

第8章 非参数回归模型

李高荣

北京师范大学统计学院

E-mail: ligaorong@bnu.edu.cn



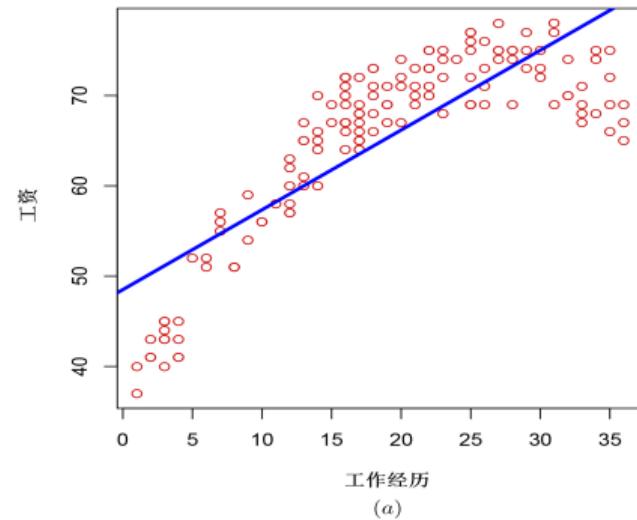
本章纲要

- 1 多项式回归
- 2 回归样条
 - d 阶回归样条
 - 线性样条
 - 三次样条
 - 自然三次样条
 - 节点个数和位置的选择
- 3 光滑样条
- 4 局部非参数光滑方法
 - N-W核光滑方法
 - 局部多项式光滑方法
- 5 广义可加模型
- 6 半参数回归模型
- 7 参考文献
- 8 作业

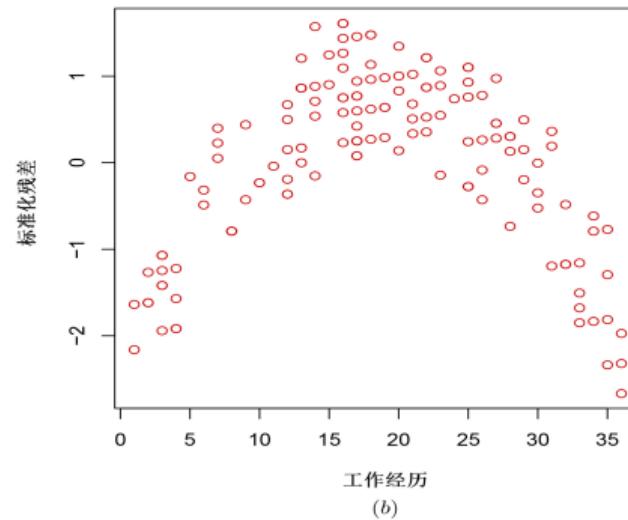


- 扫二维码获取在线课件和相关教学电子资源
- 请遵守电子资源使用协议

■ 考虑 Tryfos (1998) 提供的工资曲线数据, 讨论工作经历(单位: 年)对工资的影响.



(a) 工资曲线数据的散点图和拟合直线;



(b)

■ 为了放松线性回归模型的假设，考虑更加灵活和更好拟合数据的非参数回归模型

$$Y = g(\mathbf{X}) + \varepsilon,$$

其中

- ▶ Y 为响应变量
- ▶ $\mathbf{X} = (X_1, \dots, X_p)^T$ 为影响 Y 的 p 维协变量向量
- ▶ $g(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$ 为未知的 p 元连续光滑回归函数
- ▶ 模型误差 ε 满足: $E(\varepsilon|\mathbf{X}) = 0$ 和 $\text{Var}(\varepsilon) = \sigma^2 < \infty$

- 非参数回归模型的**优点**: $g(\cdot)$ 的形式任意, 而且模型的假设少, 可以很好地拟合实际数据并减少建模偏差.
- 非参数回归模型的**缺点**: 当 X 的维数 p 较高时, 对非参数回归模型进行估计和统计推断时, 会遇到所谓的“**维数灾难**”问题.
- 非参数回归模型的著作:
 - Fan 和Gijbels (1996): *Local Polynomial Modelling and Its Applications*
 - Hastie 等(2009): *The Elements of Statistical Learning*
 - Li 和 Racine (2007): *Nonparametric Econometrics: Theory and Practice*
 - 薛留根(2015): 现代非参数统计
 - 李高荣等(2016): 现代测量误差模型

■ 重点考虑维数 $p = 1$ 时，一元非参数回归模型的估计问题.

■ 主要估计方法：

- 多项式回归
- 回归样条
- 自然样条
- 光滑样条
- 局部非参数光滑方法

本章纲要

1 多项式回归

2 回归样条

- d 阶回归样条
- 线性样条
- 三次样条
- 自然三次样条
- 节点个数和位置的选择

3 光滑样条

4 局部非参数光滑方法

- N-W核光滑方法
- 局部多项式光滑方法

5 广义可加模型

6 半参数回归模型

7 参考文献

8 作业

多项式回归拟合

■ 假设i.i.d. 的观测样本 $\{(x_i, y_i), i = 1, \dots, n\}$ 来自下面的回归模型

$$y_i = g(x_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

- ▶ $g(\cdot)$ 是一元未知的连续光滑函数
- ▶ 模型误差 ε_i 满足: $E(\varepsilon_i) = 0$ 和 $\text{Var}(\varepsilon_i) = \sigma^2$

■ 回归函数 $g(x)$ 可用 d 阶多项式函数进行逼近, 即

$$y_i \approx \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_d x_i^d + \varepsilon_i, \quad i = 1, \dots, n,$$

- ▶ $1, x_i, x_i^2, \dots, x_i^d$ 称为多项式基函数
- ▶ $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_d)^T$ 可视为 $d+1$ 维的未知待估参数向量

多项式回归拟合

■ 令

$$\mathbf{X}_d = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^d \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^d \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

■ β 的最小二乘估计: $\hat{\boldsymbol{\beta}}_d = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_d)^T = (\mathbf{X}_d^T \mathbf{X}_d)^{-1} \mathbf{X}_d^T \mathbf{Y}$

■ 回归函数 $g(x)$ 的 d 阶多项式回归估计为

$$\hat{g}_d(x) = \hat{\beta}_0 + \hat{\beta}_1 x + \cdots + \hat{\beta}_d x^d.$$

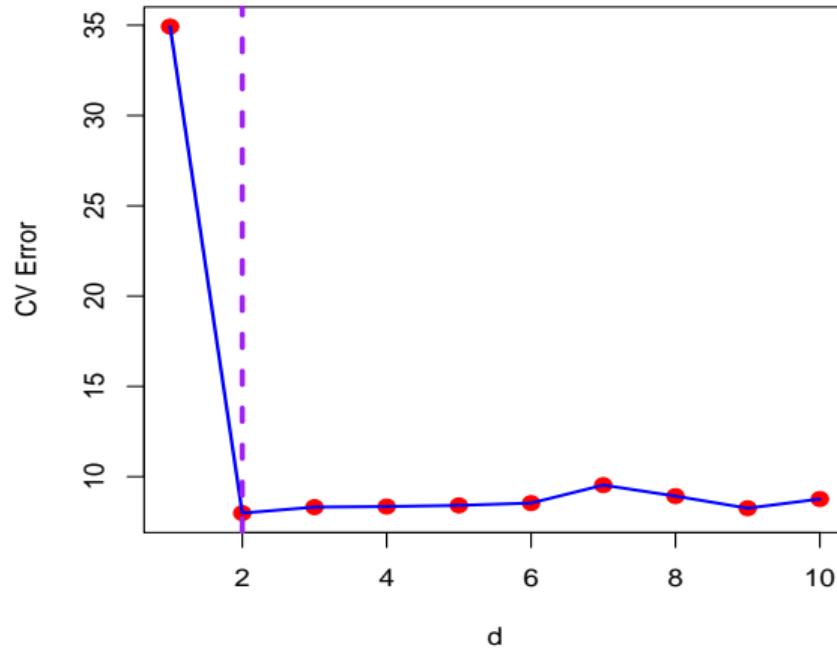
■ 多项式阶数 d 的大小非常关键, 控制着模型的自由度和复杂度

- ★ 如果 d 太大, 自由度变大, 模型变得复杂, 容易过拟合, 即偏差小而方差大
- ★ 如果 d 太小, 自由度变小, 模型变得非常光滑, 则容易欠拟合, 即偏差大而方差小

■ 多项式阶数 d 的选择: 利用第4章的LOOCV、GCV或 k 折CV方法进行选取

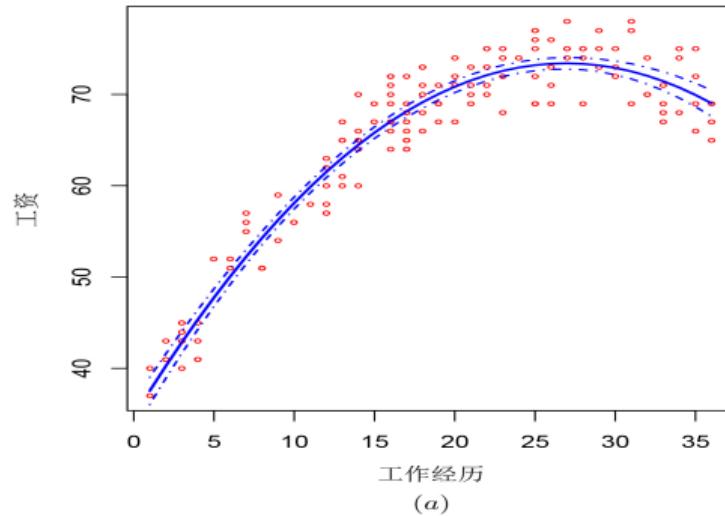
多项式回归拟合

■ 采用10折CV方法确定多项式的阶数 d

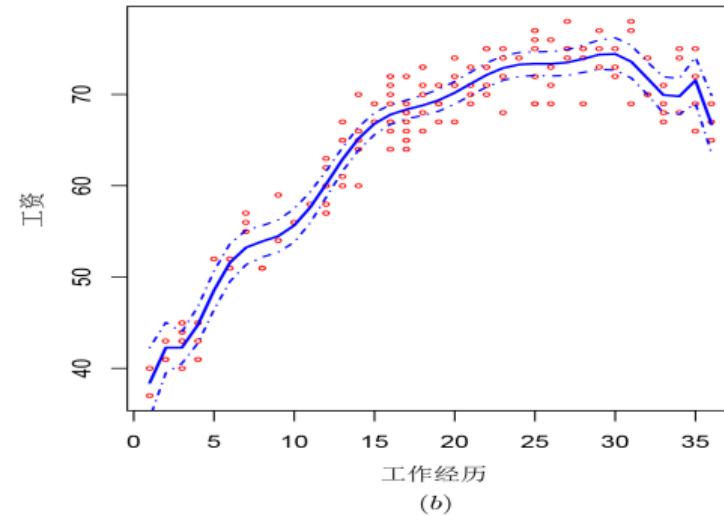


多项式回归拟合

■ 在R语言中, 用函数poly()拟合多项式回归.



(a)



(b)

工资曲线数据的散点图, 拟合曲线和95%置信带. (a) 蓝色实线表示工资关于工作经历的 $d = 2$ 阶多项式回归拟合曲线, 虚线表示95%置信带; (b) 蓝色实线表示工资关于工作经历的 $d = 12$ 阶多项式回归拟合曲线, 虚线表示95%置信带

- 多项式回归是一个全局逼近的非参数回归模型的估计方法, 很难捕捉具有局部特征的非线性数据.
- 解决方案: 分段多项式回归方法
- 假设协变量 X 的支撑集为 $[0, 1]$, 存在 K 个点 $\{\xi_k\}_{k=1}^K$, 满足 $0 = \xi_0 < \xi_1 < \cdots < \xi_K < \xi_{K+1} = 1$;
- 把支撑集 $[0, 1]$ 分为 $K + 1$ 个区间: $[\xi_0, \xi_1], (\xi_1, \xi_2], \dots, (\xi_K, \xi_{K+1}]$, 其中 $\xi_0 = 0$ 和 $\xi_{K+1} = 1$.

分段多项式回归拟合

■ 在每个区间上用 d 阶的多项式回归拟合数据, 即

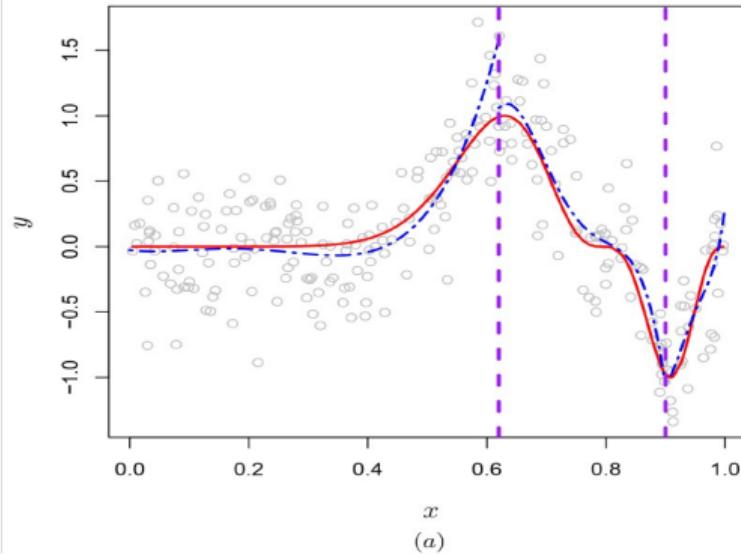
$$y_i = \begin{cases} \beta_{00} + \beta_{10}x_i + \beta_{20}x_i^2 + \cdots + \beta_{d0}x_i^d + \varepsilon_i, & x_i \in [0, \xi_1], \\ \beta_{01} + \beta_{11}x_i + \beta_{21}x_i^2 + \cdots + \beta_{d1}x_i^d + \varepsilon_i, & x_i \in (\xi_1, \xi_2], \\ \vdots \\ \beta_{0K} + \beta_{1K}x_i + \beta_{2K}x_i^2 + \cdots + \beta_{dK}x_i^d + \varepsilon_i, & x_i \in (\xi_K, 1]. \end{cases}$$

► 这里, $\boldsymbol{\beta}_{(k)} = (\beta_{0k}, \beta_{1k}, \dots, \beta_{dk})^T$ 是区间 $(\xi_k, \xi_{k+1}]$ 内 d 阶多项式模型的 $d+1$ 维未知参数向量, 且 $k = 0, 1, \dots, K$.

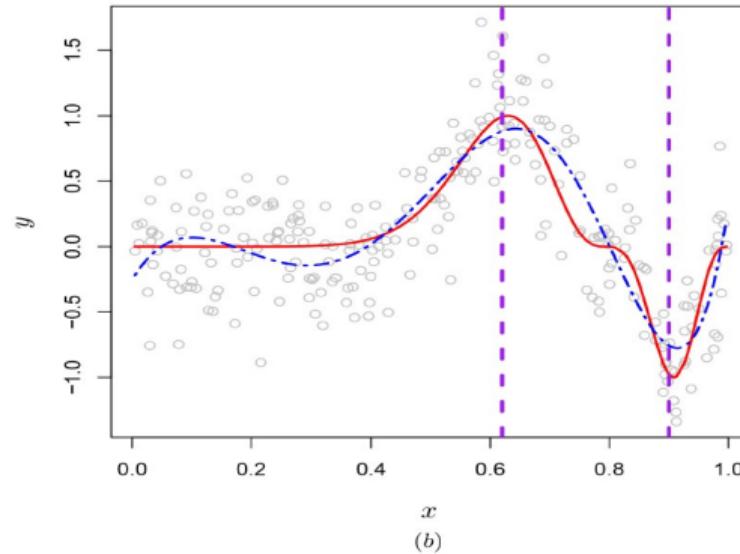
■ 在每个区间内, 参数向量 $\boldsymbol{\beta}_{(k)}$ 都不相同.

- 把系数发生变化的临界点称为节点(knot), 故 ξ_1, \dots, ξ_K 为节点.
 - 模型中共需要估计 $(K + 1) \times (d + 1)$ 个参数, 所以构建分段多项式回归模型的自由度为: $df = (K + 1) \times (d + 1)$.
 - 考虑程序包`faraway`中的`exa`数据, 该数据集包含256个样本, 是来自于模型: $Y = \sin^3(2\pi X^3) + \varepsilon$.
 - 取节点为 $x = 0.62$ 和 $x = 0.9$, 在3个区间上用四阶分段多项式回归拟合曲线, 拟合的自由度为15.
- ♠ **问题:** 如何解决分段多项式回归方法在节点处不连续的问题?

分段多项式回归拟合



(a)



(b)

程序包 [faraway](#) 中 exa 数据的拟合, 真实的回归函数为 $g(x) = \sin^3(2\pi x^3)$. (a) 红色实线表示真实曲线, 蓝色点断线表示在 3 个区间上的四阶分段多项式回归拟合曲线, 节点取 $x = 0.62$ 和 $x = 0.9$; (b) 红色实线表示真实曲线, 蓝色点断线表示具有约束的拟合曲线, 限定四阶多项式在节点 $x = 0.62$ 和 $x = 0.9$ 处连续光滑, 且一阶导数、二阶导数和三阶导数都存在的四次多项式

本章纲要

1 多项式回归

2 回归样条

- d 阶回归样条
- 线性样条
- 三次样条
- 自然三次样条
- 节点个数和位置的选择

3 光滑样条

4 局部非参数光滑方法

- N-W核光滑方法
- 局部多项式光滑方法

5 广义可加模型

6 半参数回归模型

7 参考文献

8 作业

本章纲要

1 多项式回归

2 回归样条

- d 阶回归样条
- 线性样条
- 三次样条
- 自然三次样条
- 节点个数和位置的选择

3 光滑样条

4 局部非参数光滑方法

- N-W核光滑方法
- 局部多项式光滑方法

5 广义可加模型

6 半参数回归模型

7 参考文献

8 作业

■ 假设 X 的支撑集为 $[0, 1]$, 存在 K 个节点 $\{\xi_k\}_{k=1}^K$ 满足: $0 = \xi_0 < \xi_1 < \cdots < \xi_K < \xi_{K+1} = 1$.

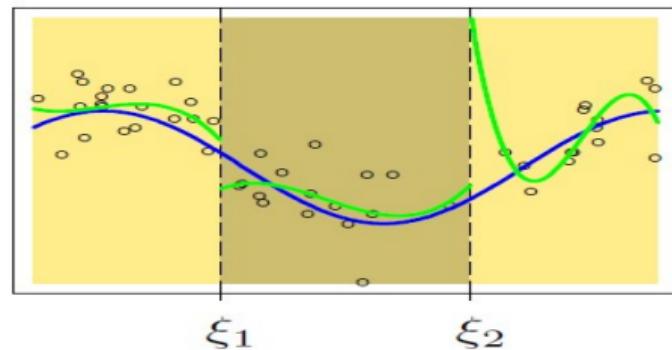
■ 把支撑集 $[0, 1]$ 分为 $K + 1$ 个区间: $[\xi_0, \xi_1], (\xi_1, \xi_2], \cdots, (\xi_K, \xi_{K+1}]$.

■ 回归样条的思想是:

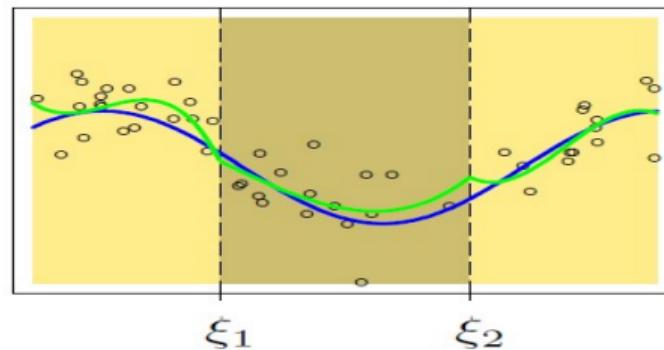
- 用 d 阶样条函数逼近未知的回归函数 $g(x)$, 它对每一个区间 $(\xi_k, \xi_{k+1}]$ 为 $d - 1$ 次连续可微函数, 其中 $k = 0, \cdots, K$
- 在节点 $\{\xi_k\}_{k=1}^K$ 处只存在前 $d - 1$ 阶导数, 而不存在 d 阶导数

d 阶回归样条

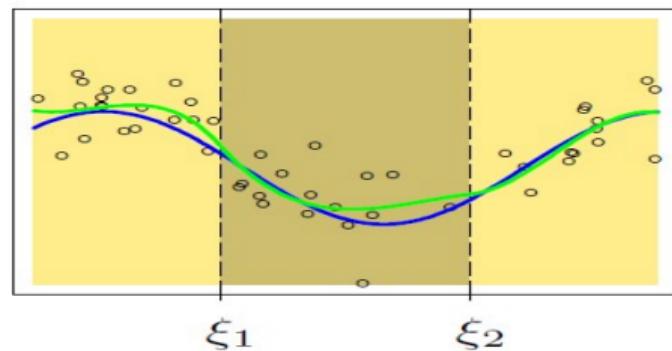
Discontinuous



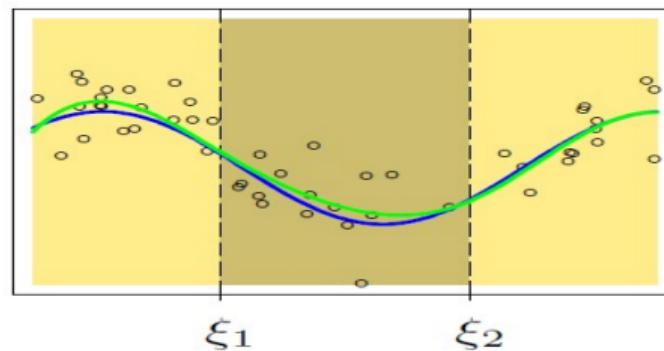
Continuous



Continuous First Derivative



Continuous Second Derivative



d 阶回归样条

■ 为了定义回归样条, 首先定义具有 K 个节点 ξ_1, \dots, ξ_K 的样条基函数

$$b_0(x) = 1, \quad b_1(x) = x, \quad b_2(x) = x^2, \quad \dots, \quad b_d(x) = x^d,$$

$$b_{d+1}(x) = (x - \xi_1)_+^d, \quad \dots, \quad b_{d+K}(x) = (x - \xi_K)_+^d,$$

■ 其中

$$(x - \xi)_+^d = \begin{cases} (x - \xi)^d, & x > \xi \\ 0, & \text{否则} \end{cases}$$

称为截断幂基函数(truncated power basis function).

■ 利用样条基函数, 回归函数 $g(x)$ 能被逼近为

$$g(x) \approx \beta_0 + \beta_1 b_1(x) + \cdots + \beta_d b_d(x) + \beta_{d+1} b_{d+1}(x) + \cdots + \beta_{d+K} b_{d+K}(x)$$

$$= \sum_{s=0}^d \beta_s x^s + \sum_{k=1}^K \beta_{d+k} (x - \xi_k)_+^d,$$

其中 $\beta_0, \beta_1, \dots, \beta_{d+K}$ 为 $K+d+1$ 个未知参数.

■ 上式称为具有 K 个节点 ξ_1, \dots, ξ_K 的 d 阶**回归样条**.

■ 对 $k = 1, \dots, K$, 有

$$g^{(d)}(\xi_k-) = d! \left(\beta_d + \sum_{l=1}^{k-1} \beta_{d+l} \right), \quad g^{(d)}(\xi_k+) = d! \left(\beta_d + \sum_{l=1}^k \beta_{d+l} \right).$$

■ 因此, 有

$$g^{(d)}(\xi_k+) - g^{(d)}(\xi_k-) = d! \beta_{d+k}.$$

■ 表明: $g^{(d)}(x)$ 在节点 ξ_k 处有一个跳 $d! \beta_{d+k}$, 且第 k 个截断幂基函数的系数 β_{d+k} 表示跳的大小(乘以 $d!$), 其中 $k = 1, \dots, K$.

- 具有 K 个节点 ξ_1, \dots, ξ_K 的 d 阶回归样条有连续的 $d - 1$ 阶导数, 但是 d 阶导数不连续.
- 这时, 非参数回归模型可写为

$$y_i \approx \underbrace{\beta_0 + \beta_1 b_1(x_i) + \cdots + \beta_d b_d(x_i)}_{\approx g(x_i)} + \underbrace{\beta_{d+1} b_{d+1}(x_i) + \cdots + \beta_{d+K} b_{d+K}(x_i)} + \varepsilon_i.$$

- 利用最小二乘方法, 可得 $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_{d+K})^T$, 则 $g(x)$ 的估计为

$$\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 b_1(x) + \cdots + \hat{\beta}_d b_d(x) + \hat{\beta}_{d+1} b_{d+1}(x) + \cdots + \hat{\beta}_{d+K} b_{d+K}(x).$$

值得注意的是：

- 由前可知，具有 K 个节点的 d 阶回归样条，需要在 $K + 1$ 个区间上共估计的参数个数为： $(K + 1)(d + 1)$ ；
- 为了保证在每个节点上连续光滑，并且前 $d - 1$ 阶导数连续，则在每个节点上需要施加 d 个约束；
- 因此，具有 K 个节点的 d 阶回归样条的自由度为：
$$(K + 1)(d + 1) - Kd = K + d + 1.$$
- 下面就具有两个节点的线性样条和立方样条为例介绍回归样条。

本章纲要

1 多项式回归

2 回归样条

- d 阶回归样条
- 线性样条
- 三次样条
- 自然三次样条
- 节点个数和位置的选择

3 光滑样条

4 局部非参数光滑方法

- N-W核光滑方法
- 局部多项式光滑方法

5 广义可加模型

6 半参数回归模型

7 参考文献

8 作业

线性样条 ($d = 1$)

- 假设存在两个节点 ξ_1 和 ξ_2 , 把区间 $[0, 1]$ 分成三个区间: $[0, \xi_1]$, $(\xi_1, \xi_2]$ 和 $(\xi_2, 1]$.
- 假设在 $[0, \xi_1]$ 上, 回归函数 $g(x)$ 可用连续的分段线性函数 $l(x)$ 进行逼近

$$l(x) = \beta_0 + \beta_1 x, \quad x \in [0, \xi_1].$$

- 既然要求 $l(x)$ 在节点 ξ_1 处必须连续, 在区间 $[\xi_1, \xi_2)$ 上新增一个线性函数, 满足在节点 ξ_1 处的截距为0, 则有

$$l(x) = \beta_0 + \beta_1 x + \beta_2(x - \xi_1)_+, \quad x \in [\xi_1, \xi_2).$$

■ 在区间 $(\xi_2, 1]$ 上, 有

$$l(x) = \beta_0 + \beta_1 x + \beta_2(x - \xi_1)_+ + \beta_3(x - \xi_2)_+, \quad x \in (\xi_2, 1].$$

■ 明显可知:

- 当 $0 \leq x \leq \xi_1$ 时, $l(x) = \beta_0 + \beta_1 x$;
- 当 $\xi_1 < x \leq \xi_2$ 时, $l(x) = (\beta_0 - \beta_2 \xi_1) + (\beta_1 + \beta_2)x$;
- 当 $\xi_2 < x \leq 1$ 时, $l(x) = (\beta_0 - \beta_2 \xi_1 - \beta_3 \xi_2) + (\beta_1 + \beta_2 + \beta_3)x$;
- $l(x)$ 在节点 ξ_1 和 ξ_2 处具有连续性, 但是在不同区间内, $l(x)$ 有不同的截距项和斜率.

■ 线性样条基函数定义为

$$b_0(x) = 1, \quad b_1(x) = x, \quad b_2(x) = (x - \xi_1)_+, \quad b_3(x) = (x - \xi_2)_+.$$

■ 用线性样条基函数逼近的非参数回归模型为

$$y_i \approx \underbrace{\beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \beta_3 b_3(x_i)}_{\approx g(x_i)} + \varepsilon_i, \quad i = 1, \dots, n.$$

线性样条

■ 令 $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T$, \mathbf{B}_1 为 $n \times 4$ 的设计矩阵和 \mathbf{Y} 为 $n \times 1$ 向量, 即

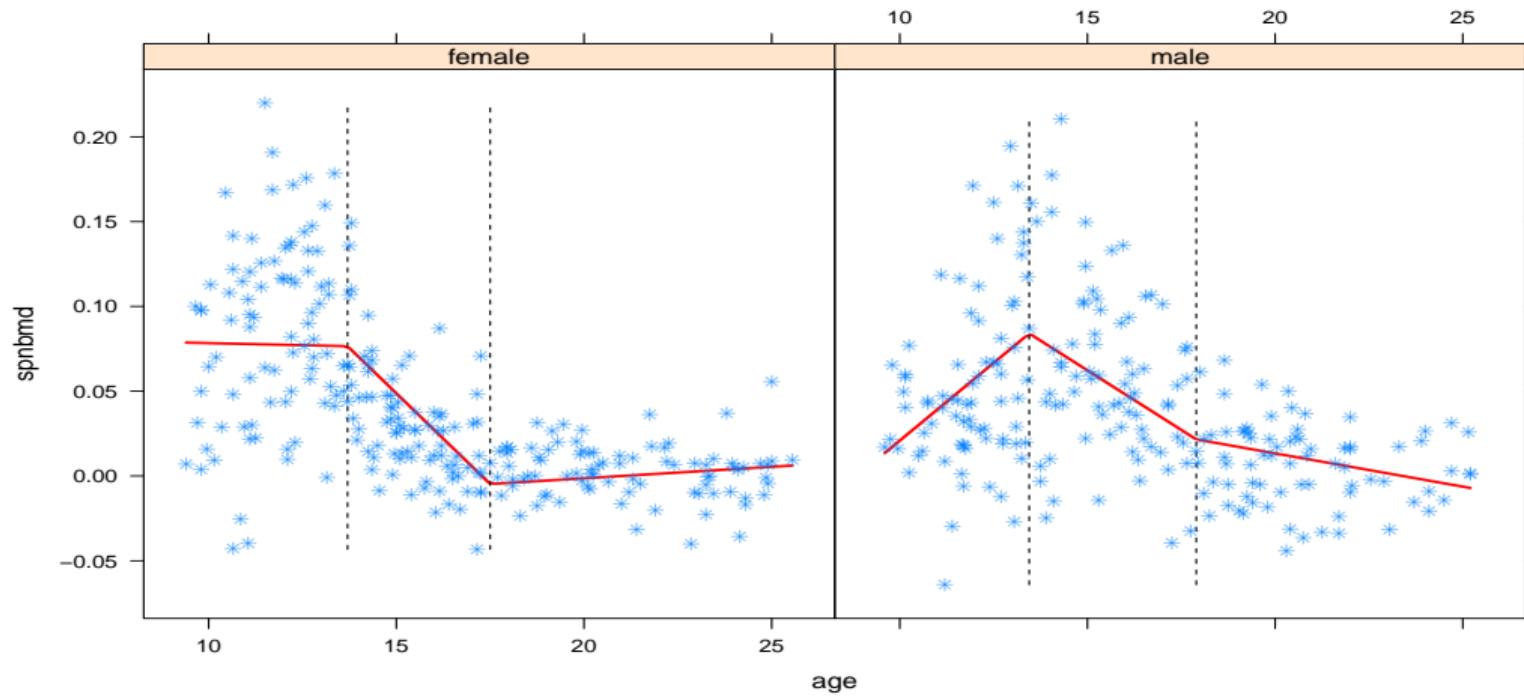
$$\mathbf{B}_1 = \begin{pmatrix} 1 & x_1 & (x_1 - \xi_1)_+ & (x_1 - \xi_2)_+ \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & (x_n - \xi_1)_+ & (x_n - \xi_2)_+ \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

■ 回归函数 $g(\cdot)$ 的估计为

$$\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 (x - \xi_1)_+ + \hat{\beta}_3 (x - \xi_2)_+,$$

其中 $\hat{\beta} = (\mathbf{B}_1^T \mathbf{B}_1)^{-1} \mathbf{B}_1^T \mathbf{Y}$ 为 β 的最小二乘估计.

线性样条



骨密度数据的线性样条拟合

本章纲要

1 多项式回归

2 回归样条

- d 阶回归样条
- 线性样条
- 三次样条
- 自然三次样条
- 节点个数和位置的选择

3 光滑样条

4 局部非参数光滑方法

- N-W核光滑方法
- 局部多项式光滑方法

5 广义可加模型

6 半参数回归模型

7 参考文献

8 作业

三次样条 ($d = 3$)

- 假设存在两个节点 ξ_1 和 ξ_2 , 把区间 $[0, 1]$ 分成三个区间: $[0, \xi_1]$, $(\xi_1, \xi_2]$ 和 $(\xi_2, 1]$.
- 假设在 $[0, \xi_1]$ 上, 回归函数 $g(x)$ 可用连续的分段三阶多项式函数 $c(x)$ 进行逼近

$$c(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3, \quad x \in [0, \xi_1].$$

- 在区间 $(\xi_1, \xi_2]$ 上, 新增一个三次函数, 即

$$c(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x - \xi_1)_+^3,$$

其中 $(x - \xi_1)_+^3$ 为截断幂基函数, 并且 $\beta_4 (x - \xi_1)_+^3$ 在节点 $x = \xi_1$ 处, 一阶导数和二阶导数为 0.

三次样条

■ 在区间 $[\xi_2, 1]$ 上, 有

$$c(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x - \xi_1)_+^3 + \beta_5 (x - \xi_2)_+^3, \quad x \in [\xi_2, 1].$$

■ 具有节点 ξ_1 和 ξ_2 的三次样条函数 $c(x)$ 有连续的一阶导数和二阶导数, 在节点处连续, 但是三阶导数不连续.

■ 三次样条基函数定义为

$$b_0(x) = 1, \quad b_1(x) = x, \quad b_2(x) = x^2, \quad b_3(x) = x^3,$$

$$b_4(x) = (x - \xi_1)_+^3, \quad b_5(x) = (x - \xi_2)_+^3.$$

三次样条

■ 因此, 可用三次样条基函数逼近非参数回归模型

$$y_i \approx \underbrace{\beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \beta_3 b_3(x_i) + \beta_4 b_4(x_i) + \beta_5 b_5(x_i)}_{\approx g(x_i)} + \varepsilon_i.$$

■ 令 $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_5)^T$, \mathbf{B}_2 为 $n \times 6$ 的设计矩阵和 \mathbf{Y} 为 $n \times 1$ 向量, 即

$$\mathbf{B}_2 = \begin{pmatrix} 1 & x_1 & x_1^2 & x_1^3 & (x_1 - \xi_1)_+^3 & (x_1 - \xi_2)_+^3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & x_n^3 & (x_n - \xi_1)_+^3 & (x_n - \xi_2)_+^3 \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

■ 回归函数 $g(x)$ 的估计为

$$\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 b_1(x) + \hat{\beta}_2 b_2(x) + \hat{\beta}_3 b_3(x) + \hat{\beta}_4 b_4(x) + \hat{\beta}_5 b_5(x),$$

其中 $\hat{\beta} = (\mathbf{B}_2^T \mathbf{B}_2)^{-1} \mathbf{B}_2^T \mathbf{Y}$ 为 β 的最小二乘估计.

■ 可见, 拟合具有两个节点的三次样条共需要6个自由度.

三次样条

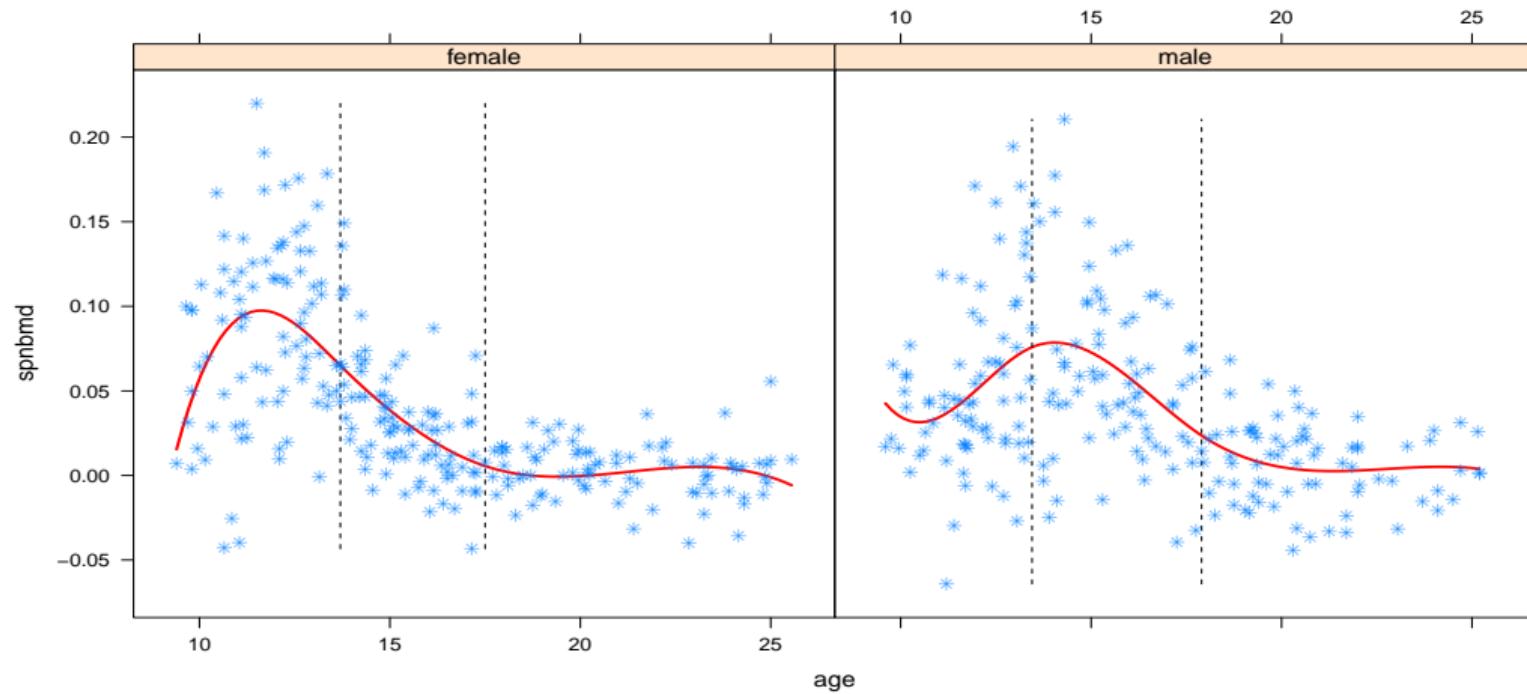
■ 一般情况下, 考虑 K 个节点 ξ_1, \dots, ξ_K , 则三次样条的基函数为

$$b_0(x) = 1, \quad b_1(x) = x, \quad b_2(x) = x^2, \quad b_3(x) = x^3,$$

$$b_4(x) = (x - \xi_1)_+^3, \quad \dots, \quad b_{K+3}(x) = (x - \xi_K)_+^3.$$

■ 因此, 拟合具有 K 个节点的三次样条共需要 $K + 4$ 个自由度.

三次样条



骨密度数据的三次样条拟合

本章纲要

1 多项式回归

2 回归样条

- d 阶回归样条
- 线性样条
- 三次样条
- 自然三次样条
- 节点个数和位置的选择

3 光滑样条

4 局部非参数光滑方法

- N-W核光滑方法
- 局部多项式光滑方法

5 广义可加模型

6 半参数回归模型

7 参考文献

8 作业

自然三次样条

- 自然三次样条是具有约束三次回归样条的一个特殊情况，自然三次样条用 $\text{NC}(x)$ 表示。
- 考虑 K 个节点 ξ_1, \dots, ξ_K ，由三次回归样条，有

$$\text{NC}(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{k=1}^K \beta_{3+k} (x - \xi_k)_+^3.$$

- 自然三次样条主要限制：在区间 $[0, \xi_1]$ 和 $(\xi_K, 1]$ 上，要求 $\text{NC}(x)$ 是线性的。

自然三次样条

- 当 $0 \leq x \leq \xi_1$ 时, $\text{NC}(x)$ 是线性的, 即要求 $\beta_2 = \beta_3 = 0$.
- 当 $\xi_K < x \leq 1$ 时, $\text{NC}(x)$ 是线性的, 除了要求 $\beta_2 = \beta_3 = 0$, 还要
求 $\sum_{k=1}^K \beta_{3+k} (x - \xi_k)^3$ 中的二次项和三次项系数为 0, 即

$$\sum_{k=1}^K \beta_{3+k} = 0, \quad \sum_{k=1}^K \xi_k \beta_{3+k} = 0.$$

自然三次样条

■ 满足上面要求的自然三次样条基函数为

$$b_0(x) = 1, \quad b_1(x) = x, \quad b_{k+1}(x) = \frac{(x - \xi_k)_+^3 - (x - \xi_K)_+^3}{\xi_K - \xi_k} - \frac{(x - \xi_{K-1})_+^3 - (x - \xi_K)_+^3}{\xi_K - \xi_{K-1}},$$

其中 $k = 1, \dots, K-2$.

■ 这时, 自然三次样条函数为

$$\text{NC}(x) = \sum_{k=0}^{K-1} \beta_k b_k(x),$$

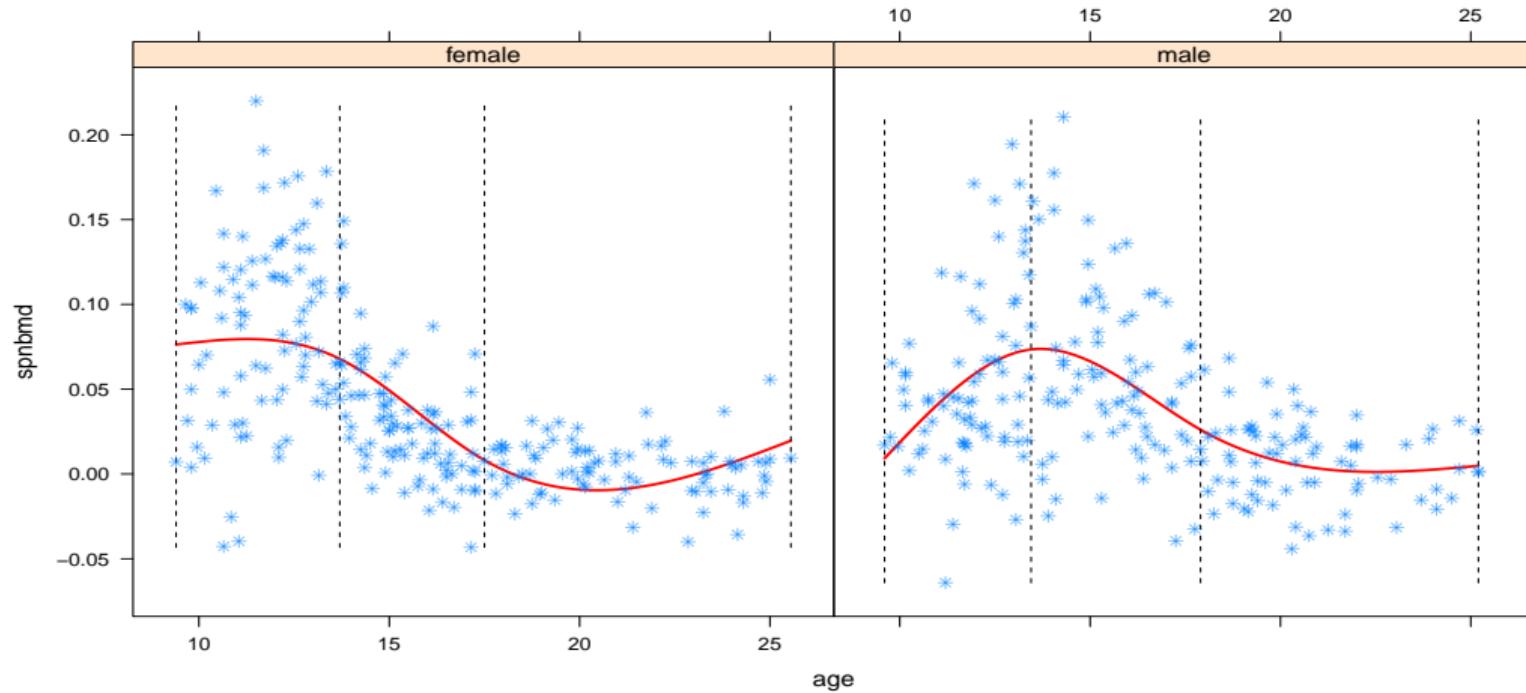
其中 $\boldsymbol{\beta} = (\beta_0, \dots, \beta_{K-1})^T$ 为 K 维的未知参数向量.

- 非参数回归模型能够被自然三次样条基函数逼近

$$y_i \approx \sum_{k=0}^{K-1} \beta_k b_k(x_i) + \varepsilon_i, \quad i = 1, \dots, n.$$

- 同样, 可用最小二乘方法估计 K 维参数向量 $\boldsymbol{\beta} = (\beta_0, \dots, \beta_{K-1})^T$.
- 因此, 拟合具有 K 个节点的自然三次样条共需要 **K 个自由度**.

自然三次样条



骨密度数据的自然三次样条拟合

本章纲要

1 多项式回归

2 回归样条

- d 阶回归样条
- 线性样条
- 三次样条
- 自然三次样条
- 节点个数和位置的选择

3 光滑样条

4 局部非参数光滑方法

- N-W核光滑方法
- 局部多项式光滑方法

5 广义可加模型

6 半参数回归模型

7 参考文献

8 作业

节点个数和位置的选择

多项式阶数 d 的影响:

- 当 d 太大, 将产生灵活的模型, 并容易过拟合, 导致偏差小, 而方差大;
- 当 d 太小, 容易导致欠拟合, 使得偏差大, 而方差小.
- 在实际应用中, 通常采用三次样条对数据进行拟合.

节点个数 K 和位置的影响:

⑤ **思考:** 在数据分析中, 节点个数 K 和位置对回归函数拟合有何影响?

♠ **问题:** 在实际应用中, 如何确定节点的个数和位置?

解决的办法有：

- ① 等间距方法
- ② 等间距样本分位数方法
- ③ 变量选择方法

1. 等间距方法

- 等间距方法是在协变量X的支撑集 $[0, 1]$ 上取等间距的K个点作为节点, 即

$$\xi_k = \frac{1}{K+1}k, \quad k = 1, \dots, K.$$

- 在实际应用中, 通常把节点放在曲率变化比较大的位置上.
- 可使用**节点删除法**选取节点, 具体步骤如下:

步骤1: 设 ξ_1, \dots, ξ_K 是在节点选择过程中可以删除的初始等间距节点, 其中节点数 K 通常取为 $\lfloor n/2 \rfloor$ 或 $\lfloor n/3 \rfloor$.

步骤2: 为确定节点个数, 可删除最小绝对 t 值 $|\hat{\beta}_{d+k}|/\text{SE}(\hat{\beta}_{d+k})$ ($1 \leq k \leq K$) 的第 k_0 ($1 \leq k_0 \leq K$) 个节点, 其中

- ▶ $\hat{\beta}_{d+k}$ 为第 k 个节点截断幂基函数 $(x - \xi_k)_+^3$ 对应的系数估计
- ▶ $\text{SE}(\hat{\beta}_{d+k})$ 为 $\hat{\beta}_{d+k}$ 的标准误差
- ▶ d 为多项式的阶数, 通常取 $d = 3$

步骤3: 每次删除一个节点, 重复上述删除过程.

2. 等间距样本分位数方法

■ 令 $x_{(1)}, \dots, x_{(n)}$ 表示样本 x_1, \dots, x_n 的次序统计量，则 K 个节点定义为

$$\xi_k = x_{(1+[kn/(K+1)])}, \quad k = 1, \dots, K.$$

■ 对于节点的个数选择，类似可采用上述提出的**节点删除法**。

★ 注意：

- ① 等间距样本分位数方法是**数据自适应的**，它会自动选取更多的节点在数据比较集中的位置；
- ② 当样本点均匀分布时，等间距样本分位数方法近似等价于等间距方法。

3. 变量选择方法

- 首先, 使用等间距方法或等间距样本分位数方法选取更多的节点, 作为候选的节点.
- 然后, 利用惩罚变量选择方法同时选取节点个数和位置, 即

$$\frac{1}{2n} \sum_{i=1}^n [y_i - \beta_0 - \beta_1 b_1(x_i) - \cdots - \beta_{K+3} b_{K+3}(x_i)]^2 + \sum_{k=1}^K p_\lambda(|\beta_{k+3}|).$$

★ 注意: 上面惩罚最小二乘目标函数中, 仅对带有节点 ξ_1, \dots, ξ_K 的截断幂基函数所对应的系数进行惩罚.

■ 程序包**splines**提供了拟合回归样条和自然样条的函数: **bs()**和**ns()**

```
bs(x, df = NULL, knots = NULL, degree = 3, intercept = FALSE,  
    Boundary.knots = range(x))  
  
ns(x, df = NULL, knots = NULL, intercept = FALSE,  
    Boundary.knots = range(x))
```

其中x表示协变量数据, df表示自由度, knots表示节点, 默认参数intercept=FALSE, 它会忽略基函数中的常数项.

(1) 函数bs()

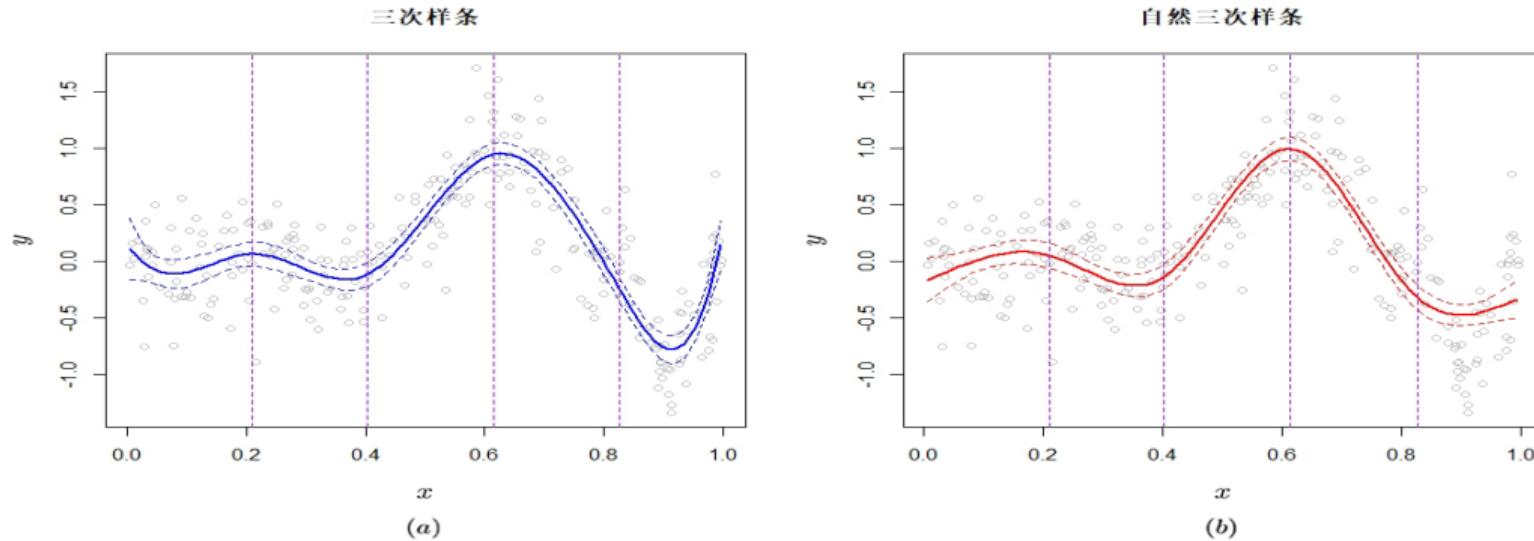
- 如果`intercept=FALSE`, 则自由度为 $K + 3$, 即输出的基函数矩阵中不包含截距项;
- 参数`df`表示自由度, 该参数将等间距样本分位数设为样条的节点;
- 如果`bs(x, df=7)`, 则表示具有 $K = 4$ 个节点的三次样条, 节点位置对应于 x 的20%, 40%, 60% 和80% 的样本分位数;
- 如果`bs(x, df=7, intercept=TRUE)`, 则表示具有 $K = 3$ 个节点的三次样条, 节点位置对应于 x 的25%, 50% 和75% 的样本分位数, 这时在基函数矩阵中包含了截距项.

(2) 函数ns()

- 如果参数intercept缺省, 表示输出的基函数矩阵中不包含截距项;
- 参数df表示自由度, 如果 $ns(x, df=5)$, 则生成一个不包含截距项的自然三次样条的 $n \times 5$ 基函数矩阵, 节点位置对应于x 的20%, 40%, 60% 和80% 的样本分位数;
- 如果 $bs(x, df=5, intercept=TRUE)$, 则生成一个包含截距项的自然三次样条的 $n \times 5$ 基函数矩阵, 节点位置对应于x 的25%, 50% 和75% 的样本分位数.

- 对程序包`faraway`中的`exa`数据进行三次样条和自然三次样条拟合
- 通过自由度限制节点位置为 x 的20%, 40%, 60% 和80% 的等间距样本分位数，并绘制拟合曲线和95%的置信带

案例分析与应用



程序包 [faraway](#) 中 exa 数据的三次样条和自然三次样条拟合，节点取 x 的 20%, 40%, 60% 和 80% 的等间距样本分位数。 (a) 三次样条拟合曲线和 95% 的置信带；(b) 自然三次样条拟合曲线和 95% 的置信带

本章纲要

1 多项式回归

2 回归样条

- d 阶回归样条
- 线性样条
- 三次样条
- 自然三次样条
- 节点个数和位置的选择

3 光滑样条

4 局部非参数光滑方法

- N-W核光滑方法
- 局部多项式光滑方法

5 广义可加模型

6 半参数回归模型

7 参考文献

8 作业

- 假设 X 的支撑集为 $[0, 1]$, 并假设 \mathcal{G}_2 是支撑集 $[0, 1]$ 上所有连续可微、二次可积且具有二阶导数的函数构成的类.
- 对 \mathcal{G}_2 中的任何函数 $g(\cdot)$, 定义如下惩罚最小二乘目标函数

$$\text{RSS}(g, \lambda) = \sum_{i=1}^n (y_i - g(x_i))^2 + \lambda \int_0^1 [g''(x)]^2 dx,$$

其中 $\lambda \geq 0$ 称为 **光滑参数**, $g''(x)$ 是 $g(x)$ 的 **二阶导数**.

- 第一项提供了 $g(\cdot)$ 对数据拟合程度的度量

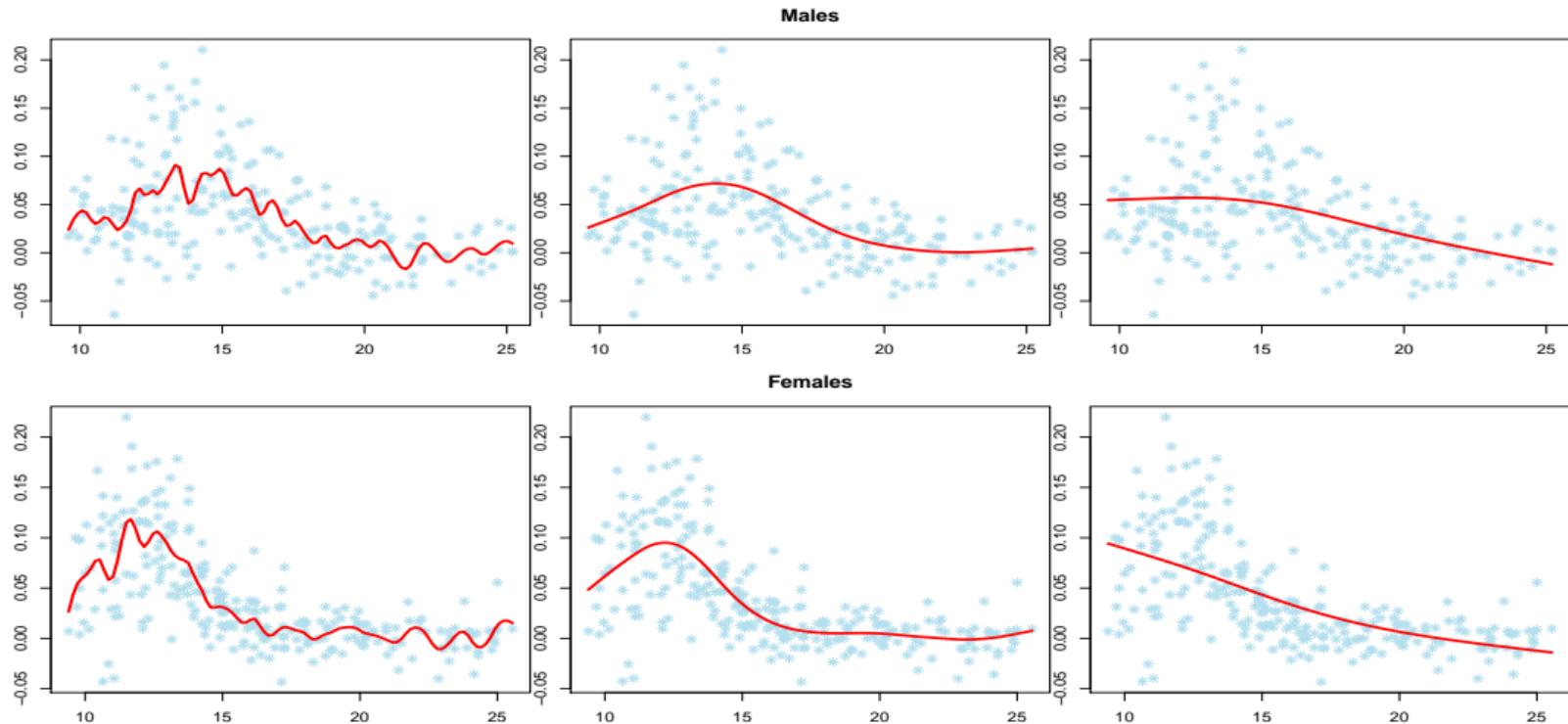
■ 第二项为惩罚项, 用来度量 $g(\cdot)$ 的光滑度, 是对 $g(\cdot)$ 的波动性进行惩罚, 因为二阶导数 $g''(\cdot)$ 是衡量函数 $g(\cdot)$ 的**粗糙度(roughness)**.

- 如果函数 $g(\cdot)$ 的粗糙度大, 则会有大的惩罚.
- 如果函数 $g(\cdot)$ 非常光滑, 则惩罚变小.

■ 极小化 $\text{RSS}(g, \lambda)$, 得到估计 $\hat{g}(\cdot)$, 称它为**光滑样条估计量**.

- 当 $\lambda = 0$ 时, 相应的估计是对数据的任意插值函数, 它可以在每一个训练数据点上作插值, 即 $\hat{g}(x_i) = y_i$;
- 当 $\lambda = \infty$ 时, $\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$, 即变成一条尽可能接近所有训练样本的直线;
- 当 λ 在0到 ∞ 范围内变化时, $g(x)$ 的估计在最复杂模型和最简单模型(线性模型)之间变化, 实际自由度也从 n 降至2.

光滑样条



采用不同光滑参数的光滑样条对骨密度数据进行拟合

♠ **问题:** 如何得到光滑样条估计量?

♠ **问题:** 如何得到光滑样条估计量?

■ 令回归函数 $g(x)$ 可被自然三次样条基函数逼近, 即

$$g(x) \approx \sum_{j=1}^n \beta_j b_j(x),$$

其中 $b_j(x)$ 是自然样条的基函数, 且 $j = 1, \dots, n$.

■ 这时, 目标函数RSS(β, λ)可以写为

$$\text{RSS}(\beta, \lambda) = (\mathbf{Y} - \mathbf{B}\beta)^T (\mathbf{Y} - \mathbf{B}\beta) + \lambda \beta^T \Omega_n \beta.$$

这里,

- ▶ $\mathbf{Y} = (y_1, \dots, y_n)^T$ 为 $n \times 1$ 向量;
- ▶ $\boldsymbol{\beta} = (\beta_1, \dots, \beta_n)^T$ 为 $n \times 1$ 的未知参数向量;
- ▶ \mathbf{B} 为 $n \times n$ 的矩阵, 其第 (i, j) 个元素为 $\{\mathbf{B}\}_{ij} = b_j(x_i)$;
- ▶ Ω_n 为 $n \times n$ 的矩阵, 其第 (j, k) 个元素为 $\{\Omega_n\}_{jk} = \int_0^1 b_j''(x) b_k''(x) dx$.

■ 极小化RSS(β, λ), 可得 β 的估计为

$$\hat{\beta} = (\mathbf{B}^T \mathbf{B} + \lambda \Omega_n)^{-1} \mathbf{B}^T \mathbf{Y}.$$

■ 因此, 拟合的光滑样条为

$$\hat{g}(x) = \sum_{j=1}^n \hat{\beta}_j b_j(x).$$

光滑参数 λ 的选取

♠ **问题:** 光滑参数 λ 决定着模型的复杂度或自由度, 实际应用中, 如何选取最优的 λ ?

■ 由上讨论可知, $\mathbf{g} = (g(x_1), \dots, g(x_n))^T$ 的拟合值为

$$\hat{\mathbf{g}} = \mathbf{B}(\mathbf{B}^T \mathbf{B} + \lambda \Omega_n)^{-1} \mathbf{B}^T \mathbf{Y} = \mathbf{S}_\lambda \mathbf{Y},$$

其中 $\mathbf{S}_\lambda = \mathbf{B}(\mathbf{B}^T \mathbf{B} + \lambda \Omega_n)^{-1} \mathbf{B}^T$ 为光滑矩阵.

■ 光滑样条拟合的有效自由度为: $df_\lambda = \text{tr}(\mathbf{S}_\lambda) = \sum_{i=1}^n \{\mathbf{S}_\lambda\}_{ii}$.

光滑参数 λ 的选取

① LOOCV方法：极小化下面的LOOCV目标函数选取 λ , 即

$$\hat{\lambda} = \arg \min_{\lambda} \frac{1}{n} \sum_{i=1}^n \left(y_i - \hat{g}_{\lambda}^{(-i)}(x_i) \right)^2 = \arg \min_{\lambda} \frac{1}{n} \sum_{i=1}^n \left[\frac{y_i - \hat{g}_{\lambda}(x_i)}{1 - \{\mathbf{S}_{\lambda}\}_{ii}} \right]^2,$$

其中 $\hat{g}_{\lambda}^{(-i)}(x_i)$ 为去掉第*i*个样本点 (x_i, y_i) 后光滑样条的拟合值.

② GCV方法：极小化下面的GCV目标函数选取 λ , 即

$$\hat{\lambda} = \arg \min_{\lambda} \text{GCV}(\lambda) = \arg \min_{\lambda} \frac{1}{n} \sum_{i=1}^n \left[\frac{y_i - \hat{g}_{\lambda}(x_i)}{1 - df_{\lambda}/n} \right]^2.$$

■ 在R语言中，可使用函数smooth.spline() 进行光滑样条拟合，调用格式为

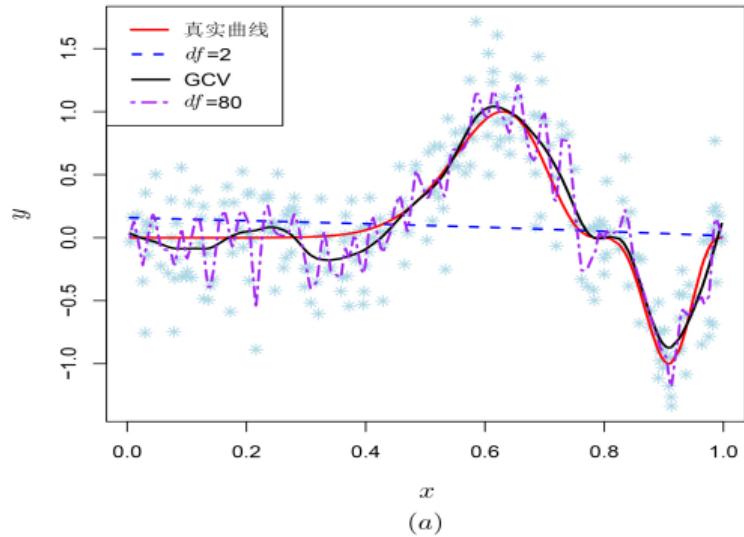
```
smooth.spline(x, y = NULL, w = NULL, df, spar = NULL, lambda = NULL,  
cv = FALSE, all.knots = FALSE, nknots = .nknots.smspl,  
keep.data = TRUE, df.offset = 0, penalty = 1,  
control.spar = list(), tol = 1e-6 * IQR(x),  
keep.stuff = FALSE)
```

其中x表示协变量观测数据，也可以是包含x和y的两列矩阵数据；y表示响应变量的观测数据；df表示自由度，定义为光滑矩阵的迹；spar表示光滑参数，取值可以指定为(0,1]；lambda也表示光滑参数，可以通过CV准则选取；参数cv 表示LOOCV和GCV方法，当cv=TRUE时，表示LOOCV 方法，当cv=FALSE时，表示GCV方法，当df和spar没有指定时，可使用LOOCV或GCV方法选取光滑参数；其余参数见在线帮助。

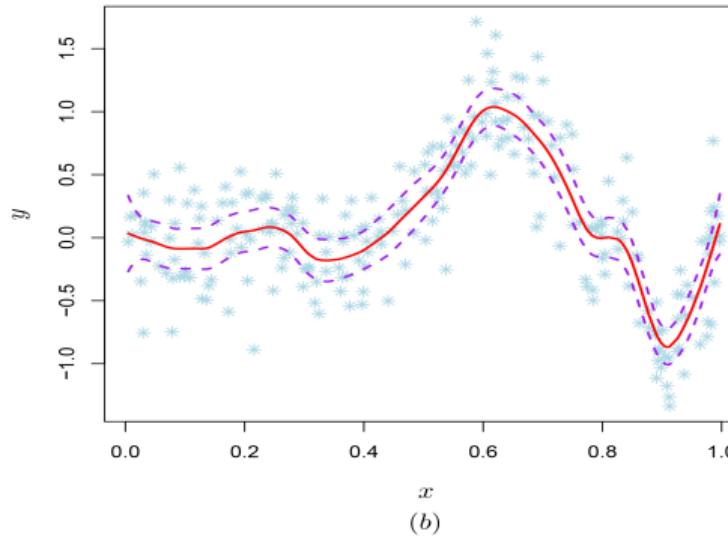
- 对于光滑样条而言，因为自由度 $df_\lambda = \text{tr}(\mathbf{S}_\lambda)$ 是关于光滑参数 λ 的单调函数，这个关系也是可逆的。
- 当光滑参数 $\lambda = \infty$ 时，有效自由度为 $df = 2$ 。
- 当光滑参数 λ 逐渐减小时，则有效自由度将逐渐变大。
- 因此，在函数**smooth.spline()**中，可通过指定参数**df**来确定拟合的光滑程度。

光滑样条的案例分析与应用

- 对程序包 **faraway** 中的 **exa** 数据, 分别取自由度 $df = 2, df = 80$ 和利用 GCV 方法选取最优的自由度, 对数据进行光滑样条拟合.



(a)

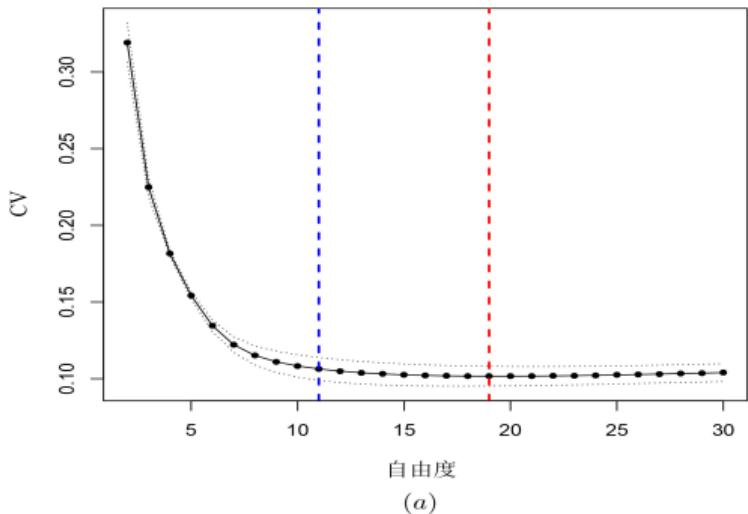


(b)

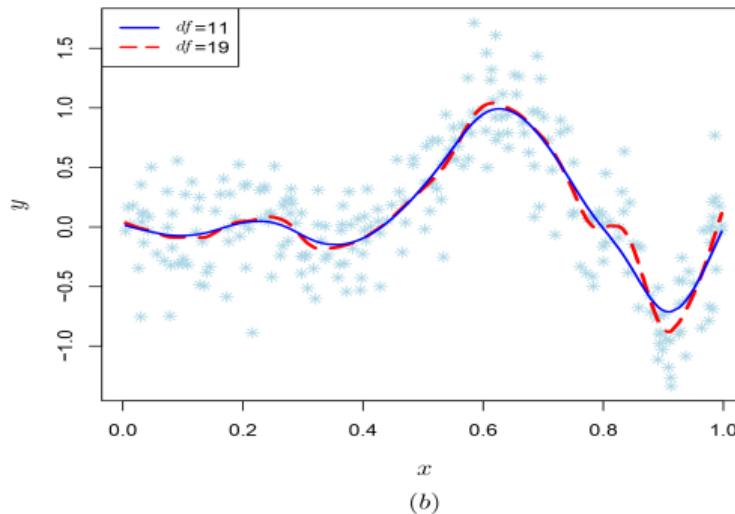
(a) 自由度分别取 $df = 2$ 和 $df = 80$, 以及 GCV 方法选取最优的自由度; (b) 自由度为 $df = 18$ 时的光滑样条拟合曲线和 95% 的置信带

光滑样条的案例分析与应用

■ 进一步, 利用5折CV方法选择最优的自由度 $df = 19$, 然后利用“一个标准差”准则确定自由度为 $df = 11$, 最后利用函数smooth.spline()进行光滑样条拟合.



(a)



(b)

(a) 5折CV误差图; (b) 基于自由度 $df = 11$ 和 $df = 19$, 光滑样条的拟合曲线

本章纲要

1 多项式回归

2 回归样条

- d 阶回归样条
- 线性样条
- 三次样条
- 自然三次样条
- 节点个数和位置的选择

3 光滑样条

4 局部非参数光滑方法

- N-W核光滑方法
- 局部多项式光滑方法

5 广义可加模型

6 半参数回归模型

7 参考文献

8 作业

三种经典的局部非参数光滑方法：

- ① Nadaraya-Watson (N-W)核光滑方法
- ② Gasser-Müller光滑方法
- ③ 局部多项式光滑方法

本章纲要

1 多项式回归

2 回归样条

- d 阶回归样条
- 线性样条
- 三次样条
- 自然三次样条
- 节点个数和位置的选择

3 光滑样条

4 局部非参数光滑方法

- N-W核光滑方法
- 局部多项式光滑方法

5 广义可加模型

6 半参数回归模型

7 参考文献

8 作业

- 假设 $\{(y_i, x_i), i = 1, \dots, n\}$ 是来自模型 $Y = g(X) + \varepsilon$ 的i.i.d.随机样本.
- 为了简单, 假设协变量X的支撑集为 $[0,1]$.
- 令 $K(\cdot)$ 是一个核函数, h 为窗宽, 主要用于控制局部区域的大小.
- Nadaraya (1964) 与 Watson (1964) 分别提出了回归函数 $g(\cdot)$ 的核光滑(kernel smoothing) 估计方法.
- 主要思想是: 对于变量 x_i 支撑集内任意给定的一点 x , 在 x 的一个邻域 $[x - h, x + h]$ 中, 假定回归函数 $g(\cdot)$ 为一个常数 θ .

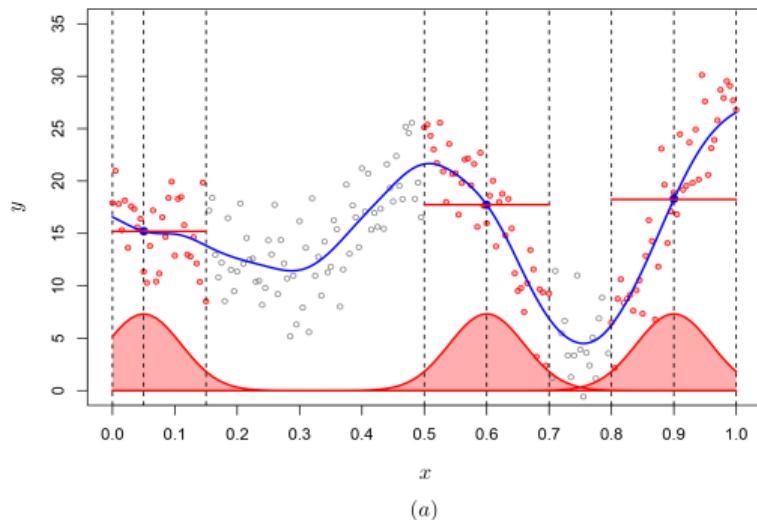
- 如果 $h \rightarrow 0$, 在局部邻域 $[x - h, x + h]$ 内, 极小化加权最小二乘目标函数, 则可得到 Nadaraya-Watson(N-W)核估计

$$\hat{g}_{\text{NW}}(x) = \arg \min_{\theta} \sum_{i=1}^n (y_i - \theta)^2 K\left(\frac{x_i - x}{h}\right) = \sum_{i=1}^n W_{ni}(x) y_i,$$

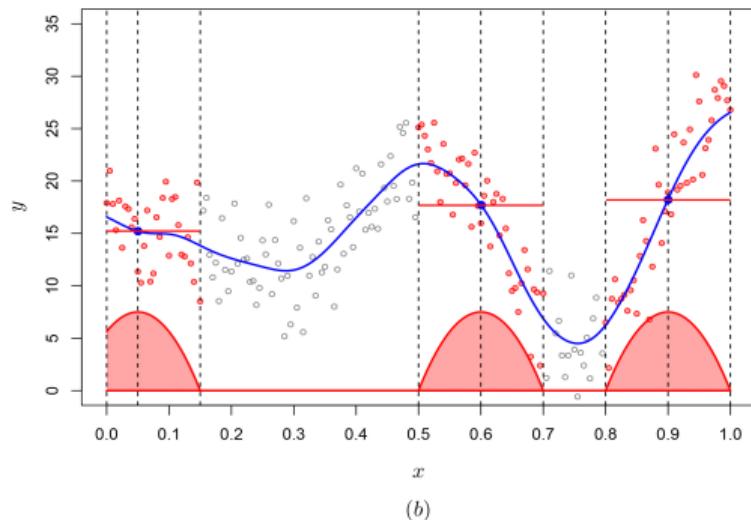
其中 $W_{ni}(x) = \frac{K\left(\frac{x_i - x}{h}\right)}{\sum_{j=1}^n K\left(\frac{x_j - x}{h}\right)}$ 为核权函数.

N-W核光滑方法

■ N-W核估计 $\hat{g}_{NW}(x)$ 是响应变量 y_i 的加权平均值.



(a)



(b)

N-W核估计的原理示意图. (a) 高斯核; (b) Epanechnikov核

核函数 $K(\cdot)$ 通常取为某个概率密度函数，常用的核函数有：

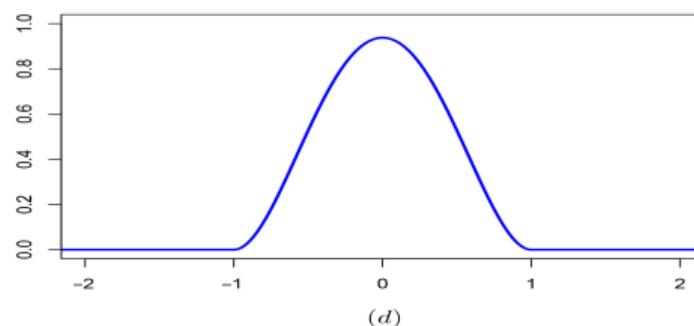
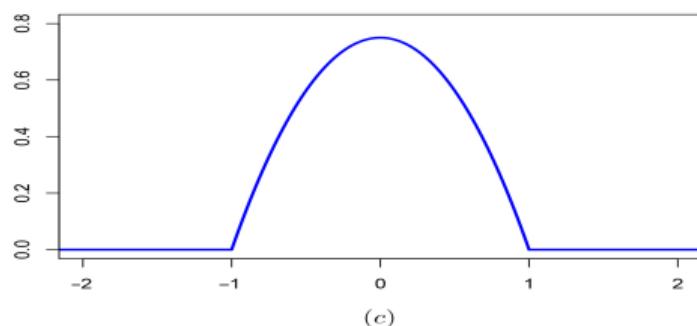
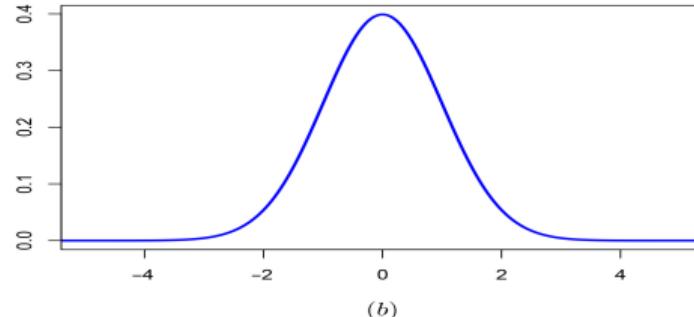
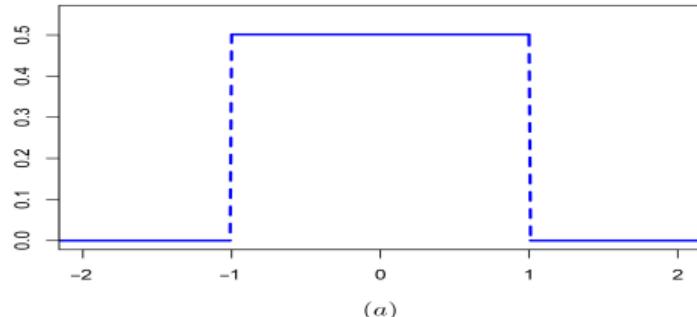
① 均匀核 $K(u) = \frac{1}{2}I(|u| \leq 1);$

② 高斯核 $K(u) = \frac{1}{\sqrt{2\pi}} \exp(-u^2/2);$

③ Epanechnikov核 $K(u) = \frac{3}{4}(1 - u^2)I(|u| \leq 1);$

④ 四次核(quartic kernel) $K(u) = \frac{15}{16}(1 - u^2)^2I(|u| \leq 1).$

N-W核光滑方法



核函数曲线图. (a) 均匀核; (b) 高斯核; (c) Epanechnikov核; (d) 四次核

■ 在一定条件下，可以推得N-W核估计 $\hat{g}_{\text{NW}}(x)$ 的渐近偏差为

$$\text{bias}(\hat{g}_{\text{NW}}(x)) = \frac{1}{2} \left(g''(x) + \frac{2g'(x)f'(x)}{f(x)} \right) d_K h^2 + o(h^2);$$

■ 在一定条件下，可以推得N-W核估计 $\hat{g}_{\text{NW}}(x)$ 的渐近方差为

$$\text{Var}(\hat{g}_{\text{NW}}(x)) = \frac{\sigma^2(x)c_K}{nhf(x)} + o((nh)^{-1}),$$

- ▶ $f(x)$ 表示 X 的密度函数, $f'(x)$ 是 $f(x)$ 的导数
- ▶ $g'(x)$ 和 $g''(x)$ 分别是 $g(x)$ 的一阶和二阶导数
- ▶ $d_K = \int_{-\infty}^{\infty} u^2 K(u) du, \quad c_K = \int_{-\infty}^{\infty} K^2(u) du$

窗宽 h 的选择

窗宽 h 对N-W核估计 $\hat{g}_{\text{NW}}(x)$ 的影响：

- ① 当窗宽 h 变小时，参与加权平均的样本就变少，这时偏差变小，而方差则变大；
- ② 当窗宽 h 变大时，参与加权平均的样本变多，则偏差变大，而方差变小。

♠ 问题：如何选取最优的窗宽 h ？

窗宽 h 的选择

窗宽 h 对N-W核估计 $\hat{g}_{\text{NW}}(x)$ 的影响：

- ① 当窗宽 h 变小时，参与加权平均的样本就变少，这时偏差变小，而方差则变大；
- ② 当窗宽 h 变大时，参与加权平均的样本变多，则偏差变大，而方差变小。

♠ 问题：如何选取最优的窗宽 h ？

- ① 理论上，平衡偏差和方差，即极小化均方误差(MSE)
- ② 应用上，利用LOOCV和GCV等数据驱动的方法

窗宽 h 的选择

■ 简单计算可得均方误差(MSE)

$$\begin{aligned}\text{MSE}(\hat{g}_{\text{NW}}(x)) &= \text{E}[\hat{g}_{\text{NW}}(x) - g(x)]^2 = [\text{bias}(\hat{g}_{\text{NW}}(x))]^2 + \text{Var}(\hat{g}_{\text{NW}}(x)) \\ &\approx \frac{1}{4} C_B^2 d_K^2 h^4 + C_V c_K (nh)^{-1},\end{aligned}$$

其中 $C_B = g''(x) + \frac{2g'(x)f'(x)}{f(x)}$ 和 $C_V = \frac{\sigma^2(x)}{f(x)}$ 是与核函数 $K(\cdot)$ 和窗宽 h 无关的量.

■ 极小化MSE, 可获得如下理论上的最优窗宽

$$h_{\text{opt}} = \left[\frac{C_V c_K}{C_B^2 d_K^2} \right]^{1/5} n^{-1/5} =: c n^{-1/5}.$$

窗宽 h 的选择

- 如果取最优窗宽 h_{opt} , 则容易看到N-W核估计 $\hat{g}_{\text{NW}}(x)$ 在内点处的最优收敛速度为

$$\text{MSE}(\hat{g}_{\text{NW}}(x)) = O(n^{-4/5}).$$

- 最优窗宽 h_{opt} 依赖于 c 的大小, 而 c 与 $f(x), f'(x), g'(x), g''(x)$ 和 $\sigma^2(x)$ 有关, 这些量都是未知的.
 - 在实际应用中, c 的估计是非常困难的事情.
- ♠ **问题:** 在实际应用中, 如何获得最优窗宽 h ?

窗宽 h 的选择

■ LOOCV方法：极小化下面的CV 目标函数获得最优窗宽 \hat{h}_{cv} , 即

$$\hat{h}_{cv} = \arg \min_{h>0} CV(h) = \arg \min_{h>0} \frac{1}{n} \sum_{i=1}^n [y_i - \hat{g}_h^{(-i)}(x_i)]^2 \omega(x_i),$$

其中

- ▶ $\hat{g}_h^{(-i)}(x_i)$ 表示去掉第 i 个观测样本 (y_i, x_i) 后得到的带有窗宽 h 的N-W核估计
- ▶ $\omega(x)$ 为非负的权函数, 可取为 $\omega(x) = I(|x - 0.5| \leq 0.4)$, 即去掉了两个边界点附近的样本点

窗宽 h 的选择

■ GCV 方法: 记 $(\hat{g}_h(x_1), \dots, \hat{g}_h(x_n))^T =: \mathbf{S}_h Y$, 其中

- ▶ $\hat{g}_h(x_i)$ 是 $g(x)$ 的具有窗宽 h 的任意一个非参数函数的拟合曲线
- ▶ \mathbf{S}_h 表示仅依赖于变量 X 的 $n \times n$ 的帽子矩阵
- ▶ $Y = (y_1, \dots, y_n)^T$

■ GCV方法选取的最优窗宽为

$$\hat{h}_{\text{gcv}} = \arg \min_{h>0} \text{GCV}(h) = \arg \min_{h>0} \frac{\frac{1}{n} \sum_{i=1}^n [y_i - \hat{g}_h(x_i)]^2}{[n^{-1} \text{tr}(\mathbf{I}_n - \mathbf{S}_h)]^2}.$$

核函数的选取

■ 把最优窗宽 h_{opt} 代入到MSE中，计算得到

$$\text{MSE}(K) = \frac{5}{4} (C_V^2 C_B c_K^2 d_K)^{2/5} n^{-4/5}.$$

■ 从上式可看到， $\text{MSE}(K)$ 关于核函数仅依赖于 $c_K^2 d_K$ ，即

$$\left(\int_{-\infty}^{\infty} K^2(u) du \right)^2 \int_{-\infty}^{\infty} u^2 K(u) du.$$

■ 极小化 $\text{MSE}(K)$ ，最优核函数为： $K(u) = \frac{3}{4}(1 - u^2)I(|u| \leq 1)$ ，即为**Epanechnikov核**.

本章纲要

1 多项式回归

2 回归样条

- d 阶回归样条
- 线性样条
- 三次样条
- 自然三次样条
- 节点个数和位置的选择

3 光滑样条

4 局部非参数光滑方法

- N-W核光滑方法
- 局部多项式光滑方法

5 广义可加模型

6 半参数回归模型

7 参考文献

8 作业

■ Fan (1993), 与Fan 和Gijbels (1996)提出的局部多项式光滑方法具有良好的性质:

- ① 可以减小N-W核估计的渐近偏差;
- ② 可以减小Gasser-Müller 估计的渐近方差;
- ③ 对边界效应具有自适应性, 即在边界处的收敛速度和内点处的收敛速度相同, 具有非参数最优的收敛速度;
- ④ 具有良好的最小最大有效性;
- ⑤ 具有容易解释和计算的优点, 并适应于导数的估计.

局部多项式估计

■ 假设 $g(\cdot)$ 在 x 的邻域内有连续的 d 阶导数, Taylor展式可近似 $g(\cdot)$ 为

$$g(u) \approx \sum_{j=0}^d \frac{g^{(j)}(x)}{j!} (u - x)^j =: \sum_{j=0}^d \beta_j (u - x)^j,$$

其中 $\beta_j = g^{(j)}(x)/j!$, u 为 x 邻域内的点.

■ 构造如下的加权最小二乘目标函数

$$\sum_{i=1}^n \left[y_i - \sum_{j=0}^d \beta_j (x_i - x)^j \right]^2 K_h(x_i - x),$$

其中 $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ 为核函数, h 为窗宽.

- 令 $\mathbf{W} = \text{diag}\{K_h(x_1 - x), \dots, K_h(x_n - x)\}$ 为 $n \times n$ 的对角矩阵, $\mathbf{Y} = (y_1, \dots, y_n)^T$ 且

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 - x & \cdots & (x_1 - x)^d \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n - x & \cdots & (x_n - x)^d \end{pmatrix}.$$

- 加权最小二乘目标函数可重新写成下面矩阵形式

$$(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{W} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

■ 极小化加权最小二乘目标函数，可得 β 的估计为

$$\hat{\beta} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y}.$$

■ 由 $\beta_j = g^{(j)}(x)/j!$ 可知，可得 $g^{(j)}(x)$ 的估计为

$$\hat{g}^{(j)}(x) = j! \hat{\beta}_j(x), \quad j = 0, 1, \dots, d.$$

■ 回归函数 $g(x)$ 的局部多项式估计为

$$\hat{g}(x) = \hat{\beta}_0(x) = \mathbf{e}_{1,d+1}^T (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y},$$

其中 $\mathbf{e}_{1,d+1}$ 表示第1个元素为1，其余元素均为0的 $d+1$ 维单位列向量。

- 下面重点讨论一种特殊情况，**局部线性估计**，即取 $d = 1$.
- **主要的原理是：**在 x 的一个邻域 $[x - h, x + h]$ 内，用一个线性函数去逼近非参数函数 $g(x)$ ，即

$$g(x) \approx g(t) + g'(t)(x - t) =: a + b(x - t),$$

其中 x 为 t 邻域内的点。当 x 在适当的范围内变化时，通过实施局部线性估计可以得到整个曲线 $g(\cdot)$ 的估计曲线。

■ 非参数函数 $g(\cdot)$ 的局部线性估计定义为

$$\hat{g}_{\text{LL}}(x) = \sum_{i=1}^n W_{ni}^{\text{LL}}(x)y_i,$$

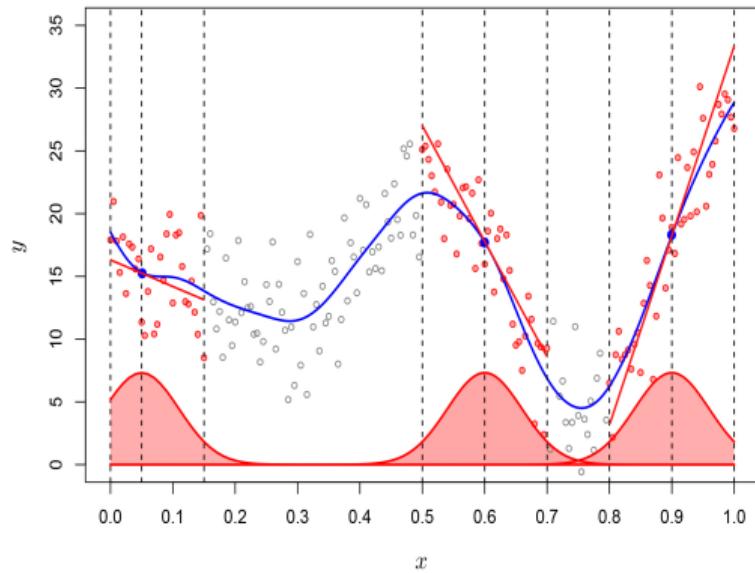
其中

$$W_{ni}^{\text{LL}}(x) = \frac{K_h(x_i - x)[S_{n,2}(x) - (x_i - x)S_{n,1}(x)]}{S_{n,0}(x)S_{n,2}(x) - S_{n,1}^2(x)},$$

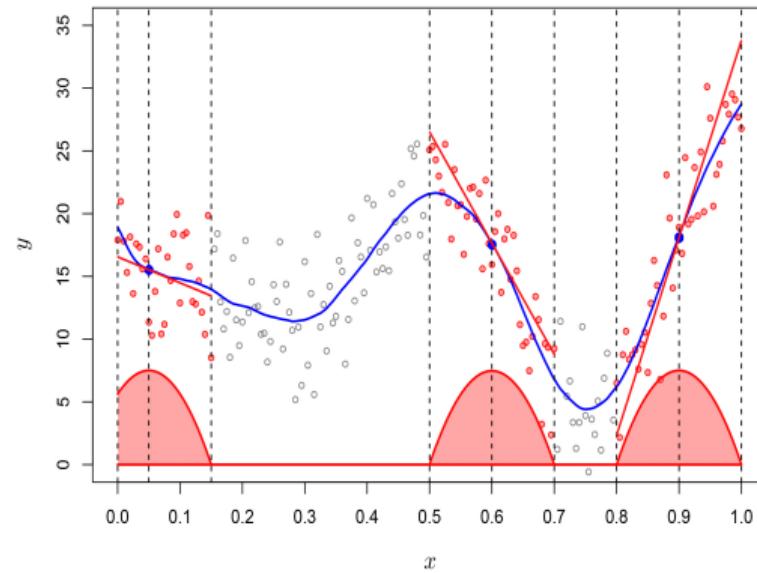
且 $S_{n,k}(x) = \sum_{i=1}^n K_h(x_i - x)(x_i - x)^k, k = 0, 1, 2.$

■ 权重满足: $\sum_{i=1}^n W_{ni}^{\text{LL}}(x) = 1.$

局部线性估计



(a)



(b)

局部线性估计的原理示意图. (a) 高斯核; (b) Epanechnikov核

- 当 $n \rightarrow \infty$ 时, $h \rightarrow 0$, $nh \rightarrow \infty$, 可推得局部线性估计 $\hat{g}_{LL}(x)$ 的渐近条件偏差为

$$\text{bias}(\hat{g}_{LL}(x)|\mathcal{F}_n) = \frac{1}{2}d_K g''(x)h^2 + o_P(h^2);$$

- 局部线性估计 $\hat{g}_{LL}(x)$ 的渐近条件方差为

$$\text{Var}(\hat{g}_{LL}(x)|\mathcal{F}_n) = \frac{c_K \sigma^2(x)}{f(x)} \frac{1}{nh} + o_P((nh)^{-1}),$$

- \mathcal{F}_n 表示由 $\{x_i, i = 1, \dots, n\}$ 产生的 σ 代数
- $f(x)$ 是 X 的密度函数, $g''(x)$ 是 $g(x)$ 的二阶导数
- $d_K = \int_{-\infty}^{\infty} u^2 K(u) du$, $c_K = \int_{-\infty}^{\infty} K^2(u) du$

■ 局部线性估计 $\hat{g}_{LL}(x)$ 的渐近条件MSE 为

$$\text{MSE}(\hat{g}_{LL}(x)|\mathcal{F}_n) = \left\{ \frac{1}{4} d_K^2 [g''(x)]^2 h^4 + \frac{c_K \sigma^2(x)}{f(x)} \frac{1}{nh} \right\} [1 + o_P(1)].$$

■ 极小化 $\text{MSE}(\hat{g}_{LL}(x)|\mathcal{F}_n)$, 可得理论的最优窗宽为

$$h_{\text{opt}} = \left\{ \frac{c_K \sigma^2(x)}{d_K^2 [g''(x)]^2 f(x)} \right\}^{1/5} n^{-1/5} =: c n^{-1/5}.$$

■ 在实际应用中, 同样可采用LOOCV或GCV等数据驱动方法选取最优窗宽.

■ 在R语言中，可用于非参数光滑方法的函数有：

- ① 函数`ksmooth()`
- ② 程序包`KernSmooth`中的函数`locpoly()`
- ③ 程序包`PLRModels`中的函数`np.est()` ——重点推荐
- ④ 函数`loess()`
- ⑤ 程序包`np`中的函数`npreg()`
- ⑥ 程序包`locfit`中的函数`locfit()`
- ⑦ 程序包`locpol`中的函数`locpol()`
- ⑧ 程序包`ggplot2` 中的函数`geom_smooth()`
- ⑨ 程序包`sm`

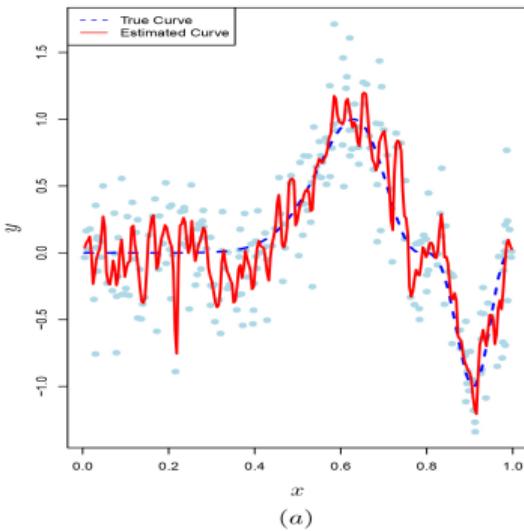
■ 函数ksmooth()是用于非参数回归模型的N-W核估计问题, 其调用格式为

```
ksmooth(x, y, kernel=c("box", "normal"), bandwidth=0.5,  
        range.x=range(x), n.points=max(100L, length(x)),  
        x.points)
```

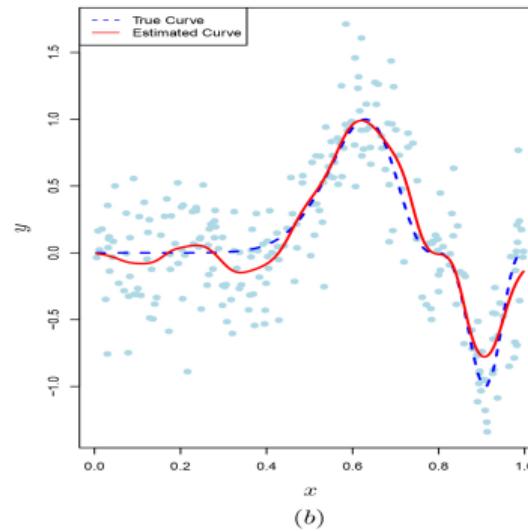
其中x表示协变量观测数据, y表示响应变量的观测数据; kernel表示核函数, normal表示高斯核函数, box表示矩形盒子核函数; bandwidth表示窗宽, 缺省为0.5.

案例分析与应用

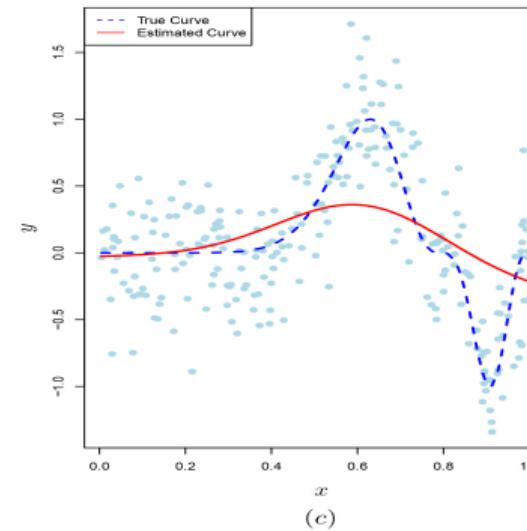
■ 对程序包`faraway`中的`exa`数据，在三种不同窗宽 $h = 0.01, 0.08, 0.5$ 下，利用函数`ksmooth()`进行N-W核估计.



(a)



(b)



(c)

程序包`faraway`中`exa`数据的N-W核估计. (a) 窗宽 $h = 0.01$; (b) 窗宽 $h = 0.08$; (c) 窗宽 $h = 0.5$

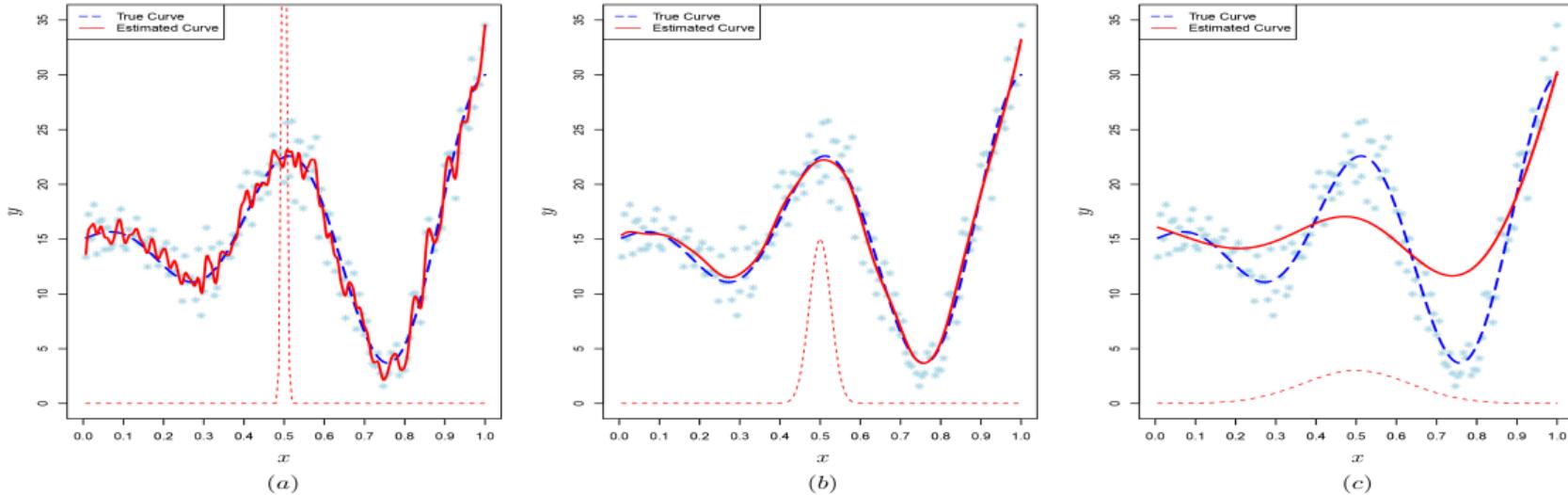
■ 程序包KernSmooth中的函数locpoly() 主要利用局部多项式光滑方法估计密度函数, 回归函数和它们的导数, 调用格式为

```
locpoly(x, y, drv = 0L, degree, kernel = "normal",
        bandwidth, gridsize = 401L, bwdisc = 25,
        range.x, binned = FALSE, truncate = TRUE)
```

其中x表示协变量观测数据, y表示响应变量的观测数据; drv表示需要估计回归函数导数的阶数; degree表示局部多项式的阶数, 要求大于参数drv; kernel表示核函数, 取"normal", "box", "epanech", "biweight"和"triweight"; bandwidth表示窗宽; 其余参数见在线帮助.

- 对Sheather (2009)中使用的curve数据集进行分析, 该数据集包含150个样本, 来自于模型 $y_i = 15(1 + x_i \cos(4\pi x_i)) + \varepsilon_i$, 其中 $\varepsilon_i \sim N(0, 4)$, 且 x_i 均匀分布于区间 $[0, 1]$.
- 首先, 利用函数`dpill()`选取最优窗宽 $h_{\text{opt}} = 0.026$;
- 其次, 为了比较, 考虑另外两种不同窗宽情形, 即 $h = h_{\text{opt}}/5 = 0.005$ 和 $h = 5h_{\text{opt}} = 0.132$;
- 最后, 基于高斯核函数, 用函数`locpoly()`对数据进行局部线性拟合.

案例分析与应用



Curve数据集的局部线性估计拟合曲线. (a) 窗宽 $h = h_{\text{opt}}/5 = 0.005$; (b) 窗宽 $h_{\text{opt}} = 0.026$; (c) 窗宽 $h = 5h_{\text{opt}} = 0.132$

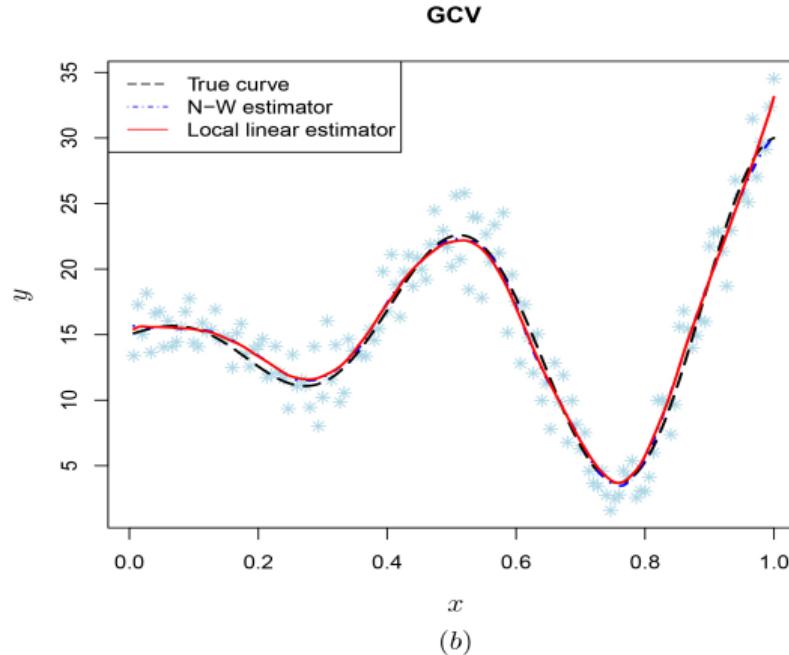
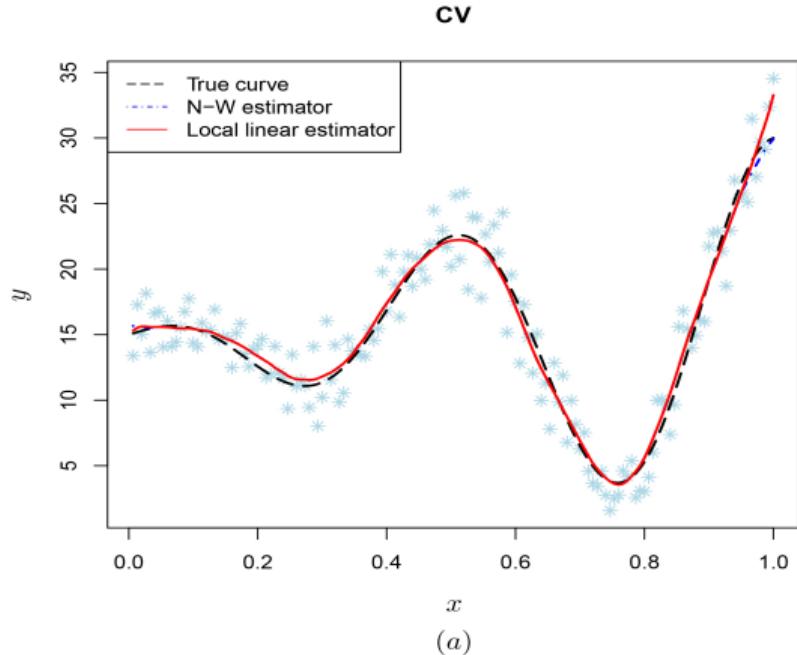
■ 程序包 **PLRModels** 中的函数 **np.est()** 提供了非参数回归模型的 N-W 核估计和局部线性估计，更适合本节介绍的局部非参数估计方法，调用格式为

```
np.est(data = data, h.seq = NULL, newt = NULL,  
       estimator = "NW", kernel = "quadratic")
```

其中 **data** 表示包含响应变量和协变量的数据； **h.seq** 表示窗宽，缺省为 CV 准则进行选取； **newt** 表示新的协变量数据； **estimator** 取 "NW" 或 "LLP"，其中 "NW" 表示 N-W 核估计， "LLP" 表示局部线性估计，缺省为 N-W 核估计； **kernel** 取 "gaussian" (高斯核函数)， "quadratic" (Epanechnikov 核函数)， "triweight" (triweight 核函数) 或者 "uniform" (均匀核函数)，缺省为 Epanechnikov 核函数。

- 程序包 **PLRModels** 中还提供了窗宽选取的函数, 如基于CV准则的函数 `np.cv()` 和基于GCV 准则的函数 `np.gcv()`.
- 针对 `curve` 数据集, 取 Epanechnikov 核函数, 分别用 CV 准则和 GCV 准则选取最优窗宽.
- 利用函数 `np.est()` 进行 N-W 核估计和局部线性估计拟合.

案例分析与应用



Curve数据集的N-W核估计和局部线性估计的拟合曲线. (a) 基于CV准则选取最优窗宽的拟合曲线;
(b) 基于GCV准则选取最优窗宽的拟合曲线

本章纲要

1 多项式回归

2 回归样条

- d 阶回归样条
- 线性样条
- 三次样条
- 自然三次样条
- 节点个数和位置的选择

3 光滑样条

4 局部非参数光滑方法

- N-W核光滑方法
- 局部多项式光滑方法

5 广义可加模型

6 半参数回归模型

7 参考文献

8 作业

广义可加模型(generalized additive model, GAM)

- 对于多元非参数回归模型: $Y = g(\mathbf{X}) + \varepsilon$, 其中协变量向量 $\mathbf{X} = (X_1, \dots, X_p)^T$.
- **优点:** 形式任意, 模型灵活, 可更好的拟合数据.
- **缺点:** 当 $p > 2$ 时, 将会遭遇“维数灾祸”问题, 估计的收敛速度慢, 且估计精度会降低.
- **解决办法:** 广义可加模型 (generalized additive model, GAM)
- **估计方法:** 后移算法 (backfitting algorithm)

■ 广义可加模型具有如下形式

$$Y = \beta_0 + g_1(X_1) + \cdots + g_p(X_p) + \varepsilon,$$

- ▶ β_0 是截距项, $g_1(\cdot), \dots, g_p(\cdot)$ 是 p 个未知的一元连续光滑函数
- ▶ 模型误差 ε 满足: $E(\varepsilon|X) = 0$ 和 $\text{Var}(\varepsilon|X) = \sigma^2 < \infty$

■ 为了保证函数 $g_1(\cdot), \dots, g_p(\cdot)$ 的可识别性, 需要假设

$$E[g_k(X_k)] = 0, \quad k = 1, \dots, p.$$

- 假设 $\{(x_i, y_i), i = 1, \dots, n\}$ 是来自 GAM 的 i.i.d. 的随机样本, 其中 $x_i = (x_{i1}, \dots, x_{ip})^T$.
- 由可识别性条件和 $E(\varepsilon|X) = 0$, 则有 $E(Y) = \beta_0$. 因此, 截距项 β_0 的估计为: $\hat{\beta}_0 = \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$.
- 为了简单, 假设 X_1, \dots, X_p 的支撑集都为 $[0, 1]$, 下面以光滑样条方法为例介绍 **后移算法**, 则 $g_1(\cdot), \dots, g_p(\cdot)$ 的可加三次光滑样条估计为

$$\hat{g}_1, \dots, \hat{g}_p = \arg \min_{g_1, \dots, g_p} \left\{ \sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \sum_{k=1}^p g_k(x_{ik}) \right)^2 + \sum_{k=1}^p \lambda_k \int_0^1 [g_k''(x)]^2 dx \right\}.$$

广义可加模型—后移算法

- 假设 $\hat{\beta}_0 = \bar{y}$, 并初始化 $\hat{g}_1, \dots, \hat{g}_p$, 不妨假设所有的初始估计为0.
- 对 $k = 1, \dots, p$, 重复迭代下面的步骤, 直到收敛, 具体**后移算法**为:
 - ① 令 $\hat{r}_i = y_i - \hat{\beta}_0 - \sum_{s \neq k} \hat{g}_s(x_{is})$, 其中 $i = 1, \dots, n$;
 - ② 对 $k = 1, \dots, p$, 利用光滑样条方法估计

$$\hat{g}_k = \arg \min_{g_k} \left\{ \sum_{i=1}^n (\hat{r}_i - g_k(x_{ik}))^2 + \lambda_k \int_0^1 [g_k''(x)]^2 dx \right\};$$

- ③ (中心化): $\hat{g}_k = \hat{g}_k - \frac{1}{n} \sum_{i=1}^n \hat{g}_k(x_{ik})$, 其中 $k = 1, \dots, p$.

在R语言中，拟合GAM常用的函数有：

- ① 函数lm()
- ② 程序包gam中的函数gam()
- ③ 程序包mgcv中的函数gam()
- ④ 程序包SemiPar中的函数spm()

— 配套Ruppert, Wand和Carroll(2003)的专著 *Semiparametric Regression*

■ 以James 等(2021)中的Wage数据集为例进行说明，该数据集包含3000个样本和11个变量，此处仅考虑协变量year, age和education对响应变量wage的影响，建立下面的广义可加模型

$$\text{wage} = \beta_0 + g_1(\text{year}) + g_2(\text{age}) + g_3(\text{education}) + \varepsilon,$$

其中

- ▶ year和age是定量变量
- ▶ education是取五个水平的定性变量，即<HS Grad, HS Grad, Some College, College Grad, Advanced Degree, 指的是个体高中和大学的履历记录

■ 用函数lm()进行GAM估计：

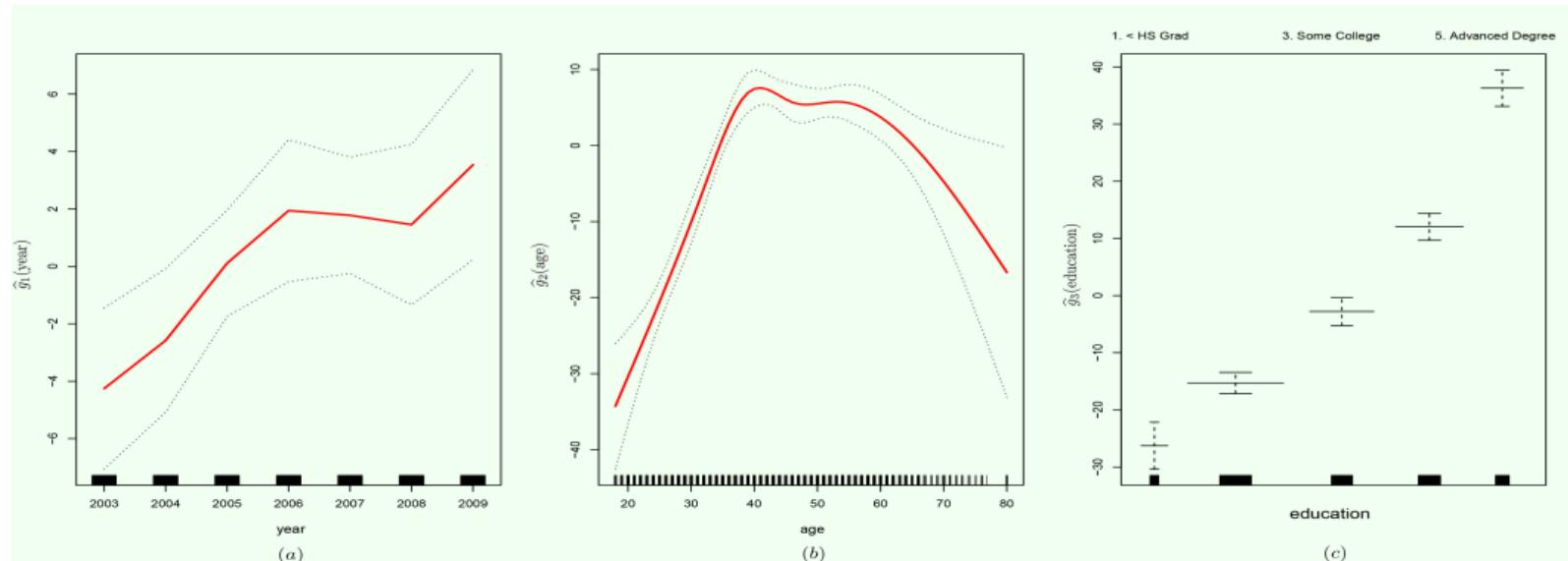
```
library(ISLR2); library(splines); library(gam)
attach(Wage)

gam1=lm(wage~bs(year,df=4)+ns(age,df=5)+education,data=Wage)
summary(gam1)

par(mfrow = c(1, 3))

plot.Gam(gam1, se = TRUE, lwd = 2, col = "red")
```

GAM的案例与应用



Wage数据集的广义可加模型拟合. (a) 函数 $g_1(\text{year})$ 的自由度为4 的三次样条拟合曲线和95% 置信带; (b) 函数 $g_2(\text{age})$ 的自由度为5的自然样条拟合曲线和95% 置信带; (c) 函数 $g_3(\text{education})$ 估计的阶梯函数和95% 置信区间

GAM的案例与应用

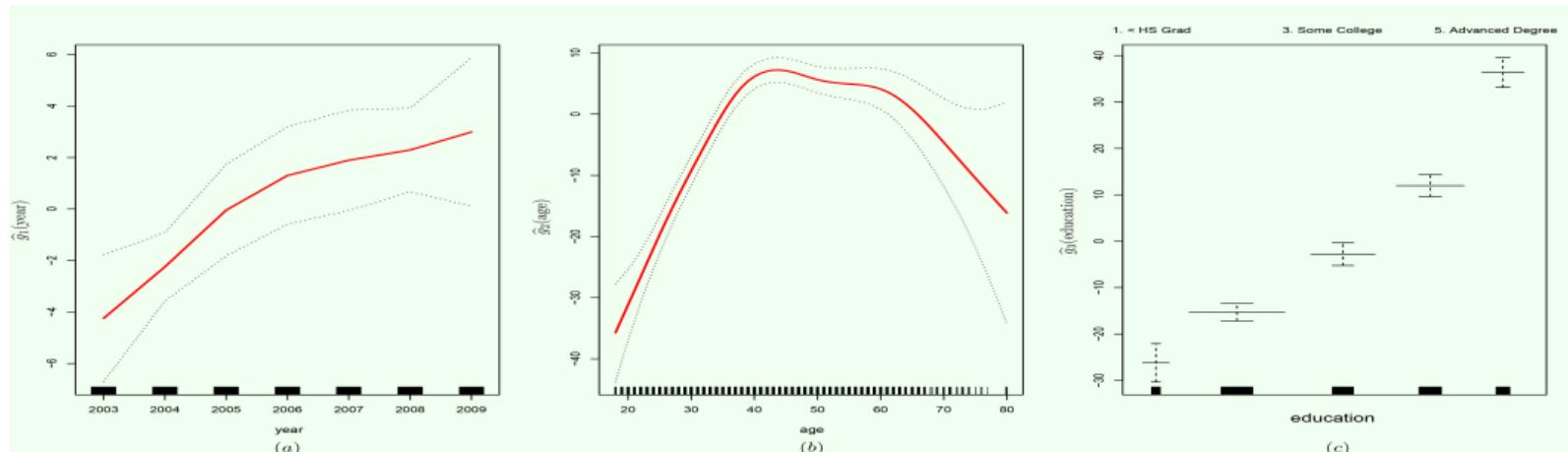
■ 程序包`gam`中的函数`gam()`提供了后移算法对非参数函数的拟合，调用格式为

```
gam(formula, family = gaussian, data, weights, subset,  
na.action, start, etastart, mustart,  
control = gam.control(...),  
model=TRUE, method, x=FALSE, y=TRUE, ...)
```

其中`formula`为模型公式，类似于函数`lm()`或`glm()`；`family`表示拟合模型的误差分布或联系函数，缺省时表示`gaussian`，更多有`binomial`，`Gamma`，`poisson`和`quasi`等；其余参数见在线帮助。

GAM的案例与应用

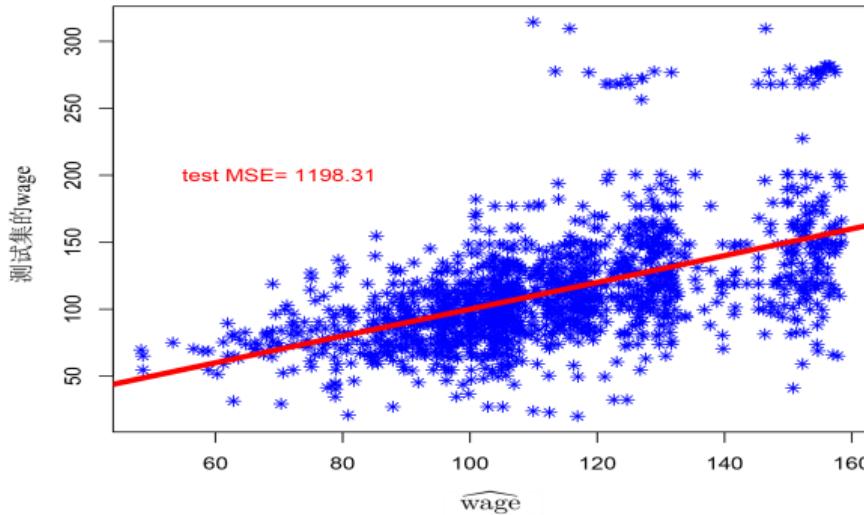
```
gam2=gam(wage~lo(year, span=0.8)+s(age, df=5)+education, data=Wage)
summary(gam2)
par(mfrow = c(1, 3))
plot.Gam(gam2, se = TRUE, lwd = 2, col = "red")
```



利用程序包`gam`中的函数`gam()`对Wage数据集的GAM拟合

GAM的案例与应用

- 把3000个样本随机分成1500个样本的训练集和1500个样本的测试集
- 在训练集上拟合模型，然后用函数predict()在测试集上进行预测



预测的散点图和 45° 线，测试均方误差为：test MSE = 1198.31

■ 程序包mgcv中的函数gam()同样提供了后移算法对非参数函数的拟合, 拟合效果比程序包gam中的函数gam()更优, 调用格式为

```
gam(formula, family=gaussian(), data=list(), weights=NULL,  
subset=NULL, na.action, offset=NULL, method="GCV.Cp",  
optimizer=c("outer", "newton"), control=list(), scale=0,  
select=F, knots=NULL, sp=NULL, min.sp=NULL, H=NULL, gamma=1,  
fit=T, paraPen=NULL, G=NULL, in.out, drop.unused.levels=T,  
drop.intercept=NULL, discrete=F, ...)
```

其中formula为模型公式, 类似于函数lm()或glm(); family表示拟合模型的误差分布或联系函数, 缺省时表示gaussian, 其他类似于函数glm(); method表示光滑参数的估计方法, 其中GCV.Cp表示用GCV准则和Cp准则进行估计, 其他估计方法见在线帮助; 其余参数见在线帮助.

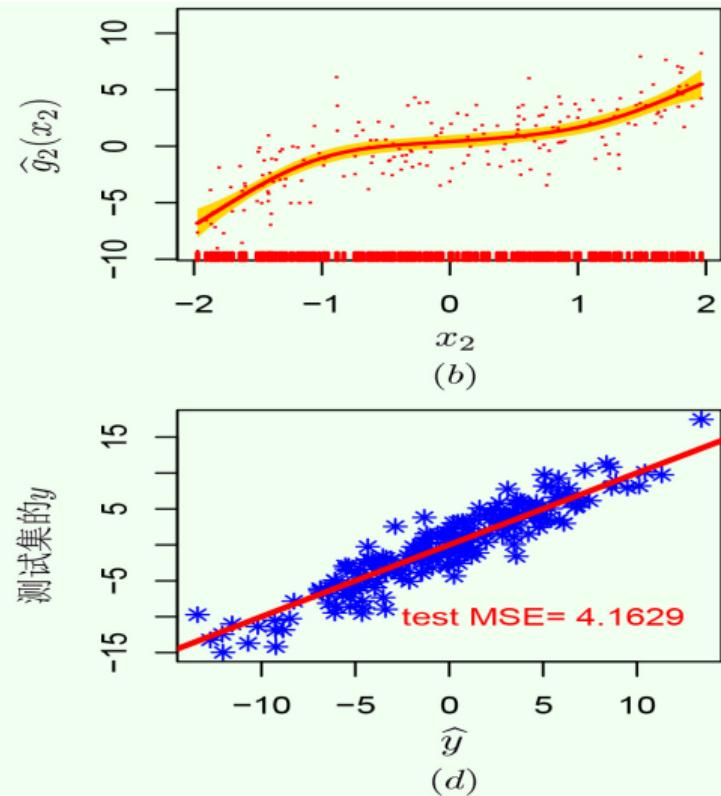
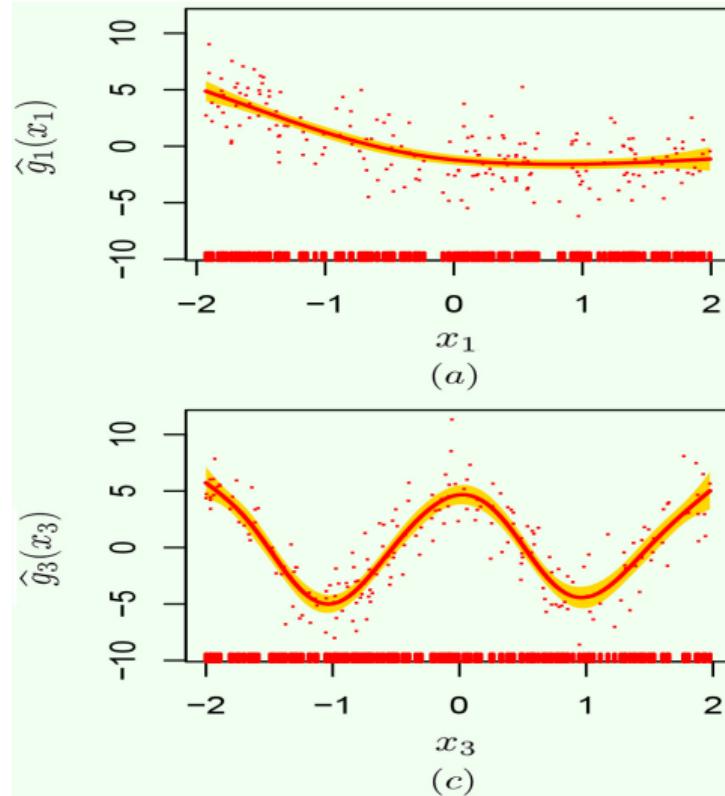
■ 模拟例子：从下面的GAM中生成随机数据进行分析，即

$$y_i = g_1(x_{i1}) + g_2(x_{i2}) + g_3(x_{i3}) + \varepsilon_i, \quad i = 1, \dots, 400,$$

- ▶ $g_1(x_{i1}) = \exp(-x_{i1}) - 1.63$, $g_2(x_{i2}) = (x_{i2} - 0.1)^3$, $g_3(x_{i3}) = 5 \cos(\pi x_{i3})$
- ▶ 模型误差 $\varepsilon_i \sim N(0, 4)$
- ▶ $x_{ik} \sim U(-2, 2)$, $k = 1, 2, 3$

■ 固定种子 `set.seed(2022)`, 把 400 个样本随机分成相同样本量大小的训练集和测试集. 在训练集上拟合 GAM, 然后在测试集上预测.

GAM的案例与应用

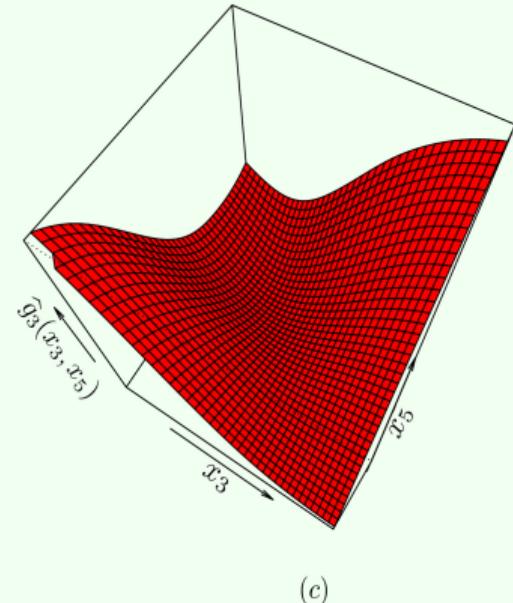
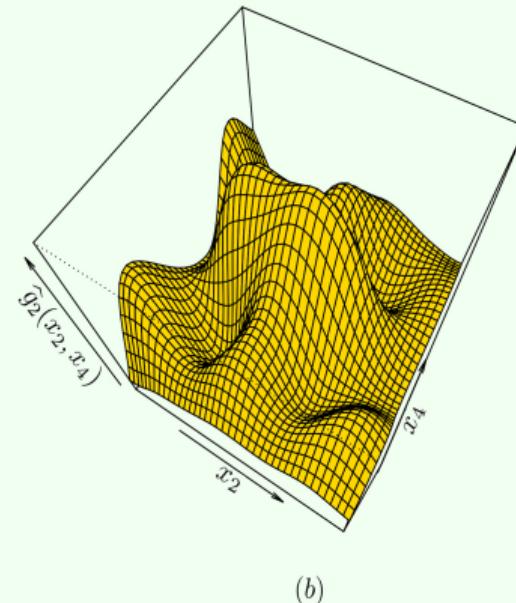
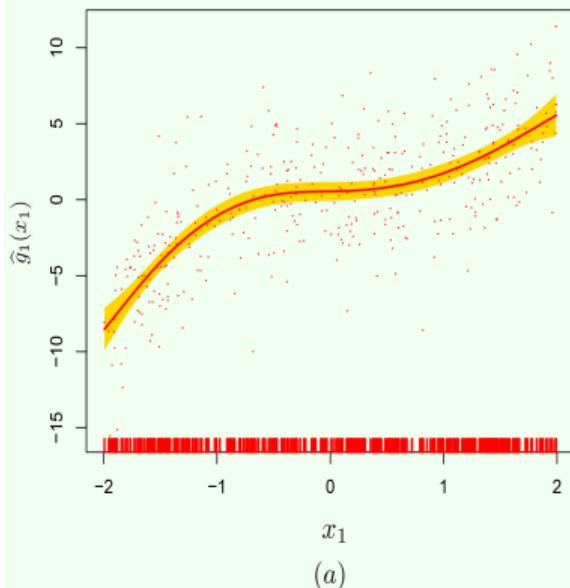


■ 程序包`mgcv`中的函数`gam()`, 也能对多元函数进行拟合. 考虑下面的广义可加模型

$$y_i = g_1(x_{i1}) + g_2(x_{i2}, x_{i4}) + g_3(x_{i3}, x_{i5}) + \varepsilon_i, \quad i = 1, \dots, 400,$$

- ▶ $g_1(x_{i1}) = (x_{i1} - 0.1)^3$ 为一元函数
- ▶ $g_2(x_{i2}, x_{i4}) = 5 \cos(\pi x_{i2} x_{i4})$ 为二元函数
- ▶ $g_3(x_{i3}, x_{i5}) = (x_{i3} + x_{i5})^2$ 为二元函数
- ▶ 模型误差 $\varepsilon_i \sim N(0, 4)$
- ▶ $x_{ik} \sim U(-2, 2), \quad k = 1, \dots, 5$

GAM的案例与应用



(a) 函数 $g_1(x_1)$ 的拟合曲线和 95% 置信带; (b) 函数 $g_2(x_2, x_4)$ 的拟合曲面; (c) 函数 $g_3(x_3, x_5)$ 的拟合曲面

本章纲要

1 多项式回归

2 回归样条

- d 阶回归样条
- 线性样条
- 三次样条
- 自然三次样条
- 节点个数和位置的选择

3 光滑样条

4 局部非参数光滑方法

- N-W核光滑方法
- 局部多项式光滑方法

5 广义可加模型

6 半参数回归模型

7 参考文献

8 作业

■ 为了解决多元非参数回归模型的“维数灾祸”问题，也发展了一些经典
的半参数模型：

- ① 部分线性模型
- ② 单指标模型
- ③ 变系数模型

■ 考虑如下的部分线性模型

$$Y = \mathbf{X}^T \boldsymbol{\beta} + g(T) + \varepsilon,$$

- ▶ $\mathbf{X} = (X_1, \dots, X_p)^T$ 为 p 维的协变量向量
- ▶ $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ 是 p 维的未知参数向量
- ▶ $g(\cdot)$ 是一元未知的光滑函数
- 柴根象和洪圣岩(1995, 半参数回归模型)
- Härdle, Liang 和 Gao (2000, Partially Linear Models)
- 薛留根(2012, 现代统计模型)
- 李高荣和杨宜平(2015, 纵向数据半参数模型)
- 李高荣, 张君和冯三营(2016, 现代测量误差模型)

■ 程序包PLRModels中的函数plrm.est(), 调用格式为

```
plrm.est(data = data, b = NULL, h = NULL, newt = NULL,  
         estimator = "NW", kernel = "quadratic")
```

其中data表示包含数据集, data[,1]为响应变量数据, data[,2:p+1]为线性部分协变量数据, data[,p+2]为非参数部分协变量数据; b为估计参数分量时的窗宽; h为估计非参数函数的窗宽, b和h缺省, 表示用CV准则选取窗宽; newt表示新的非参数部分协变量数据; 参数estimator和kernel相同于函数np.est()。

部分线性模型的模拟分析

■ 从下面的部分线性模型生成随机数据进行分析, 即

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + g(t_i) + \varepsilon_i, \quad i = 1, \dots, 400,$$

其中

- ▶ $\boldsymbol{\beta} = (-0.5, 1, -1, 1.5)^T$ 为参数分量的回归系数向量;
- ▶ $g(t_i) = 2(1 + t_i \cos(4\pi t_i))$ 为非参数函数;
- ▶ 模型误差 $\varepsilon_i \sim N(0, 1)$;
- ▶ 协变量 t 从区间 $[0, 1]$ 中等间距生成 400 个样本;
- ▶ 协变量向量 $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})^T \sim N_4(\mathbf{0}, \boldsymbol{\Sigma})$, 这里 $\boldsymbol{\Sigma} = (\sigma_{ij})_{1 \leq i, j \leq 4}$, 且 $\sigma_{ij} = 0.3^{|i-j|}$;
- ▶ 响应变量 y 可由模型产生, 则数据集为 $D = \{(y_i, \mathbf{x}_i, t_i), i = 1, \dots, 400\}$.

部分线性模型的模拟分析

- 用N-W核估计方法或局部线性光滑方法估计非参数函数 $g(\cdot)$, 取 Epanechnikov核函数, 并用CV准则选取最优窗宽;
- 用profile最小二乘方法估计回归系数向量 β ;
- 评价准则: 计算均方误差(MSE), 即

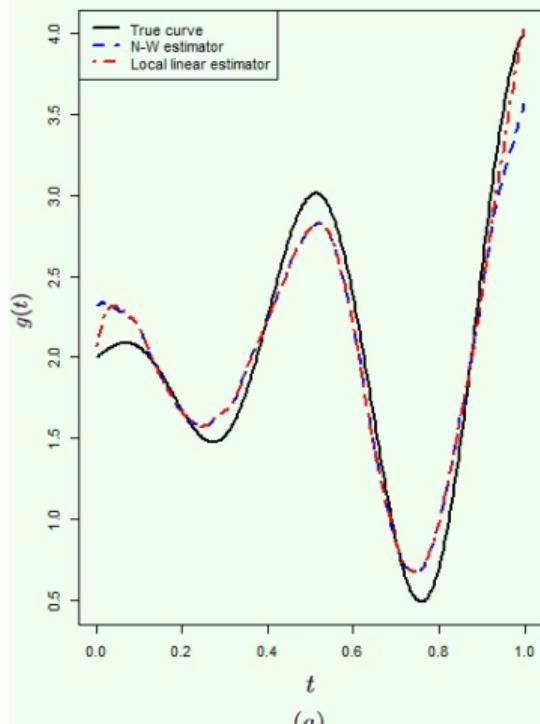
$$\text{MSE} = \frac{1}{400} \sum_{i=1}^{400} (y_i - \hat{y}_i)^2,$$

其中 $\hat{y}_i = \mathbf{x}_i^T \hat{\boldsymbol{\beta}} + \hat{g}(t_i), i = 1, \dots, 400.$

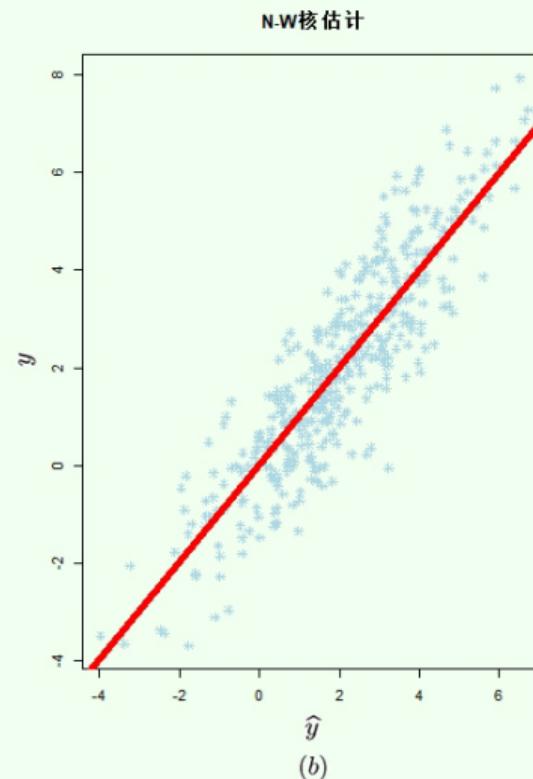
部分线性模型的模拟分析

```
library(PLRModels); library(MASS); set.seed(2022)
n = 400; p = 4; rho = 0.3; mu = rep(0, p)
g = function(t) { 2 + 2*t*cos(4*pi*t) }           ## g的定义
ar1mat = rho^outer(1:p, 1:p, function(x,y) abs(x-y))
x = mvrnorm(n, mu, ar1mat); beta = c(-0.5, 1, -1, 1.5)
t = ((1:n)-0.5)/n; y = x%*%beta + g(t) + rnorm(n, 0, 1)
D = matrix(c(y, x, t), nrow=n)
fit.nw = plrm.est(data = D)                      ## N-W核估计
fit.llp = plrm.est(data = D, estimator = "LLP") ## 局部线性估计
> fit.nw$beta                                     > fit.llp$beta
      [,1]                                         [,1]
[1,] -0.3796017                                 [1,] -0.3763588
[2,] 0.9878191                                  [2,] 0.9891673
[3,] -0.9606024                                 [3,] -0.9573624
[4,] 1.5356926                                  [4,] 1.5328420
> mean(fit.nw$residuals^2)                     > mean(fit.llp$residuals^2)
[1] 1.005253                                    [1] 0.9952479
```

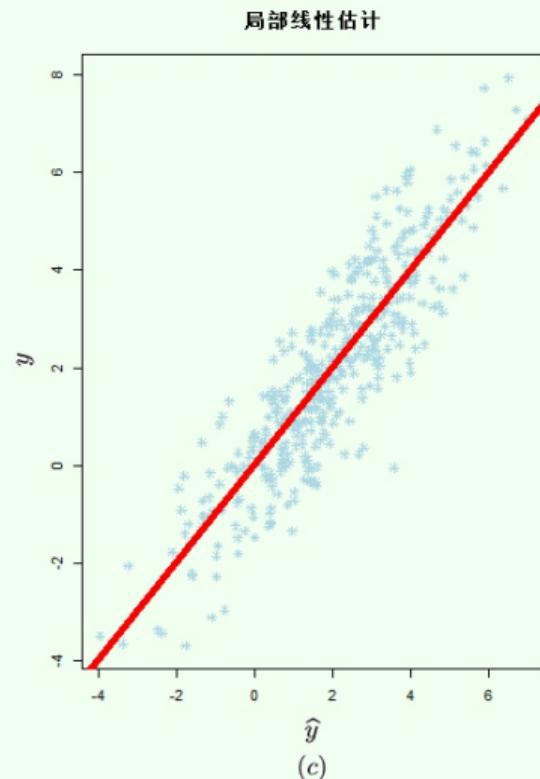
部分线性模型的模拟分析



(a)



(b)



(c)

■ 考虑部分线性单指标模型

$$Y = X^T \beta + g(Z^T \theta) + \varepsilon,$$

其中 β 是 p 维的未知参数向量, θ 为 q 维未知参数向量, $g(\cdot)$ 为一元未知联系函数. 为了模型的可识别性, 假定 $\|\theta\|_2 = 1$. 部分线性单指标模型包含了许多重要的统计模型, 例如

- 若 θ 为一维变量, 则模型就退化为部分线性模型
- 若 $X = \mathbf{0}$ 且 θ 为一维变量, 则模型就简化为一元非参数模型
- 若 $X = \mathbf{0}$, $g(\cdot)$ 为正态分布函数或 logistic 分布函数时, 则模型就成为 probit 模型或 logistic 模型
- 若 $g(\cdot) = 0$, 则模型就成为经典的线性模型
- 若 $X = \mathbf{0}$, 则模型就退化为如下的单指标模型: $Y = g(Z^T \theta) + \varepsilon.$

■ 单指标模型的综述论文见薛留根(2012, 数理统计与管理)

■ 为了平衡建模偏差和高维数据的“维数灾祸”等问题, 提出了变系数模型

$$Y = \mathbf{X}^T \boldsymbol{\beta} + \mathbf{Z}^T \boldsymbol{\alpha}(T) + \varepsilon,$$

其中 $\boldsymbol{\alpha}(\cdot) = (\alpha_1(\cdot), \dots, \alpha_q(\cdot))^T$ 是一个 q 维的未知回归系数函数, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ 是 p 维的未知参数向量. 变系数模型包含了许多重要的统计模型, 例如

- 当 $\mathbf{Z} = \mathbf{0}$, 模型退化为经典的线性回归模型
- 当 $q = 1$ 且 $Z = 1$, 模型退化为部分线性回归模型
- 当 $\mathbf{X} = \mathbf{0}$, 模型变成了著名的变系数模型: $Y = \mathbf{Z}^T \boldsymbol{\alpha}(T) + \varepsilon$
- 当 $\mathbf{X} = \mathbf{0}$ 且 $\mathbf{Z} = (1, \dots, 1)^T$ 时, 模型退化为广义可加模型

■ 详细讨论见Hastie 和Tibshirani (1993), Fan 和Zhang (2008), 张日权和卢一强(2004).

本章纲要

1 多项式回归

2 回归样条

- d 阶回归样条
- 线性样条
- 三次样条
- 自然三次样条
- 节点个数和位置的选择

3 光滑样条

4 局部非参数光滑方法

- N-W核光滑方法
- 局部多项式光滑方法

5 广义可加模型

6 半参数回归模型

7 参考文献

8 作业

- 柴根象, 洪圣岩 (1995). 半参数回归模型. 合肥: 安徽教育出版社.
- 李高荣, 杨宜平 (2015). 纵向数据半参数模型. 北京: 科学出版社.
- 李高荣, 张君, 冯三营 (2016). 现代测量误差模型. 北京: 科学出版社.
- 薛留根 (2012a). 现代统计模型. 北京: 科学出版社.
- 薛留根 (2012b). 单指标模型的统计推断. 数理统计与管理, 31(1): 55–78.
- 薛留根 (2012c). 单指标模型的统计推断. 数理统计与管理, 31(2): 226–246.
- 薛留根 (2015). 现代非参数统计. 北京: 科学出版社.
- 张日权, 卢一强 (2004). 变系数模型. 北京: 科学出版社.
- Fan, J. Q. (1993). Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics*, 21(1): 196–216.
- Fan, J. and Gijbels, I. (1996). Local Polynomial Modelling and Its Applications. London: Chapman and Hall.

- Fan, J. Q. and Zhang, W. Y. (2008). Statistical methods with varying coefficient models. *Statistics and Its Inference*, 1: 179–195.
- Gasser, T. and Müller, H. G. (1979). Kernel estimation of regression function. In *Smoothing Techniques for Curve Estimation*, Lecture Notes in Mathematics, 757: 23–68. New York: Springer-Verlag.
- Härdle, W., Liang, H. and Gao, J. T. (2000). *Partially Linear Models*. Heidelberg: Physica Verlag.
- Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B*, 55(4): 757–796.
- Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd Edition). New York: Springer-Verlag.
- James, G., Witten, D., Hastie, T. and Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R*. 2nd Ed. New York: Springer-Verlag.

- Li, Q. and Racine, J. (2007). Nonparametric Econometrics: Theory and Practice. Princeton: Princeton University Press.
- Nadaraya, E. A. (1964). On estimating regression. Theory of Probability and Its Application, 9(1): 141–142.
- Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). Semiparametric Regression. Cambridge: Cambridge University Press.
- Sheather, S. J. (2009). A Modern Approach to Regression with R. New York: Springer.
- Tryfos, P. (1998). Methods for Business Analysis and Forecasting: Text & Cases. New York: John Wiley & Son.
- Watson, G. S. (1964). Smooth regression analysis. Sankhya: The Indian Journal of Statistics, Series A, 26(4): 359–372.

本章纲要

1 多项式回归

2 回归样条

- d 阶回归样条
- 线性样条
- 三次样条
- 自然三次样条
- 节点个数和位置的选择

3 光滑样条

4 局部非参数光滑方法

- N-W核光滑方法
- 局部多项式光滑方法

5 广义可加模型

6 半参数回归模型

7 参考文献

8 作业

作业

[习题见教材: 统计学习(R语言版) — 习题8]

- **课后思考题:** 第1题、第3题、第5题、第6题
- **需要完成的课后作业:** 第2题、第7题、第9题
- **应用:** 第11题、第16题、第18题. 具体要求:
 - ① 能使用R语言把数据读入, 并对数据中的每个变量进行了解;
 - ② 能用学过的一些统计方法, 按照题目要求, 利用R语言对数据进行一些简单的分析, 并思考数据分析的结果.



谢谢，请多提宝贵意见！