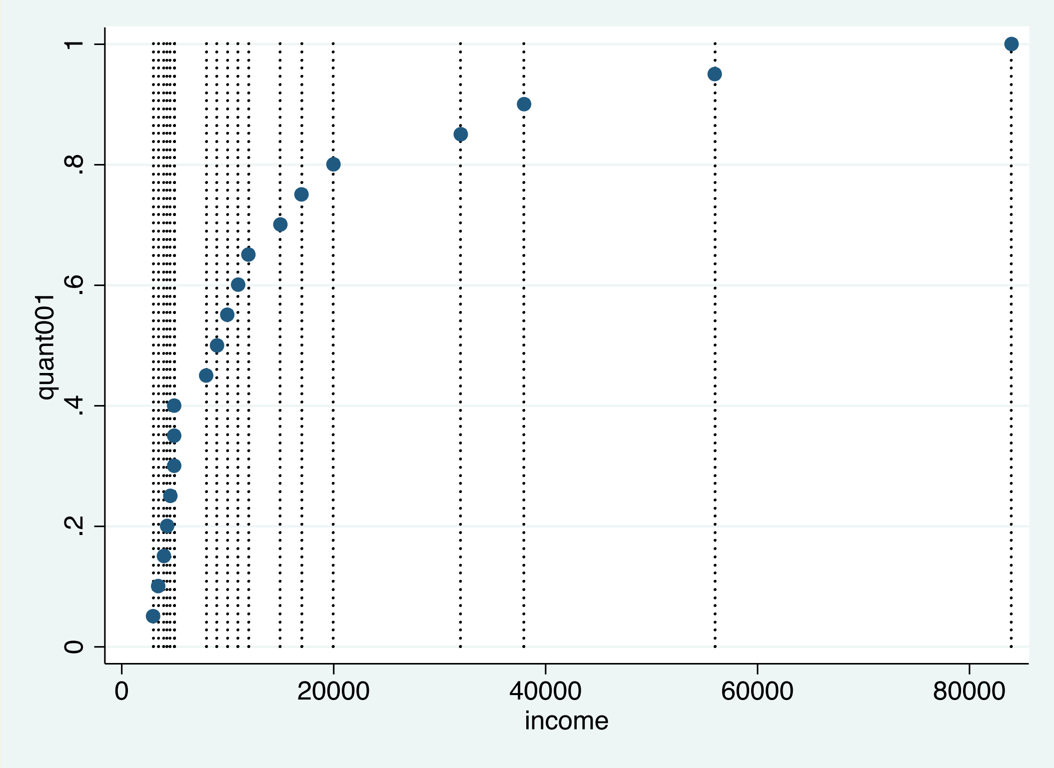
denoted by $Q^{(p)}(F)$ (or simply $Q^{(p)}$ when it is clear what distribution is being discussed), **is** the value of the inverse of the CDF at p, that is, a value of y such that $F(y) = p$.

Or simply the .5th quantile is the median.

If $F(.)$ is strictly increasing and continuous, then $F^{-1}(\theta)$ is unique real number y, such that $F(y)=\theta$.



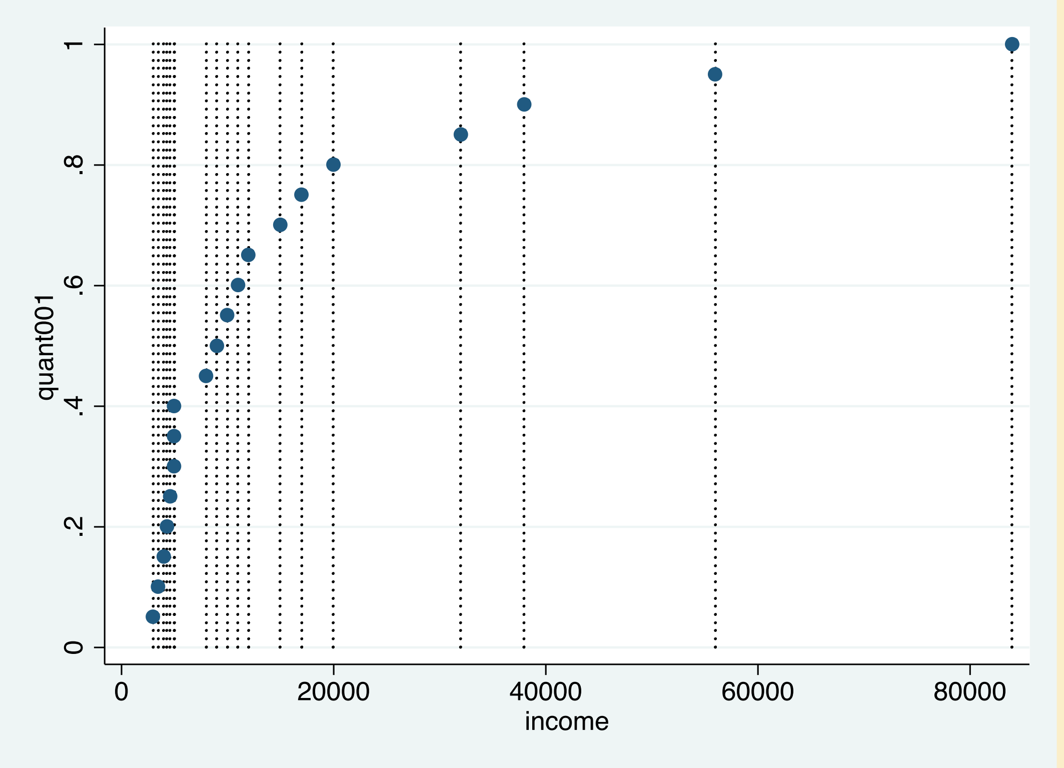
For example, in the standard normal case, $F(1.28)=0.9$, so $Q^{(0.9)}=1.28$, that is, the proportion of the population with the attribute below 1.28 is 0.9 or 90%.



Consider a sample consisting of 20 households with income of 3,000, 3,500, 4,000, 4,300, 4,600, 5,000, 5,000, 5,000, 8,000, 9,000, 10,000, 11,000, 12,000, 15,000, 17,000, 20,000, 32,000, 38,000, 56,000, 84,000.

A plot of the empirical CDF is in the right figure, which consists of one jump and several flat parts.

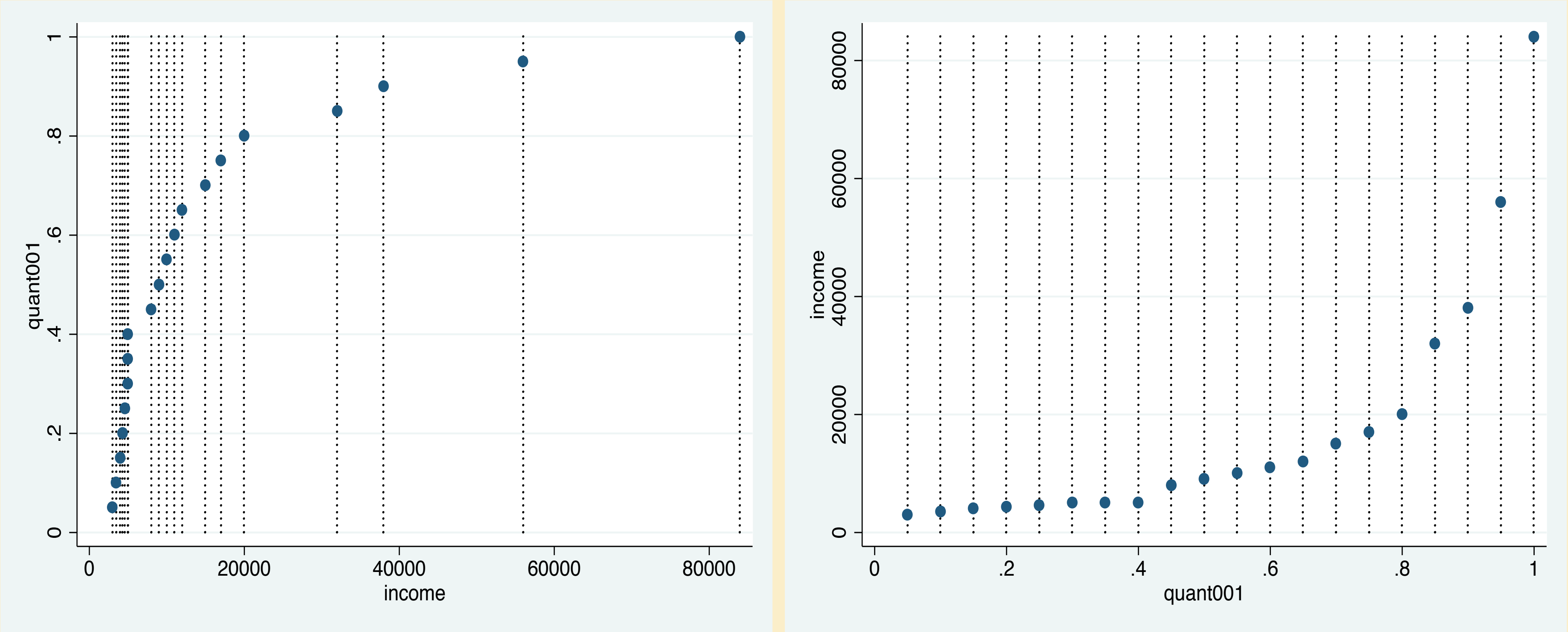
For example, there is a jump of size 3/20 at 5,000, indicating that the value of 5,000 appears three times in the sample.



There are also flat parts such as the portion between 56,000 and 84,000, indicating that there are no sample values in the interior of this interval.

Definition. The p^{th} quantile $Q^{(\mathrm{p})}$ of a CDF F is the minimum of the set of values y such that $F(y) \geq p$.

The function $Q^{(\mathrm{p})}$ (as a function of p) is referred to as the quantile function of F.



The above figures shows an example of a quantile function and the corresponding CDF.

Observe that the quantile function is a monotonic non-decreasing function that is continuous from below.

Sampling Distribution of a Sample Quantile

It is important to note how sample quantiles behave in large samples.

For a large sample $y_1, \ldots, y_n$ drawn from a distribution with quantile function $Q^{(p)}$ and probability density function $f = F'$, the distribution of $\hat{Q}^{(p)}$ is approximately normal with ***mean*** $Q^{(p)}$ and ***variance***

$$\frac{p(1-p)}{n} \cdot \frac{1}{f(Q^{(p)})^2}$$

<u>In particular, this variance of the sample distribution is completely determined by **the probability density evaluated *at the quantile***.</u>

The dependence on the density at the quantile has a simple intuitive explanation:

If there are more data nearby (higher density), the sample quantile is less variable; conversely there are fewer data nearby (low density), the sample quantile is more variable.

To estimate the quantile sampling variability, we make use of the variance approximation above, which requires a way of estimating the unknown probability density function.

Quantile as a Solution to a Certain Minimization Problem

A quantile can also be considered as a solution to a certain minimization problem.

We start with the median, the 50th quantile.

To motivate the minimization problem, we first consider the familiar mean, μ, of the y distribution.

We can measure how far a given data point of y is from the value μ using the square deviation $(y-\mu)^2$, and then how far y is from μ, ***on average***, by the ***mean squared deviation*** $E[(y-\mu)^2]$.

One way to think about how to define the ***centre*** of a distribution is to ask for the point μ at which the average squared deviation from y is minimized.

$$\mu = argmin_c E[(Y-c)^2]$$

Therefore, we can write $E[(y-\mu)^2] = E[y]^2 - 2E[y]\mu + \mu^2$

$$= (\mu - E[y])^2 - (E[y^2] - (E[y])^2)$$

$$= (\mu - E[y])^2 + \mathrm{Var}(y).$$

Because the second term $\mathrm{Var}(y)$ is constant, we minimize the above equation by minimizing the first term $(\mu - E[y])^2$.

Taking $\mu = E[y]$ makes the first term zero and minimizes the equation, because any other values of μ make the first term positive and cause the equation to depart from the minimum.

The sample analog is:

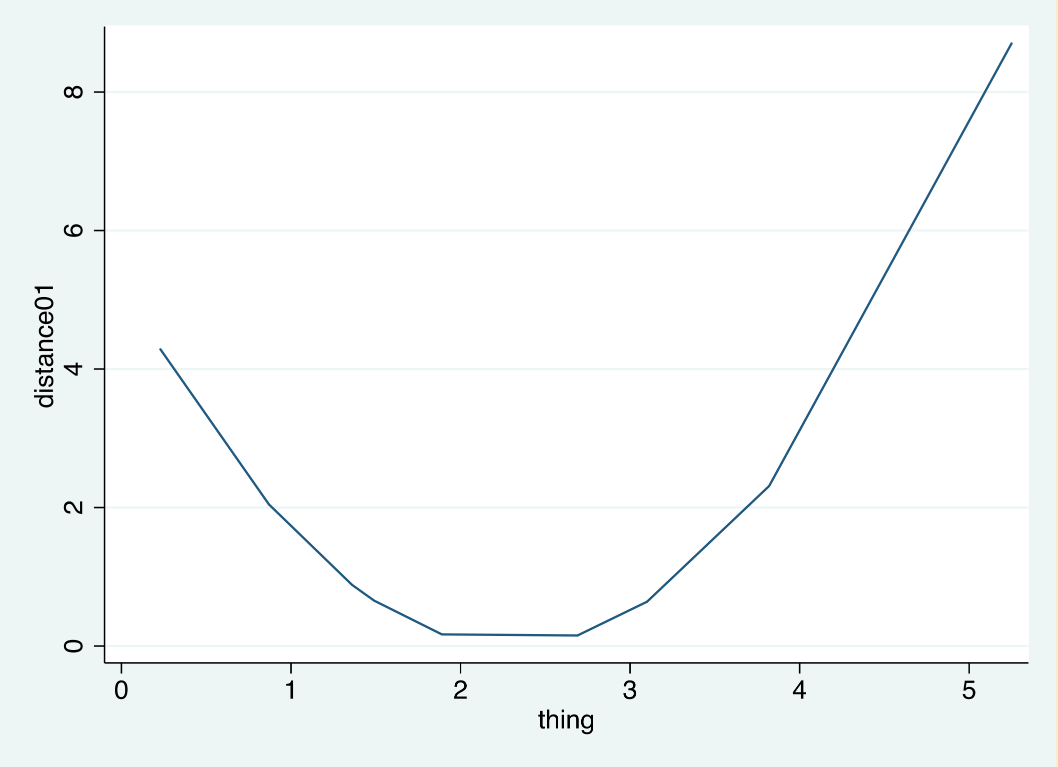
$$\frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)^2 = \frac{1}{n}\sum_{i=1}^{n}(\mu-\bar{y})^2 + \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2$$

$$= (\mu-\bar{y})^2 + s_y^2,$$

where $\bar{y}$ denotes the sample mean, and s_y^2 the sample variance.

Consider a sample of the following 9 values: 0.23, 0.87, 1.36, 1.49, 1.89, 2.69, 3.10, 3.82, and 5.25.

A plot of the mean squared distance of sample points from ***a given point*** μ is shown in the right figure.



Mean squared deviation from μ

Note that ***the function to minimize is convex, with a parabolic form***.

The median m also has a similar minimizing property.

The median *m* has a similar minimizing property.

Instead of using squared distance, we can measure **how far** *y* is from *m* by the absolute distance $|y\text{-}m|$ and measure the average distance in the population from *m* by the mean absolute distance $E|y\text{-}m|$.

$$q\theta = argmin_m E[\rho\theta(Y - m)] \quad (1.5)$$

Again we can solve for the value *m* by minimizing $E|y\text{-}m|$.

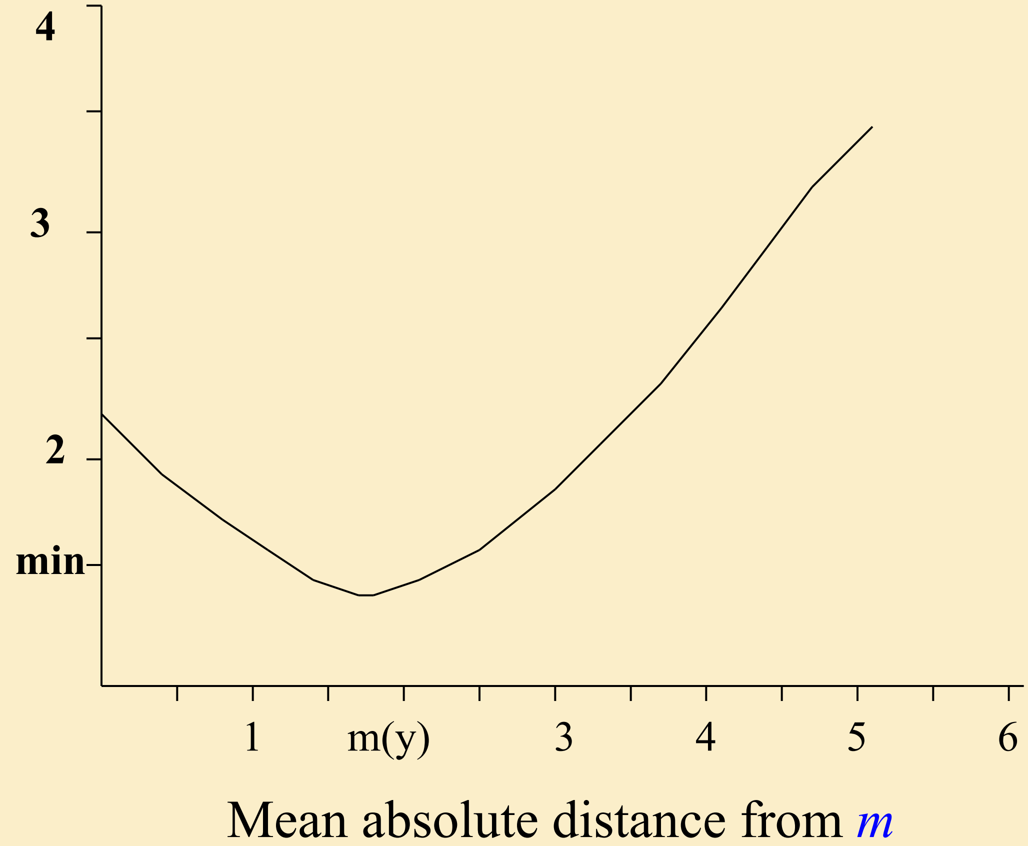
As we shall see, the function of $|y\text{-}m|$ is **convex**, **so that** the minimization solution is to find a point **where** the derivative with respect to *m* is zero (the first order derivative is zero) **or where** the two *directional* derivatives disagree in sign.

The solution is the median of the distribution.

The sample analog is $\frac{1}{n}\sum_{i=1}^{n}|y_i - m|.$

A plot of this function for that nine values is given in the right figure.

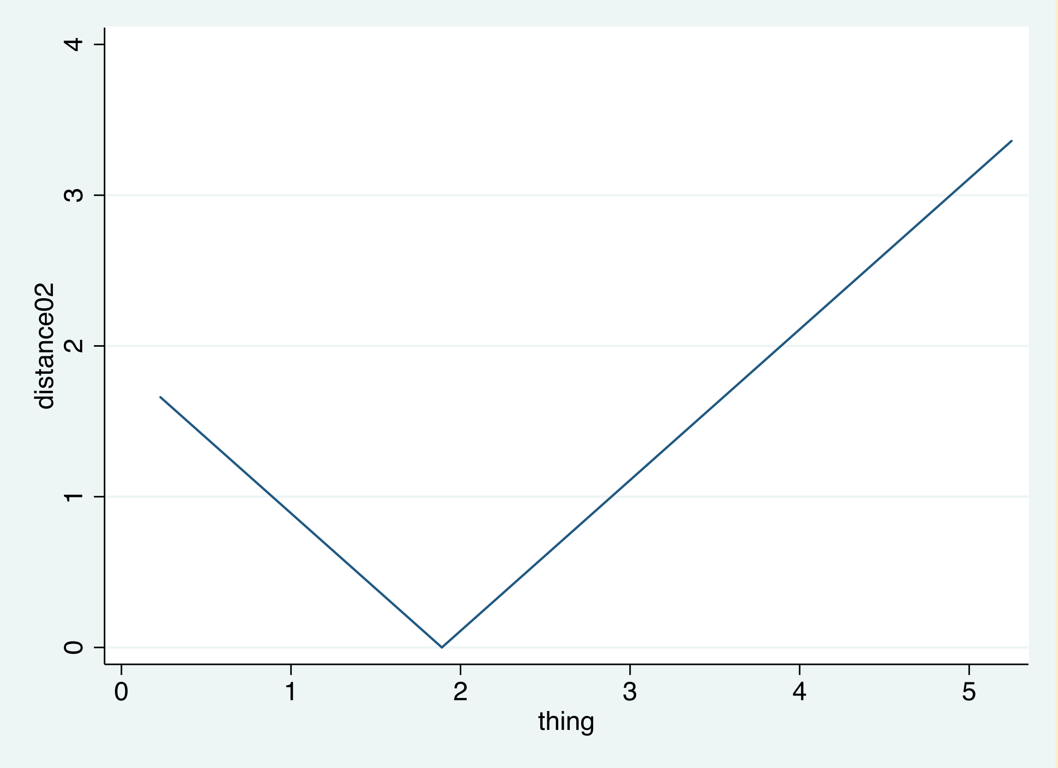
Compared with the function plotted in last figure (the mean squared deviation), this figure remains convex and parabolic in appearance.



However, rather than being smooth, the function here is piecewise linear, with the slope changing precisely at each sample point.

The minimum value of the function shown in this figure coincides with the median sample value of 1.89.

The above is special case of a more general phenomenon.



Mean absolute distance from m

For any sample, the function defined by

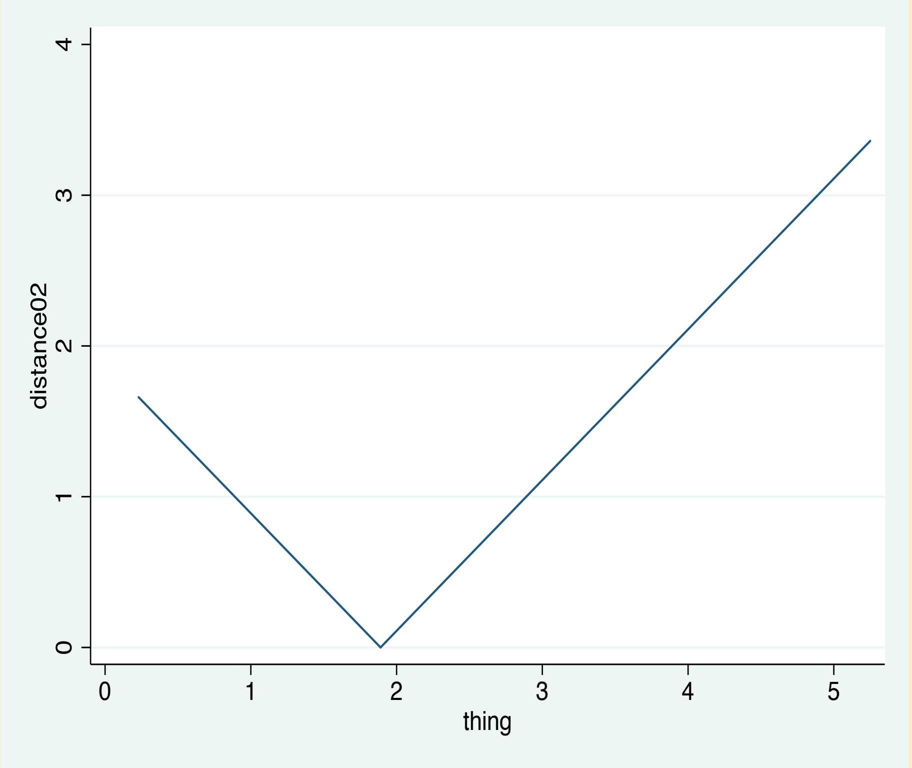
$$f(m) = \frac{1}{n}\sum_{i=1}^{n}|y_i - m|$$

is the sum of V-shaped functions $f_i(m) = \frac{1}{n}|y_i - m|$.

The function f_i takes a minimum value of zero **when** $m = y_i$, has a derivative of $-1/n$ for $m < y_i$ and $1/n$ for $m > y_i$.

While the function is not differentiable at $m = y_i$, it does have a directional derivative there of -1/n in the negative direction and 1/n in the positive direction.

Being the sum of these functions, the directional derivative of f at m is (r-s)/n in the negative direction and (s-r)/n in the positive direction,

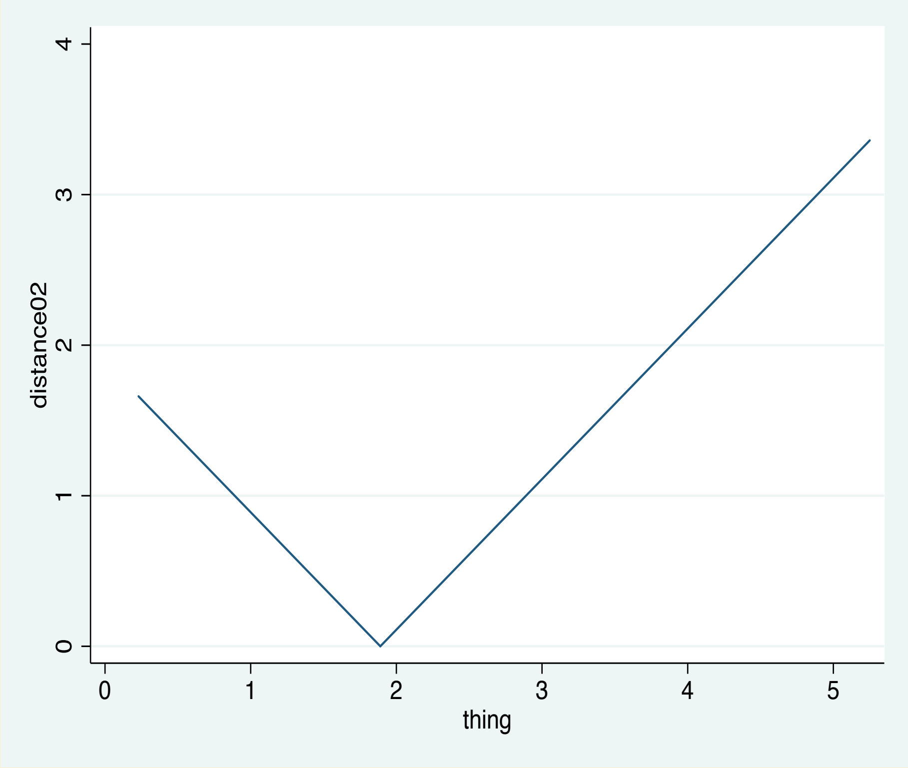


Mean absolute distance from m

where s is the number of data points to the right of m **and** r is the number of data points to the left of m.

It follows **that** the minimum of f occurs **when** m has the same number of data points to its left and right, **that is**, when m is the sample median.

This representation of the median generalizes to the other quantiles as follows.



Mean absolute distance from m

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:

$$q\theta = argmin_m E[\rho\theta(Y - m)] \quad (1.5)$$

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

$$\rho\theta\,(y) = [\theta - I(y < 0)]y$$

$$= [(1 - \theta)I(y \leq 0) + \theta I(y > 0)]|y|.$$

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:

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

$$= \underset{c}{\operatorname{argmin}} \left\{ (1-\theta) \sum_{y \le c} |y - c| f(y) + \theta \sum_{y > c} |y - c| f(y) \right\}.$$

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

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

$$= \underset{c}{\operatorname{argmin}} \left\{ (1-\theta) \int_{-\infty}^{c} |y - c| f(y) d(y) + \theta \int_{c}^{+\infty} |y - c| f(y) d(y) \right\}$$

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

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) Median = $\text{argmin}_c E|Y-c|$.

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.

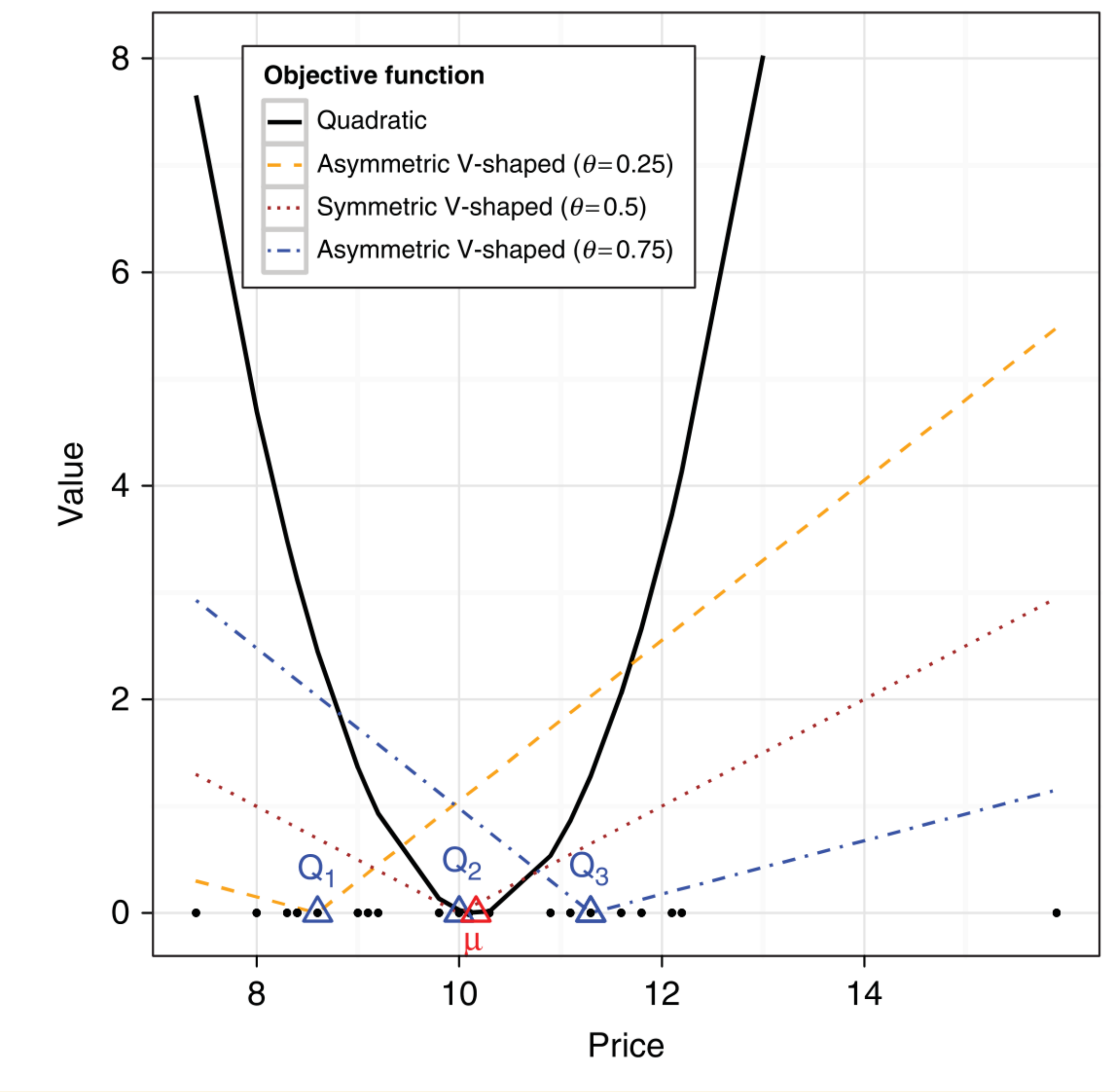
The different objective function for the mean and the three quartiles are shown on the *y*-axis.

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.

Figure 1.2 Comparison of mean and quartiles as location indexes of a univariate distribution.

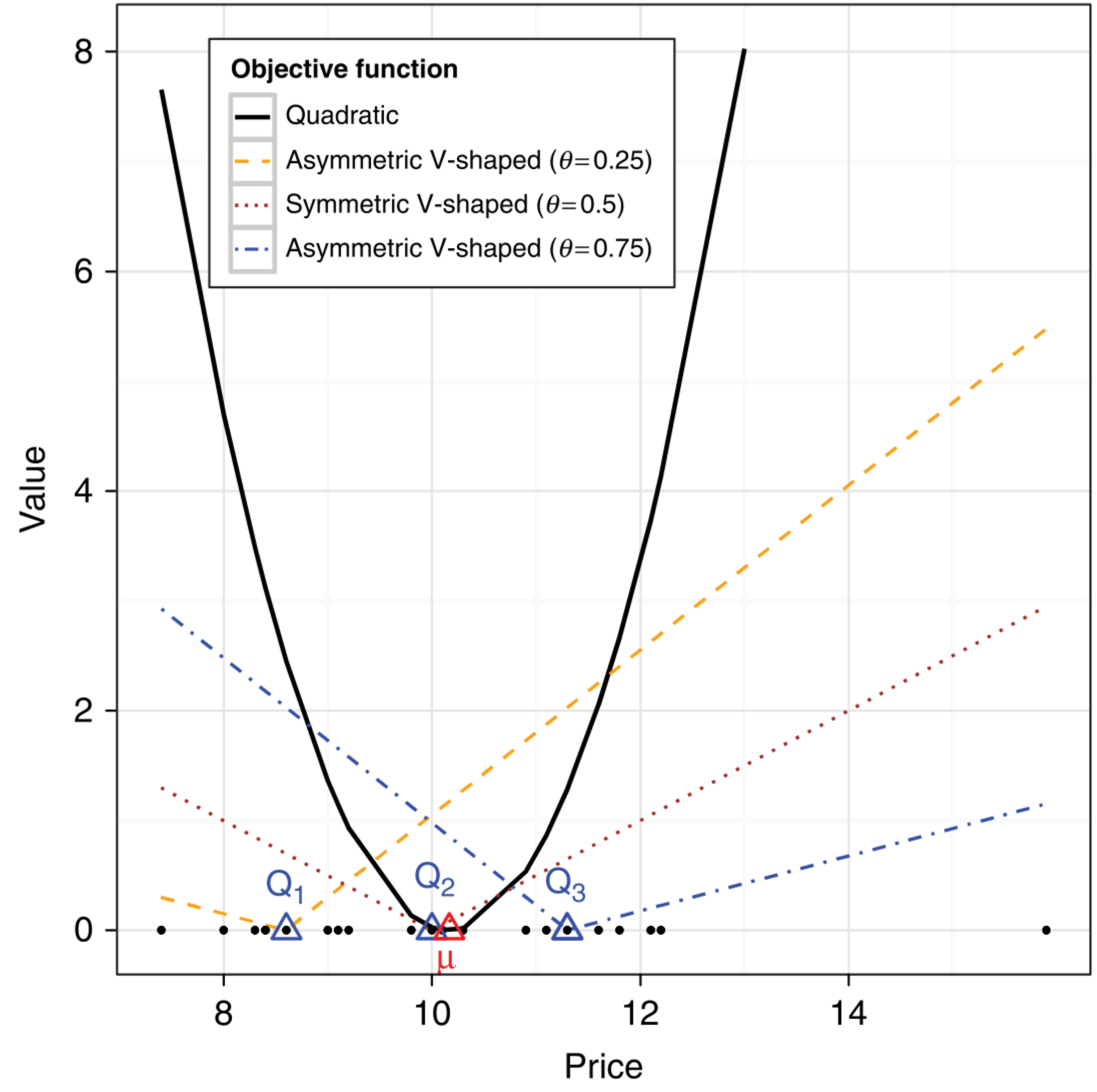
Data refer to the Price of small cars as defined by the Type variable (Cars93 dataset).

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.



For any $p \in (0, 1)$, the distance from y to a given $Q^{(p)}$ is measured by the absolute distance, **but** we apply different weight depending on **whether** y is to the left or to the right of $Q^{(p)}$.

Thus, we define the distance from y to a given $Q^{(p)}$ as

$$d_p(y, Q^{(p)}) = \begin{cases} (1-p)\left|y - Q^{(p)}\right| & \text{for } y < Q^{(p)} \\ p\left|y - Q^{(p)}\right| & \text{for } y \geq Q^{(p)} \end{cases}.$$

We look for the value $Q^{(p)}$ **that** minimizes the mean distance from y: $E[d_p(y, Q^{(p)})]$.

The minimum occurs **when** $Q^{(p)}$ is the p^{th} quantile.

Similarly, the p^{th} sample quantile is the value of $Q^{(p)}$ **that** minimizes the average (weighted) distance:

$$\frac{1}{n}\sum_{i=1}^{n} d_p(y_i, Q^{(p)}) = \frac{1-p}{n}\sum_{y_i<q}\left|y_i - Q^{(p)}\right| + \frac{p}{n}\sum_{y_i>q}\left|y_i - Q^{(p)}\right|.$$

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) $q\theta = argmin_m E[\rho\theta(Y - m)]$, the presentation of the solution for the median case, Equation(1.2) (Median = $\text{argmin}_c E|Y\text{-}c|$), 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:

$$E|Y - c| = \int_{y \in R} |y - c| f(y) dx$$

$$= \int_{y<c} |y - c| f(y) dy + \int_{y>c} |y - c| f(y) dy$$

$$= \int_{y<c} (c - y) f(y) dy + \int_{y>c} (y - c) f(y) dy$$

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: $\frac{\partial}{\partial c}E|Y-c|=0$.
The solution can then be obtained applying the derivative and integrating per part as follows:

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

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 $y=c$:
$\frac{\partial}{\partial c}(c-Y)=1$ and $\frac{\partial}{\partial c}(Y-c)=-1$.

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

Using then the CDF definition, Equation (1.3) $F_Y(y) = F(y) = P(Y \leq y)$, the previous equation reduces to: $F(c)-[1-F(c)] = 0$ and thus:

$2F(c)-1=0 \Rightarrow F(c)= 1/2 \Rightarrow c=Me$.

The solution of the minimization problem formulated in Equation Median $= \text{argmin}_c E|Y\text{-}c|$ (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)

$q\theta = argmin_m E[\rho\theta(Y - m)]$, 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 $q\theta$ as the solution of the minimization problem:

$$F(c) - \theta F(c) - \theta + \theta F(c) = 0 \Rightarrow F(c) = \theta \Rightarrow c = q\theta.$$

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) $\mu = \text{argmin}_c\ E(Y\text{-}c)^2$ can be extended to the estimation of the conditional mean function:

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

where $\mu(\mathbf{x}_i, \beta) = E[Y|\mathbf{X} = \mathbf{x}_i]$ is the conditional mean function.

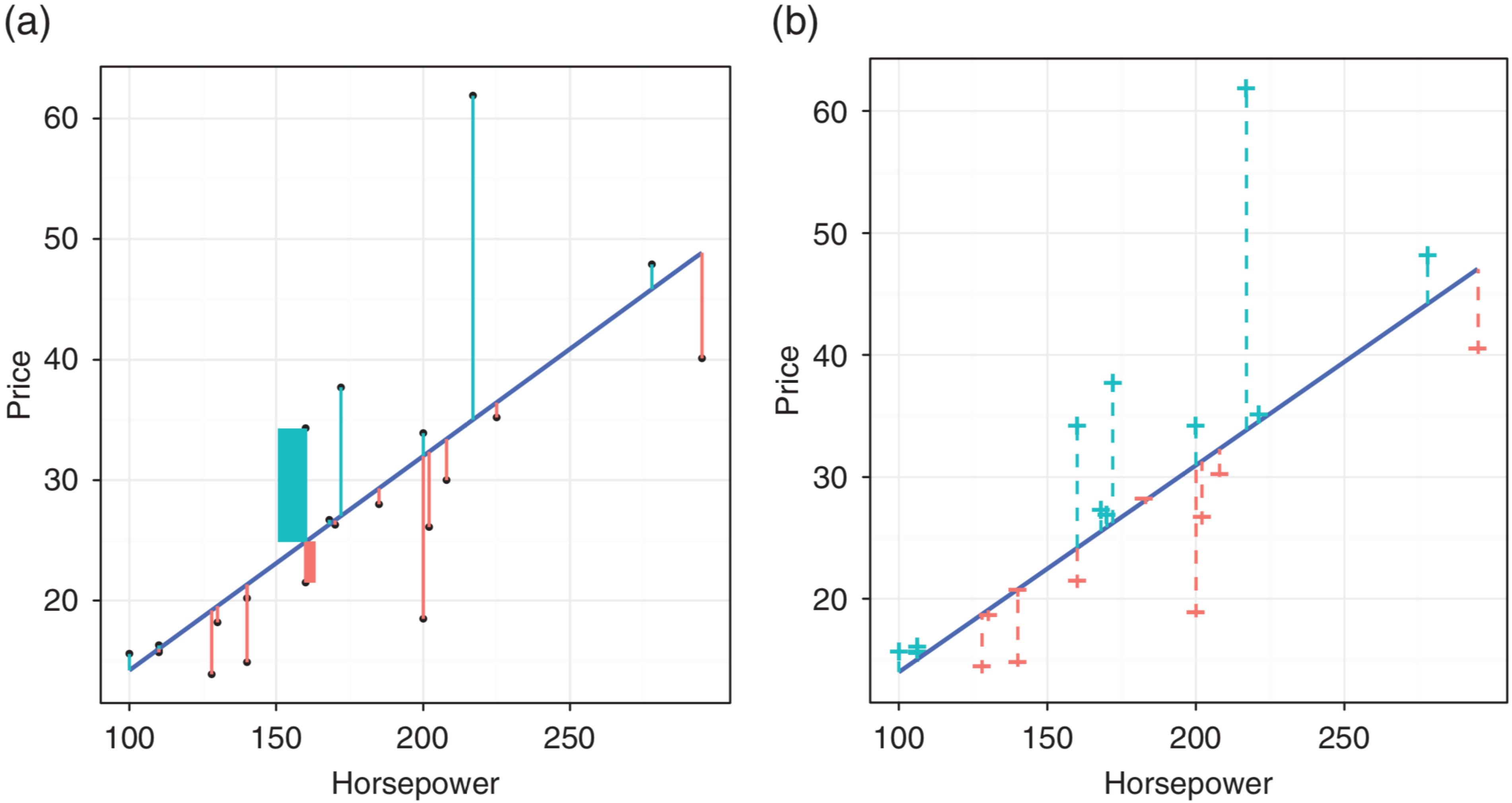
In the case of a linear mean function, $\mu(\mathbf{x}_i, \beta) = \mathbf{x}_i^\top\beta$, the previous equation becomes: $\widehat{\boldsymbol{\beta}} = argmin_{\beta} E[Y - \boldsymbol{x}_i'\boldsymbol{\beta}]^2$ yielding the least squares linear regression model.

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 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.


(a)
60
50
40
30
20
Price
100
150
200
250
Horsepower
(b)
60
50
40
30
20
Price
100
150
200
250
Horsepower

The same approach can be used to extend Equation (1.2) Median = $\text{argmin}_c E|Y\text{-}c|$ to the median, or the more general Equation (1.5) $q\theta = argmin_m E[\rho\theta(Y - m)]$ to the generic θ^{th} quantile.

In this latter case, we obtain:

$$\hat{q}_Y(\theta, \boldsymbol{X}) = argmin_{Q_Y(\theta,X)} E[\rho\theta(Y - Q_Y(\theta, \boldsymbol{X}))]$$

where $Q_Y(\theta, \boldsymbol{X}) = Q_\theta[Y|\boldsymbol{X} = \boldsymbol{x}]$ denotes the generic conditional quantile function.

Likewise, for the linear model case the previous equation becomes:

$$\hat{\beta}(\theta) = argmin_\beta E[\rho\theta(Y - \boldsymbol{X}\beta)]$$

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.

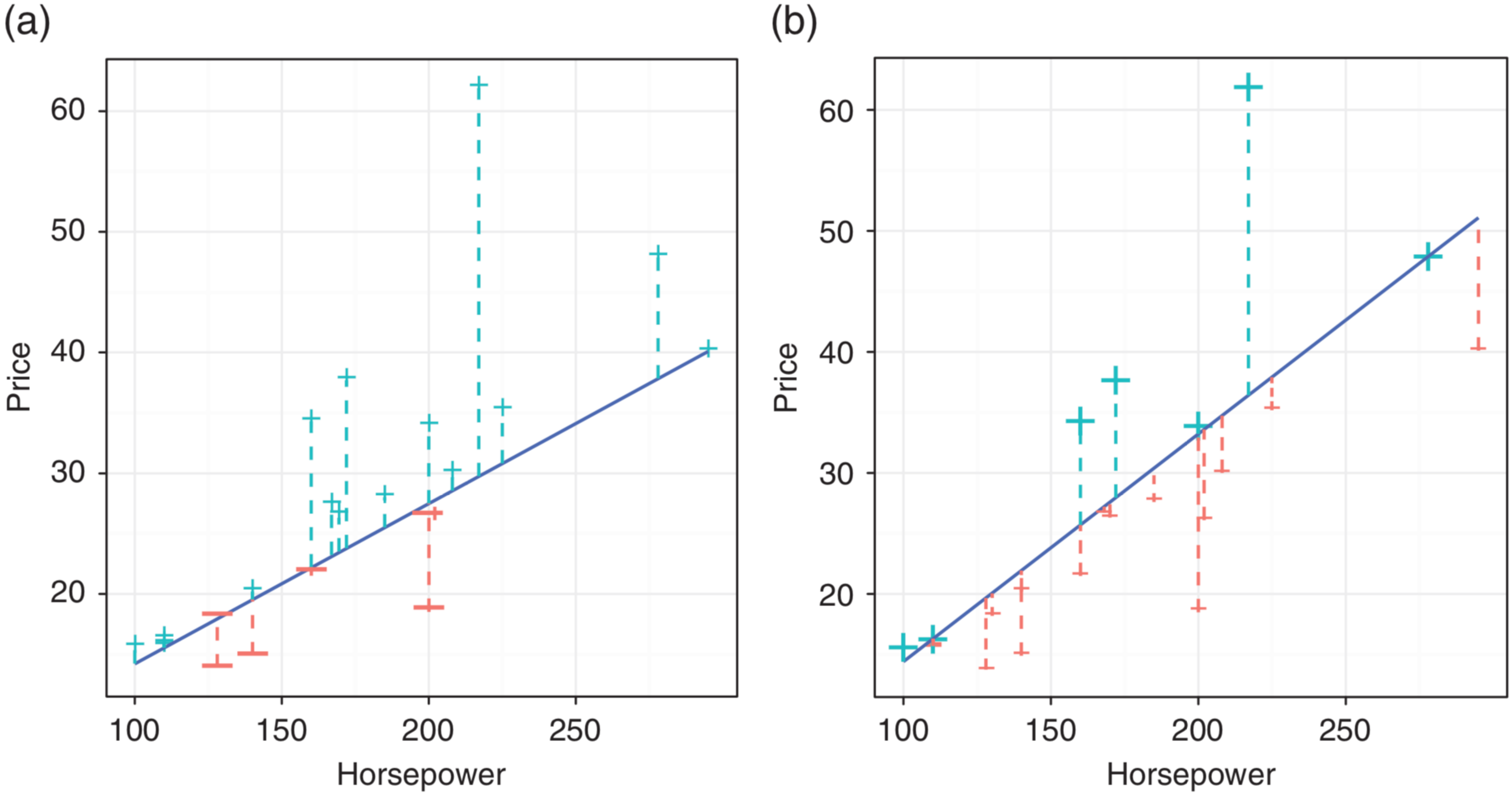
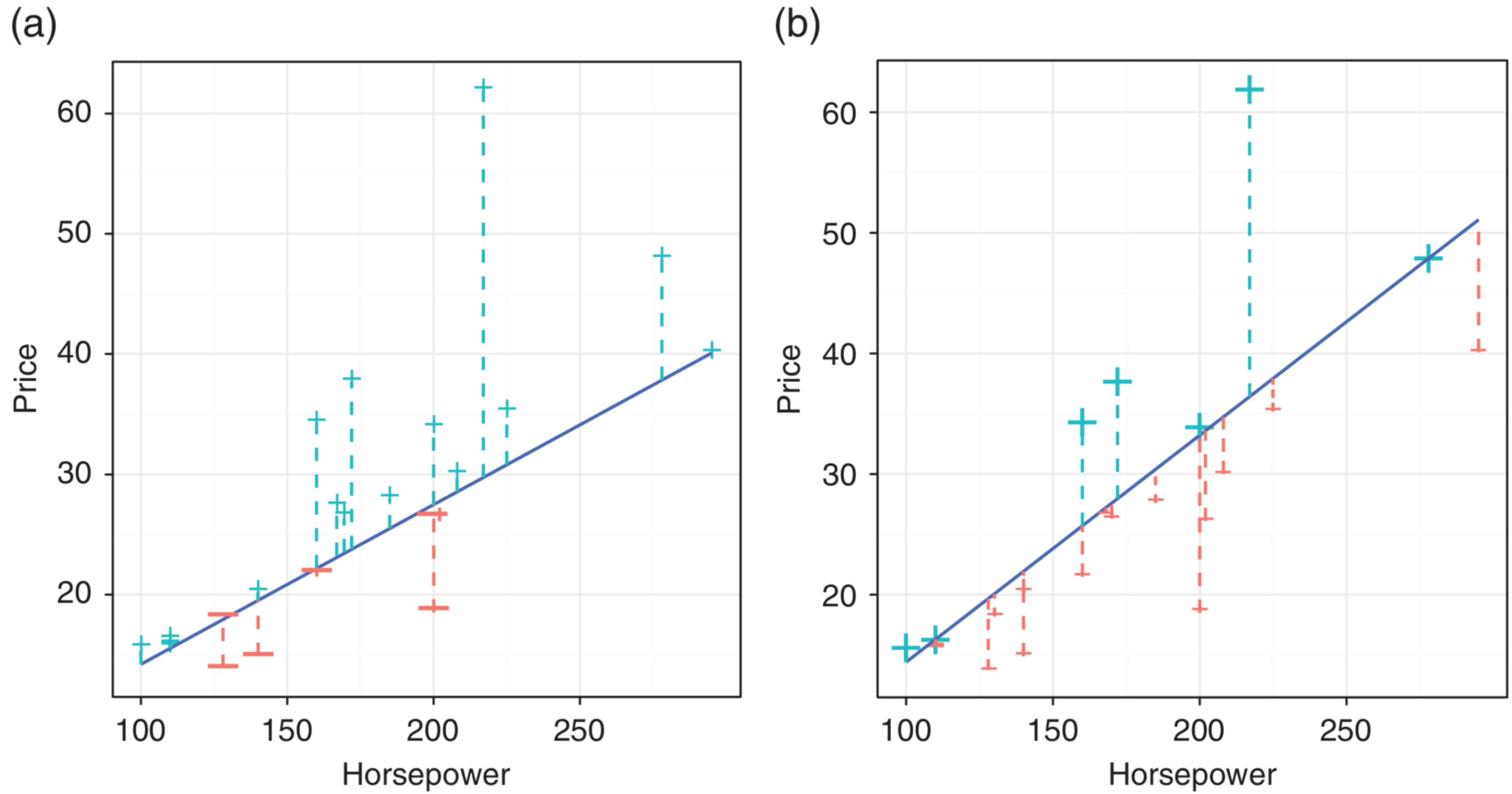
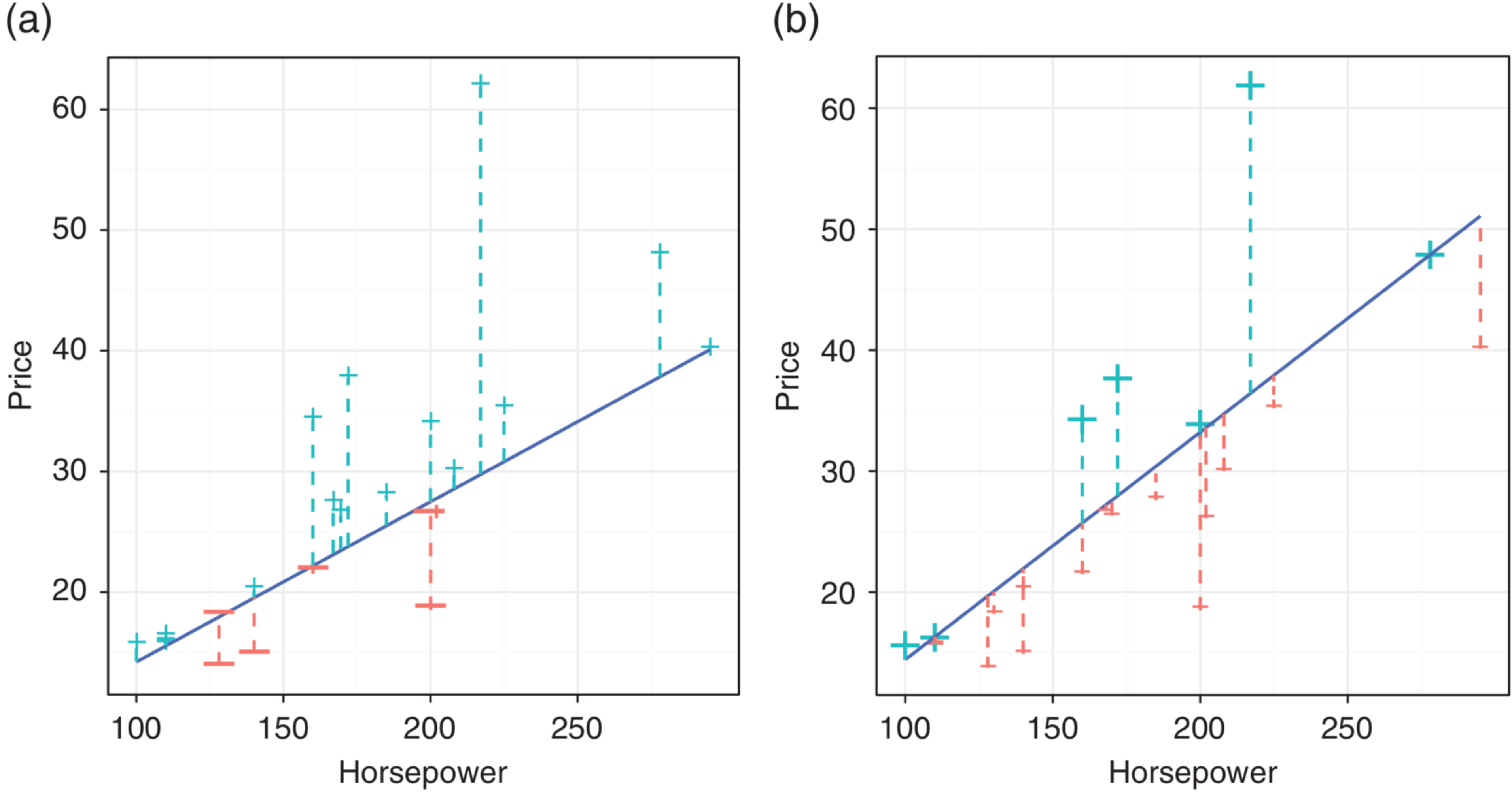


Figure 1.4 shows the geometric interpretation of the QR solution for the case of θ=0.25 (first quartile) in Figure 1.4(a), and of θ=0.75 (third quartile) in Figure 1.4(b), respectively.



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.

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

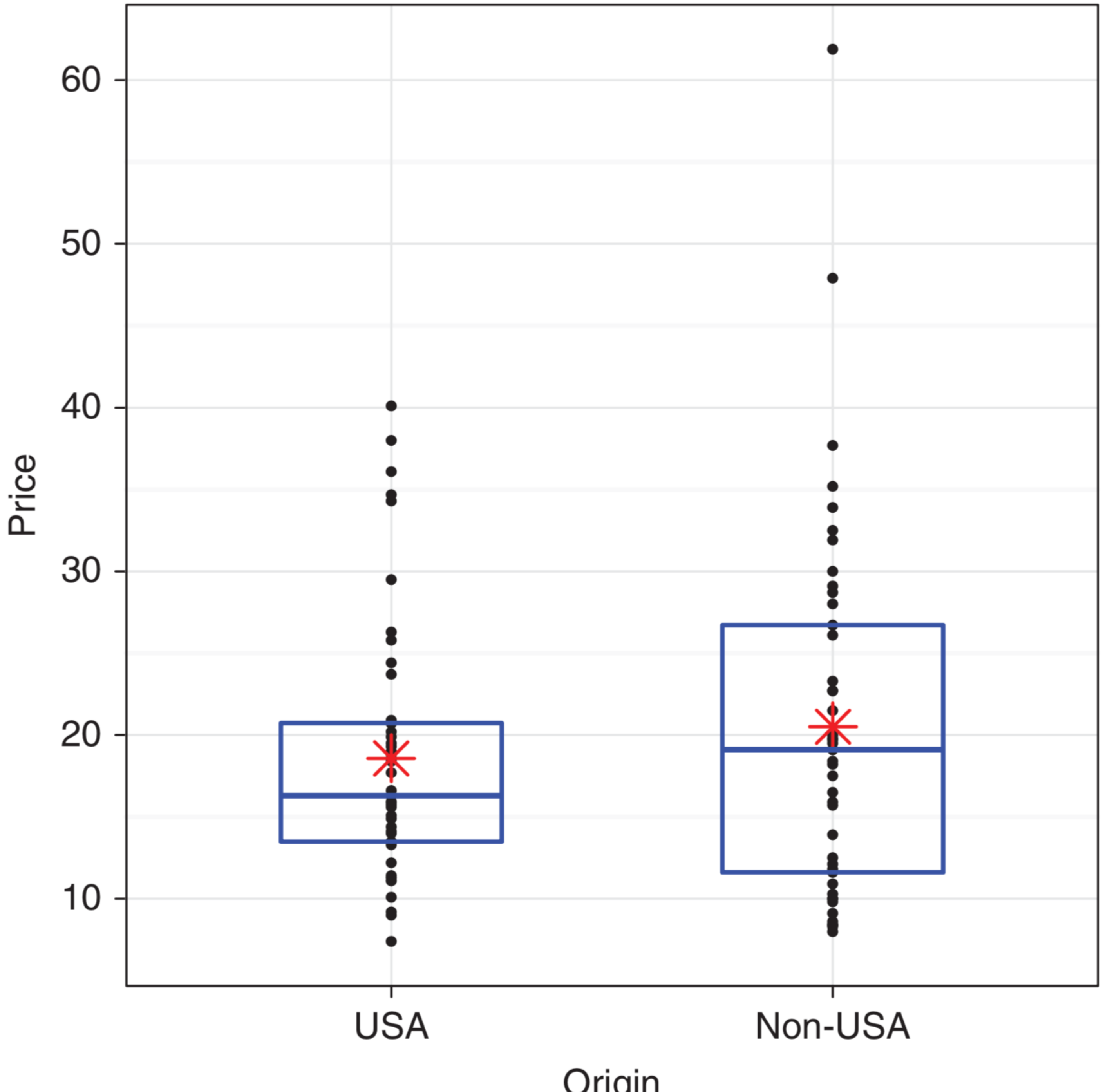
In order to introduce quantile regression, it is useful to begin by illustrating the simplest form of linear model; that is a model with a quantitative response variable and a dummy predictor variable.

This simple setting aims to study differences in the response variable between the two groups as determined by the dichotomous predictor variable.

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.

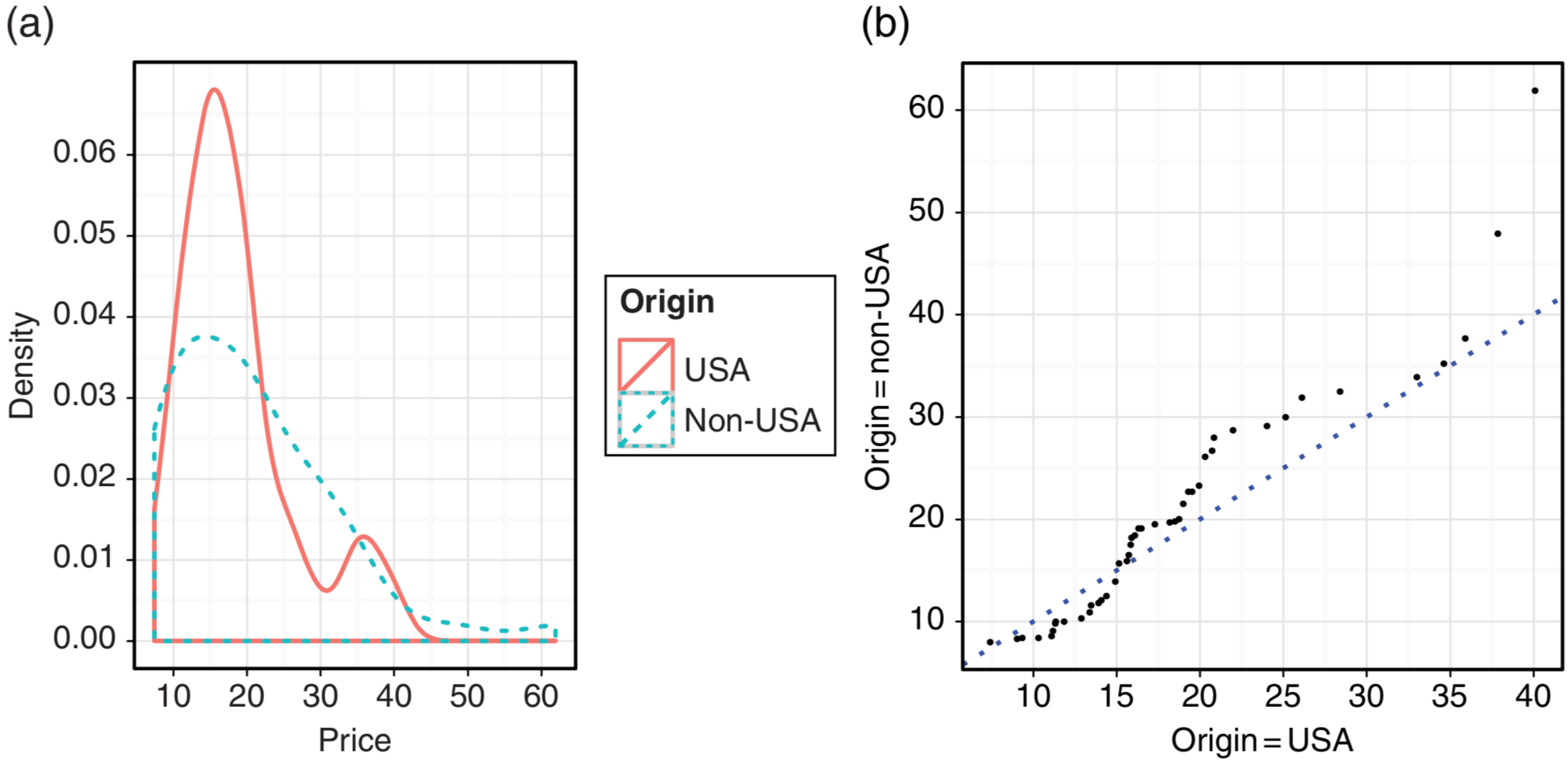
Figure 1.5 shows the dotplots for the two groups, USA cars represented on the left-hand side and non-USA cars on the right-hand side.


60
50
40
30
20
10
Price
USA
Non-USA
Origin

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)
Density
0.06
0.05
0.04
0.03
0.02
0.01
0.00
10
20
30
40
50
60
Price
Origin
USA
Non-USA
(b)
Origin = non-USA
60
50
40
30
20
10
10
15
20
25
30
35
40
Origin = USA

A third graphical representation frequently used to 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^{o} 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:

$\hat{}Price = \hat{}\beta_0 + \hat{}\beta_1 Origin$

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:

Likewise, the estimation of the QR model:

$\hat{}Price\theta = \hat{}\ \beta_0(\theta) + \hat{}\ \beta_1(\theta) Origin$ (1.6)

for different values of θ ∈ [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 straightforward that the estimated price for the USA cars is then:

$$\hat{}Price\theta = \hat{}\beta_0(\theta) + \hat{}\beta_1(\theta) \times \mathbf{0} = \beta_0(\theta),$$

while for the non-USA subset the model becomes:

$$\hat{}Price\theta = \hat{}\beta_0(\theta) + \hat{}\beta_1(\theta) \times \mathbf{1} = \hat{}\beta_0(\theta) + \hat{}\beta_1(\theta).$$

$\hat{}\beta_0(\theta)$ 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

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

1.5 Properties of Quantiles

One basic property of quantiles is ***monotone equivalence*** property.

It states that if we apply a monotone transformation h (for example, the exponential or logarithmic function) to a random variable, the quantiles are obtained by applying the same transformation to the quantile function.

In other words, if q is the p^{th} quantile of y, then $h(q)$ is the p^{th} quantile of $h(y)$.

Another <u>basic property of sample quantile relates to their insensitivity to the influence of ***outliers***</u>.

This feature, which has an analog in quantile regression, helps make quantiles and quantile-based procedures useful in many contexts.

2.3 Quantile Estimation

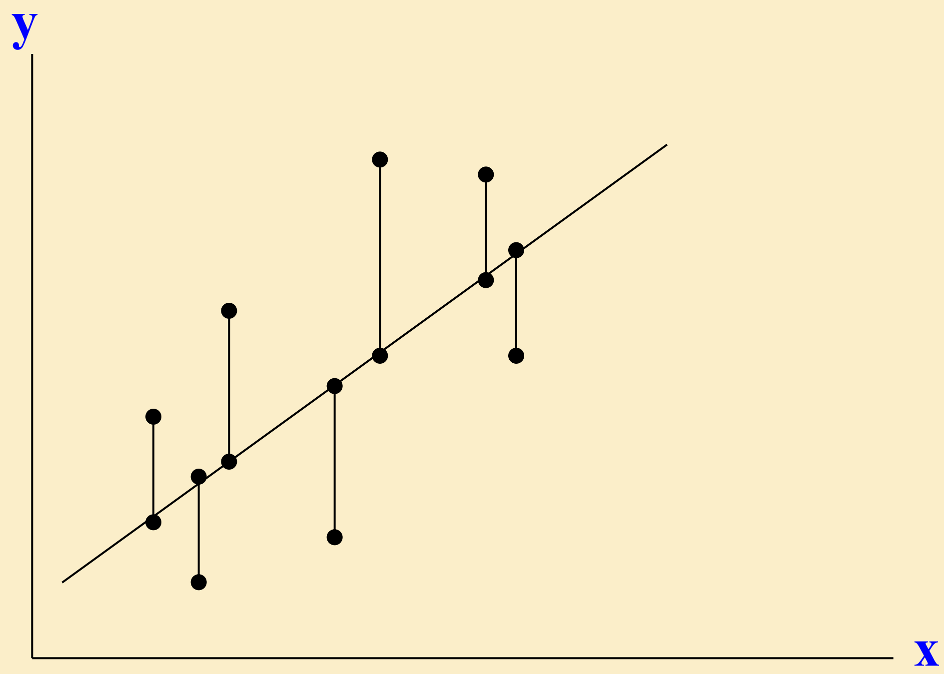
We review least squares estimation so as to place Quantile Regression (QR) estimation in a familiar context.

The least squares estimator solves for the parameter estimates $\hat{}\beta_0$ and $\hat{}\beta_1$ by taking those values of the parameters that minimize the sum of squared residuals:

$$\mathrm{Min}\Sigma_i[y_i-(\beta_0+\beta_1 x_i)]^2$$

If the LRM assumptions are correct, the fitted dependent function $\hat{}\beta_0+\hat{}\beta_1 x_i$ approaches the population conditional mean $E[y|x]$ as the sample size goes to infinity.

In $\mathrm{Min}\Sigma_i[y_i-(\beta_0+\beta_1 x_i)]^2$, the expression minimized is the ***sum of squared vertical distance*** between data points (x_i, y_i) and the fitted line $y_i=\hat{}\beta_0+\hat{}\beta_1 x_i$.


y
x

A closed form solution to the minimization problem is obtained by

(a) taking partial derivatives of $\text{Min}\Sigma_i[y_i-(\beta_0+\beta_1 x_i)]^2$ with respect to β_0 and β_1, respectively;

(b) Setting each partial derivatives equal to zero;

(c) Solving the resulting system of two equations with two unknowns.

We then arrive at the two estimators:

$$\hat{\beta}_1 = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^n (x_i - \bar{x})^2}, \qquad \beta_0 = \bar{y} - \beta_1 \bar{x}.$$

For a start, we first consider the estimator for the median-regression model.

We described how the median (m) of y can be viewed as the minimizing value of $E|y-m|$.

For an analogous prescription in median regression case, we choose to minimize the sum of absolute residuals.

In other words, we find the coefficients that minimize the ***sum of absolute residuals*** (the ***absolute distance from an observed point to its fitted line***).

The estimator solves for the βs by minimizing: $\Sigma_i|y_i - \beta_0 - \beta_1 x_i|$.

Under appropriate model assumptions, as the sample size goes to infinity, we obtain the conditional median of y given x at the population level.

When expression $\Sigma_i|y_i - \beta_0 - \beta_1 x_i|$ is minimized, the resulting solution, **which** we refer to as the median regression line, must pass through a pair of data points with half of the remaining data lying above the regression line and the other half falling below.

That is, roughly half of the residuals are positive and half are negative.

There are typically multiple lines with this property, and among these lines, the one that minimizes $\Sigma_i|y_i-\beta_0-\beta_1 x_i|$ is the solution.

The median regression estimator can be generalized to allow for p^{th} quantile regression estimators (Koenker & d'Orey, 1987).

Recall that the p^{th} quantile of a univariate sample $y_1, \ldots, y_n$ distribution is the value q **that** minimizes the sum of weighted distances from the sample points, **where** points below q receive a weight of 1-p and points above q receive a weight of p.

In a similar manner, we define the p^{th} quantile regression estimators $\hat{\beta}_0^{(p)}$ and $\hat{\beta}_1^{(p)}$ as the values **that** minimize the weighted sum of distances between fitted line $\hat{y}_i = \hat{\beta}_0^{(p)} + \hat{\beta}_1^{(p)} x_i$ and the (x_i, y_i), **where** we use a weight of 1-p if the fitted value under-predicts the observed value y_i and a weight of p otherwise.

In other words, we seek to minimize a weighted sum of residuals $y_i - \hat{y}_i$ **where** positive residuals receive a weight of p and negative residuals receive a weight of 1-p.

Formally, the p^{th} quantile regression estimator $\hat{\beta}_0^{(p)}$ and $\hat{\beta}_1^{(p)}$ are chosen to minimize:

$$\sum_{i=1}^{n} d_p(y_i, \hat{y}_i) = p \sum_{y_i \geq \beta_0^{(p)} + \beta_1^{(p)} x_i} \left| y_i - \beta_0^{(p)} - \beta_1^{(p)} x \right|$$
$$+ (1 - p) \sum_{y_i < \beta_0^{(p)} + \beta_1^{(p)} x_i} \left| y_i - \beta_0^{(p)} - \beta_1^{(p)} x \right|,$$

where d_p is the distance introduced before.

Thus, unlike $\Sigma_i |y_i - \beta_0 - \beta_1 x_i|$, **which** states that the negative residuals are given the same importance as the positive residuals, the above expression assigns different weights to positive and negative residuals.

Observe that in the above expression, the first term is the sum of vertical distance of data points from the line $y_i = \beta_0^{(p)} + \beta_1^{(p)} x_i$, for points lying above the line.

$$\sum_{i=1}^{n} d_p(y_i, \hat{y}_i) = p \sum_{y_i \geq \beta_0^{(p)} + \beta_1^{(p)} x_i} \left| y_i - \beta_0^{(p)} - \beta_1^{(p)} x \right|$$

$$+ (1 - p) \sum_{y_i < \beta_0^{(p)} + \beta_1^{(p)} x_i} \left| y_i - \beta_0^{(p)} - \beta_1^{(p)} x \right|,$$

The second is a similar sum over all data points lying below the line.

Observe that, contrary to a common misconception, the estimation of coefficient for each quantile regression is based on the weighted data of the whole sample, not just the portion of the sample at that quantile.