分位数回归以及R实现笔记

马学俊

副教授，苏州大学数学科学学院统计系，主要从事海量数据分析、高维数据分析、统计计算、非参数回归等统计模型及其应用等研究。个人主页https://xuejunma.github.io.

缩写与记号

- s.t. ：约束
- $\boldsymbol{Y}$：向量
- $\mathbf{X}$ ：矩阵（不包含向量）
- $\mathbb{R}^k$：k维欧式空间
- $\top$：转置

目　　录

目录 …… **3**

第1章　分位数回归 …… **1**

1.1　总体分位数定义 …… 1

1.2　分位数回归 …… 3

1.2.1　引言 …… 3

1.2.2　模型表达 …… 4

1.2.3　分位数回归求解 …… 4

1.2.4　大样本理论 …… 6

1.2.5　实例分析 …… 10

1.3　参考文献 …… 13

第2章　局部线性回归 …… **15**

2.1　局部线性回归 …… 15

2.2　带宽的选择 …… 17

2.3　R实现 …… 18

2.4　实例分析 …… 19

2.5　参考文献 …… 22

表　　格

2.1　核函数 ·· 15
2.2　常用的h_τ和h_{mean}之间的关系 ···························· 18

第1章　分位数回归

分位数回归(Quantile Regression)由Koenker和Bassett(1978)提出。相比均值回归(Mean Regression)，它可以全面刻画自变量对条件因变量的分布，对异常值有很强的抵抗性。Yu 等(2003)对分位数回归的发展做了综述。R包`quantreg` (Koenker等, 2015)可以实现分位数回归。

1.1　总体分位数定义

设Y是实值随机变量，$F(y)=P(Y\le y)$是其分布函数，对任意$\tau\in(0,1)$，则它的τ分位数(无条件)定义为

$$Q_\tau(Y)=\text{Arginf}\{y\in\mathbb{R},F(y)\ge\tau\}.$$

若将分布函数$F(x)$的逆定义为$F_Y^{-1}(\tau)=\inf\{y\in\mathbb{R},F(y)\ge\tau\}$，则$Q_\tau(Y)=F_Y^{-1}(\tau)$。这种定义的分位数具有唯一性。对于样本实现值可以使用`quantile`函数得到。

R 代码

```
> y <- rnorm(1000)
> quantile(y,  probs = c(0.1, 0.5,0.9))
       10%          50%          90%
-1.29480525  0.01854632  1.27305081
```

分位数的实现可以转化为最小化问题。为了更好的说明这个问题，我们先讨论大家熟悉的均值(总体)问题。对于随机变量Y，其均值μ可以通过最小化$\mathrm{E}[(Y-\theta)^2]$实现。因为

$$\begin{aligned}\mathrm{E}[(Y-\theta)^2]&=E[Y^2]-2E[Y]\theta+\theta^2\\&=(\theta-E[Y])^2+\{E[Y]^2-(E[Y])^2\}\\&=(\theta-E[Y])^2+\text{Var}(Y)\end{aligned}$$

由于第二项$\text{Var}(Y)$是固定的，所以$\theta=E(Y)=\mu$。这也是为什么均值回归使用均方误差(Average Squared Deviation) 的原因。换句话说这也是最小二乘得到

结果是均值的原因。对于给定的样本$y_1, y_2, \ldots, y_n$，它的均值可以通过最小化

$$\frac{1}{n}\sum_{i=1}^{n}(y_i-\theta)^2$$

实现。

假设Y是连续型随机变量，且分布函数是F，密度函数f，则分位数可以通过最小化$E[\rho_\tau(Y-\theta)]$实现。因为

$$E[\rho_\tau(Y-\theta)] = \int_{-\infty}^{\theta}(\tau-1)(Y-\theta)f(y)dy + \int_{\theta}^{+\infty}\tau(Y-\theta)f(y)dy$$

其中$\rho_\tau(t) = \tau t I(t \geq 0) + (\tau-1)tI(t<0)$是检验函数(Check Function)，$I(\bullet)$是示性函数。不包含示性函数的检验函数可以表示为

$$\rho_\tau(t) = \begin{cases} \tau t & t \geq 0 \\ (\tau-1)t & t < 0 \end{cases} \tag{1.1}$$

由于

$$\begin{aligned}
&\frac{d}{d\theta}\int_{-\infty}^{\theta}(\tau-1)(Y-\theta)f(y)dy \\
&= \int_{-\infty}^{\theta}\frac{\partial}{\partial\theta}(\tau-1)(Y-\theta)f(y)dy + \frac{d\theta}{d\theta}\times(\tau-1)(Y-\theta)f(y)|_{y=\theta} \\
&= (1-\tau)\int_{-\infty}^{\theta}f(y)dy \\
&= (1-\tau)F(\theta) \\
&\frac{d}{d\theta}\int_{\theta}^{+\infty}\tau(Y-\theta)f(y)dy \\
&= \int_{\theta}^{+\infty}\frac{\partial}{\partial\theta}\tau(Y-\theta)f(y)dy - \tau(Y-\theta)f(y)|_{y=\theta} \\
&= -\tau\int_{\theta}^{+\infty}f(y)dy \\
&= -\tau(1-F(\theta))
\end{aligned}$$

所以，

$$\frac{d}{d\theta}E[\rho_\tau(Y-\theta)] = (1-\tau)F(\theta) - \tau(1-F(\theta))$$

$\frac{d}{dt}\int_{a(t)}^{b(t)} f(x,t)dt = \int_{a(t)}^{b(t)}\frac{\partial f(x,t)}{\partial t}dt + f[a(t),t]\frac{da(t)}{dt} - f[b(t),t]\frac{db(t)}{dt}$

进而得到：

$$F(\theta) = \tau$$

对于给定的样本$y_1, y_2, \ldots, y_n$，它的τ分位数可以通过最小化

$$\sum_{i=1}^{n} \rho_\tau(y_i - \theta)$$

得到。

1.2 分位数回归

1.2.1 引言

介绍分位数回归前，我们先直观比较均值回归和分位数回归。以一元均值回归为例说明，$E(Y|X = x) = \beta x$。给定自变量x，均值回归得到的是随机变量$Y|X = x$的均值。也就是图1.1中的实直线。类似的，分位数回归$Q_\tau(Y|X = x) = \beta_\tau x$ 得到的是随机变量$Y|X = x$的τ分位数(图1.1中的虚直线)。

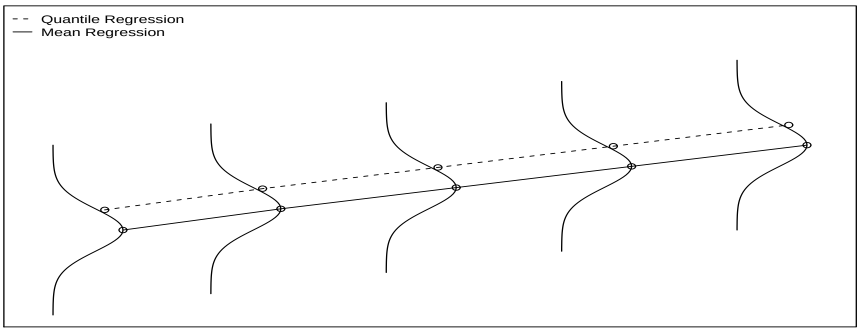

图 1.1: 局部线性回归

图1.1 的R 代码

```
win.graph(width=13,height=6,pointsize=10)
y <- seq(-4,4,0.1)
```

```
x <- dnorm(y)
plot((x+0.1),y,xlim=c(0,4.5), ylim=c(-4,10),xlab="",
     type="l", ylab="", xaxt="n", yaxt="n")
#abline(h=-4)#画竖线
#abline(v=0)#画竖线
points((x+1),(y+1),type="l")
points((x+2),(y+2),type="l")
points((x+3),(y+3),type="l")
points((x+4),(y+4),type="l")
index <- which(x==max(x))
meanx <- x[index] + c(0.1, seq(1:4))
meany <- y[index] + c(0, seq(1:4))
points(meanx,meany, type="o")
points(meanx - dnorm(1.64), meany + 0.95, type="o",lty=2)
legend("topleft",legend=c("Quantile Regression","Mean Regression"),
       lty=c(2,1),cex=0.9,box.lty=0)
```

1.2.2 模型表达

设$\boldsymbol{Y}=(Y_1,Y_2,\ldots,Y_n)^\top$, $\boldsymbol{X}_i=(X_{i1},X_{i2},\ldots,X_{ip})^\top$ ，线性模型：

$$Y_i=\boldsymbol{X}_i^\top\boldsymbol{\beta}+\varepsilon_i \;\; i=1,2,\ldots,n$$

其中τ是分位点(Quantile)，当$\tau=0.5$时，分位数回归是中位数回归。$\boldsymbol{\beta}$是p维参数向量。ε_i 是随机误差项，且满足$Q_\tau(\varepsilon_i|\boldsymbol{X}_i)=0$；则分位数回归是

$$Q_\tau(Y_i|\boldsymbol{X}_i=\boldsymbol{x}_i)=\boldsymbol{x}_i^\top\boldsymbol{\beta}_\tau \tag{1.2}$$

其中$\boldsymbol{x}_i$是$\boldsymbol{X}_i$的实现值，$\boldsymbol{\beta}_\tau$与τ有关。它可以最小化下面式子实现：

$$\sum_i^n\rho_\tau(Y_i-\boldsymbol{x}_i^\top\boldsymbol{\beta}_\tau) \tag{1.3}$$

1.2.3 分位数回归求解

我们引入松弛因子$\boldsymbol{u}=(u_1,u_2,\ldots,u_n)^\top$和$\boldsymbol{v}=(v_1,v_2,\ldots,v_n)^\top$，将(1.3)式转化为线性规划问题。令

$$Y_i-\boldsymbol{X}_i^\top\beta=u_i-v_i$$

其中$u_i = \max(0, Y_i - X_i^\top \boldsymbol{\beta})$表示正部，$v_i = \max(0, -(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}))$表示负部；则

$$
\begin{aligned}
\sum_i^n \rho_\tau(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}) &= \sum_{i=1}^n (\tau u_i - (\tau - 1) v_i) \\
&= \tau \boldsymbol{I}_n^\top \boldsymbol{u} + (1-\tau) \boldsymbol{I}_n^\top \boldsymbol{v} \\
&= (\mathbf{0}_{p\times 1}^\top \boldsymbol{\beta} + \tau \boldsymbol{I}_n^\top \boldsymbol{u} + (1-\tau) \boldsymbol{I}_n^\top \boldsymbol{v} \\
&= (\mathbf{0}_{p\times 1}^\top, \tau \boldsymbol{I}_n^\top, (1-\tau) \boldsymbol{I}_n^\top)(\boldsymbol{\beta}^\top, \boldsymbol{u}^\top, \boldsymbol{v}^\top)^\top \\
&\doteq \mathbf{A}^\top \boldsymbol{\gamma}
\end{aligned}
$$

其中$\boldsymbol{I}_n$是n维单位向量，$\mathbf{A} = (\mathbf{0}_{p\times 1}^\top, \tau \boldsymbol{I}_n^\top, (1-\tau) \boldsymbol{I}_n^\top)^\top$，$\boldsymbol{\gamma} = (\boldsymbol{\beta}^\top, \boldsymbol{u}^\top, \boldsymbol{v}^\top)^\top$。约束条件是

$$\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta} = \boldsymbol{u} - \boldsymbol{v}$$

即

$$\mathbf{X}\boldsymbol{\beta} - \boldsymbol{u} - \boldsymbol{v} = \boldsymbol{Y}$$

从而

$$\begin{bmatrix} \mathbf{X} & \mathbf{E}_n & -\mathbf{E}_n \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \\ \boldsymbol{v} \end{bmatrix} = \boldsymbol{Y}$$

令$\mathbf{B} = (\mathbf{X}, \mathbf{E}_n, -\mathbf{E}_n)$，其中$\mathbf{E}_n$是单位阵，则约束条件转化为$\mathbf{B}\boldsymbol{\gamma} = \boldsymbol{Y}$，从而分位数求解转化为

$$
\begin{aligned}
&\min \mathbf{A}\boldsymbol{\gamma} \\
&s.t. \mathbf{B}\boldsymbol{\gamma} = \boldsymbol{Y}
\end{aligned}
$$

R包`quantreg`中的函数`rq`可以实现。

模拟的R代码

```
rm(list=ls(all=TRUE))#清空所有对象
####产生数据
n <- 100
p <- 2
tau <- 0.5
x <- rnorm(n)
y <- 1 + 2 * x + rnorm(n)
```

```
library(quantreg)
fit.rq <- rq(y~x)
fit.rq$coefficients#rq的解
(Intercept)           x
  0.9304538   2.0448851
```

注1.2.1 (1)分位数回归的结果(参数估计值)与均值回归的结果没有可比性。因为分位数回归得到的分位数回归线，而均值回归得到均值回归线。一般来说，如果误差项的分布中位数和均值相等(如标准正态分布)时，中位数回归与均值回归几乎一样。有些文章中将中位数回归和均值回归相比较。这可以说明数据是否含有异常值。如果两个回归线差距比较大，那么数据有可能存在异常值。(2)分位数回归是求解(1.3)式。读者注意$\boldsymbol{\beta}$ 实际与τ 有关。不同的τ，得到的不同$\boldsymbol{\beta}$。高分位点的分位数回归线理论上应该高于低分位点的分位数回归线，但利用(1.3)式却不能保证。关于如何解决分位数回归相交的方法，详见Yu和Jones(1998)。

1.2.4 大样本理论

为了简单方便，我们假设$Y_1, Y_2, \ldots, Y_n$是独立的随机变量，并且分布函数分别是$F_1, F_2, \ldots, F_n$，假设τ分位条件函数是

$$Q_\tau(Y_i|\boldsymbol{X}_i = \boldsymbol{x}_i) = \boldsymbol{x}_i\boldsymbol{\beta}_\tau$$

Y_i的条件分位数函数也可以写成

$$P(Y_i < y|\boldsymbol{X}_i = \boldsymbol{x}_i) = F_{Y_i}(y|\boldsymbol{x}_i) = F_i(y)$$

所以

$$Q_\tau(Y_i|\boldsymbol{X}_i = \boldsymbol{x}_i) = F_{Y_i}^{-1}(\tau|\boldsymbol{x}_i) \equiv \xi_i(\tau).$$

另外，我们记(1.3)的解为

$$\widehat{\boldsymbol{\beta}}_\tau = \operatorname{argmin}_{\boldsymbol{\beta}_\tau \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau)$$

大样本理论需要下面条件:

C1 $\{F_i\}$绝对连续，有连续密度函数$f_i(\xi)$，它们在点$\xi_i(\tau)$上一致不为0和∞。

C2 存在正定矩阵D_0和$D_1(\tau)$，使得

(1) $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top = D_0$

(2) $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}f_i(\xi_i(\tau))\boldsymbol{x}_i\boldsymbol{x}_i^\top = D_1(\tau)$

(3) $\max_{i=1,2,\ldots,n}\frac{\|\boldsymbol{x}_i\|}{\sqrt{n}}\to 0$

定理1.2.2 *在C1和C2下，*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_\tau) \sim \mathcal{N}(0, \tau(1-\tau)D_1^{-1}D_0D_1^{-1})$$

在独立同分布误差模型下，有

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_\tau) \sim \mathcal{N}(0, \omega^2 D_0^{-1})$$

其中$\omega^2 = \frac{\tau(1-\tau)}{f_i^2(\xi_i(\tau))}$。

证明 我们考虑下面目标函数

$$Z_n(\boldsymbol{\delta}) = \sum_{i=1}^{n}\left[\rho_\tau(u_i - \boldsymbol{x}_i^\top\frac{\delta}{\sqrt{n}}) - \rho_\tau(u_i)\right]$$

其中$u_i = Y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}_\tau$。显然函数$Z_n(\delta)$是凸的(Convex)，并且在

$$\widehat{\boldsymbol{\delta}}_n = \sqrt{n}(\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_\tau)$$

取得最小值。根据Knight(1998)可以得到$\widehat{\boldsymbol{\delta}}_n$的极限决定于$Z_n(\boldsymbol{\delta})$ 的极限。根据Knight等式，有

$$\rho_\tau(u-v) - \rho_\tau(u) = -v\psi_\tau(u) + \int_0^v [I(u\le s) - I(u\le 0)]ds$$

其中$\psi_\tau(u) = \tau - I(u<0)$，所以，$Z_n(\boldsymbol{\delta})$可以重写为

$$Z_n(\boldsymbol{\delta}) = Z_{1n}(\boldsymbol{\delta}) + Z_{2n}(\boldsymbol{\delta})$$

其中

$$Z_{1n}(\boldsymbol{\delta}) = -\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i^\top\boldsymbol{\delta}\psi_\tau(u_i)$$

$$
Z_{2n}(\boldsymbol{\delta}) = \sum_{i=1}^{n} \int_0^{v_{ni}} [I(u_i \le s) - I(u_i \le 0)] ds \equiv \sum_{i=1}^{n} Z_{2ni}(\boldsymbol{\delta})
$$

其中$v_{ni} = \frac{\boldsymbol{x}_i^\top \boldsymbol{\delta}}{\sqrt{n}}$。根据Lindeberg - Feller中心极限定理和条件C2，

$$
Z_{1n}(\boldsymbol{\delta}) \xrightarrow{\mathrm{d}} -\boldsymbol{\delta}^\top \boldsymbol{W},
$$

其中$\boldsymbol{W} \sim \mathcal{N}(0, \tau(1-\tau)D_0)$。

对于$Z_{2n}(\boldsymbol{\delta})$，我们有

$$
Z_{2n}(\boldsymbol{\delta}) = \sum_{i=1}^{n} \mathrm{E} Z_{2ni}(\boldsymbol{\delta}) + \sum_{i=1}^{n} \Big[Z_{2ni}(\boldsymbol{\delta}) - \mathrm{E} Z_{2ni}(\boldsymbol{\delta}) \Big]
$$

由于

$$
\begin{aligned}
\sum_{i=1}^{n} \mathrm{E} Z_{2ni}(\boldsymbol{\delta}) &= \sum_{i=1}^{n} \int_0^{v_{ni}} \Big[F_i(\xi_i + s) - F_i(\xi_i) \Big] \\
&= \frac{1}{n} \sum_{i=1}^{n} \int_0^{\boldsymbol{x}_i^\top \boldsymbol{\delta}} \Big[F_i(\xi_i + t/sqrtn) - F_i(\xi_i) \Big] dt \\
&= n^{-1} \sum_{i=1}^{n} \int_0^{\boldsymbol{x}_i^\top \boldsymbol{\delta}} \sqrt{n} \Big[F_i(\xi_i + t/sqrtn) - F_i(\xi_i) \Big] dt \\
&= n^{-1} \sum_{i=1}^{n} \int_0^{\boldsymbol{x}_i^\top \boldsymbol{\delta}} f_i(\xi_i) dt + o(1) \\
&= (2n)^{-1} \sum_{i=1}^{n} f_i(\xi_i) \boldsymbol{\delta}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{\delta} + o(1) \\
&\to \frac{1}{2} \boldsymbol{\delta}^\top D_1 \boldsymbol{\delta}
\end{aligned}
$$

对于$Z_{2ni}(\boldsymbol{\delta})$，我们有

$$
\begin{aligned}
\mathrm{Var}[Z_{2n}(\boldsymbol{\delta})] &= \sum_{i=1}^{n} \mathrm{E} \Big[Z_{2ni}(\boldsymbol{\delta}) - \mathrm{E}[Z_{2ni}(\boldsymbol{\delta})] \Big] \\
&\le \frac{2}{n} \max_{1 \le i \le n} |\boldsymbol{x}_i^\top \boldsymbol{\delta}| \sum_{i=1}^{n} \mathrm{E}(Z_{2ni})(\boldsymbol{\delta}) \\
&= \frac{2}{n} \max_{1 \le i \le n} |\boldsymbol{x}_i^\top \boldsymbol{\delta}| \sum_{i=1}^{n} \mathrm{E}(Z_{2n})(\boldsymbol{\delta})
\end{aligned}
$$

所以，如果$\boldsymbol{\delta}^\top D_1\boldsymbol{\delta} < \infty$，则

$$Z_{2n}(\delta) - \mathrm{E}(Z_{2n}(\delta)) \xrightarrow{\mathrm{p}} 0, n \to \infty$$

从而$Z_{2n}(\delta) \xrightarrow{\mathrm{p}} \frac{1}{2}\boldsymbol{\delta} D_1 \boldsymbol{\delta}$。如果$\boldsymbol{\delta} D_1 \boldsymbol{\delta} = \infty$，则

$$\begin{aligned}
&P\Big(|Z_{2n}(\delta) - \mathrm{E}(Z_{2n}(\delta)) > \varepsilon \mathrm{E}(Z_{2n}(\delta))|\Big)\\
&\le \frac{\mathrm{Var}(Z_{2n}(\boldsymbol{\delta}))}{\varepsilon^2 \mathrm{E}^2(Z_{2n}(\boldsymbol{\delta}))}\\
&\le 2\frac{\max_{1\le i\le n}|\boldsymbol{x}_i^\top \boldsymbol{\delta}|/\sqrt{n}}{\varepsilon^2 \mathrm{E}^2(Z_{2n}(\boldsymbol{\delta}))}\\
&\to 0
\end{aligned}$$

$$Z_{2n}(\delta) - \mathrm{E}(Z_{2n}(\delta)) \xrightarrow{\mathrm{p}} 0, n \to \infty$$

从而$Z_{2n}(\delta) \xrightarrow{\mathrm{p}} \infty = \frac{1}{2}\boldsymbol{\delta}^\top D_1 \boldsymbol{\delta}$.

由此，我们可以得到

$$Z_n(\boldsymbol{\delta}) \xrightarrow{\mathrm{d}} Z_0(\boldsymbol{\delta}) = -\boldsymbol{\delta}^\top \boldsymbol{W} + \frac{1}{2}\boldsymbol{\delta}^\top D_1 \boldsymbol{\delta}$$

由于$Z(\delta)$有唯一的最小值，所以

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_\tau) = \widehat{\boldsymbol{\delta}}_n = \mathrm{argmin} Z_n(\delta) \xrightarrow{\mathrm{d}} \widehat{\boldsymbol{\delta}}_0 = \mathrm{argmin} Z_0(\boldsymbol{\delta})$$

(参见Pollard(1991)；Hjϕrt 和Pollard(1993)；Knight(1998))。最后，我们可以看出$\widehat{\boldsymbol{\delta}}_0 = D^{-1}\boldsymbol{W}$。

由于$\boldsymbol{W} \sim \mathcal{N}(0, \tau(1-\tau)D_0)$，所以

$$\mathrm{Var}(D_1^{-1}W) = \tau(1-\tau)D_1^{-1}D_0D_1^{-1}$$

从而

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_\tau) \sim \mathcal{N}(0, \tau(1-\tau)D_1^{-1}D_0D_1^{-1}).$$

注1.2.3 sumarry.rq可以实现对于参数估计值得检验，除了可以实现定理*1.2.2*的方法外*(*se="nid"、se="iid"*)*，还可以实现秩方法*(Koenker, 1994;* se="rank"*)*、核方法*(Powell,1990;* se="ker"*)*和自助法*(Bootstrap,* se="boot"*)*。这些方法均在*Koenker(2005)* 有详细阐述。其中核方法也是nid一种方法。nid和ker区别在于对于$f_i(\xi_i(\tau))$ 估计方法。秩方法和自助法适合于小样本，尤其是后者。summery.rq默然秩的方法。需要注意的是秩方法得到秩置信区间，没有参数检验。自助法由于是重复抽样，每一次抽的样本可能不同，所以每一次结果不一样。一般使用boot或nid。

1.2.5 实例分析

例1.2.4 (Engel数据) *Engel*数据包含*235*个观测值，*2*个变量：

- income：家庭收入连续变量自变量
- fooexp：食物支出连续变量因变量

该数据主要用来研究家庭支出和家庭收入之间的关系，可以在quantreg中engel找到。

图*1.2*中长虚线是均值回归的拟合线，短虚线是中位数回归的拟合线，自下而上的实线分别是分位点为$\{0.05, 0.1, 0.25, 0.75, 0.9, 0.95\}$分位数回归拟合线。随着家庭收入水平的提高，食物支出的增长呈现出扩散的趋势。分位数回归拟合线之间的空寂表明食物支出的条件分位数是左偏的*(*田茂再*,2014)*。

R 代码

```
library(quantreg)
data(engel)
attach(engel)
fit.rq <- rq(foodexp~income, tau=0.5)
summary(fit.rq, se="rank")
summary(fit.rq, se="iid")
summary(fit.rq, se="nid")
summary(fit.rq, se="ker")
summary(fit.rq, se="boot")
```

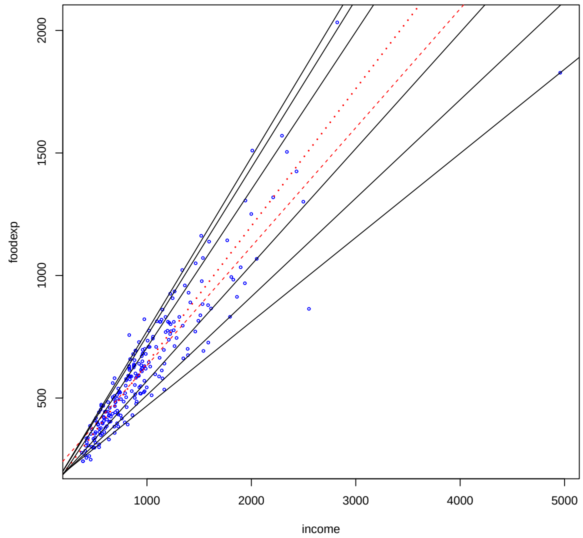

图 1.2: Engle的拟合曲线

```
#画图
win.graph(width=13, height=13,pointsize=10)
plot(income, foodexp, cex=.25,type="n",xlab="income", ylab="foodexp")
points(income, foodexp, cex=0.5,col="blue")
#中位数回归
abline(rq(foodexp~income, tau=0.5), lwd =2,lty=3, col="red")
 #均值回归
abline(lm(foodexp~income), lty=2, col="red")
taus <- c(0.05, 0.1, 0.25, 0.75, 0.90, 0.95)
for( i in 1:length(taus)) {
  abline(rq(foodexp~income, tau=taus[i])) #分位数回归
}
```

输出结果

```
> summary(fit.rq, se="rank")
```

```
Call: rq(formula = foodexp ~ income, tau = 0.5)
tau: [1] 0.5
Coefficients:
            coefficients lower bd  upper bd
(Intercept)  81.48225     53.25915 114.01156
income        0.56018      0.48702   0.60199
> summary(fit.rq, se="iid")
Call: rq(formula = foodexp ~ income, tau = 0.5)
tau: [1] 0.5
Coefficients:
            Value    Std. Error t value  Pr(>|t|)
(Intercept) 81.48225 13.23908    6.15468  0.00000
income       0.56018  0.01192   46.99766  0.00000
> summary(fit.rq, se="nid")
Call: rq(formula = foodexp ~ income, tau = 0.5)
tau: [1] 0.5
Coefficients:
            Value    Std. Error t value  Pr(>|t|)
(Intercept) 81.48225 19.25066    4.23270  0.00003
income       0.56018  0.02828   19.81032  0.00000
> summary(fit.rq, se="ker")
Call: rq(formula = foodexp ~ income, tau = 0.5)
tau: [1] 0.5
Coefficients:
            Value    Std. Error t value  Pr(>|t|)
(Intercept) 81.48225 30.21532    2.69672  0.00751
income       0.56018  0.03732   15.01139  0.00000
> summary(fit.rq, se="boot")
Call: rq(formula = foodexp ~ income, tau = 0.5)
tau: [1] 0.5
Coefficients:
            Value    Std. Error t value  Pr(>|t|)
```

```
(Intercept) 81.48225 28.42273    2.86680  0.00453
income       0.56018  0.03771   14.85645  0.00000
```

1.3 参考文献

1. 田茂再. 复杂数据统计推断理论、方法及应用. 科学出版社, 2014.
2. Hjϕrt N. , Pollard D.(1993). Asymptotics for Minimizers of Convex Processes. Statistical Research Report.
3. Koenker R. Quantile Regression. Cambridge, 2005.
4. Knight K. (1998). Limiting Distributions for L1 Regression Estimators under General Conditions. Annals of Statistics, 26, 755 - 770.
5. Koenker R., Bassett G. (1978). Regression quantiles. Econometrica, 46, 33 - 50.
6. Koenker R.等(2015). quantreg. R package version 5.19, `http://CRAN.R-project.org/package=quantreg`.
7. Pollard D. (1991). Asymptotics for Least Absolute Deviation Regression Estimators. Econometric Theory, 7, 186 - 199.
8. Yu, K., Jones, M. (1998). Local linear quantile regression. Journal of the American statistical Association, 93(441), 228-237.
9. Yu K., Lu Z., Stander J. (2003). Quantile regression: applications and current research areas. Journal of the Royal Statistical Society: Series D (The Statistician), 52, 331–350.
10. Zou H., Yuan M.(2008). Composite quantile regression and the oracle model selection theory. The Annals of Statistics,36(3): 1108 - 1126.

第2章 局部线性回归

局部线性回归(Locally Linear Regression)是非参数回归最基本的理论之一，它的思想广泛应用于非参数回归求解，如变系数模型等。本文主要介绍局部线性回归的思想、软件实现和应用，至于大样本理论性质，详见(田茂再, 2014; Fan等,1994)

2.1 局部线性回归

设$\{Y_i, X_i\}_{i=1}^n$是样本序列，我们考虑下面模型：

$$Y_i = m(X_i) + \varepsilon_i \tag{2.1}$$

其中ε_i是独立同分布，且分布未知。$m(\bullet)$是未知函数。如果$m(\bullet)$是线性函数，那么模型2.1就是线性模型。

目前比较流行的拟合方式是局部线性拟合，即对于任意一点x

$$(\hat{a}, \hat{b}) = \text{argmin}_{a,b\in\mathbb{R}} \sum_{i=1}^{n} l\Big(Y_i - a - b(X_i - x)\Big) K\Big(\frac{X_i - x}{h}\Big) \tag{2.2}$$

其中$l(\bullet)$是一个凸函数，并且在原点有唯一的最小值。$\hat{m}(x) = \hat{a}$，$\hat{m}'(x) = \hat{b}$。当$l(u) = u^2$得到的回归是均值回归(locally Linear Mean Regression)；当$l(u) = \rho_\tau(u)$ 得到的回归是分位数回归(locally Linear Quantile Regression)。h 带宽(Bandwidth)。它决定估计曲线的光滑程度，其越大越光滑。$K(\bullet)$ 是核函数(Kernel Function)，它满足$\int_{-\infty}^{+\infty} K(z)dz = 1$且$\int_{-\infty}^{+\infty} zK(z)dz = 0$。经常使用的核函数详见表2.1

表 2.1: 核函数

名称	$K(\bullet)$
均匀核(Uniform)	$\frac{1}{2}I(\|z\| \le 1)$
Epanechnikov	$\frac{3}{4}(1-z^2)I(\|z\| \le 1)$
高斯核(Gaussian)	$\frac{1}{2\pi}\exp\Big(-\frac{1}{2}z^2\Big)$

下面简单介绍局部线性回归的思想(根据田茂再老师上课内容整理)：

如图2.1所示，要估计$m(x)$ 时，需要利用$(x, m(x))$ 附近(X_i, Y_i) 信息使得

$$\min \sum_{i=1}^{n} l(Y_i - m(X_i)) \tag{2.3}$$

由于$m(X_i)$未知，通常用$\widehat{Y}_i$代替。$(X_i, \widehat{Y}_i)$ 在过$(x, m(x))$ 的切线上(图2.1中实直线)。可以证明$\widehat{Y}_i$ 其实是$m(X_i)$在x的一阶Taylor展开式，即

$$m(X_i) \approx m(x) + m'(x)(X_i - x) \tag{2.4}$$

所以(2.3)可以转化为

$$\sum_{i=1}^{n} l\Big(Y_i - m(x) - m'(x)(X_i - x)\Big) \tag{2.5}$$

考虑到(X_i, Y_i) 与$(x, m(x))$ 距离不同，起的作用不同，越近的作用越大，所以采用局部加权的方法，(2.5)式进一步演变为

$$\sum_{i=1}^{n} l\Big(Y_i - m(x) - m'(x)(X_i - x)\Big) K\Big(\frac{X_i - x}{h}\Big) \tag{2.6}$$

(2.6)即是局部线性回归的求解式。比较(2.6)式和(2.2)式，可以得到$a = m(x)$，$b = m'(x)$。

$\frac{\widehat{Y}_i - m(x)}{X_i - x} = m'(x)$

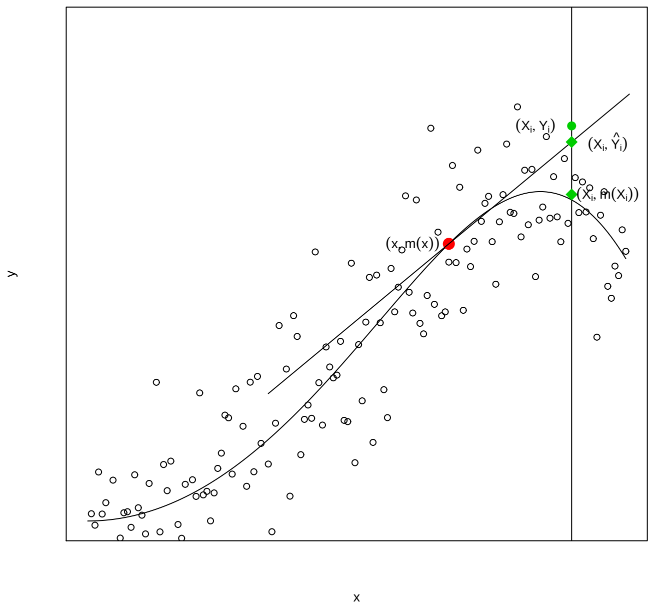

图 2.1: 局部线性回归

2.2　带宽的选择

h的选择方法非常多，可以详见Härdle等(2004)。下面我们介绍经常使用的方法：交叉核实(Cross Validation, CV)。交叉核实被广泛应用于核密度估计、核回归以及样条光滑等，其定义为：

$$CV(h)=\sum_{i=1}^{n}\left(Y_i-\widehat{Y}_{(-i)}\right)^2$$

其中$\widehat{Y}_{(-i)}$ 是去掉Y_i的估计。h 可以最小$CV(h)$得到。

相对于局部线性分位数回归的h_τ，局部线性均值回归的h_{mean}比较容易获得。R包`KernSmooth` 中的函数`dpill` 可以实现Ruppert等(1995)提出的插入方法(Plug-in)。Yu和Jone(1998) 得到h_τ和h_{mean}的关系是

$$h_\tau=\frac{\tau(1-\tau)}{\phi^2(\Phi^{-1}(\tau))}h_{mean}$$

其中ϕ和Φ分别是标准正态分布的密度函数和分布函数。

表 2.2: 常用的h_τ和h_{mean}之间的关系

τ	0.05或0.95	0.25或0.75	0.5
h_τ	$1.34h_{mean}$	$1.13h_{mean}$	$1.10h_{mean}$

2.3 R实现

局部线性均值回归可以使用KernSmooth包中的locpoly实现；局部线性分位数回归可以用R包quantreg中的lprq实现，它使用高斯核函数。本文主要介绍局部线性分位数回归实现。

```
lprq(x, y, h, tau = .5, m = 50)

用法
x 自变量
y 因变量
h 带宽
tau 分为点
m 需要拟合的点数

输出结果
xx 需要估计的自变量
fv xx拟合值
dev xx拟合值一阶导数
```

lprq得到不是x的拟合值，而是xx = seq(min(x),max(x), length=m)的拟合值，也即x最小值和最大值之间m个等距插值的拟合值。?lprq查看lprq的源代码，修改部分参数实现x 的拟合值。修改后的函数命名为lprq.my。

```
lprq.my <- function (x, y, h, tau = 0.5, m = 50) {
  #xx <- seq(min(x), max(x), length = m) 去掉
  xx <- x #修改的参数
  fv <- xx
```

```
  dv <- xx
  for (i in 1:length(xx)) {
    z <- x - xx[i]
    wx <- dnorm(z/h)
    r <- rq(y ~ z, weights = wx, tau = tau, ci = FALSE)
    fv[i] <- r$coef[1]
    dv[i] <- r$coef[2]
  }
  list(xx = xx, fv = fv, dv = dv)
}
```

2.4　实例分析

数据`fossil`来自于R包`SemiPar`，它有106个观测值，两个变量年龄(age，单位是百万年)和锶同位素的百分比(strontium.ratio) (Bralower 等,1997; Chaudhuri 和Marron, 1999)。锶同位素的百分比随着年龄而变化。它们的散点图如图2.2所示。

我们将对其进行局部线性分位数回归拟合。取分位点分别是0.05，0.5，0.95。从图2.2 可以看出，局部线性分位数回归拟合的效果是非常好的，它能刻画锶同位素的百分比随年龄的变化。另外，利用局部线性分位数回归，我们可以得到某年的锶同位素的百分比的概率值表，这有利于地质学家查阅年龄和锶同位素的百分比(见输出结果，本文只是粗糙的列举了5个，读者可以试试全部的)。如当年龄为112.33691，即112.33691百万年前，锶同位素的百分比不超过0.7072158519的概率为5%，不超过0.7072603496的概率为50%，不超过0.70732的概率为95%。需要注意的是从`summary`得到`strontium.ratio` 数据前几位相同，即都是0.707，所以我们利用`(strontium.ratio - 0.707) * 10^3`命令对`strontium.ratio` 进行变换。

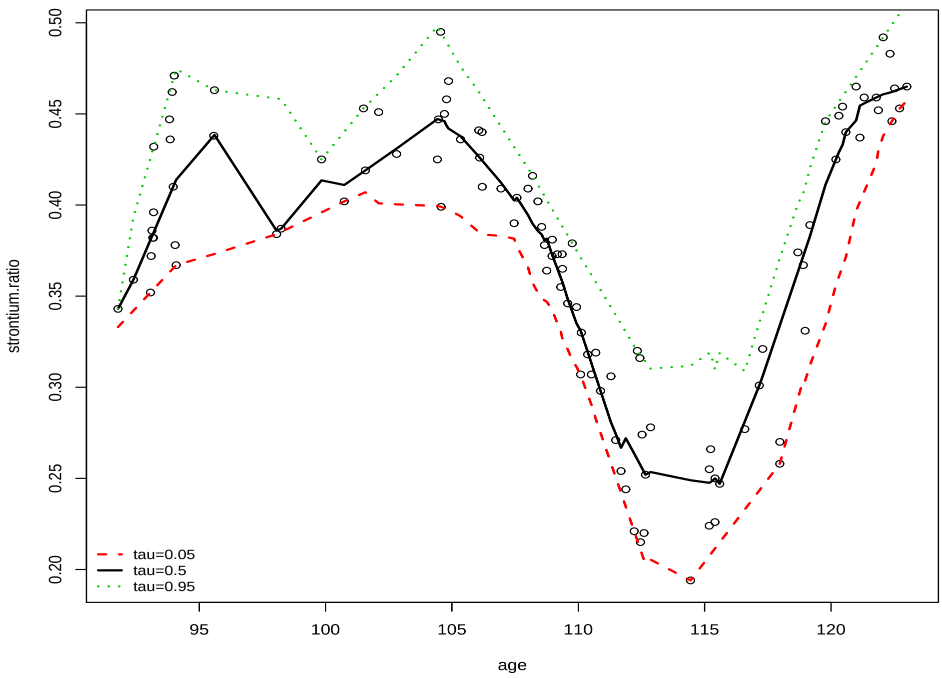

图 2.2: 数据fossil拟合图

R 代码

```
library(SemiPar)
library(KernSmooth)
library(quantreg)
data(fossil)
attach(fossil)
summary(fossil)
strontium.ratio <- (strontium.ratio - 0.707) * 10^3
#画图
win.graph(width=13, height=10,pointsize=10)
plot(age,strontium.ratio)
#带宽
h.mean <- dpill(x=age, y=strontium.ratio)
```

```
h.05 <- 1.34 * h.mean
h.95 <- h.05
h.50 <- 1.10 * h.mean
fit.05 <- lprq.my(age, strontium.ratio, h=h.05, tau=.05)
index <- order(fit.05$xx)
lines(fit.05$xx[index], fit.05$fv[index], col=2, lwd=2, pch=2, lty=2)
fit.50 <- lprq.my(age, strontium.ratio, h=h.50, tau=0.5)
lines(fit.50$xx[index], fit.50$fv[index], col=1, lwd=2, pch=1, lty=1)
fit.95 <- lprq.my(age, strontium.ratio, h=h.95, tau=0.95)
lines(fit.95$xx[index], fit.95$fv[index], col=3, lwd=2, pch=3, lty=3)
legend("bottomleft",legend=c("tau=0.05","tau=0.5","tau=0.95"),
       col=c(2,1,3),lty=c(2,1,3),lwd=c(2,2,2),cex=0.9, box.lty=0)
#查看部分拟合值
head(cbind(fit.05$xx,fit.05$fv,fit.50$fv,fit.95$fv))
```

输出结果

```
> summary(fossil)
      age          strontium.ratio
 Min.   : 91.79   Min.   :0.7072
 1st Qu.:104.43   1st Qu.:0.7073
 Median :109.48   Median :0.7074
 Mean   :108.78   Mean   :0.7074
 3rd Qu.:115.41   3rd Qu.:0.7074
 Max.   :123.00   Max.   :0.7075

> head(cbind(fit.05$xx,fit.05$fv,fit.50$fv,fit.95$fv))
          [,1]      [,2]      [,3]      [,4]
[1,]  91.78525 0.3330150 0.3430000 0.3430000
[2,]  92.39579 0.3420259 0.3590000 0.3933725
[3,]  93.97061 0.3652687 0.4100689 0.4690135
```

```
[4,]  95.57577 0.3730354 0.4380000 0.4631363
[5,]  95.60286 0.3731550 0.4384726 0.4630000
[6,] 112.33691 0.2158519 0.2603496 0.3200000
```

2.5 参考文献

1. 田茂再. 复杂数据统计推断理论、方法及应用. 科学出版社, 2014.

2. Bralower T.等.(1997). Mid-Cretaceous Strontium-Isotope Stratigraphy of Deep-Sea Sections. Geological Society of America Bulletin, 109(11), 1421-1442

3. Chaudhuri P., Marron J. (1999). Sizer for Exploration of Structures in Curves. Journal of the American Statistical Association, 94(447), 807-823

4. Fan J., Hu T., Truong Y.(1994) Robust Non-Parametric Function Estimation. Scandinavian Journal of Statistics, Vol. 21(4), 433-446.

5. Hjϕrt N. , Pollard D.(1993). Asymptotics for Minimizers of Convex Processes. Statistical Research Report.

6. Härdle W.等. Nonparametric and Semiparametric Models. Springer, 2004.

7. Ruppert D., Sheather, S., Wand M. (1995). An effective bandwidth selector for local least squares regression. Journal of the American Statistical Association, 90, 1257 - 1270.

8. Yu K. , Jones M. (1998).Local Linear Quantile Regression. Journal of the American Statistical Association, 93(441):228–237.