

第 27 章 非参数与半参数估计

27.1 为什么需要非参数与半参数估计

“参数估计法”(parametric estimation)假设总体服从带未知参数的某个分布(比如正态), 或具体的回归函数, 然后估计这些参数。

其缺点是, 对模型设定所作的假定较强, 可能导致较大的设定误差, 不够稳健。

“非参数估计法” (nonparametric estimation)一般不对模型的具体分布或函数形式作任何假定，更为稳健。

缺点是要要求样本容量较大，且估计量收敛的速度较慢。

作为折衷，同时包含参数部分与非参数部分的“半参数方法” (semiparametric estimation)，降低对样本容量的要求，又有一定稳健性。

非参及半参方法与传统的参数法互补；后者不太适用时，可考虑前者。

27.2 对密度函数的非参数估计

考虑根据样本数据来推断总体的分布，即密度函数。

如用参数估计法，则先对总体分布的具体形式进行假定。

比如，假设总体服从正态分布 $N(\mu, \sigma^2)$ ，然后估计参数 (μ, σ^2) 。
如果真实总体与正态分布相去甚远，则统计推断有较大偏差。

如不假设总体分布的具体形式，则为非参数方法。

最原始的非参数方法是画直方图，即将数据的取值范围等分为若干组，计算数据落入每组的频率，以此画图，作为对密度函数的估计。

直方图的缺点是，即使随机变量连续，直方图始终是不连续的阶梯函数。

为得到对密度函数的光滑估计，Rosenblatt(1956)提出“核密度估计法”(kernel density estimation)。

首先考察直方图的数学本质。假设要估计连续型随机变量 x 在 x_0 处的概率密度 $f(x_0)$ 。

概率密度 $f(x_0)$ 是累积分布函数 $F(x)$ 在 x_0 处的导数：

$$\begin{aligned} f(x_0) &= \lim_{h \rightarrow 0} \frac{F(x_0 + h) - F(x_0 - h)}{2h} \\ &= \lim_{h \rightarrow 0} \frac{P(x_0 - h < x < x_0 + h)}{2h} \end{aligned}$$

对于样本 $\{x_1, x_2, \dots, x_n\}$ ，用数据落入区间 $(x_0 - h, x_0 + h)$ 的频率来估计概率 $P(x_0 - h < x < x_0 + h)$ ，得到直方图估计量：

$$\begin{aligned}\hat{f}_{\text{HIST}}(x_0) &= \frac{\sum_{i=1}^n \mathbf{1}(x_0 - h < x_i < x_0 + h) / n}{2h} \\ &= \frac{1}{nh} \sum_{i=1}^n \frac{1}{2} \cdot \mathbf{1}\left\{\left|\frac{x_i - x_0}{h}\right| < 1\right\}\end{aligned}$$

$\hat{f}_{\text{HIST}}(x_0)$ 对于区间 $(x_0 - h, x_0 + h)$ 内的观测值给予相同权重，而区间外的观测值权重为 0。

区间半径 h 定义了“在 x_0 附近邻域的大小”，称为“带宽”(bandwidth)。 $2h$ 称为“窗宽”(window width)。

直方图得不到光滑的密度估计，根本原因在于使用示性函数作为“权重函数”(weighting function)，以及各组间不允许交叠。

核密度估计法使用更一般的权重函数，并允许各组之间交叠。

核密度估计量为

$$\hat{f}(x_0) = \frac{1}{nh} \sum_{i=1}^n K[(x_i - x_0)/h]$$

函数 $K(\cdot)$ 称为“核函数”(kernel function)，本质上就是权重函数。

带宽 h 越大，在 x_0 附近邻域越大，则估计的密度函数 $\hat{f}(x)$ 越光滑，故称带宽 h 为“光滑参数”(smoothing parameter)。

一般假设核函数 $K(z)$ 满足以下性质：

(i) $K(z)$ 连续且关于原点对称(偶函数)；

$$(ii) \int_{-\infty}^{+\infty} K(z) dz = 1, \int_{-\infty}^{+\infty} zK(z) dz = 0, \int_{-\infty}^{+\infty} |K(z)| dz < +\infty;$$

(iii) 或者①存在 $z_0 > 0$ ，使得当 $|z| > z_0$ 时， $K(z) = 0$ ；或者②当 $|z| \rightarrow +\infty$ 时， $|z|K(z) \rightarrow 0$ ；

$$(iv) \int_{-\infty}^{+\infty} z^2 K(z) dz = \gamma, \text{ 其中 } \gamma \text{ 为常数。}$$

条件(ii)要求核函数的曲线下面积为 1，并满足一些有界条件。

条件(iii)①比条件(iii)②更强，实践中常采用条件(iii)①。常将邻

域 $[-z_0, z_0]$ 标准化为 $[-1, 1]$ 。条件(iv)也是有界条件。

常见核函数见表 27.1。这些核函数的共同特点是，离原点越近，则核函数取值越大，并在原点达到最大；即越近的点权重越大。

其中，均匀核也用于直方图，只是在用均匀核进行核密度估计时并不固定分组，而在每个点上进行估计。

最流行的核函数为二次核(也称 Epanechnikov 核)与高斯核。

表 27.1 常用的核函数

核函数名称	核函数的数学形式	δ
均匀核 (uniform or rectangular)	$\frac{1}{2} \cdot \mathbf{1}(z < 1)$	1.3510
三角核 (triangular or Bartlett)	$(1 - z) \cdot \mathbf{1}(z < 1)$	—
伊 番 科 尼 可 夫 核 (Epanechnikov) 或二次核(quadratic)	$\frac{3}{4}(1 - z^2) \cdot \mathbf{1}(z < 1)$	1.7188
四次核(quartic)	$\frac{15}{16}(1 - z^2)^2 \cdot \mathbf{1}(z < 1)$	2.0362

或双权核(biweight)		
三权核(Triweight)	$\frac{35}{32}(1-z^2)^3 \cdot \mathbf{1}(z < 1)$	2.3122
三三核(Tricubic)	$\frac{70}{81}(1- z ^3)^3 \cdot \mathbf{1}(z < 1)$	—
高 斯 核 (Gaussian or Normal)	$\frac{1}{\sqrt{2\pi}} \exp\{-z^2/2\}$	0.7764

注：其中 δ 为用来计算“Silverman 嵌入估计”的常数。

给定核函数 $K(\cdot)$ 与带宽 h ，可估计核密度 $\hat{f}(x_0)$ 。在 Stata 中，默认设置为在等距离的 $\min(n, 50)$ 个点来计算 $\hat{f}(x_0)$ ，然后连成光滑的密度函数。

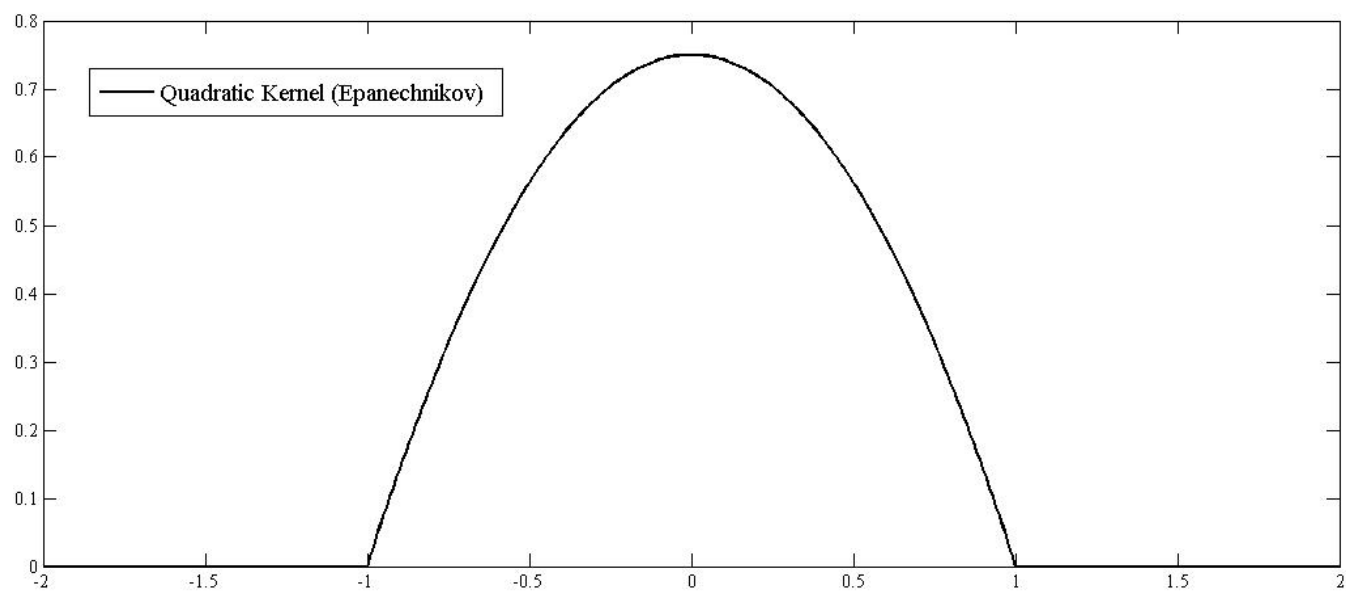


图 27.1 二次核(Epanechnikov 核)

27.3 核密度估计的性质

由于核密度估计使用了在 x_0 附近的点 x 来估计 $\hat{f}(x_0)$ ，而一般地，如果 $x \neq x_0$ ，则 $f(x) \neq f(x_0)$ ，故核密度估计通常是有偏的：

$$\text{Bias}(x_0) \equiv E[\hat{f}(x_0)] - f(x_0) \approx \frac{1}{2} h^2 f''(x_0) \int_{-\infty}^{+\infty} z^2 K(z) dz$$

即偏差与 h^2 成正比，为 h^2 的同阶无穷小，记为 $O(h^2)$ 。

带宽 h 越大，则将使用离 x_0 更远的点在估计 $f(x_0)$ ，导致偏差增大（以 h^2 的速度迅速上升）。

当 $n \rightarrow \infty$ 时，让带宽 $h \rightarrow 0$ ，则偏差将在大样本中消失。

密度函数的二阶导数 $f''(x_0)$ 越大，即在 x_0 处的曲率越大，则 x_0 附近的函数值波动越大，也会引起偏差增大。

偏差还取决于核函数 $K(z)$ 。

核密度估计的方差为：

$$\text{Var}[\hat{f}(x_0)] = \frac{1}{nh} f(x_0) \int_{-\infty}^{+\infty} K(z)^2 dz + o(1/nh)$$

故 $\text{Var}[\hat{f}(x_0)] = O(1/nh)$ ，是 $(1/nh)$ 的同阶无穷小。

样本容量 n 越大，则方差越小；

带宽 h 越大，由于使用了更多观测点来估计 $f(x_0)$ ，故方差越小。

当 $n \rightarrow \infty$ 时，让 $nh \rightarrow \infty$ (虽然 $h \rightarrow 0$ ，但 h 趋于 0 的速度比样本容量 $n \rightarrow \infty$ 的速度更慢)，则此方差将在大样本中消失。

核密度估计的一致性

当 $n \rightarrow \infty$ 时，让带宽 $h \rightarrow 0$ 且 $nh \rightarrow \infty$ ，则偏差 $\text{Bias}(x_0)$ 与方差 $\text{Var}[\hat{f}(x_0)]$ 在大样本下都趋于 0。根据均方收敛可知， $\hat{f}(x_0)$ 是 $f(x_0)$ 的一致估计量。

核密度估计的渐近正态性

如果核函数 $K(z)$ 的条件(iv)满足, 则 $\hat{f}(x_0)$ 服从渐近正态分布:

$$\sqrt{nh} \left[\hat{f}(x_0) - f(x_0) - \text{Bias}(x_0) \right] \xrightarrow{d} N \left(0, f(x_0) \int_{-\infty}^{+\infty} K(z)^2 dz \right)$$

据此可进行区间估计。

核密度估计量的收敛速度为 \sqrt{nh} 。

由于最优带宽 h^* 与 $n^{-0.2}$ 成正比(参见下节), 故

$$\sqrt{nh} = \sqrt{n \cdot n^{-0.2}} = \sqrt{n^{0.8}} = n^{0.4} < n^{0.5} = \sqrt{n}$$

这意味着非参估计量的收敛速度 $n^{0.4}$ 慢于参数估计量的通常收敛速度 $n^{0.5}$ 。

27.4 最优带宽

如果带宽 h 越大, 则 x_0 附近的邻域越大, 故偏差也越大(偏差与 h^2 成正比); 而带宽 h 越大, 则 $\hat{f}(x_0)$ 越光滑, 即方差 $\text{Var}[\hat{f}(x_0)]$ 越小。

在选择“最优带宽”(optimal bandwidth) h^* 时, 希望最小化均方误差(MSE), 即方差与偏差平方之和:

$$\min_h \text{MSE}[\hat{f}(x_0)] = [\text{Bias}(x_0)]^2 + \text{Var}[\hat{f}(x_0)]$$

由于 $\text{Bias}(x_0) = O(h^2)$, 故 $[\text{Bias}(x_0)]^2 = O(h^4)$, 而 $\text{Var}[\hat{f}(x_0)] = O(1/nh)$, 故此最小化问题可大致写为

$$\min_h \text{MSE}[\hat{f}(x_0)] = k_1 h^4 + (k_2/nh)$$

其中, k_1, k_2 为常数。对 h 求导, 可得一阶条件为

$$4k_1 h^3 + k_2 \frac{1}{n} (-1/h^2) = 0$$

$$h = (4k_1/k_2)^{-0.2} n^{-0.2}$$

故最优带宽为 $h^* = O(n^{-0.2})$ 。

随着 n 增大, $n^{-0.2} = 1/\sqrt[5]{n}$ 的下降速度远慢于 $n^{-1} = 1/n$ 。

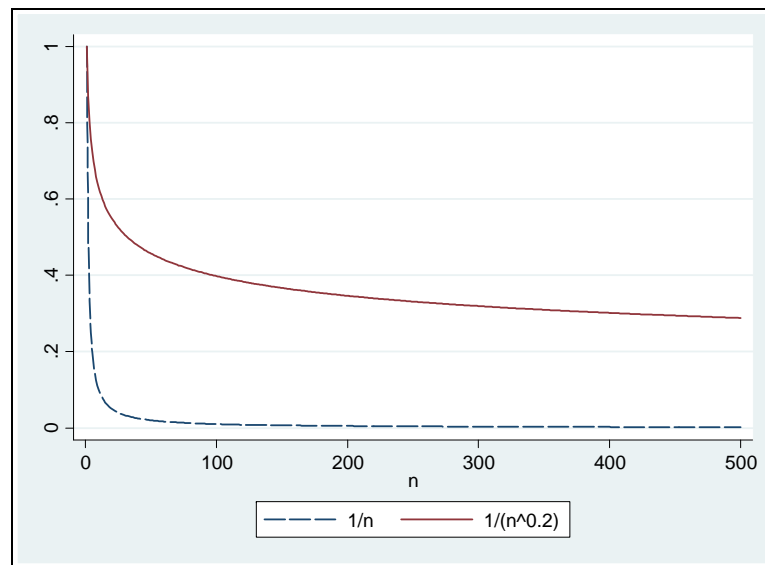


图 27.2 对比 $n^{-0.2}$ 与 n^{-1} 的下降速度

当 $n \rightarrow \infty$ 时, $h^* \rightarrow 0$, 而 $nh^* = n \cdot O(n^{-0.2}) = O(n^{0.8}) \rightarrow \infty$ 。

选择最优带宽 h^* , 就能保证核密度估计的一致性。

均方误差 $\text{MSE}[\hat{f}(x_0)]$ 仍取决于 x_0 。为得到对于 x_0 所有可能取值的整体度量, 可最小化“积分均方误差”(Integrated Mean Squared Error, 简记 IMSE):

$$\min_h \text{IMSE} \equiv \int_{-\infty}^{+\infty} \text{MSE}[\hat{f}(x_0)] dx_0$$

Silverman(1986)证明最优带宽为:

$$h^* = \delta \left[\int_{-\infty}^{+\infty} f''(x_0)^2 dx_0 \right]^{-0.2} n^{-0.2}$$

其中，常数 $\delta \equiv \left[\int_{-\infty}^{+\infty} K(z)^2 dz / \left(\int_{-\infty}^{+\infty} z^2 K(z) dz \right)^2 \right]^{0.2}$ 仅依赖于核函数。

最优带宽 h^* 还取决于密度函数的曲率 ($f''(x_0)$)。

当密度函数波动较大时，将带来较大偏差，故最优带宽 h^* 较小。

由于 δ 依赖于核函数，故最优带宽 h^* 也依赖于核函数。

对于不同的核函数分别使用相应的最优带宽，则积分均方误差 $\text{IMSE}(h^*)$ 差别不大。

能使 $\text{IMSE}(h^*)$ 最小化的核函数为 “伊 番 科 尼 可 夫 核 ” (Epanechnikov), 是 Stata 默认核函数, 但只有微弱优势。

对于最优带宽的选择远比核函数的选择更重要。使用不同核函数得到的密度估计一般非常接近。

最优带宽 h^* 仍依赖于 $f''(x_0)$ 。如果样本来自正态总体, 则 $\int_{-\infty}^{+\infty} f''(x_0)^2 dx_0 = 3/(8\sqrt{\pi}\sigma^5) = 0.2116/\sigma^5$, 故

$$h^* = 1.3643 \delta n^{-0.2} s$$

其中, s 为样本标准差。为了防止样本标准差受极端值的影响, 常使用 “Silverman 嵌入估计” (Silverman’s plug-in estimate):

$$h^* = 1.3643 \delta n^{-0.2} \min(s, iqr/1.349)$$

其中, “*iqr*” 为样本四分位距(sample interquartile range), 即样本3/4分位数与1/4分位数之间的距离。

为保险起见, 可比较两倍嵌入估计与一半嵌入估计的效果。

实践中也常使用“眼球法”(eyeball method):

用肉眼对带宽进行判断, 是否密度函数“过度光滑”(oversmoothed)或“不够光滑”(undersmoothed), 再微调到合适的带宽。

27.5 多元密度函数的核估计

对于 k 维随机变量 \mathbf{x} ，可进行“多元密度函数的核估计”：

$$\hat{f}(x_0) = \frac{1}{nh} \sum_{i=1}^n K[(\mathbf{x}_i - \mathbf{x}_0)/h]$$

其中， $K(\cdot)$ 是 k 维核函数，即权重函数。 $K(\cdot)$ 通常为一维核函数的乘积，也可使用多维正态的密度函数。

多元密度函数核估计的性质与一元情形相似。但最优带宽为 $h^* = O(n^{-1/(k+4)})$ (大于一元情形下的最优带宽)，而 $\hat{f}(x_0)$ 的收敛速度也更慢。

在多维情况下，易出现“数据稀疏”问题(sparseness of data)，即在 \mathbf{x}_0 附近的观测点很少。

估计多维密度函数的用途之一是估计条件密度函数(conditional density function)。

由于条件密度 $f(y|x) = f(x,y)/f(x)$,

故可用 $\hat{f}(y|x) = \hat{f}(x,y)/\hat{f}(x)$ 作为条件密度的估计量，

其中， $\hat{f}(x,y)$ 与 $\hat{f}(x)$ 分别为二维与一维的密度函数核估计。

27.6 非参数核回归

考虑以下非参数一元回归模型：

$$y_i = m(x_i) + \varepsilon_i$$

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

其中， $m(\cdot)$ 是未知函数(连函数形式也未知)。

对于每一个 i ($i = 1, \dots, n$)，分别估计 $m(x_i)$ ，从而得到对回归函数 $m(x)$ 的估计。不寻求 $m(x)$ 的解析解，而是寻找其数值解。

假设对于 x 的某个特定取值，比如 x_0 ，都有若干个 y 的观测值，比如 n_0 个。则可把这 n_0 个 y 观测值的平均值作为 $m(x_0)$ 的估计量。

现实数据中， n_0 可能很小(对于连续变量，可能仅为 1)，导致估计量的方差过大。

解决方法是，对 x_0 附近邻域中的观测值也进行加权平均，即“局部加权平均估计量”(local weighted average estimator)：

$$\hat{m}(x_0) = \sum_{i=1}^n w_{i0,h} y_i$$

其中，权重 $w_{i0,h}$ 是 (x_i, x_0, h) 的函数，即 $w_{i0,h} = w(x_i, x_0, h)$ ，且满足 $\sum_{i=1}^n w_{i0,h} = 1$ 。 x_i 是 x_0 附近的点，而 h 是带宽。

Nadaraya(1964)与 Watson(1964) 使用核函数来定义以下权重, 得到“核回归估计量”(kernel regression estimator):

$$w_{i0,h} = \frac{K[(x_i - x_0)/h]}{\sum_{i=1}^n K[(x_i - x_0)/h]}$$

故核回归估计量可写为

$$\hat{m}(x_0) = \frac{\sum_{i=1}^n K[(x_i - x_0)/h] y_i}{\sum_{i=1}^n K[(x_i - x_0)/h]}$$

由于使用了 x_0 附近邻域的信息，核回归估计量 $\hat{m}(x_0)$ 有偏：

$$\text{Bias}(x_0) \equiv \mathbb{E}[\hat{m}(x_0)] - m(x_0) = h^2 \left[m'(x_0) \frac{f'(x_0)}{f(x_0)} + \frac{1}{2} m''(x_0) \right] \int_{-\infty}^{+\infty} z^2 K(z) \mathrm{d}z$$

故 $\text{Bias}(x_0) = O(h^2)$ 。核回归估计的方差为

$$\text{Var}[\hat{m}(x_0)] = \frac{1}{nh} \frac{\sigma_\varepsilon^2}{f(x_0)} \int_{-\infty}^{+\infty} K(z)^2 \mathrm{d}z + o(1/nh)$$

故 $\text{Var}[\hat{m}(x_0)] = O(1/nh)$ 。

当 $n \rightarrow \infty$ 时，让带宽 $h \rightarrow 0$ ，且 $nh \rightarrow \infty$ ，则根据均方收敛，核回

归估计是一致的。

如果 $\{x_1, x_2, \dots, x_n\}$ 为 iid, 则核回归估计量 $\hat{m}(x_0)$ 服从渐近正态:

$$\sqrt{nh}[\hat{m}(x_0) - m(x_0) - \text{Bias}(x_0)] \xrightarrow{d} N\left(0, \frac{\sigma_\varepsilon^2}{f(x_0)} \int_{-\infty}^{+\infty} K(z)^2 dz\right)$$

由于 $\text{Bias}(x_0) = O(h^2)$ 且 $\text{Var}[\hat{m}(x_0)] = O(1/nh)$, 最小化 IMSE 的结果显示, 最优带宽为 $h^* = O(n^{-0.2})$ 。

最优带宽 h^* 取决于待估计回归函数的导数(因为 $m'(x_0), m''(x_0)$ 出现在偏差的表达式中), 而估计 $m'(x_0), m''(x_0)$ 又需指定最优带宽 h^* 。

实践中常使用“交叉核实”(Cross Validation, 简记 CV)的方法来确定最优带宽 h^* 。

基本思想：在估计 $\hat{m}(x_i)$ 时，不使用 y_i 的信息，看其余观测值预测 y_i 的能力有多强；而这个能力又取决于带宽 h 。故选择带宽 h ，使得此预测能力最强，即最小化以下目标函数：

$$\min_h \text{CV}(h) \equiv \sum_{i=1}^n [y_i - \hat{m}_{-1}(x_i)]^2 \pi(x_i)$$

其中， $\hat{m}_{-1}(x_i) \equiv \frac{\sum_{j \neq i} w_{ji, h} y_j}{\sum_{j \neq i} w_{ji, h}}$ 是对 $m(x_i)$ 的“去掉一个观测值”估计量(leave-one-out estimate)，即 $j = 1, \dots, n$ ，但 $j \neq i$ 。

$\pi(x_i)$ 是权重函数(weighting function)，主要是为了给边界附近的

端点更小的权重，以避免扭曲。比如，不考虑 x_i 的 5% 分位数以下与 95% 分位数以上的观测值，即对这些观测值令 $\pi(x_i) = 0$ 。

之所以去掉自身第 i 个观测值是因为，如果将它保留，则总可以选择足够小的带宽 h 使得对于任何 i ，都有 $\hat{m}(x_i) = y_i$ ，故 $CV(h) = 0$ 得到最小化。

可以证明，最小化 $CV(h)$ 与最小化 IMSE 是渐近等价的。

交叉核实并非决定最优带宽的完美方法，仍常辅之以眼球法。

能使 $IMSE(h^*)$ 最小化的核函数为“伊潘科尼可夫核”(Epanechnikov)，但只有微弱优势。

27.7 多元核回归

对于 k 维解释向量 \mathbf{x} ，考虑如下非参数多元回归模型：

$$y_i = m(\mathbf{x}_i) + \varepsilon_i = m(x_{1i}, x_{2i}, \dots, x_{ki}) + \varepsilon_i$$

其中， $m(\cdot)$ 是未知的多元函数。在 \mathbf{x}_0 处的核回归估计量为

$$\hat{m}(\mathbf{x}_0) = \frac{\sum_{i=1}^n K[(\mathbf{x}_i - \mathbf{x}_0)/h] y_i}{\sum_{i=1}^n K[(\mathbf{x}_i - \mathbf{x}_0)/h]}$$

其中， $K(\cdot)$ 为 k 维核函数。

多元核回归估计量的性质与一元核回归相似。

但最优带宽为 $h^* = O(n^{-1/(k+4)})$ (大于一元情形下的最优带宽)，而 $\hat{m}(\mathbf{x}_0)$ 的收敛速度也更慢。

多元回归的解释变量越多，则收敛的速度越慢，对样本容量的要求也就越大。

这种“维度的诅咒” (curse of dimensionality) 限制了多元非参数回归的应用。

一种解决方法是，使用半参数估计降低模型中非参数部分的维度。

27.8 k 近邻回归

核回归估计是局部加权平均估计量(local weighted average estimator)的一个特例，使用一个特别的权重。

选择权重的另一方式是，对于最靠近 x_0 的 k 个 x_i 的观测值都给予相同的权重，而对其余观测值则给予权重 0。

记 $N_k(x_0)$ 为最靠近 x_0 的 k 个 x_i 观测值的集合(包括 x_0 自身)，则 k 近邻估计量(k -nearest neighbor estimator)定义为：

$$\hat{m}_{\text{KNN}}(x_0) \equiv \frac{1}{k} \sum_{i=1}^n \mathbf{1}\{x_i \in N_k(x_0)\} \cdot y_i$$

此估计量可看成是使用“均匀核”(uniform kernel)的核估计，但带宽可变，且可以不对称(左边的带宽不等于右边的带宽)。

“对称化”(symmetrized)的 k 近邻估计量则对小于 x_0 的 $(k-1)/2$ 观测值与大于 x_0 的 $(k-1)/2$ 观测值进行简单算术平均。

k 近邻估计量相当于移动平均(moving average)。

由于 k 近邻估计量使用的是简单算术平均，而不是核回归估计所使用的加权平均，故前者可能不如后者光滑(正如直方图不如核密度估计光滑)。

越靠近端点，可用于移动平均的样本点越少，估计量越不准确。这种边界问题(boundary problem)可通过“局部线性回归”缓解。

27.9 局部线性回归

核回归估计量实际上是“局部常数估计”(local constant estimator)，假定在 x_0 附近的某个邻域里， $m(x)$ 均等于一个常数。

局部线性回归假定 $m(x)$ 在 x_0 附近的某个邻域里为线性函数，即在该邻域里， $m(x) = a_0 + b_0(x - x_0)$ ，然后使用加权最小二乘法(WLS)来估计此线性函数：

$$\min_{\{a_0, b_0\}} \sum_{i=1}^n K[(x_i - x_0)/h] [y_i - a_0 - b_0(x_i - x_0)]^2$$

其中， $K(\cdot)$ 为核函数。离 x_0 越近，则权重越大(除非使用均匀核，则权重一样，等价于 OLS 回归)。

在 x_0 附近的小邻域里, $\hat{m}(x) = \hat{a}_0 + \hat{b}_0(x - x_0)$ 。

这种方法称为“局部线性回归”(local linear regression), 由 Fan (1992) 首倡, 也称“范回归”(Fan regression)。

局部线性回归不仅能较好地解决“边界问题”, 而且比核回归更有效率且适用于更多数据类型。

如果带宽足够小, 则在此小邻域内, 一般的函数都可以很好地用线性函数来近似, 故局部线性回归具有较好的性质。

更一般地, 假定 $m(x)$ 在 x_0 附近的某个邻域为 p 级多项式。“局部 p 级多项式估计量”(local polynomial estimator of degree p) 最小化以下目标函数:

$$\min_{\{a_0, b_0\}} \sum_{i=1}^n K[(x - x_0)/h] \left[y_i - a_{0,0} - a_{0,1}(x_i - x_0) - \dots - \frac{a_{0,p}}{p!}(x_i - x_0)^p \right]^2$$

一个常用的局部回归估计量为 Cleveland(1979)所提出的“局部加权散点光滑估计量”(Locally weighted scatterplot smoothing, 简记 Lowess), 是局部多项式估计量的变种或升级版。

该估计量使用“三三核”(tricubic kernel), 同时使用可变带宽 $h_{0,k}$ (由 x_0 到其最近的 k 个观测值的距离所决定), 以及对较大的残差 $[y_i - \hat{m}(x_i)]$ 给予较小的权重。

Lowess 的优点是, 使用了可变带宽(依数据的稠密程度而定), 对于极端值更稳健, 且缓解了在两端估计不准的边界问题。

27.10 非参数估计的 **Stata** 命令及实例

27.11 半参数估计

非参数回归假设对回归函数 $m(x)$ 一无所知。

经济理论可能对 $m(x)$ 的具体形式有所限制，比如需求函数要满足对称性与齐次性(homogeneity)。利用这些信息可提高估计效率。

当解释变量较多时，完全的非参数方法面临“维度的诅咒”，要求很大的样本容量。

可使用同时包含“参数部分”(a parametric component)与“非参数部分”(a nonparametric component)的“半参数模型”(semiparametric model)。

最常见的半参数模型为“部分线性模型”(Partially Linear model):

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + g(\mathbf{z}_i) + \varepsilon_i$$

其中，参数部分 $\mathbf{x}_i' \boldsymbol{\beta}$ 为线性函数，而非参数部分 $g(\mathbf{z}_i)$ 为未知函数(连函数形式也不知道)

假设扰动项 ε_i 均值独立于 $\mathbf{x}_i, \mathbf{z}_i$ ，即 $E(\varepsilon_i | \mathbf{x}_i, \mathbf{z}_i) = 0$ 。

Robinson(1988)提出“罗宾逊差分估计量”(Robinson difference

estimator), 以消去 $g(\mathbf{z}_i)$ 。给定 \mathbf{z}_i , 对方程两边取条件期望可得

$$\mathbf{E}(y_i | \mathbf{z}_i) = \mathbf{E}(\mathbf{x}_i | \mathbf{z}_i)' \boldsymbol{\beta} + g(\mathbf{z}_i) + \underbrace{\mathbf{E}(\varepsilon_i | \mathbf{z}_i)}_{=0}$$

其中, 根据迭代期望定律,

$$\mathbf{E}(\varepsilon_i | \mathbf{z}_i) = \mathbf{E}_{\mathbf{x}_i} \mathbf{E}[(\varepsilon_i | \mathbf{z}_i) | \mathbf{x}_i] = \mathbf{E}_{\mathbf{x}_i} \underbrace{\mathbf{E}(\varepsilon_i | \mathbf{x}_i, \mathbf{z}_i)}_{=0} = 0$$

将两个方程相减可得:

$$y_i - \mathbf{E}(y_i | \mathbf{z}_i) = [\mathbf{x}_i - \mathbf{E}(\mathbf{x}_i | \mathbf{z}_i)]' \boldsymbol{\beta} + \varepsilon_i$$

在此“差分方程”中, 未知函数 $g(\mathbf{z}_i)$ 被消去, 而条件期望 $\mathbf{E}(y_i | \mathbf{z}_i)$

与 $E(\mathbf{x}_i | \mathbf{z}_i)$ 可用非参数方法来估计(比如, 核回归)。

假设 $\hat{E}(y_i | \mathbf{z}_i)$ 与 $\hat{E}(\mathbf{x}_i | \mathbf{z}_i)$ 分别为对 $E(y_i | \mathbf{z}_i)$ 与 $E(\mathbf{x}_i | \mathbf{z}_i)$ 的非参数估计, 可对以下线性方程进行 OLS 估计:

$$y_i - \hat{E}(y_i | \mathbf{z}_i) = \left[\mathbf{x}_i - \hat{E}(\mathbf{x}_i | \mathbf{z}_i) \right]' \boldsymbol{\beta} + u_i$$

记此估计量为 $\hat{\boldsymbol{\beta}}_{\text{PL}}$ 。由于使用了条件期望的估计量来替代条件期望本身, 故扰动项不再是 ε_i , 记新扰动项为 u_i 。

假设 ε_i 为iid($0, \sigma^2$), 可以证明:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{PL}} - \boldsymbol{\beta}) \xrightarrow{d} N \left(0, \sigma^2 \left(\text{plim}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbf{w}_i \mathbf{w}_i' \right)^{-1} \right)$$

其中, $\mathbf{w}_i \equiv \mathbf{x}_i - \mathbf{E}(\mathbf{x}_i | \mathbf{z}_i)$ 。

用 $\hat{\mathbf{E}}(\mathbf{x}_i | \mathbf{z}_i)$ 替代 $\mathbf{E}(\mathbf{x}_i | \mathbf{z}_i)$ 即可估计 $\text{Avar}(\hat{\boldsymbol{\beta}}_{\text{PL}})$ 。

如存在异方差, 可使用夹心估计量得到稳健标准误。

最后, 可得到对 $g(\mathbf{z}_i)$ 的非参数估计:

$$\hat{g}(\mathbf{z}_i) = \hat{\mathbf{E}}(y_i | \mathbf{z}_i) - \hat{\mathbf{E}}(\mathbf{x}_i | \mathbf{z}_i)' \hat{\boldsymbol{\beta}}_{\text{PL}}$$