

## 第 26 章 分位数回归

### 26.1 为什么需要分位数回归

一般的回归模型着重考察  $\mathbf{x}$  对  $y$  的条件期望  $E(y|\mathbf{x})$  的影响, 实际上是均值回归。

但我们关心  $\mathbf{x}$  对整个条件分布  $y|\mathbf{x}$  的影响, 而  $E(y|\mathbf{x})$  只是刻画条件分布  $y|\mathbf{x}$  集中趋势的一个指标而已。

如果  $y|\mathbf{x}$  不是对称分布, 则  $E(y|\mathbf{x})$  很难反映条件分布的全貌。

如能够估计条件分布  $y|\mathbf{x}$  的若干重要的条件分位数(conditional quantiles), 比如中位数(median)、1/4分位数(lower quartile)、3/4分位数(upper quartile), 能更全面认识条件分布  $y|\mathbf{x}$ 。

使用 OLS 进行“均值回归”, 由于最小化的目标函数为残差平方和( $\sum_{i=1}^n e_i^2$ ), 故易受极端值影响。

Koenker and Bassett(1978)提出“分位数回归”(Quantile Regression, 简记 QR), 使用残差绝对值的加权平均(比如,  $\sum_{i=1}^n |e_i|$ )作为最小化的目标函数, 不易受极端值影响, 较为稳健。

分位数回归还能提供关于条件分布  $y|\mathbf{x}$  的全面信息。

## 26.2 总体分位数

假设 $Y$ 为连续型随机变量，其累积分布函数为 $F_y(\cdot)$ 。

$Y$ 的“总体 $q$ 分位数”(population  $q^{\text{th}}$  quantile,  $0 < q < 1$ ), 记为 $y_q$ , 满足以下定义式:

$$q = \mathbf{P}(Y \leq y_q) = F_y(y_q)$$

总体 $q$ 分位数 $y_q$ 正好将总体分布分为两部分, 其中小于或等于 $y_q$ 的概率为 $q$ , 而大于 $y_q$ 的概率为 $(1-q)$ 。

如果 $q = 1/2$ ，则为中位数，正好将总体分为两个相等的部分。

如果 $F_y(\cdot)$ 严格单调递增，则有

$$y_q = F_y^{-1}(q)$$

其中， $F_y^{-1}(\cdot)$ 为 $F_y(\cdot)$ 的逆函数，参见图 26.1。

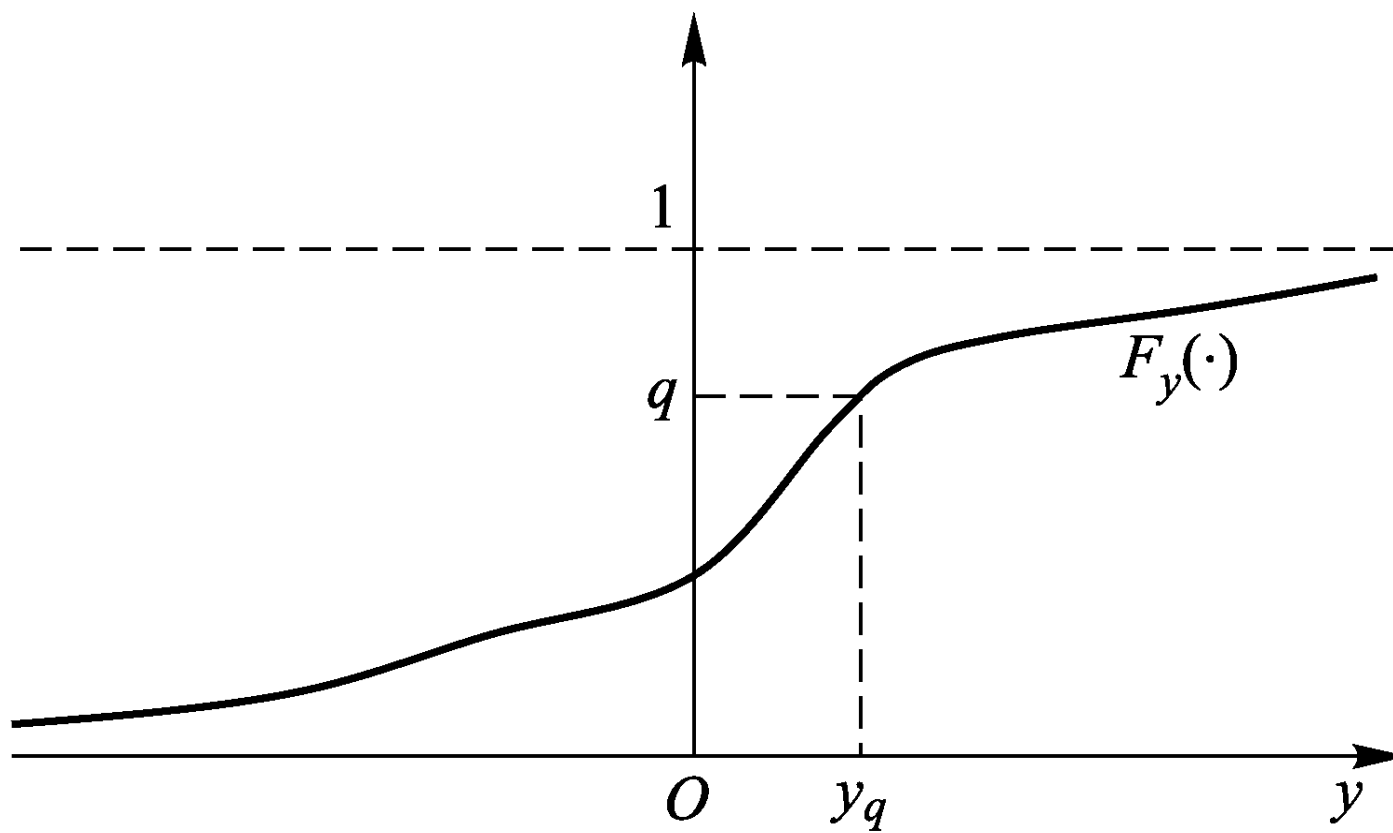


图 26.1 总体  $q$  分位数与累积分布函数

对于回归模型，记条件分布  $y|\mathbf{x}$  的累积分布函数为  $F_{y|\mathbf{x}}(\cdot)$ 。

条件分布  $y|\mathbf{x}$  的总体  $q$  分位数，记为  $y_q$ ，满足以下定义式：

$$q = F_{y|\mathbf{x}}(y_q)$$

假设  $F_{y|\mathbf{x}}(\cdot)$  严格单调递增，则有

$$y_q = F_{y|\mathbf{x}}^{-1}(q)$$

由于条件累积分布函数  $F_{y|\mathbf{x}}(\cdot)$  依赖于  $\mathbf{x}$ ，故条件分布  $y|\mathbf{x}$  的总体  $q$  分位数  $y_q$  也依赖于  $\mathbf{x}$ ，记为  $y_q(\mathbf{x})$ ，称为“条件分位数函数”(conditional quantile function)。

对于线性回归模型，如果扰动项满足同方差的假定，或扰动项异方差的形式为乘积形式，则  $y_q(\mathbf{x})$  是  $\mathbf{x}$  的线性函数。

考虑以下模型：

$$y = \mathbf{x}'\boldsymbol{\beta} + u$$

$$u = \mathbf{x}'\boldsymbol{\alpha} \cdot \varepsilon$$

$$\varepsilon \sim \text{iid}(0, \sigma^2)$$

不失一般性，假设  $\mathbf{x}'\boldsymbol{\alpha} > 0$ 。

如果  $\mathbf{x}'\boldsymbol{\alpha}$  为常数，则扰动项  $u$  为同方差；反之，则为乘积形式的异方差。

根据定义，条件分位数函数  $y_q(\mathbf{x})$  满足

$$\begin{aligned} q &= \mathbf{P}\{y \leq y_q(\mathbf{x})\} && \text{(条件分位数的定义)} \\ &= \mathbf{P}\{\mathbf{x}'\boldsymbol{\beta} + u \leq y_q(\mathbf{x})\} && \text{(代入 } y = \mathbf{x}'\boldsymbol{\beta} + u\text{)} \\ &= \mathbf{P}\{u \leq y_q(\mathbf{x}) - \mathbf{x}'\boldsymbol{\beta}\} && \text{(移项)} \\ &= \mathbf{P}\{\mathbf{x}'\boldsymbol{\alpha} \cdot \varepsilon \leq y_q(\mathbf{x}) - \mathbf{x}'\boldsymbol{\beta}\} && \text{(代入 } u = \mathbf{x}'\boldsymbol{\alpha} \cdot \varepsilon\text{)} \\ &= \mathbf{P}\left\{\varepsilon \leq \frac{y_q(\mathbf{x}) - \mathbf{x}'\boldsymbol{\beta}}{\mathbf{x}'\boldsymbol{\alpha}}\right\} && \text{(两边同除以 } \mathbf{x}'\boldsymbol{\alpha} > 0\text{)} \\ &= F_\varepsilon\left(\frac{y_q(\mathbf{x}) - \mathbf{x}'\boldsymbol{\beta}}{\mathbf{x}'\boldsymbol{\alpha}}\right) && \text{(累积分布函数的定义)} \end{aligned}$$

其中， $F_\varepsilon(\cdot)$  为  $\varepsilon$  的累积分布函数。因此，



$$\frac{y_q(\mathbf{x}) - \mathbf{x}'\boldsymbol{\beta}}{\mathbf{x}'\boldsymbol{\alpha}} = F_{\varepsilon}^{-1}(q)$$

$$y_q(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta} + \mathbf{x}'\boldsymbol{\alpha}F_{\varepsilon}^{-1}(q) = \mathbf{x}'\left[\boldsymbol{\beta} + \boldsymbol{\alpha}F_{\varepsilon}^{-1}(q)\right]$$

故  $y_q(\mathbf{x})$  是  $\mathbf{x}$  的线性函数。

在同方差的情况下， $\mathbf{x}'\boldsymbol{\alpha}$  为常数，所有条件分位数函数  $\{y_q(\mathbf{x}), 0 < q < 1\}$  的斜率都等于  $\boldsymbol{\beta}$ ，只有截距项  $\mathbf{x}'\boldsymbol{\alpha}F_{\varepsilon}^{-1}(q)$  依赖于  $q$ 。

一般地，条件分位数函数的“斜率”也依赖于  $q$ ，记为  $\boldsymbol{\beta}_q$ 。

在下文中，假设条件分位数函数是解释变量  $\mathbf{x}$  的线性函数。

## 26.3 样本分位数

对于随机变量 $Y$ ，如果总体的 $q$ 分位数 $y_q$ 未知，可使用样本 $q$ 分位数 $\hat{y}_q$ 来估计 $y_q$ 。

将样本数据 $\{y_1, y_2, \dots, y_n\}$ 按从小到大的顺序排列为 $\{y_{(1)}, y_{(2)}, \dots, y_{(n)}\}$ 。

$\hat{y}_q$ 等于第 $[nq]$ 个最小观测值，其中 $n$ 为样本容量， $[nq]$ 表示大于或等于 $nq$ 而离 $nq$ 最近的正整数。

**【例】**  $n = 97$ ， $q = 0.25$ ，则 $[nq] = [97 \times 0.25] = [24.25] = 25$ 。

但这种方法不易推广到回归模型。

一种等价方法是，将样本分位数看成是某最小化问题的解。

样本均值也可看成是最小化残差平方和的解：

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

样本中位数可视为“最小化残差绝对值之和”问题的解：

$$\min_{\mu} \sum_{i=1}^n |y_i - \mu| \Rightarrow \mu = \text{median}\{y_1, y_2, \dots, y_n\}$$

为什么求解这个最小化问题会得到样本中位数呢？

因为只要 $\mu$ 的取值偏离中位数，就会使得残差绝对值之和上升。

例 考虑一个样本容量为 99 的样本，假设其样本中位数(即第 50 个最小观测值)为 10，参见图 26.2。

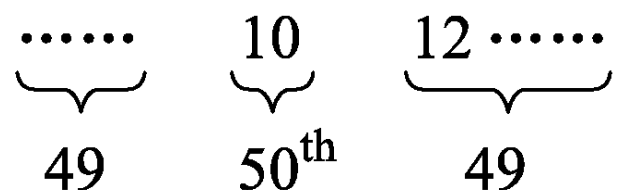


图 26.2

假设第 51 个最小观测值为 12。如让 $\mu = 12$ 而不是 10，则对于前 50 个观测值而言，其残差绝对值 $|y_i - \mu|$ 都将增加 2；

对于后 49 个观测值而言，其残差绝对值 $|y_i - \mu|$ 都将减少 2。

故总变动为 $(50 \times 2) - (49 \times 2) = 2$ ，故第 51 个最小观测值不如第 50 个最小观测值(中位数)更能使目标函数最小化。

同理，第 49 个最小观测值也不如第 50 个最小观测值。

由此可知，第 50 个最小观测值(中位数)是最优解。

**命题** 可以将样本  $q$  分位数视为以下最小化残差绝对值的加权平均问题的最优解：

$$\min_{\mu} \sum_{i: y_i \geq \mu}^n q |y_i - \mu| + \sum_{i: y_i < \mu}^n (1 - q) |y_i - \mu| \Rightarrow \mu = \hat{y}_q$$

例 如果 $q = 1/4$ ，则满足“ $y_i \geq \mu$ ”条件的观测值只得到1/4的权重，而满足“ $y_i < \mu$ ”条件的其余观测值则得到3/4的权重。

因为估计的是1/4分位数(位于总体的底部)，故较大的观测值得到的权重较小，而较小的观测值得到的权重较大。

证明：将目标函数中的绝对值去掉可得

$$\min_{\mu} \sum_{i: y_i \geq \mu}^n q(y_i - \mu) + \sum_{i: y_i < \mu}^n (1 - q)(\mu - y_i)$$

对 $\mu$ 求一阶导数可得

$$\sum_{i: y_i \geq \mu}^n q(-1) + \sum_{i: y_i < \mu}^n (1 - q) = 0$$

假设  $y_{(k)} < \mu \leq y_{(k+1)}$ ，其中  $y_{(k)}$  为第  $k$  个最小观测值，则共有  $k$  个观测值满足 “ $y_i < \mu$ ”， $(n-k)$  个观测值满足 “ $y_i \geq \mu$ ”，故

$$-(n-k)q + k(1-q) = 0$$

经整理可得

$$k = nq$$

$k$  必须是整数。故最优解  $\mu = y_{([nq])} = \hat{y}_q$ ，即样本分位数。

为证明二阶条件满足，只要说明目标函数为凸函数即可。

定义函数  $\rho_q(\cdot)$  为

$$\rho_q(y_i - \mu) \equiv \begin{cases} q|y_i - \mu|, & \text{若 } y_i \geq \mu \\ (1-q)|y_i - \mu|, & \text{若 } y_i < \mu \end{cases}$$

函数  $\rho_q(y_i - \mu)$  的形状如图 26.3。

称为“倾斜的绝对值函数” (tilted absolute value function) 或“打钩函数” (check function)。

从图形易知， $\rho_q(\cdot)$  为凸函数。而目标函数可以写为  $\sum_{i=1}^n \rho_q(y_i - \mu)$ ，即  $n$  个凸函数之和，故仍是凸函数。



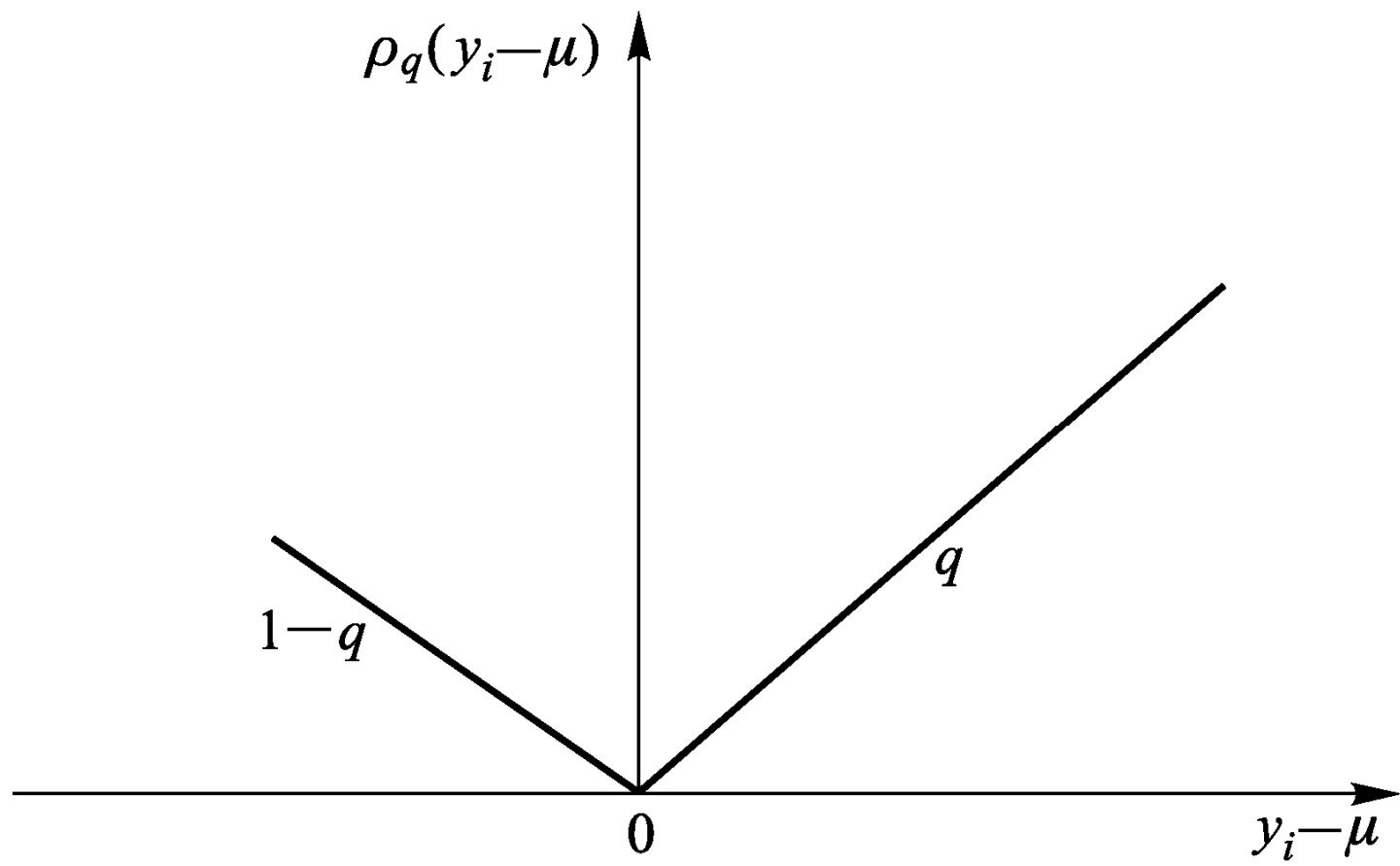


图 26.3 打钩函数  $\rho_q(\cdot)$  及其斜率

## 26.4 分位数回归的估计方法

将单变量情形下对样本分位数的估计方法推广到线性回归。

假设条件分布  $y|\mathbf{x}$  的总体  $q$  分位数  $y_q(\mathbf{x})$  是  $\mathbf{x}$  的线性函数：

$$y_q(\mathbf{x}_i) = \mathbf{x}_i' \boldsymbol{\beta}_q$$

其中， $\boldsymbol{\beta}_q$  称为“ $q$  分位数回归系数”，其估计量  $\hat{\boldsymbol{\beta}}_q$  由以下最小化问题来定义：

$$\min_{\boldsymbol{\beta}_q} \sum_{i: y_i \geq \mathbf{x}_i' \boldsymbol{\beta}_q} q |y_i - \mathbf{x}_i' \boldsymbol{\beta}_q| + \sum_{i: y_i < \mathbf{x}_i' \boldsymbol{\beta}_q} (1-q) |y_i - \mathbf{x}_i' \boldsymbol{\beta}_q|$$

如果  $q = 1/2$ ，则为“中位数回归”(median regression):

$$\min_{\beta_q} \sum_{i=1}^n |y_i - \mathbf{x}_i' \beta_q|$$

中位数回归也称为“最小绝对离差估计量”(Least Absolute Deviation Estimator, 简记 LAD)。

它比均值回归(OLS)更不易受到极端值的影响，更加稳健。

由于分位数回归的目标函数带有绝对值，不可微分，通常使用线性规划的方法来计算  $\hat{\beta}_q$ 。

样本分位数回归系数  $\hat{\boldsymbol{\beta}}_q$  是总体分位数回归系数  $\boldsymbol{\beta}_q$  的一致估计量，且服从渐近正态分布：

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_q - \boldsymbol{\beta}_q) \xrightarrow{d} N(\mathbf{0}, \text{Avar}(\hat{\boldsymbol{\beta}}_q))$$

其中，渐近方差  $\text{Avar}(\hat{\boldsymbol{\beta}}_q) = \mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-1}$ （夹心估计量）， $\mathbf{A} \equiv \text{plim}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n f_{u_q}(0 | \mathbf{x}_i) \mathbf{x}_i \mathbf{x}_i'$ ， $\mathbf{B} \equiv \text{plim}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n q(1-q) \mathbf{x}_i \mathbf{x}_i'$ ，而  $f_{u_q}(0 | \mathbf{x}_i)$  是扰动项  $u_q \equiv y - \mathbf{x}' \boldsymbol{\beta}_q$  的条件密度函数在  $u_q = 0$  处的取值。

要计算  $\text{Avar}(\hat{\boldsymbol{\beta}}_q)$ ，首先要估计  $f_{u_q}(0 | \mathbf{x}_i)$ 。这是 Stata 的默认方法。Stata 也提供自助法作为计算  $\text{Avar}(\hat{\boldsymbol{\beta}}_q)$  的另一方法。

对于  $q$  分位数回归，可使用准  $R^2$  度量其拟合优度，其定义为：

$$1 - \frac{\sum_{i: y_i \geq \mathbf{x}'_i \hat{\boldsymbol{\beta}}_q} q |y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}_q| + \sum_{i: y_i < \mathbf{x}'_i \hat{\boldsymbol{\beta}}_q} (1-q) |y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}_q|}{\sum_{i: y_i \geq \hat{y}_q} q |y_i - \hat{y}_q| + \sum_{i: y_i < \hat{y}_q} (1-q) |y_i - \hat{y}_q|}$$

其中， $\hat{y}_q$  为样本  $q$  分位数，上式第二项的分子为  $q$  分位数回归目标函数的最小值(sum of weighted deviations about estimated quantiles)，而分母为 “sum of weighted deviations about raw quantiles”。