高级计量暑期班
(Seminar of Advanced Econometrics)

胡华平

西北农林科技大学

经济管理学院数量经济教研室

huhuaping01@hotmail.com

2022-06-28

RDD PART 01 ：非参数估计

1.均值估计(Means Estimator)

2.局部回归(Local Regression)

3.估计效果(Performance Analysis)

4.群组分析(Cluster observations)

1.均值估计(Means Estimator)

1.1 箱组均值估计(Binned Means Estimator)

1.2 滚动箱组均值估计(Rolling Binned Means Estimator)

1.3 核估计(Kernel Estimator)

（引子）为什么要进行均值估计？

对于计量模型：

$$Y = m(X) + e$$

研究者首先需要关注的是

- **条件期望函数**（Conditional Expectation Function ,CEF）：

$$\mathbb{E}[Y|X = x] \equiv m(x)$$

- 此时：

$$\begin{aligned} Y &= m(X) + e \\ &= \mathbb{E}[Y|X = x] + e \end{aligned}$$

（引子）什么是非参数估计？

- 理论上，条件期望函数 $m(x)$可以表现为明确的**参数化**形式（parametric function），也可以表现为任意的**非参数化**形式（non-parametric function）。

- 常见的参数化条件期望函数，例如线性形式：

$$Y = m(x) + e = \beta_0 + \beta_1 X + e$$

（引子）什么是非参数估计？

- **非参数回归模型**（nonparametric regression model）：假定**条件期望函数**表现为任意的**非参数化**形式的回归模型。

  - **非参数回归模型**可以表达为：

$$Y = \mathbb{E}(Y|X = x) + e = m(x) + e$$
$$\mathbb{E}(e|X = x) = 0$$
$$\mathbb{E}(e^2|X = x) = \sigma^2(X)$$

- 此时，我们的目标就是估计得到条件期望函数 $\widehat{m}(x)$。

（示例）模拟数据集

为了更好地进行数据验证，我们将根据如下规则生成蒙特卡洛模拟数据集：

$$
\begin{aligned}
Y_i = m(X) + e_i &= \frac{sin(\frac{\pi}{4} \cdot (X_i - 2))}{\frac{\pi}{4} \cdot (X_i - 2)} + e_i \\
X_i &\sim U(0, 10) \\
e_i &\sim N(0, 2) \\
n &= 100
\end{aligned}
$$

- 此时，我们具有**上帝视角**，实际上已经知道**数据生成机制**（DGP）
- 此时，我们心里面已知真实模型为**非线性的**

（示例）模拟的样本数据集

模拟的样本数据集(n=100)

index	X	Y
1	2.9060	0.8120
2	8.4993	-0.4883
3	6.7846	-0.0454
4	3.2569	1.0822
5	8.8766	-0.4881
6	0.9354	1.1109
7	4.4082	0.2456
8	3.8783	0.9625

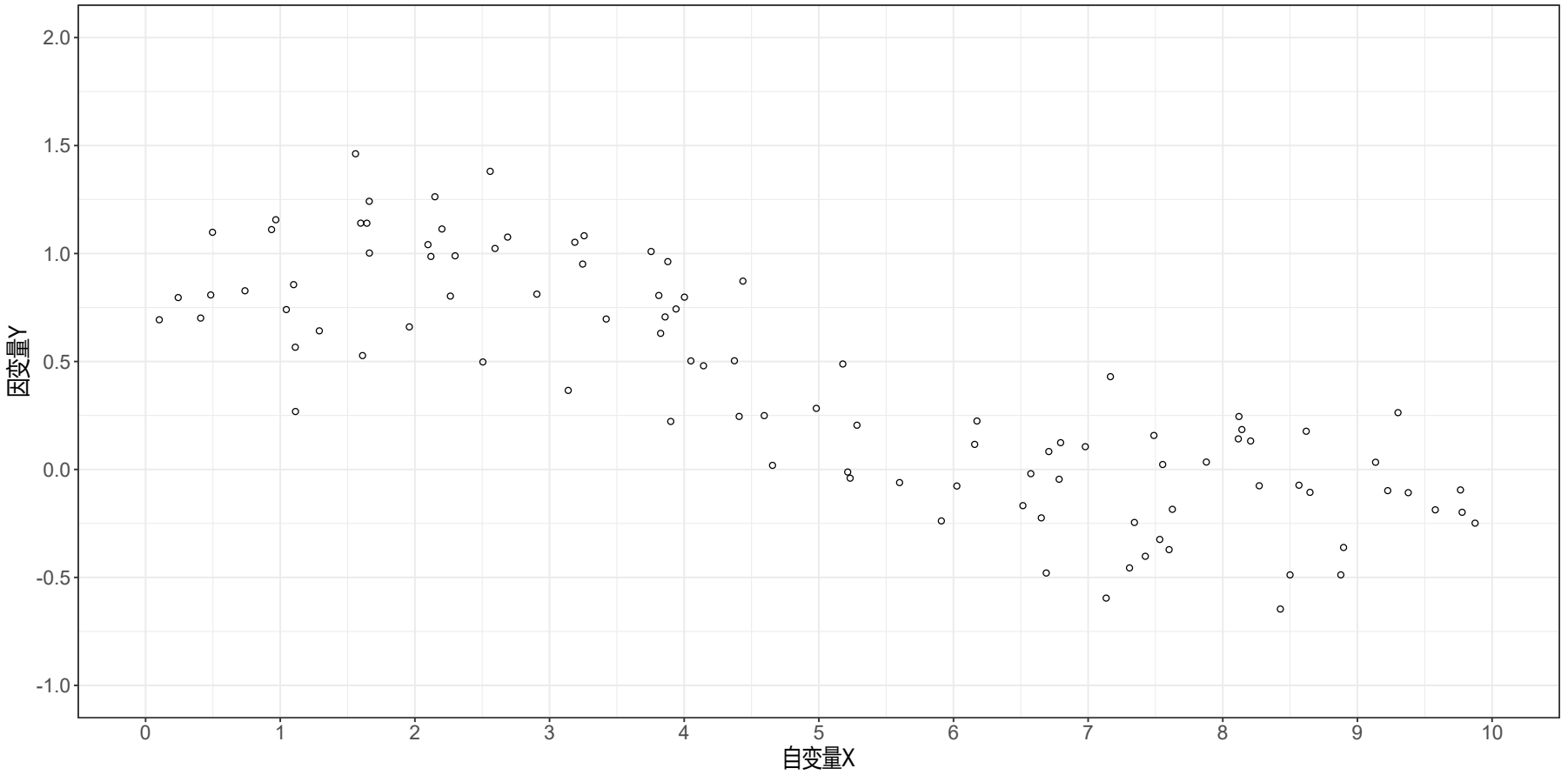
Showing 1 to 8 of 100 entries

Previous 1 2 3 4 5 … 13 Next

- 样本数据的描述性统计如下：

```
     index             X                Y
Min.   :  1   Min.   :0.1   Min.   :-0
1st Qu.: 26   1st Qu.:2.3   1st Qu.:-0
Median : 50   Median :4.6   Median : 0
Mean   : 50   Mean   :4.9   Mean   : 0
3rd Qu.: 75   3rd Qu.:7.5   3rd Qu.: 0
Max.   :100   Max.   :9.9   Max.   : 1
```

（示例）样本数据散点图



1.1 箱组均值估计：表达式

对于**非参数模型**：

$$Y = \mathbb{E}(Y|X = x) + e = m(x) + e$$

我们可以直接把数据集划分为不同**箱组**（bins），然后简单地计算各个箱组中 Y_i 的均值。

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} 1\{|X_i - x| \leq h\} \cdot Y_i}{\sum_{i=1}^{n} 1\{|X_i - x| \leq h\}}$$

其中：

- $1\{|X_i - x| \leq h\}$为**指示函数**，取值为 $\{0, 1\}$，以表明 X_i是否落在特定**箱组**内
- 以上公式可以简单视作为箱组内的**简单算数平均数**公式

1.1 箱组均值估计：操作步骤

箱组均值估计（Binned Estimator）的操作步骤如下：

- 根据计算点 $X = x_j$，按照特定谱宽 h，划分出若干**箱组**（bins）：$\{b_1, b_2, \cdots, b_q\}$

$$b_j = [x_j - h, x_j + h]$$

- 根据样本数据集，以及 X_i的实际情况，确定数据对 $\{X_i, Y_i\}$的箱组归属：

$$1\{|X_i - x_j| \leq h\}$$

- 最后计算不同箱组的 Y_i的均值 $\widehat{m}(x_{b_j}), j \in (1, 2, \cdots, q)$：

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} 1\{|X_i - x| \leq h\} \cdot Y_i}{\sum_{i=1}^{n} 1\{|X_i - x| \leq h\}}$$

（示例）箱组均值估计：设定箱组划分规则

下面我们分别设定如下**箱组**划分规则：

- 设定箱组取值中心点 $(X = x_i), x_i \in (1, 3, 5, 7, 9)$，以及谱宽 $h = 1$
- 然后得到箱组区块bins=[0,2)、[2,4)、[4,6)、[6,8)、[8,10]，箱组数为5。

（示例）箱组均值估计：数据计算表

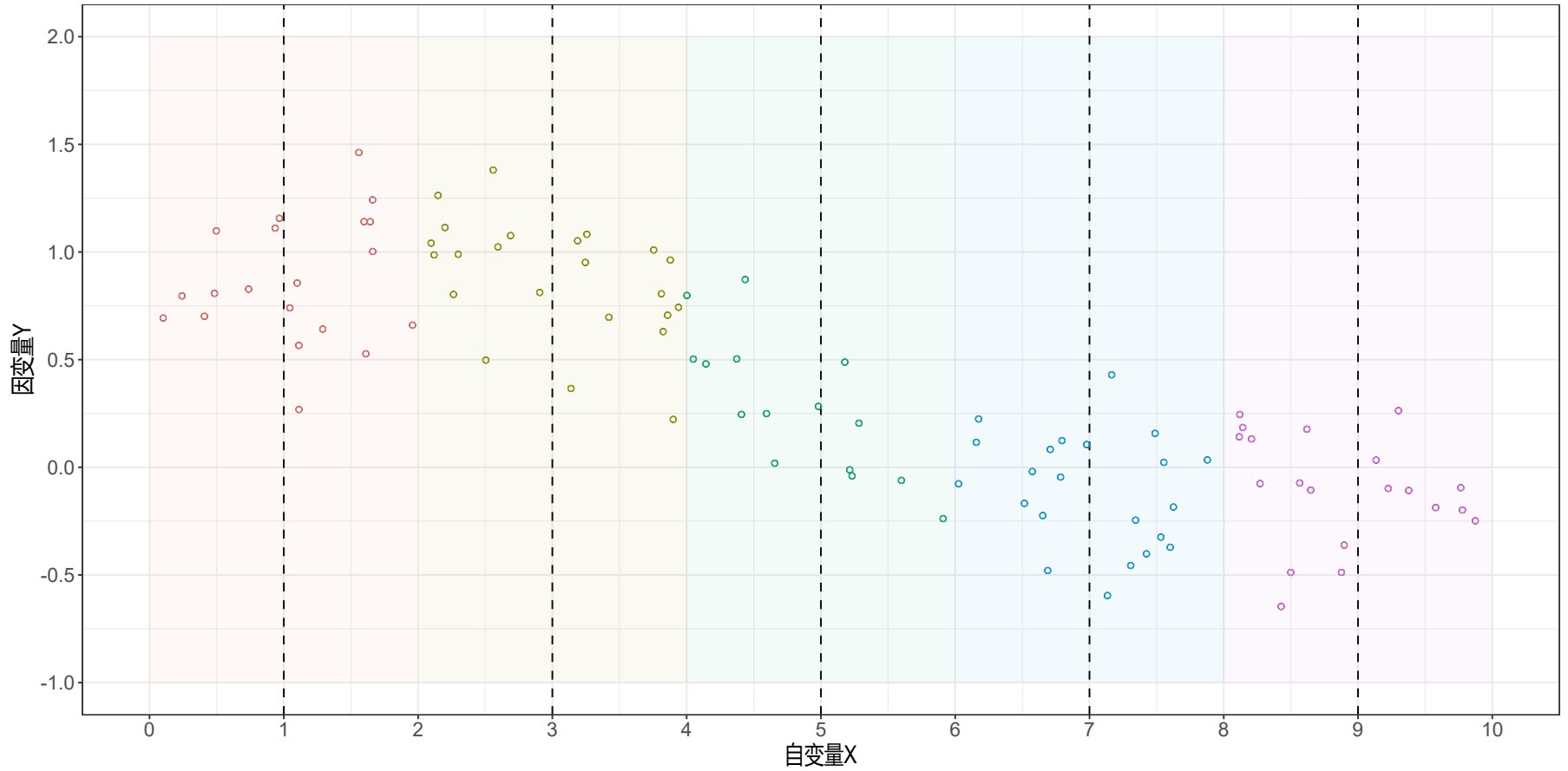
利用箱组均值估计公式，我们可以计算得到不同箱组的均值估计：

x	bins	sum_ky	sum_k	m0
1	[0,2)	17.4392	20	0.8720
3	[2,4)	20.2144	23	0.8789
5	[4,6)	4.2950	15	0.2863
7	[6,8)	-2.2950	22	-0.1043
9	[8,10]	-1.9965	20	-0.0998

- 箱组内因变量观测值的求和$\texttt{sum_ky} = \sum_{i=1}^{n} 1\{|X_i - x| \leq h\} \cdot Y_i$
- 箱组的样本数$\texttt{sum_k} = \sum_{i=1}^{n} 1\{|X_i - x| \leq h\}$
- 箱组的均值估计$\texttt{m0} = \widehat{m}(x_j), j \in (1, 2, \cdots, 5)$

（示例）箱组均值估计：图形表达1/3

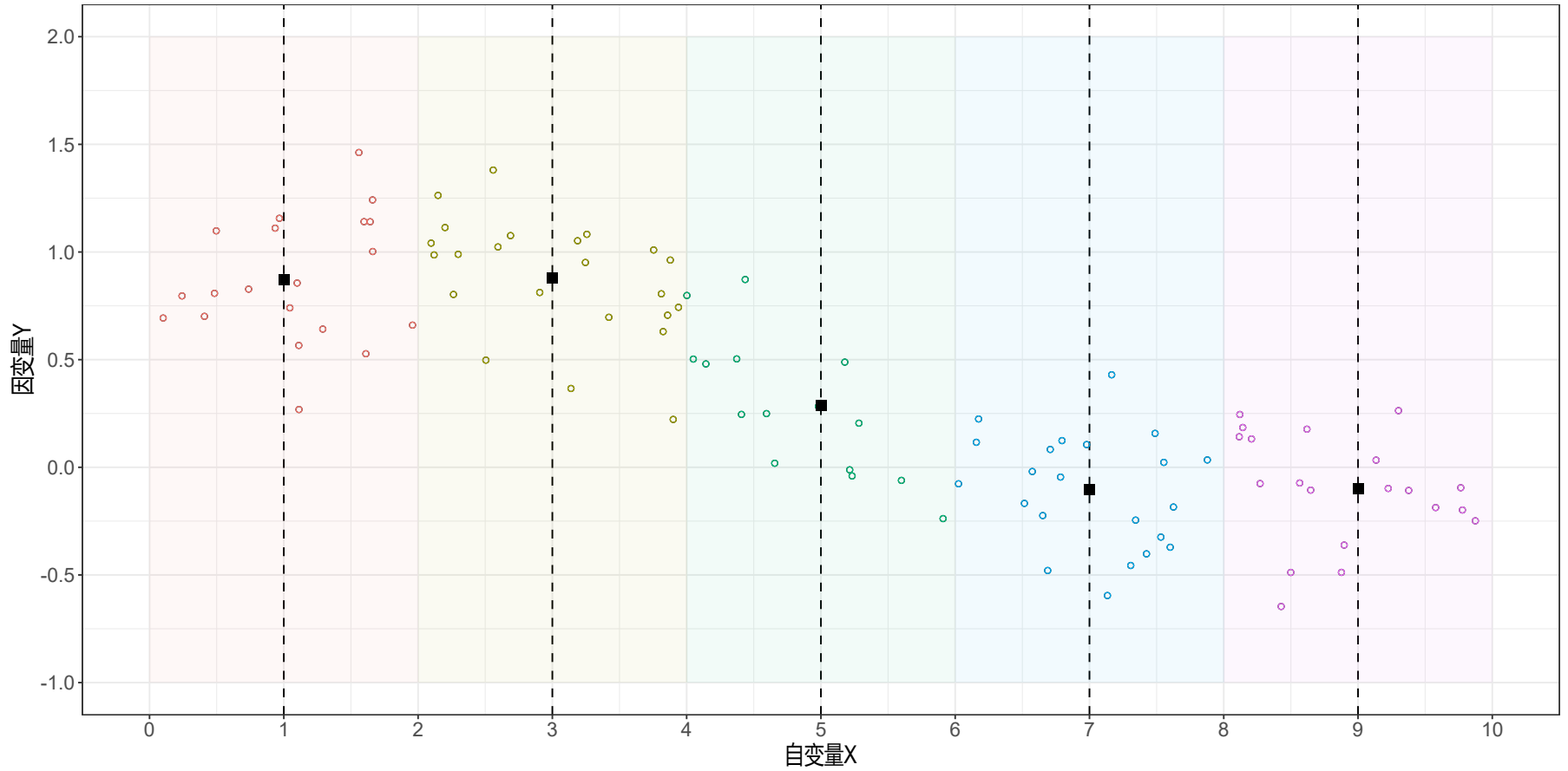
- 首先我们展示的是5个箱组的划分：



说明：a)垂直虚线表示箱组中心取值点 x_j；b)不同矩形颜色区块表示不同箱组。

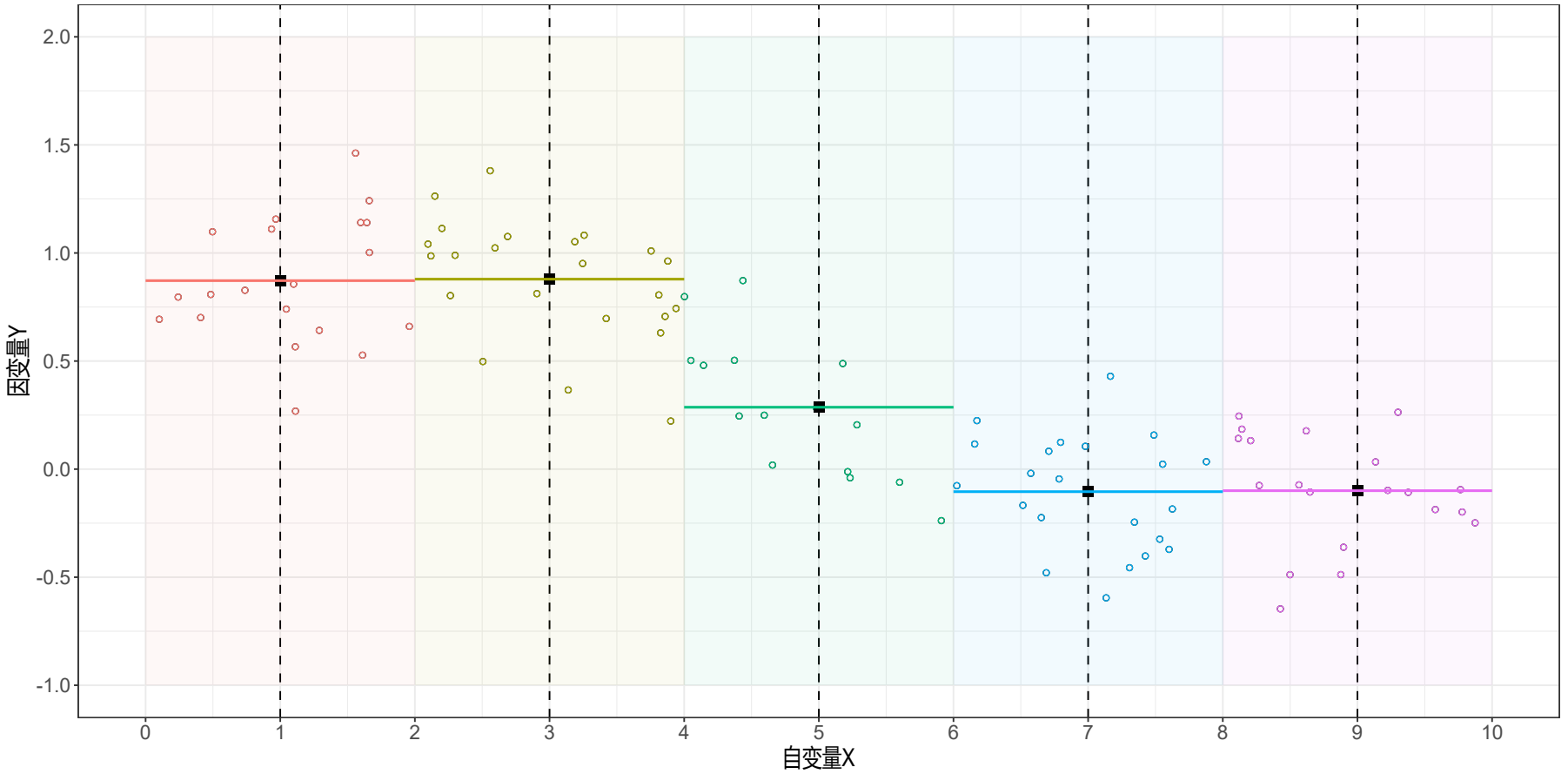
（示例）箱组均值估计：图形表达2/3

- 根据箱组均值计算值，我们展示在散点图中：



（示例）箱组均值估计：图形表达3/3

- 简单地，可将**箱组均值**作为对这一箱组**条件期望函数**CEF的近似：



1.2 滚动箱组均值估计：定义及表达式

滚动箱组均值估（The rolling binned means estimator）：以系列数值x为中心，以h为谱宽，**滚动**构建一系列箱组（箱组会有重叠），并分别计算出系列箱组的均值。

$$\hat{m}(x)=\frac{\sum_{i=1}^{n}1\{|X_i-x|\leq h\}\cdot Y_i}{\sum_{i=1}^{n}1\{|X_i-x|\leq h\}}$$

- 我们后面马上会介绍，**滚动箱组均值估**实际上是一类特殊的**核估计**（kernel）情形，具体为Nadaraya-Watson 矩形和估计（NW rectangular kernel estimator）。

1.2 滚动箱组均值估计：操作过程

滚动箱组均值估计（Rolling Binned Estimator）的操作步骤如下：

- 根据计算点 $X = x_j$，按照特定谱宽 h，划分出若干**箱组**（bins）（箱组会有重叠）：$\{b_1, b_2, \cdots, b_q\}$

$$b_j = [x_j - h, x_j + h]$$

- 根据样本数据集，以及 X_i的实际情况，确定数据对 $\{X_i, Y_i\}$的箱组归属：

$$1\{|X_i - x_j| \leq h\}$$

- 最后计算不同箱组的 Y_i的均值 $\widehat{m}(x_{b_j}), j \in (1, 2, \cdots, q)$：

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} 1\{|X_i - x| \leq h\} \cdot Y_i}{\sum_{i=1}^{n} 1\{|X_i - x| \leq h\}}$$

（示例）滚动箱组均值估计：设定箱组划分规则

下面我们分别设定如下**箱组**划分规则：

- 设定箱组取值中心点
  $(X = x_i), x_i \in (0.00, 0.01, 0.02, 0.03, 0.04, 0.05, \cdots, 9.95, 9.96, 9.97, 9.98, 9.99, 10.00)$，
  以及谱宽[a] $h = 1$
- 那么，可以得到箱组区块bins=[-1,1)、[-0.99,1.01)、[-0.98,1.02) . . .[8.98,10.98)、[8.99,10.99)、[9,11)，箱组数1001。

[a] 此时滚动箱组均值估计等价于谱宽 $h = 1$的**矩形核函数**（rectangular kernel）估计

（示例）滚动箱组均值估计：数据计算表

利用箱组均值估计公式，我们可以计算得到不同箱组的均值估计：

x	bins	sum_ky	sum_k	m1
0.00	[-1,1)	7.1916	8	0.8990
0.01	[-0.99,1.01)	7.1916	8	0.8990
0.02	[-0.98,1.02)	7.1916	8	0.8990
0.03	[-0.97,1.03)	7.1916	8	0.8990
0.04	[-0.96,1.04)	7.1916	8	0.8990

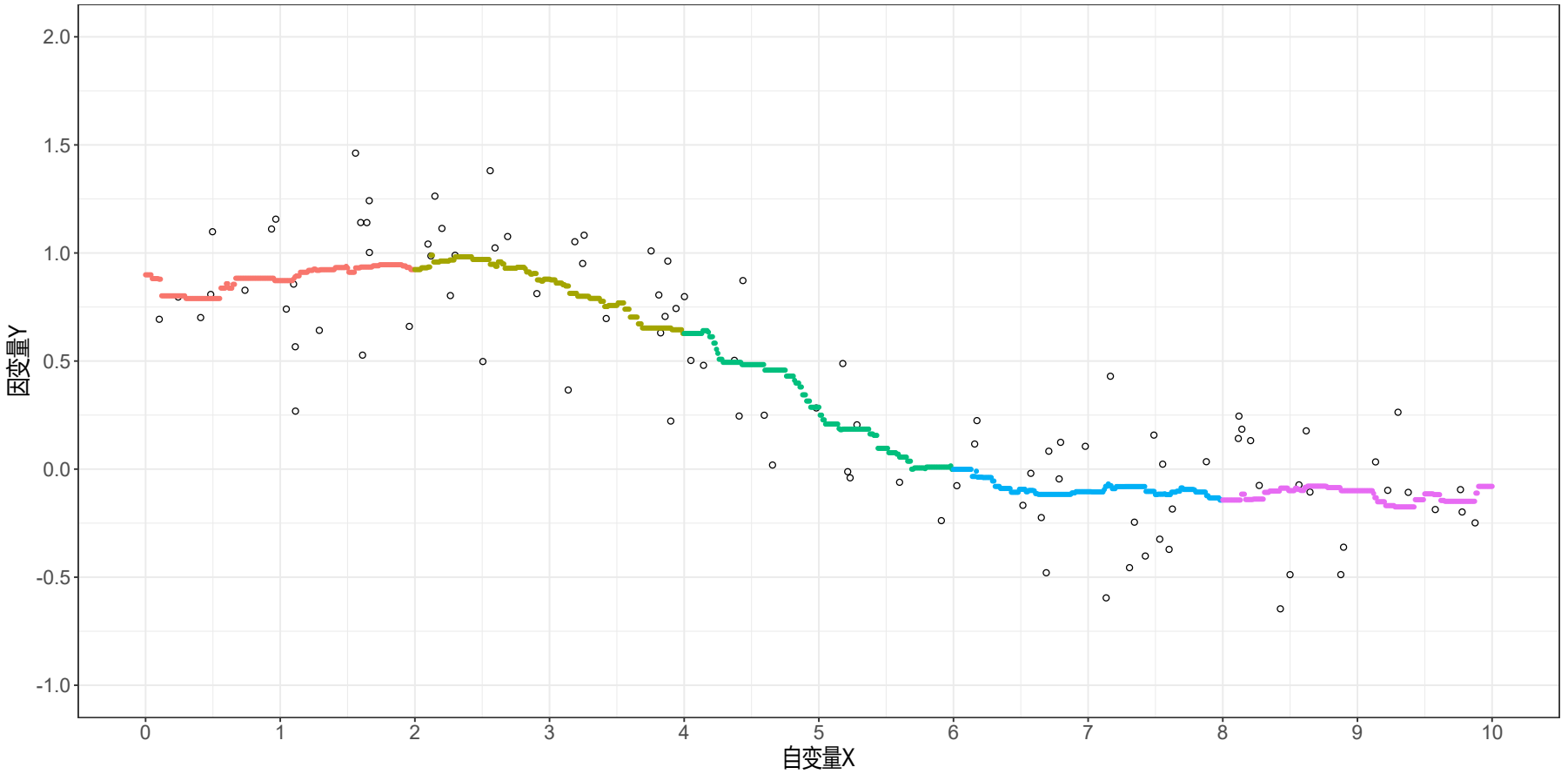
Showing 1 to 5 of 1,001 entries

Previous 1 2 3 4 5 … 201 Next

- 箱组内因变量观测值的求和 $\texttt{sum\_ky} = \sum_{i=1}^{n} 1\{|X_i - x| \le h\} \cdot Y_i$
- 箱组的样本数 $\texttt{sum\_k} = \sum_{i=1}^{n} 1\{|X_i - x| \le h\}$
- 箱组的均值估计 $\texttt{m1} = \widehat{m}(x_j), j \in (1, 2, \cdots, 1001)$

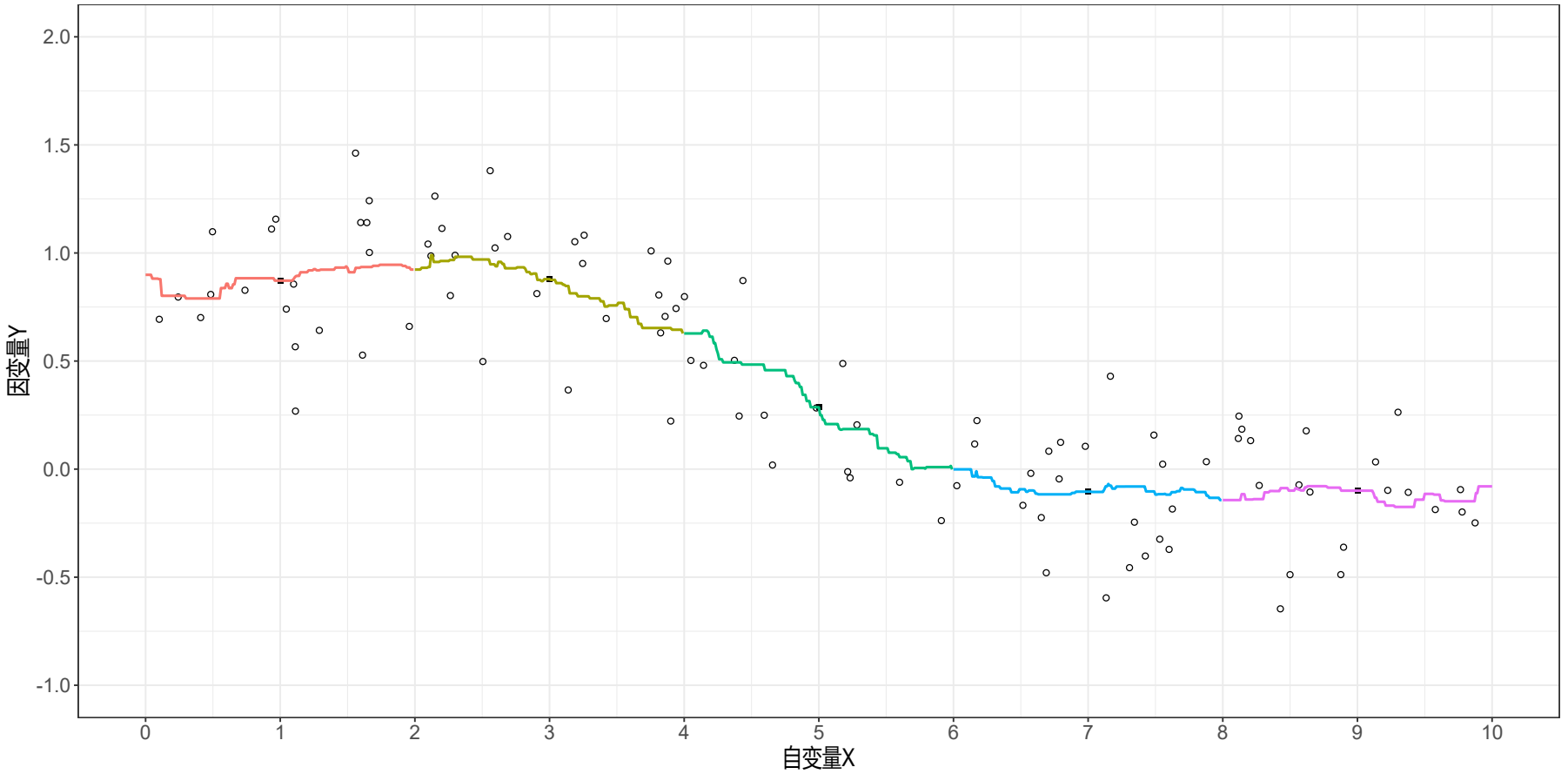
（示例）滚动箱组均值估计：图形表达

- 根据箱组均值计算值，我们展示在散点图中：



（示例）滚动箱组均值估计：图形表达

- 同样地，可将**箱组均值**作为对这一箱组**条件期望函数**CEF的近似：



上述两种箱组均值估计的公式中：

$$\hat{m}(x)=\frac{\sum_{i=1}^{n}1\{|X_i-x|\leq h\}\cdot Y_i}{\sum_{i=1}^{n}1\{|X_i-x|\leq h\}}$$

- 对条件期望的估计 $\widehat{m}(x)$结果都呈现一定的**锯齿**形态（jagged），也即估计结果不太**平滑**（smoothed）。

1.3 核估计：回顾与思考

思考：

- 为什么估计结果会呈现不太**平滑**的**锯齿**形态？
- 能不能让估计结果更加平滑呢？

回答：

- 问题的关键在于上面的估计公式使用了箱组**指示**函数（可视作为权重）$\sum_{i=1}^{n} 1\{|X_i - x| \leq h\}$，而这个**权重函数**本身是**跳跃的**。
- 有没有一种办法能够基于平滑的**权重函数**来计算CEF的估计值呢？

1.3 核估计：概念

核估计（kernel estimator）：基于多种类型的**核函数**（kernel function）作为**权重函数**——可以是**连续的**，也可以是**跳跃的**——来估计条件期望函数 $\widehat{m}(x)$的一种估计方法。

$$\hat{m}_{\mathrm{nw}}(x)=\frac{\sum_{i=1}^{n} K\left(\frac{X_i-x}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{X_i-x}{h}\right)}$$

- 其中： $K(u)$为核函数（kernel function）

1.3 核估计：特点

$$\hat{m}_{\mathrm{nw}}(x)=\frac{\sum_{i=1}^{n} K\left(\frac{X_i-x}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{X_i-x}{h}\right)}$$

- 尽管这里使用了核函数来计算权重，但是本质上上述公式还是利用了**箱组估计**的方法，也即把箱组计算值作为这一箱组的**代表值**。
- 以上估计方法也被称为**局部常数估计**（local constant estimator）或**NW估计**（Nadaraya-Watson estimator）
- 容易发现前述**箱组均值估计**和**滚动箱组均值估计**都是**局部常数估计**（local constant estimator）的两个特例。

1.3 核估计：核函数定义

定义：若满足如下条件，则可称之为**核函数**（Kernel function）$K(u)$

- $0 \leq K(u) \leq \bar{K} < \infty$,
- $K(u) = K(-u)$,
- $\int_{-\infty}^{\infty} K(u)du = 1$,
- 对所有的正整数 r都有 $\int_{-\infty}^{\infty} |u|^r K(u)du < \infty$

定义：正规化核函数（normalized kernel function）需满足

$$\int_{-\infty}^{\infty} u^2 K(u)du = 1$$

- **核函数**本质上是一种**边界约束的**概率密度函数（bounded pdf）
- **核函数**是原点对称的，且核函数为非负数（因此可用于权重）

1.3 核估计：常见的正规化核函数1/2

- **矩形核函数**（Rectangular Kernel），$R_K = \frac{1}{2\sqrt{3}}$

$$K(u) = \begin{cases} \frac{1}{2\sqrt{3}} & \text{if } |u| < \sqrt{3} \\ 0 & \text{otherwise} \end{cases}$$

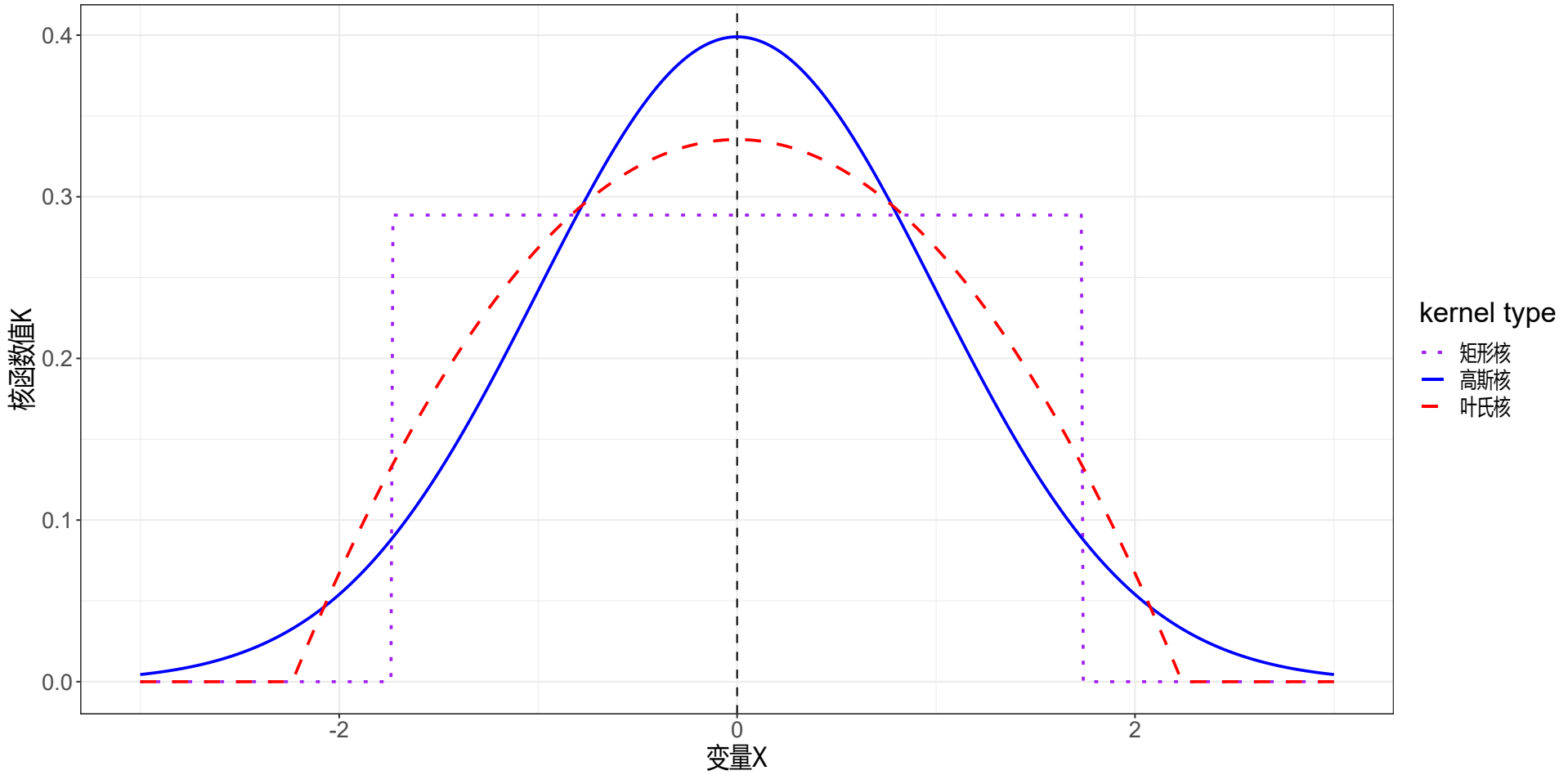
- **高斯核函数**（Gaussian Kernel），$R_K = \frac{1}{2\sqrt{\pi}}$

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$$

- **叶氏核函数**（Epanechnikov Kernel），$R_K = \frac{3\sqrt{5}}{25}$

$$K(u) = \begin{cases} \frac{3}{4\sqrt{5}}\left(1 - \frac{u^2}{5}\right) & \text{if } |u| < \sqrt{5} \\ 0 & \text{otherwise} \end{cases}$$

（示例）核函数类型I



1.3 核估计：常见的正规化核函数2/2

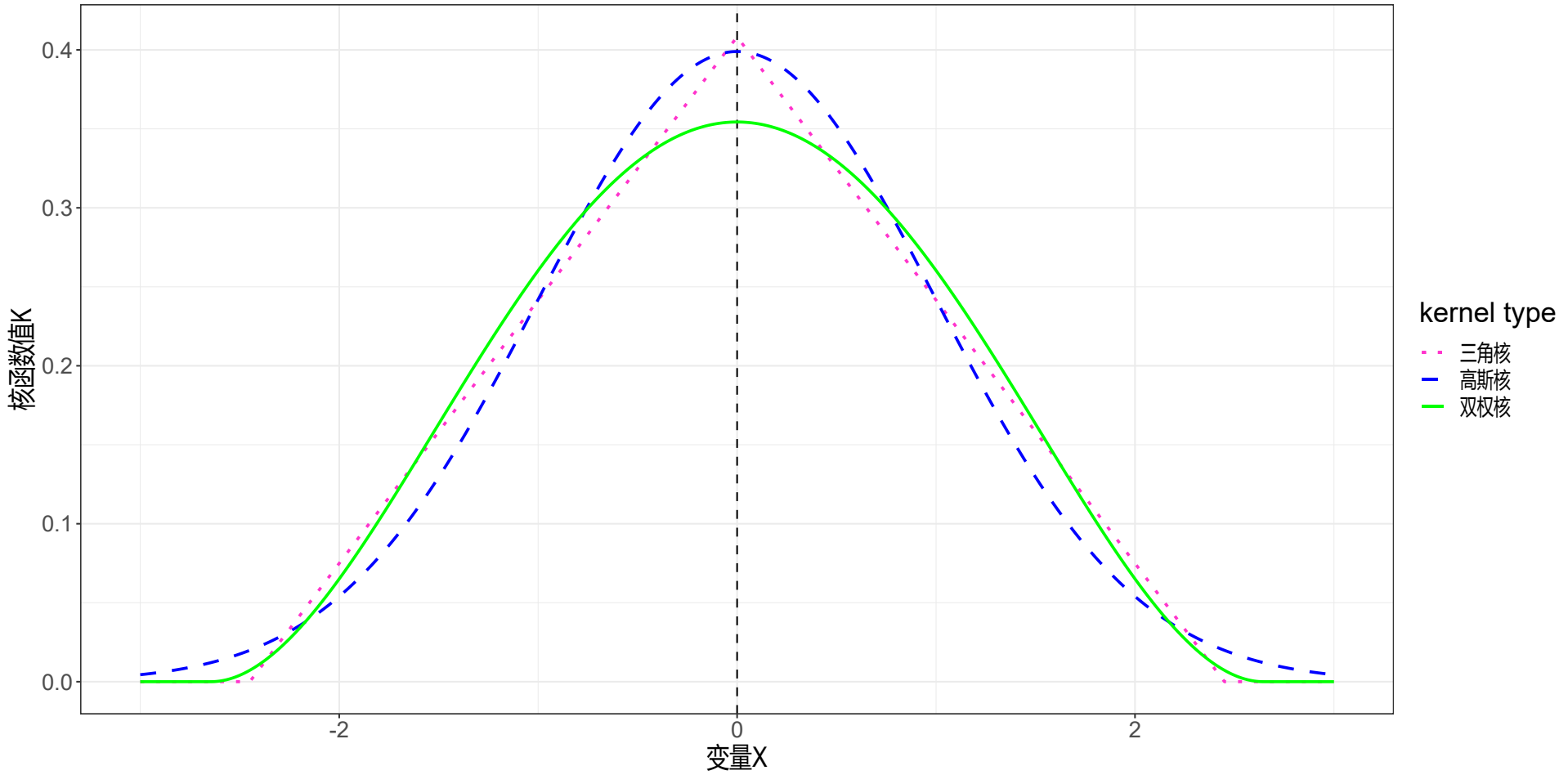
- **三角核函数**（Triangular Kernel），$R_K = \frac{\sqrt{3}}{9}$

$$K(u) = \begin{cases} \frac{1}{\sqrt{6}}\left(1 - \frac{|u|}{\sqrt{6}}\right) & \text{if } |u| < \sqrt{6} \\ 0 & \text{otherwise} \end{cases}$$

- **双权核函数**（Biweight Kernel），$R_K = \frac{5\sqrt{7}}{49}$

$$K(u) = \begin{cases} \frac{15}{16\sqrt{7}}\left(1 - \frac{u^2}{7}\right) & \text{if } |u| < \sqrt{7} \\ 0 & \text{otherwise} \end{cases}$$

（示例）核函数类型（续）



1.3 核估计：操作过程

NW核估计（NW Estimator）的操作步骤如下：

- 根据计算点 $X = x_j$，按照特定谱宽 h，划分出若干**箱组**（bins）（箱组会有重叠）：$\{b_1, b_2, \cdots, b_q\}$

$$b_j = [x_j - h, x_j + h]$$

- 根据样本数据集，以及 X_i的实际情况，计算特定**核函数化**的权重值：

$$K\left(\frac{X_i - x_j}{h}\right)$$

- 最后计算不同箱组的 Y_i的均值 $\widehat{m}(x_{b_j}), j \in (1, 2, \cdots, q)$：

$$\hat{m}_{\mathrm{nw}}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)}$$

（示例）核估计：设定箱组划分规则

下面我们分别设定如下**箱组**划分规则：

- 设定箱组取值中心点

  $(X = x_i), x_i \in (0.00, 0.01, 0.02, 0.03, 0.04, 0.05, \cdots, 9.95, 9.96, 9.97, 9.98, 9.99, 10.00)$
- 确定核函数类型为**高斯核函数**[a]，谱宽设定为 $h = 1/\sqrt{3}$。
- 那么，可以得到箱组区块bins=[-1,1)、[-0.99,1.01)、[-0.98,1.02) . . .[8.98,10.98)、[8.99,10.99)、[9,11)，箱组数1001。

（示例）核估计：数据计算表

利用箱组均值核函数估计公式，我们可以计算得到不同箱组的均值估计：

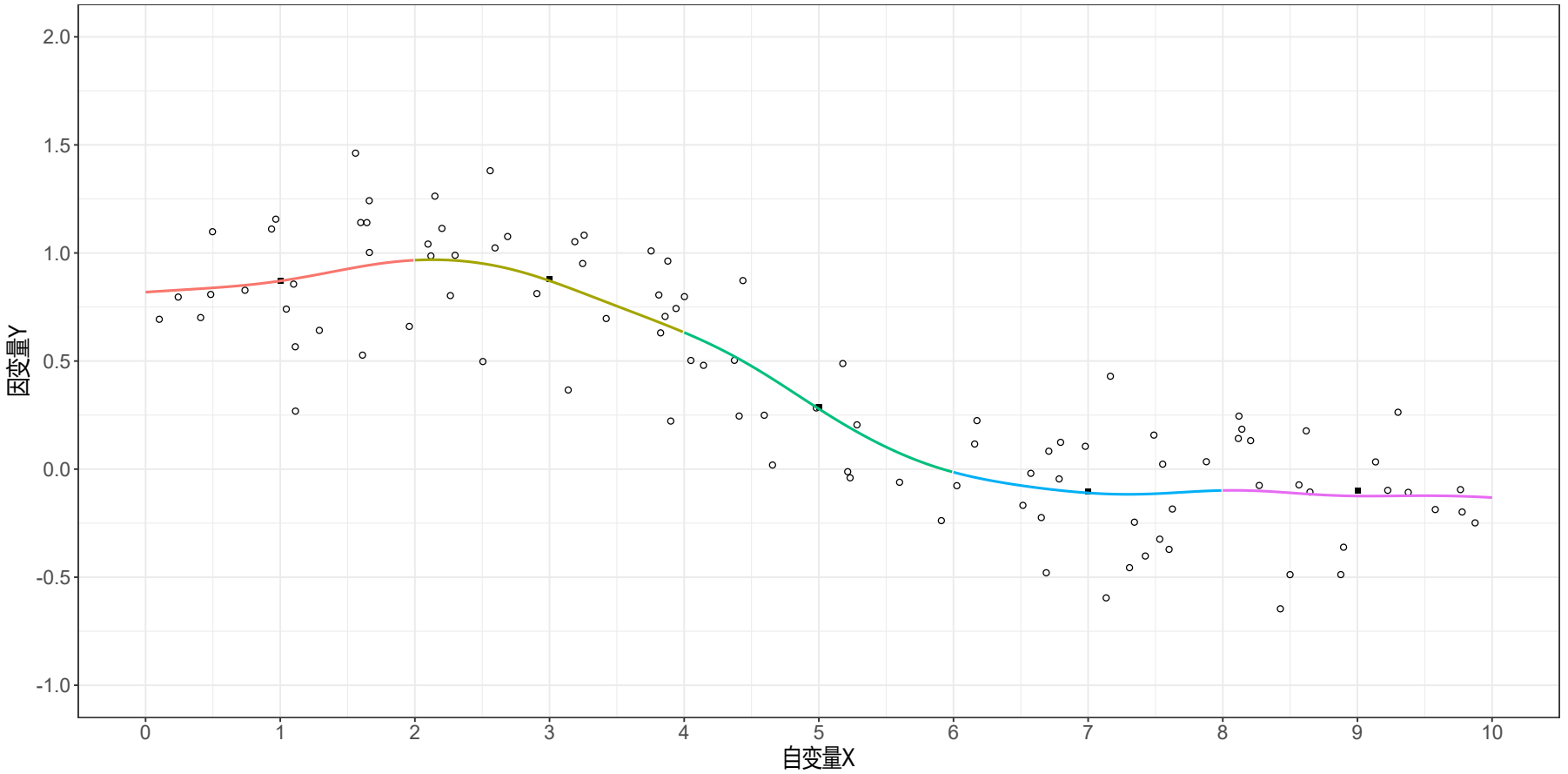
x	bins	sum_ky	sum_k	m2
0.00	[-1,1)	1.9298	2.3571	0.8187
0.01	[-0.99,1.01)	1.9621	2.3952	0.8191
0.02	[-0.98,1.02)	1.9945	2.4336	0.8196
0.03	[-0.97,1.03)	2.0271	2.4722	0.8200
0.04	[-0.96,1.04)	2.0599	2.5109	0.8204

Showing 1 to 5 of 1,001 entries

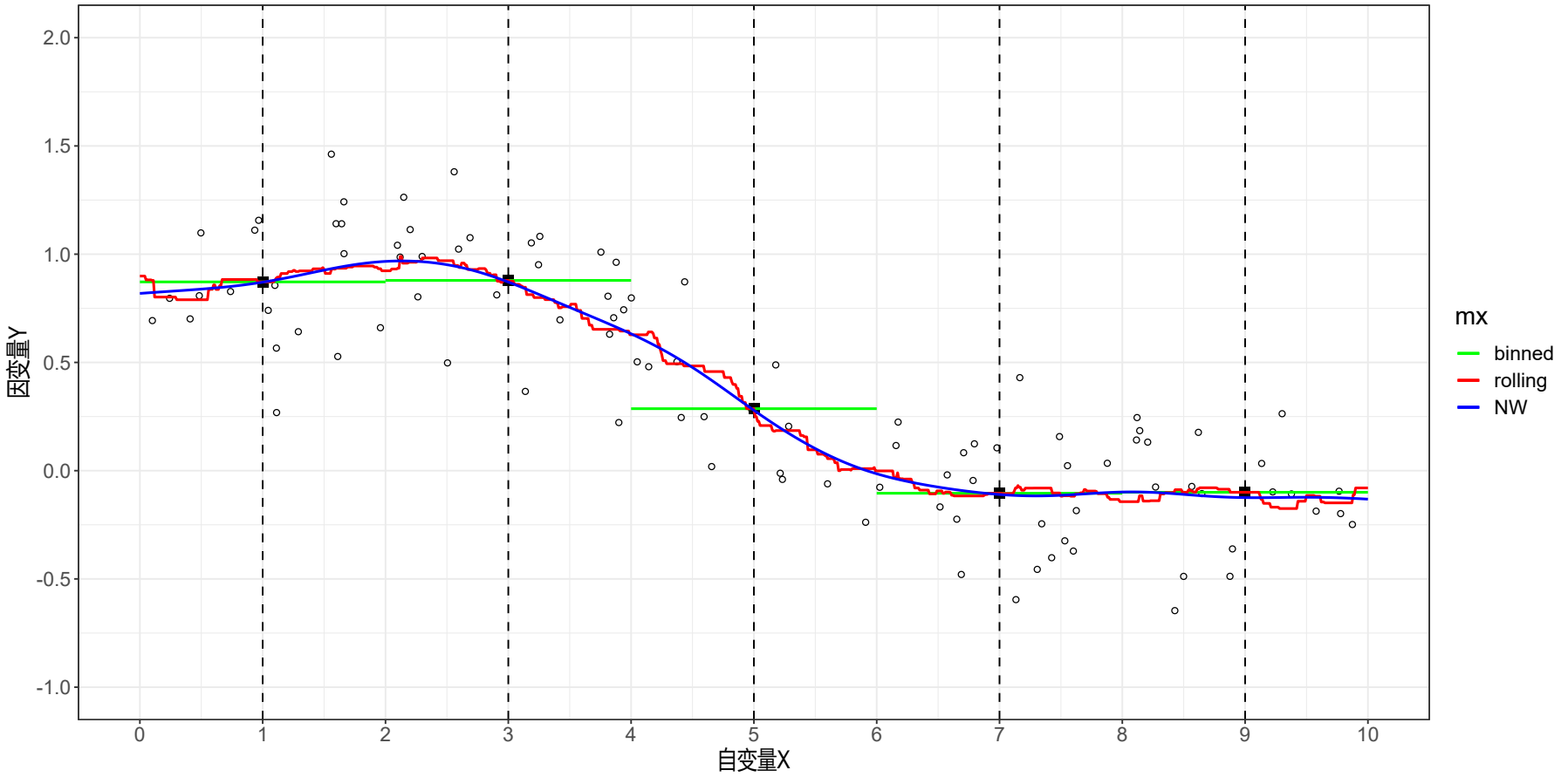
Previous 1 2 3 4 5 … 201 Next

（示例）核估计：图形表达

- 同样地，可将**箱组均值**作为对这一箱组**条件期望函数**CEF的近似：



（示例）CEF估计：三种估计方法的图形比较



- **NW核估计**方法相比**箱组均值估计**和**滚动箱组均值估计**要更加平滑。

2.局部回归(Local Regression)

2.1 箱组线性回归(Binned Regression)

2.2 滚动箱组回归(Rolling Regression)

2.3 局部线性回归(Local Linear Regression)

2.4 局部多项式回归(Local Polynomial Regression)

引子：从NW估计量说起

在一个局部区域（如 $X=x$的领域内），对 $m(x)$的Nadaraya-Watson (NW)估计量将表现为**常函数**（constant function）形态，此时也称为**局部常数估计量**（Local constant estimator）。

- 此时，Nadaraya-Watson (NW)估计量为一种局部近似，也即当在局部渐近取值 $X \simeq x$时，$m(X) \simeq m(x)$

$$Y = m(X) + e \simeq m(x) + e$$

上述模型可以视作为回归方程，我们需要估计得到 $\widehat{m}(x)$，也即：

$$\widehat{m}_{\text{nw}}(x) = \underset{m}{\text{argmin}} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)(Y_i - m)^2$$

- 本质上，以上就是一个 Y对截距项的**加权回归**（Weighted regression）估计量。

引子：CEF的回归估计方法

事实上，基于**箱组**原理的CEF估计方法，我们都可以使用回归方法来进行估计，具体包括：

- 箱组线性回归（binned regression）：箱组内直接使用OLS回归估计
- 滚动箱组核回归（NW regression）：箱组内直接使用OLS回归估计
- 局部线性回归（Local linear regression）：箱组内使用加权OLS回归估计
- 局部多项式回归（Local Polynomial regression）：箱组内使用多项式的加权OLS回归估计

2.1 箱组线性回归：原理

箱组线性回归（binned regression）：箱组内直接使用OLS回归估计

$$Y = m(X) + e \simeq m(x) + e$$

- 上述模型可以视作为回归方程，我们需要估计得到 $\widehat{m}(x)$，也即：

$$\widehat{m}_{\text{bin}}(x) = \underset{m}{\operatorname{argmin}} \sum_{i=1}^{n} \mathbf{1}\{|X_i - x| \leq h\} \cdot (Y_i - m)^2$$

- 因为在箱组内的权重都是一样的，因此可以直接在**箱组内**的子样本数据里使用OLS回归估计得到

$$\widehat{m}(x) = \widehat{Y}_i \quad \text{(OLS)}$$

2.1 箱组线性回归：操作步骤

箱组线性回归（Binned regression）的操作步骤如下：

- 根据计算点 $X = x_j$，按照特定谱宽 h，划分出若干**箱组**（bins）：$\{b_1, b_2, \cdots, b_q\}$

$$b_j = [x_j - h, x_j + h]$$

- 根据样本数据集，以及 X_i的实际情况，确定数据对 $\{X_i, Y_i\}$的箱组归属（子样本）：

$$1\{|X_i - x_j| \leq h\}$$

- 然后采用OLS方法计算不同箱组的 Y_i的拟合值 $\hat{Y}_i = \widehat{m}(x) \simeq m(X)$

$$\hat{Y}_i = \hat{\alpha}_0 + \hat{\alpha}_1 X_i$$

- 最后，我们可以给定任意值 x_i得到预测的 $\widehat{m}(x_i) = \hat{Y}_i | x_i$

（示例）箱组回归估计：设定箱组划分规则

下面我们分别设定如下**箱组**划分规则：

- 设定箱组取值中心点 $(X = x_i), x_i \in (1, 3, 5, 7, 9)$，以及谱宽 $h = 1$
- 然后得到箱组区块bins=[0,2)、[2,4)、[4,6)、[6,8)、[8,10]，箱组数为5。
- 箱组内的线性回归模型为

$$\hat{Y}_i = \hat{\alpha}_0 + \hat{\alpha}_1 X_i$$

- 基于设定的 $x_i \in (0.00, 0.01, 0.02, 0.03, 0.04, 0.05, \cdots, 9.95, 9.96, 9.97, 9.98, 9.99, 10.00)$ 计算预测值 $\widehat{m}(x_i) = \widehat{Y}_i | x_i$，共有 $N = 1001$ 个

（示例）箱组线性回归：数据计算表

$$\hat{Y}_i = \hat{\alpha}_0 + \hat{\alpha}_1 X_i$$

- 对五个箱组的区块下，我们依次对子样本数据对 (Y_i, X_i)进行线性OLS拟合，分别得到截距和斜率：

par	bd1	bd2	bd3	bd4	bd5
intercept	0.75	1.45	2.32	0.74	0.72
slope	0.11	-0.19	-0.42	-0.12	-0.09

说明： `bd1`、`bd2`等分别代表五个箱组。

（示例）箱组线性回归：数据计算表

利用箱组均值估计公式，我们可以计算得到不同箱组的均值估计：

bd	my	bins	x	m0
[0,2)	0.8622	[-1.00,1.00)	0.0000	0.7489
[0,2)	0.8622	[-0.99,1.01)	0.0100	0.7500
[0,2)	0.8622	[-0.98,1.02)	0.0200	0.7511
[0,2)	0.8622	[-0.97,1.03)	0.0300	0.7523
[0,2)	0.8622	[-0.96,1.04)	0.0400	0.7534
[0,2)	0.8622	[-0.95,1.05)	0.0500	0.7545

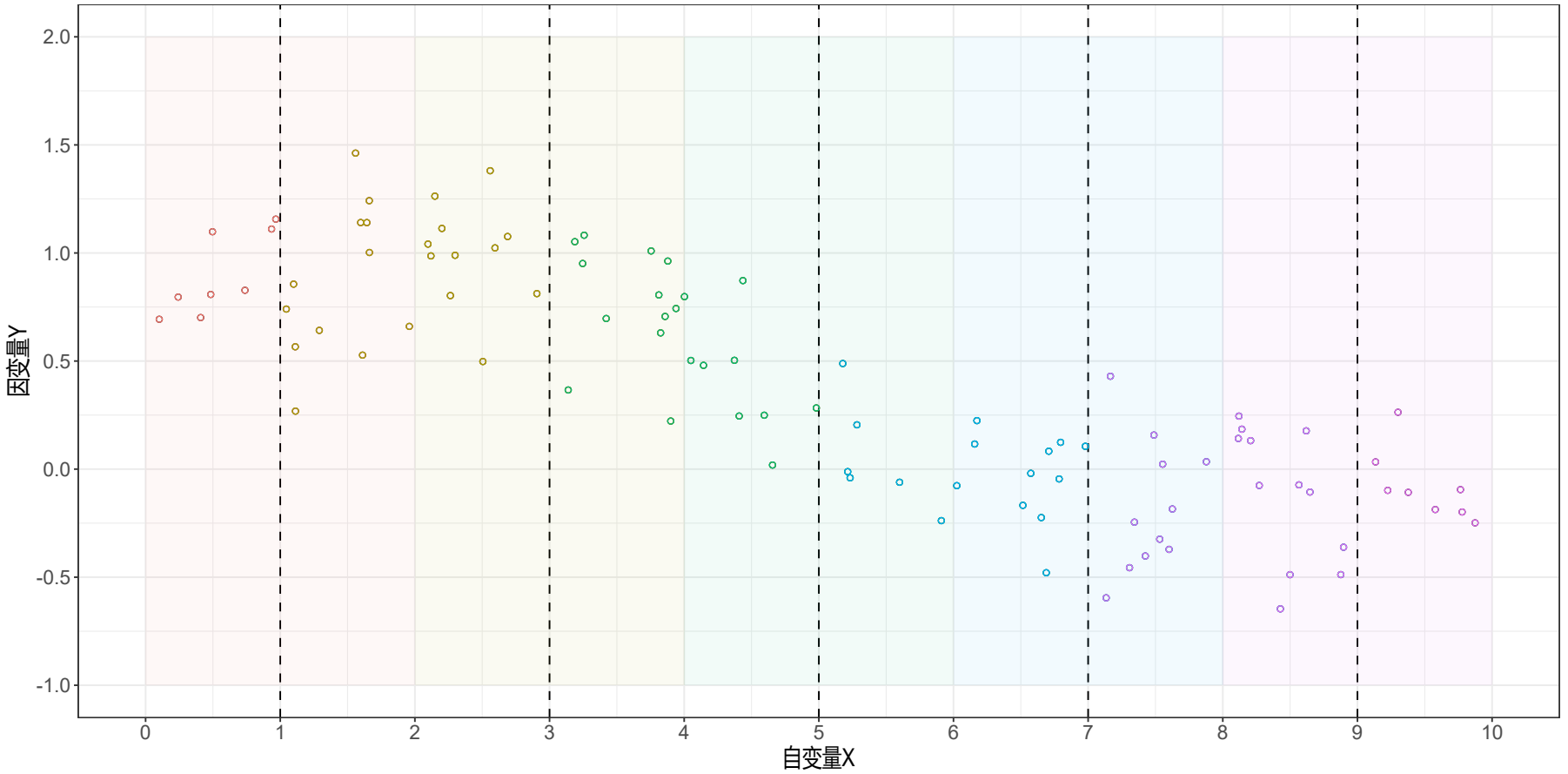
Showing 1 to 6 of 1,001 entries

Previous 1 2 3 4 5 … 167 Next

- 基于设定的`x`表示 $x_i \in (0.00, 0.01, 0.02, 0.03, 0.04, 0.05, \cdots)$，共有 $N = 1001$个
- 箱组回归估计`m0` $\widehat{m}(x_i) = \widehat{Y}_i | x_i$

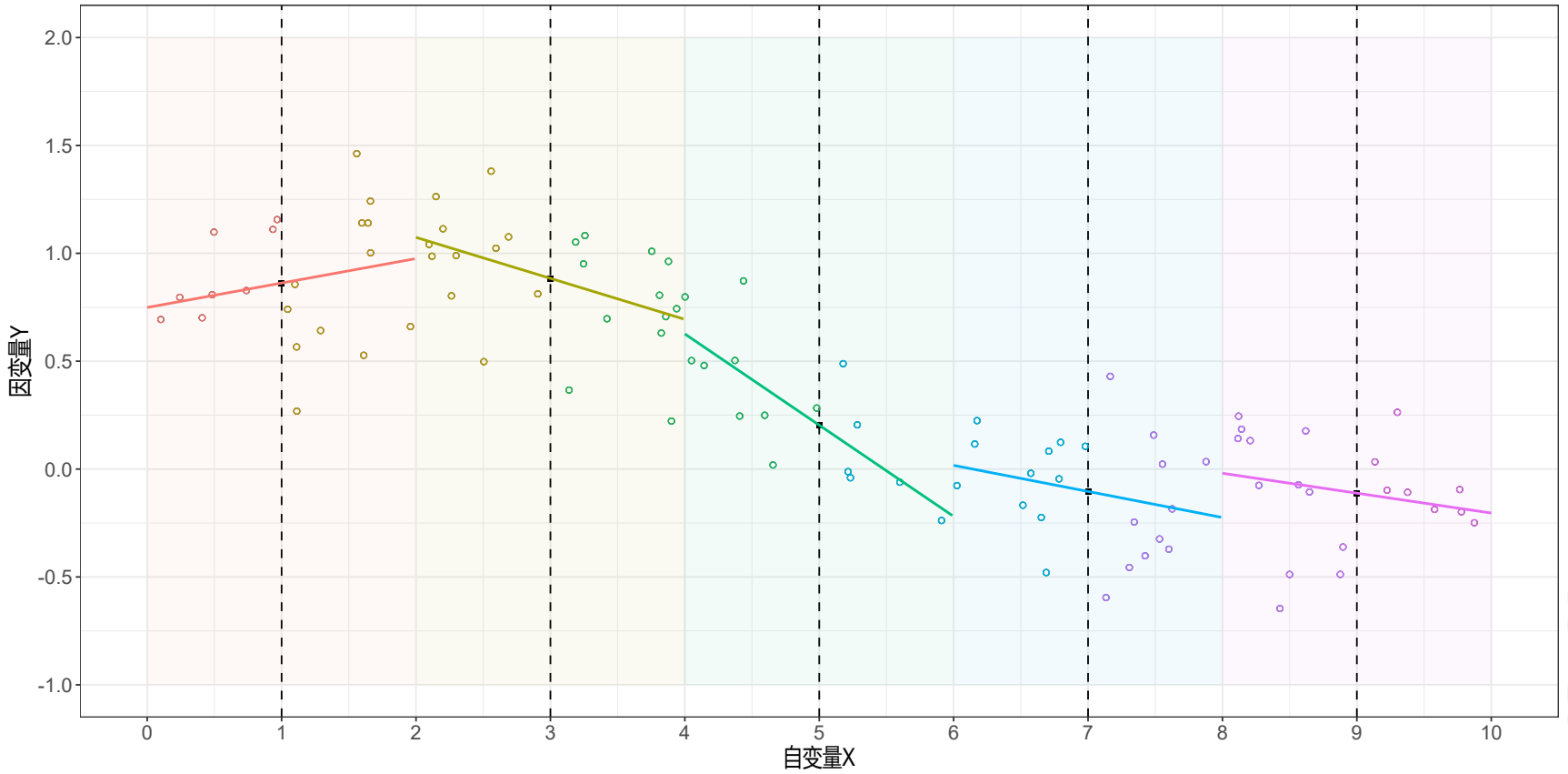
（示例）箱组线性回归：图形表达1/2

- 如前，我们设定区隔了5个箱组



（示例）箱组线性回归：图形表达2/2

- 根据前面计算表的1001个拟合数据对 $(x_i, \widehat{m}(x_i))$，我们可以得到箱组线性回归估计结果：



2.2 滚动箱组回归：原理

滚动箱组线性回归（rolling regression）：简单地，是通过构建滚动箱组（可能出现箱组重叠状态），然后再对箱组内样本数据直接使用OLS回归估计。此外，容易发现它实际上就是等价于**矩形核函数**的**NW回归**（$h=1$）。

- 此时，Nadaraya-Watson (NW)估计量为一种局部近似，也即当在局部渐近取值 $X \simeq x$时，$m(X) \simeq m(x)$

$$Y = m(X) + e \simeq m(x) + e$$

上述模型可以视作为回归方程，我们需要估计得到 $\widehat{m}(x)$，也即：

$$\widehat{m}_{\text{nw}}(x) = \underset{m}{\text{argmin}} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)(Y_i - m)^2$$

- 本质上，以上就是一个 Y对截距项的**加权回归**（Weighted regression）估计量。

2.2 滚动箱组回归：操作步骤

滚动箱组线性回归（Rolling regression）的操作步骤如下：

- 根据计算点 $X=x_j$，按照特定谱宽 h，划分出若干**箱组**（bins）（箱组会有重叠）：$\{b_1,b_2,\cdots,b_q\}$，其中 $b_j=[x_j-h,x_j+h]$。
- 根据样本数据集 X_i的实际情况，计算**矩形核函数**的权重值（ $h=1$）：

$$K\left(\frac{X_i-x_j}{h}\right)$$

- 然后采用**加权OLS**方法计算不同箱组下仅含截距项的回归模型 的截距系数 $\widehat{m}(x)\simeq m(X)$

$$Y=m(x)+e$$

（示例）滚动箱组回归：设定箱组划分规则

下面我们分别设定如下**箱组**划分规则：

- 设定箱组取值中心点
  $(X = x_i), x_i \in (0.00, 0.01, 0.02, 0.03, 0.04, 0.05, \cdots, 9.95, 9.96, 9.97, 9.98, 9.99, 10.00)$
- 确定核函数类型为**矩形核函数**[a]，谱宽设定为 $h = 1$。
- 那么，可以得到箱组区块bins=[-1,1)、[-0.99,1.01)、[-0.98,1.02) . . .[8.98,10.98)、[8.99,10.99)、[9,11)，箱组数1001。
- 采用加权OLS方法进行拟合，直接计算得到 $\widehat{m}(x)$

（示例）滚动箱组回归：数据计算表

利用箱组均值估计公式，我们可以计算得到不同箱组的均值估计：

bd	bins	x	m1
[0,2)	[-1.00,1.00)	0.0000	0.6397
[0,2)	[-0.99,1.01)	0.0100	0.6444
[0,2)	[-0.98,1.02)	0.0200	0.6492
[0,2)	[-0.97,1.03)	0.0300	0.6539
[0,2)	[-0.96,1.04)	0.0400	0.6587
[0,2)	[-0.95,1.05)	0.0500	0.7266

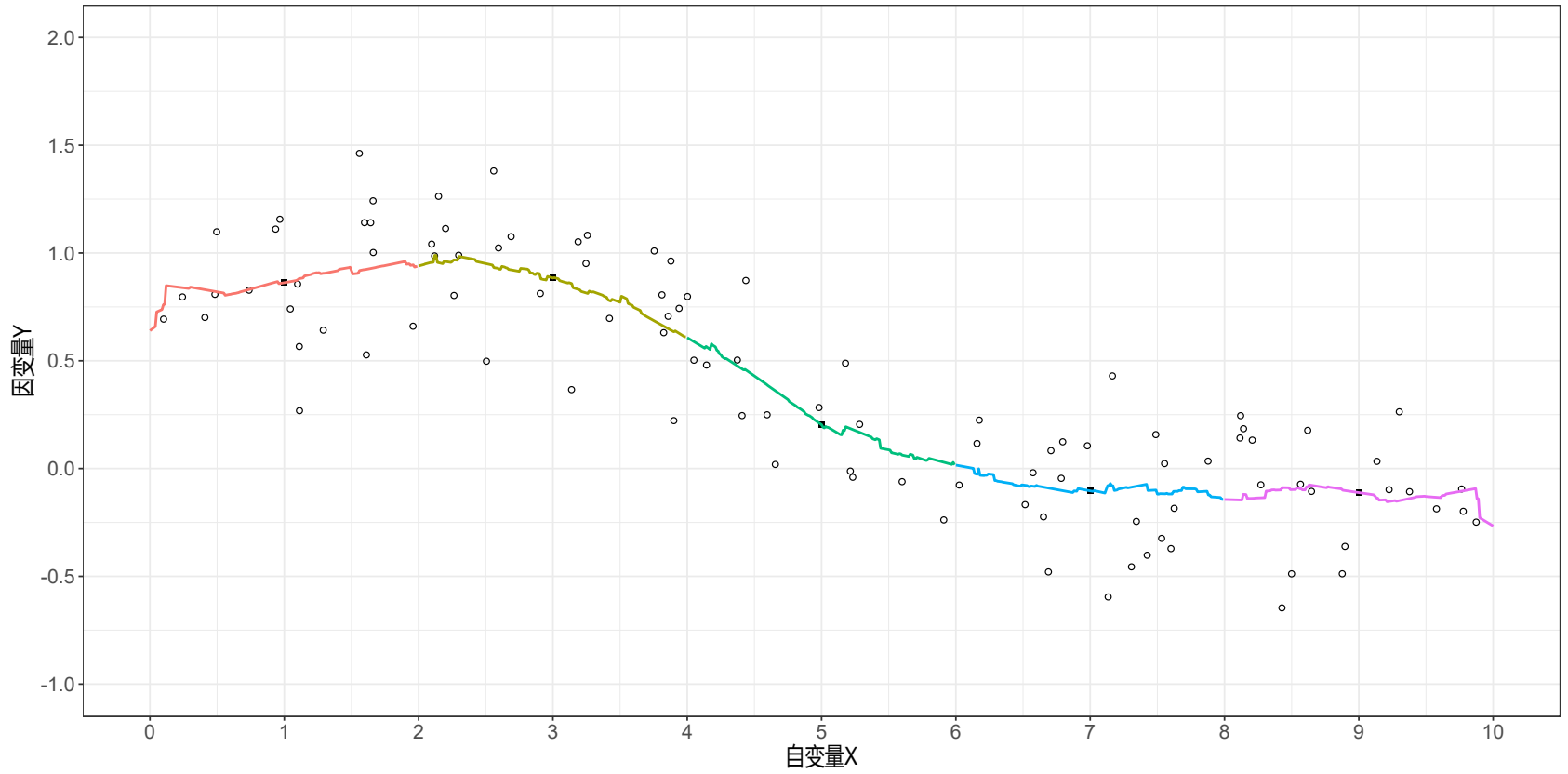
Showing 1 to 6 of 1,001 entries

Previous 1 2 3 4 5 … 167 Next

- 基于设定的`x`表示 $x_i \in (0.00, 0.01, 0.02, 0.03, 0.04, 0.05, \cdots)$，共有 $N = 1001$个
- 箱组回归估计`m1` $= \widehat{m}(x_i)$

（示例）滚动箱组回归：图形表达

- 根据前面计算表的1001个拟合数据对 $(x_i, \widehat{m}(x_i))$，我们可以得到滚动箱组线性回归估计结果：



2.3 局部线性回归：渐进近似的另一种选择

> **回顾**：Nadaraya-Watson (NW)估计量的局部近似选择为 $m(X) \simeq m(x)$

局部线性回归（Local linear regression, LLR）方法则选择了另一种局部线性近似，也即：

$$m(X) \simeq m(x) + m'(x)(X - x)$$

因此**局部线性**（LL）模型可以表达为：

$$\begin{aligned} Y &= m(X) + e \\ &\simeq m(x) + m'(x)(X - x) + e \\ &= \beta_0 + \beta_1 \cdot (X - x) + e \end{aligned}$$

- 以上模型可以视作为**一元线性回归模型**，其中 $\beta_0 = m(x); \beta_1 = m'(x)$。
- 我们的目标是估计得到 $\widehat{m}(x)$，也即上述模型的**截距项**。

2.3 局部线性回归：矩阵表达式

前述一元线性模型也可以表达为如下矩阵形式：

$$Y \simeq Z(X,x)'\boldsymbol{\beta}(x) + e$$

其中：

$$Z(X,x) = \begin{pmatrix} 1 \\ X-x \end{pmatrix}$$
$$\boldsymbol{\beta}(x) = \big(m(x), m'(x)\big)'$$

2.3 局部线性回归：最小化问题求解

对于前述渐近近似回归模型，最小化问题可以表达为：

$$\left\{\widehat{m}_{\mathrm{LL}}(x), \widehat{m}'_{\mathrm{LL}}(x)\right\} = \underset{\beta_0,\beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)\left(Y_i - \beta_0 - \beta_1\left(X_i - x\right)\right)^2$$

我们可以发现两个有意思的结论：

- 当谱宽趋近于无穷大时 $h \to \infty$，局部线性估计量将趋近于全样本OLS估计量 $\widehat{m}_{\mathrm{LL}}(x) \to \widehat{\beta}_0 + \widehat{\beta}_1 x$。因为，此时所有样本的权重都会相等。
- 局部线性估计量会同时得到在点 x处，条件期望函数CEF的估计量 $\widehat{m}(x)$，及其斜率 $\widehat{m}'(x)$。

2.3 局部线性回归：估计量

以**核函数**为权重，以及利用**加权最小二乘法**原理，可以证明局部线性回归的估计量为：

$$\begin{aligned}\widehat{\beta}_{\mathrm{LL}}(x) &= \left(\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Z_i(x) Z_i(x)'\right)^{-1} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Z_i(x) Y_i \\ &= \left(\boldsymbol{Z}'\boldsymbol{K}\boldsymbol{Z}\right)^{-1} \boldsymbol{Z}'\boldsymbol{K}\boldsymbol{Y}\end{aligned}$$

其中：

- $\boldsymbol{K} = \mathrm{diag}\{K\left(\left(X_1 - x\right)/h\right), \ldots, K\left(\left(X_n - x\right)/h\right)\}$
- $\boldsymbol{Z}$是 $Z_i(x)'$的堆栈（stacked）形态
- $\boldsymbol{Y}$是 $Y_i(x)'$的堆栈形态

2.3 局部线性回归：操作步骤

局部线性回归（Local Linear regression）的操作步骤如下：

- 根据计算点 $X=x_j$，按照特定谱宽 h，划分出若干**箱组**（bins）（箱组会有重叠）：$\{b_1,b_2,\cdots,b_q\}$，其中 $b_j=[x_j-h,x_j+h]$。
- 根据样本数据集 X_i的实际情况，计算**高斯核函数**的权重值（$h=1/\sqrt{3}$）：

$$K\left(\frac{X_i-x_j}{h}\right)$$

- 然后采用**加权OLS**方法计算不同箱组下一元回归模型的截距系数 $\widehat{m}(x)\simeq m(X)$

$$\begin{aligned} Y&=m(X)+e\\ &\simeq m(x)+m'(x)(X-x)+e \end{aligned}$$

（示例）局部线性回归：数据计算表

利用局部线性回归估计公式，我们可以计算得到不同箱组的估计：

bd	bins	x	m2
[0,2)	[-1.00,1.00)	0.0000	0.7639
[0,2)	[-0.99,1.01)	0.0100	0.7657
[0,2)	[-0.98,1.02)	0.0200	0.7674
[0,2)	[-0.97,1.03)	0.0300	0.7691
[0,2)	[-0.96,1.04)	0.0400	0.7707
[0,2)	[-0.95,1.05)	0.0500	0.7723

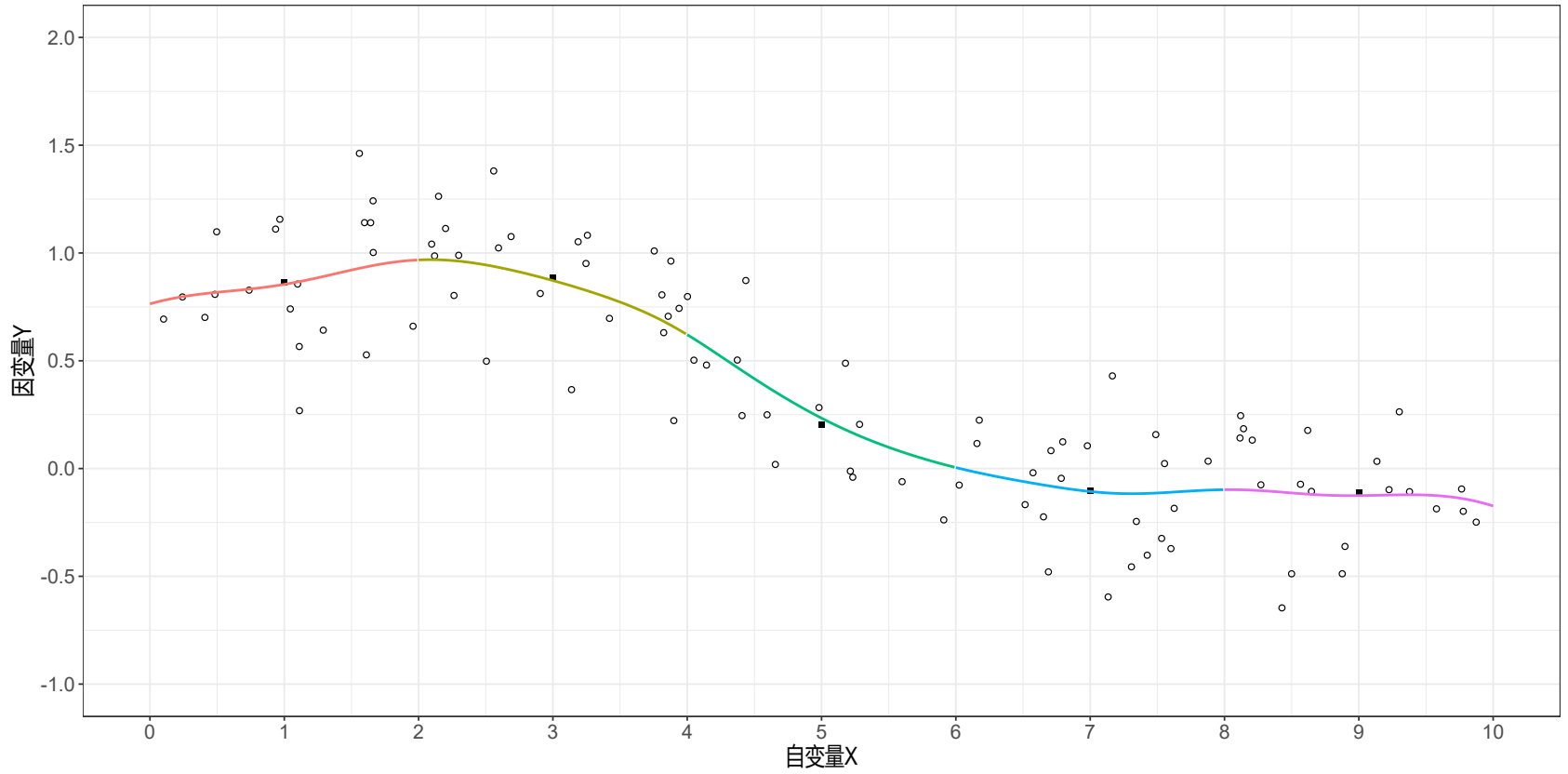
Showing 1 to 6 of 1,001 entries

Previous 1 2 3 4 5 … 167 Next

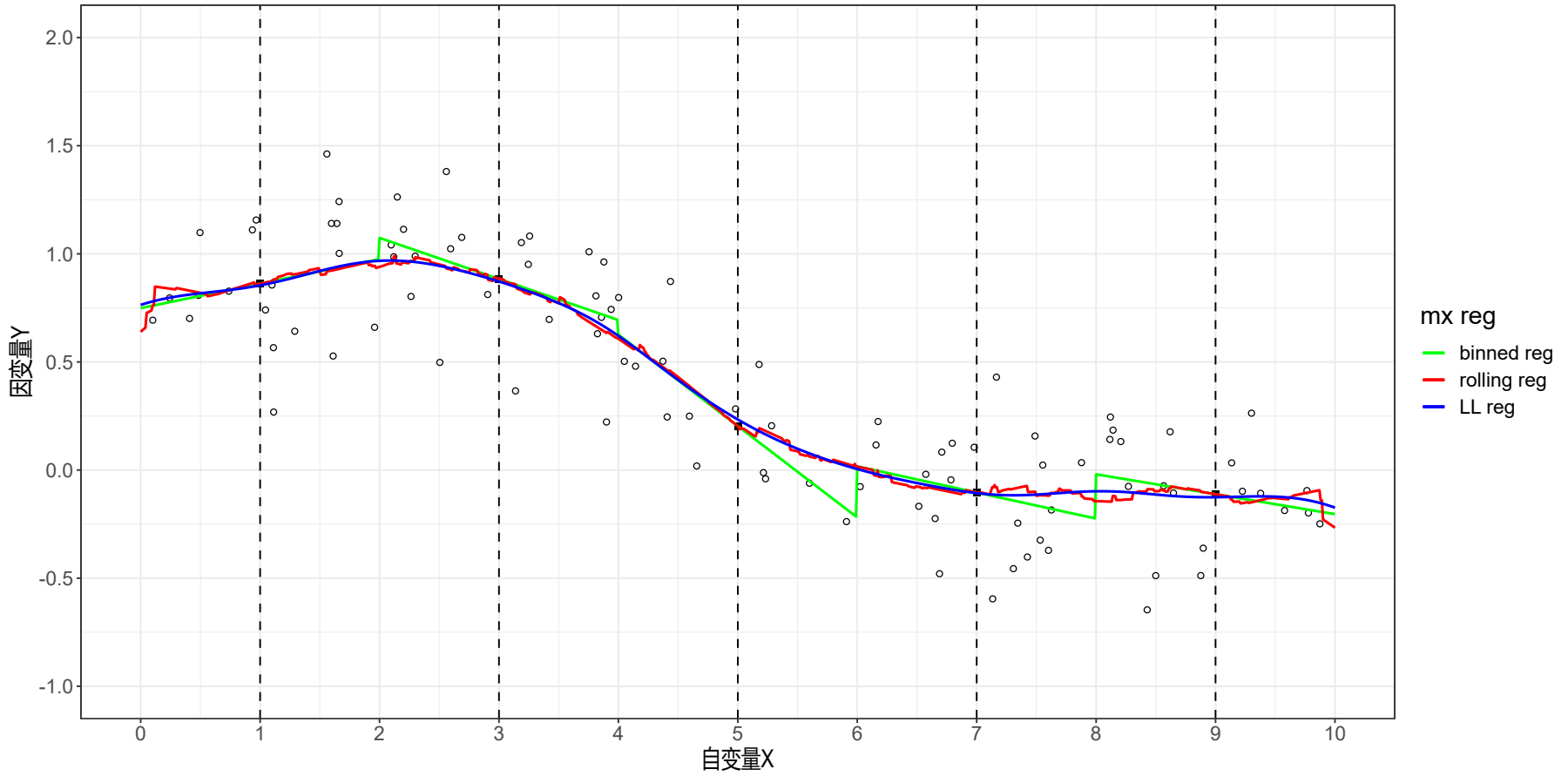
- 基于设定的`x`表示 $x_i \in (0.00, 0.01, 0.02, 0.03, 0.04, 0.05, \cdots)$，共有 $N = 1001$个
- 箱组回归估计`m2` $= \widehat{m}(x_i)$

（示例）局部线性回归估计：图形表达

- 根据前面计算表的1001个拟合数据对 $(x_i, \widehat{m}(x_i))$，我们可以得到局部线性回归估计结果：



（示例）CEF估计：三种回归估计方法的图形比较



- 尽管三种回归估计方法都较好地估计了真实CEF的趋势，但是**局部线性回归**LLR拟合方法要更平滑。

2.4 局部多项式回归：模型表达

局部多项式回归（Local Polynomial regression）模型可以表达为：

$$
\begin{aligned}
Y &= m(X) + e \\
&\simeq m(x) + m'(x)(X-x) + \cdots + m^{(p)}(x)\frac{(X-x)^p}{p!} + e \\
&= Z(X,x)'\beta(x) + e_i
\end{aligned}
$$

- 其中：

$$
Z(X,x) = \begin{pmatrix} 1 \\ X-x \\ \vdots \\ \frac{(X-x)^p}{p!} \end{pmatrix} \qquad \beta(x) = \begin{pmatrix} m(x) \\ m'(x) \\ \vdots \\ m^{(p)}(x) \end{pmatrix}
$$

2.4 局部多项式回归：估计量

同样地，以**核函数**为权重，可以证明**局部多项式回归**估计量的理论公式为：

$$\begin{aligned}\widehat{\beta}_{\mathrm{LP}}(x) &= \left(\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Z_i(x) Z_i(x)'\right)^{-1} \left(\sum_{i=1}^{n} K\left(\frac{Y_i - x}{h}\right) Z_i(x) Y_i\right) \\ &= \left(\boldsymbol{Z}'\boldsymbol{K}\boldsymbol{Z}\right)^{-1} \boldsymbol{Z}'\boldsymbol{K}\boldsymbol{Y}\end{aligned}$$

- 其中 $Z_i(x) = Z(X_i, x)$

2.4 局部多项式回归：特点

- 局部多项式回归LPR具有一般性：

  其中有两种特定情形：a）当 $p=0$时为Nadaraya-Watson回归估计；b）当 $p=1$时为局部线性回归估计（LL）

- 估计结果需要在阶数 p和局部平滑谱宽 h之间寻求平衡。

  一方面如果增加多项式阶数 p，可以改进模型渐近精度，因而倾向选择更大的谱宽 h。另一方面，增加多项式阶数 p同时也会导致估计量方差的增大，估计可靠性会降低。

3.估计效果(Performance Analysis)

3.1 渐近偏误(Asymptotic Bias)

3.2 渐近方差(Asymptotic Variance)

3.3 渐近均方误(AMSE & AIMSE)

3.4 谱宽选择(Bandwidth Selection)

3.5 渐近分布(Asymptotic Distribution)

3.6 方差估计(Variance Estimation)

3.7 工资案例

期望：

$$\mathbb{E}\left[\widehat{m}_{\mathrm{nw}}(x) \mid \boldsymbol{X}\right]=\frac{\sum_{i=1}^{n} K\left(\frac{X_i-x}{h}\right) \mathbb{E}\left[Y_i \mid X_i\right]}{\sum_{i=1}^{n} K\left(\frac{X_i-x}{h}\right)}=\frac{\sum_{i=1}^{n} K\left(\frac{X_i-x}{h}\right) m\left(X_i\right)}{\sum_{i=1}^{n} K\left(\frac{X_i-x}{h}\right)}$$

3.1 渐近偏误：NW和LL下估计量的期望

对于局部NW估计量有：

$$\mathbb{E}\left[\hat{m}_{\mathrm{nw}}(x) \mid \boldsymbol{X}\right]=m(x)+h^{2} B_{\mathrm{nw}}(x)+o_{p}\left(h^{2}\right)+O_{p}\left(\sqrt{\frac{h}{n}}\right)$$

$$B_{\mathrm{nw}}(x)=\frac{1}{2} m^{\prime \prime}(x)+f(x)^{-1} f^{\prime}(x) m^{\prime}(x)$$

- 其中：$B_{\mathrm{nw}}(x)=\frac{1}{2} m^{\prime \prime}(x)+f(x)^{-1} f^{\prime}(x) m^{\prime}(x)$

对于局部线性LL估计量有：

$$\mathbb{E}\left[\hat{m}_{\mathrm{LL}}(x) \mid \boldsymbol{X}\right]=m(x)+h^{2} B_{\mathrm{LL}}(x)+o_{p}\left(h^{2}\right)+O_{p}\left(\sqrt{\frac{h}{n}}\right)$$

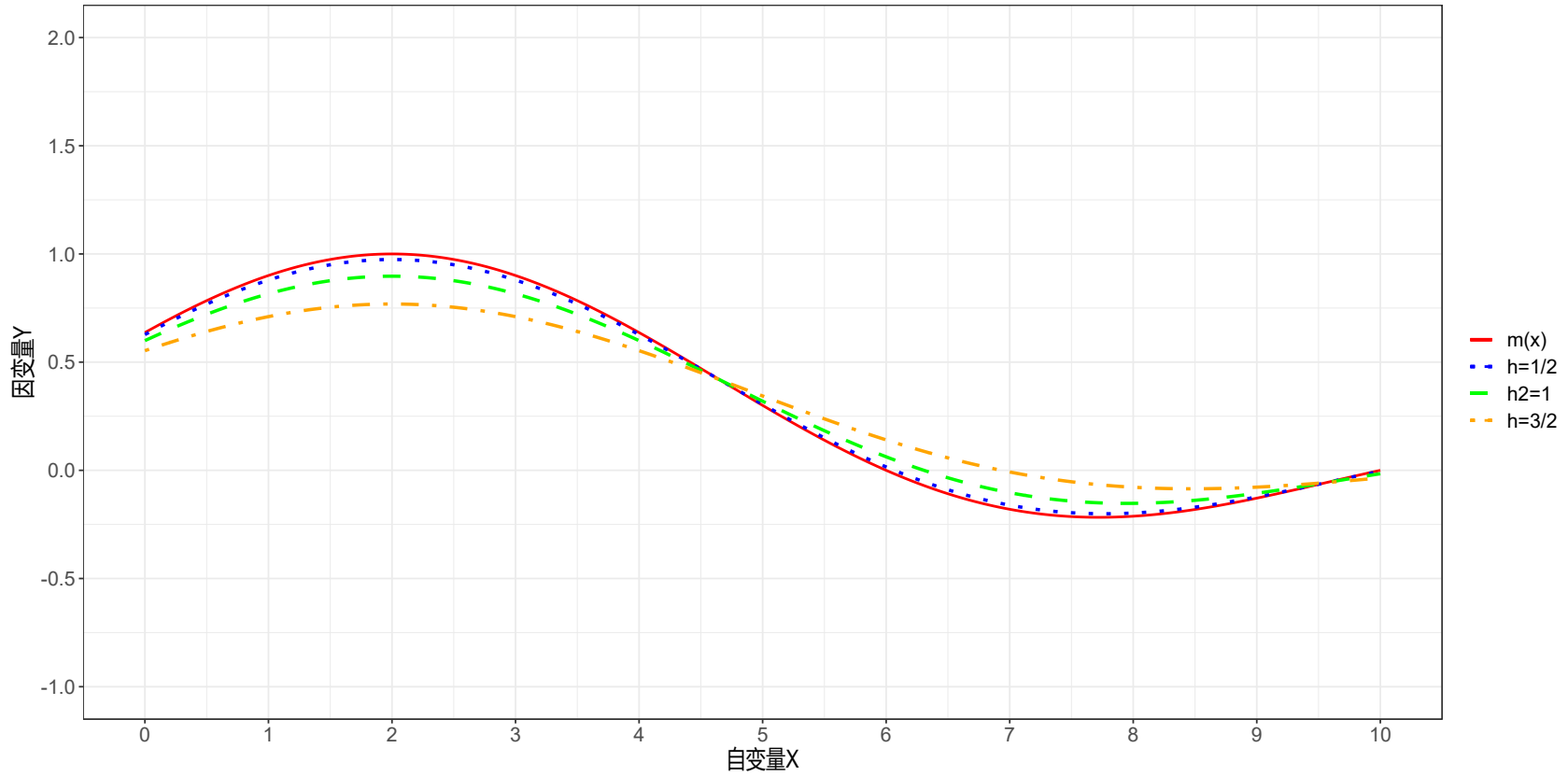
- 其中 $B_{\mathrm{LL}}(x)=\frac{1}{2} m^{\prime \prime}(x)$

3.1 渐近偏误：定义

渐近偏误（Asymptotic Bias）：我们称 $h^2 B_{\mathrm{nw}}(x)$和 $h^2 B_{\mathrm{LL}}(x)$为**渐近偏误**。

（示例）不同谱宽下渐近偏误的表现

- 根据前面计算表的1001个拟合数据对 $(x_i, \widehat{m}(x_i))$，我们可以得到滚动箱组线性回归估计结果：



3.2 渐近方差：

渐近方差（Asymptotic Variance）

$$\hat{m}_{\mathrm{nw}}(x)-\mathbb{E}\left[\hat{m}_{\mathrm{nw}}(x) \mid \boldsymbol{X}\right]=\frac{\sum_{i=1}^{n} K\left(\frac{X_i-x}{h}\right) e_i}{\sum_{i=1}^{n} K\left(\frac{X_i-x}{h}\right)}$$

$$\operatorname{var}\left[\hat{m}_{\mathrm{nw}}(x) \mid \boldsymbol{X}\right]=\frac{\sum_{i=1}^{n} K\left(\frac{X_i-x}{h}\right)^2 \sigma^2\left(X_i\right)}{\left(\sum_{i=1}^{n} K\left(\frac{X_i-x}{h}\right)\right)^2}$$

3.2 渐近方差：

$$\operatorname{var}[\hat{m}_{\mathrm{nw}}(x) \mid \boldsymbol{X}] = \frac{R_K \sigma^2(x)}{f(x) n h} + o_p\left(\frac{1}{nh}\right)$$

$$\operatorname{var}[\hat{m}_{\mathrm{LL}}(x) \mid \boldsymbol{X}] = \frac{R_K \sigma^2(x)}{f(x) n h} + o_p\left(\frac{1}{nh}\right)$$

- **核函数 $K(u)$的粗糙度**（roughness）定义为： $R_K = \int_{-\infty}^{\infty} K(u)^2 du$

3.3 渐近均方误：定义

渐近均方误（Asymptotic MSE, AMSE）：是估计量 $\widehat{m}(x)$的**平方渐近偏误**以及**渐近误差**二者之和。定义为：

$$\mathrm{AMSE}(x) \stackrel{\text{def}}{=} h^4 B(x)^2 + \frac{R_K \sigma^2(x)}{nhf(x)}$$

3.3 渐近均方误：连续情形下

渐近积分均方误(Asymptotic integrated MSE, AIMSE)：类似地定义如下：

$$\begin{aligned}\text{AIMSE} &\overset{\text{def}}{=} \int_S \text{AMSE}(x) f(x) w(x) dx \\ &= \int_S \left(h^4 B(x)^2 + \frac{R_K \sigma^2(x)}{nhf(x)} \right) f(x) w(x) dx \\ &= h^4 \bar{B} + \frac{R_K}{nh} \bar{\sigma}^2\end{aligned}$$

- 其中：

$$\begin{aligned}\bar{B} &= \int_S B(x)^2 f(x) w(x) dx \\ \bar{\sigma}^2 &= \int_S \sigma^2(x) w(x) dx\end{aligned}$$

3.4 谱宽选择：最优谱宽

最小化**渐近积分均方误**目标下，可以得到**最优谱宽**（Optimal Bandwidth）：

$$h_0 = \left(\frac{R_K \bar{\sigma}^2}{4\bar{B}} \right)^{1/5} n^{-1/5}$$

- 此时，随着 $h \sim n^{-1/5}$，则有 $\text{AIMSE}\left[\hat{m}(x)\right] = O\left(n^{-4/5}\right)$
- 因此，可以计算出上述最优谱宽下的**渐近积分均方误**：

$$\text{AIMSE}_0 \simeq 1.65 \left(R_K^4 \bar{B} \bar{\sigma}^8 \right)^{1/5} n^{-4/5}$$

3.4 谱宽选择：最优谱宽

至此，我们可以得到最优谱宽的理论计算公式：

$$h_0 = \left(\frac{R_K}{4}\right)^{1/5}\left(\frac{\bar{\sigma}^2}{n\bar{B}}\right)^{1/5} \simeq 0.58\left(\frac{\bar{\sigma}^2}{n\bar{B}}\right)^{1/5}$$

$$\bar{B} = \mathbb{E}\left[B(X)^2 w(X)\right] = \mathbb{E}\left[\left(\frac{1}{2}m''(X)\right)^2 1\{\xi_1 \le X \le \xi_2\}\right].$$

3.4 谱宽选择：参考谱宽Rot

Fan and Gijbels(1996)提出了一种**经验参考谱宽**（Rule of Thumb bandwidth,ROT）的计算办法：

$$h_{\text{rot}} = 0.58\left(\frac{\widehat{\sigma}^2\left(\xi_2-\xi_1\right)}{n\widehat{B}}\right)^{1/5}$$

- 首先构建**先验q阶多项式回归模型**，并分别估计得到 $\widehat{m}(x)$及其二阶导 $\widehat{m}''(x)$

$$m(x)=\beta_0+\beta_1 x+\beta_2 x^2+\cdots+\beta_q x^q+\epsilon$$

- 并使用 $\overline{B}$的**矩估计量**： $\widehat{B}=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}\hat{m}''\left(X_i\right)\right)^2 1\left\{\xi_1 \leq X_i \leq \xi_2\right\}$
- 其次，假定随机干扰项为同方差，也即 $\mathbb{E}\left[e^2 \mid X\right]=\sigma^2$，从而可以使用：

$$\bar{\sigma}^2=\sigma^2\left(\xi_2-\xi_1\right) \approx \hat{\sigma}^2\left(\xi_2-\xi_1\right)$$

3.4 带宽选择：参考带宽Rot（主要步骤）

参考带宽rot的主要机选步骤具体包括：

- 步骤1：确定权重取值范围 $w(x)=1\{\xi_1 \le x \le \xi_2\}$
- 步骤2：构建**先验q阶多项式回归模型**（建议为4阶），并分别估计得到 $\widehat{m}(x)$及其二阶导 $\widehat{m}''(x)$

$$
\begin{aligned}
m(x) &= \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_q x^q + \epsilon \\
\widehat{m}(x) &= \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \cdots + \hat{\beta}_q x^q \\
\widehat{m}''(x) &= 2\hat{\beta}_2 + 6\hat{\beta}_3 x + 12\hat{\beta}_4 x^2 + \cdots + q(q-1)\hat{\beta}_q x^{q-2}
\end{aligned}
$$

- 步骤3：利用上述估计结果计算

$$
\widehat{B} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}\hat{m}''(X_i)\right)^2 1\{\xi_1 \le X_i \le \xi_2\}
$$

3.4 谱宽选择：参考谱宽Rot（主要步骤）

参考谱宽rot的主要机选步骤具体包括：

- 步骤4：计算上述**先验q阶多项式回归模型**的回归误差方差 $\hat{\sigma}^2$

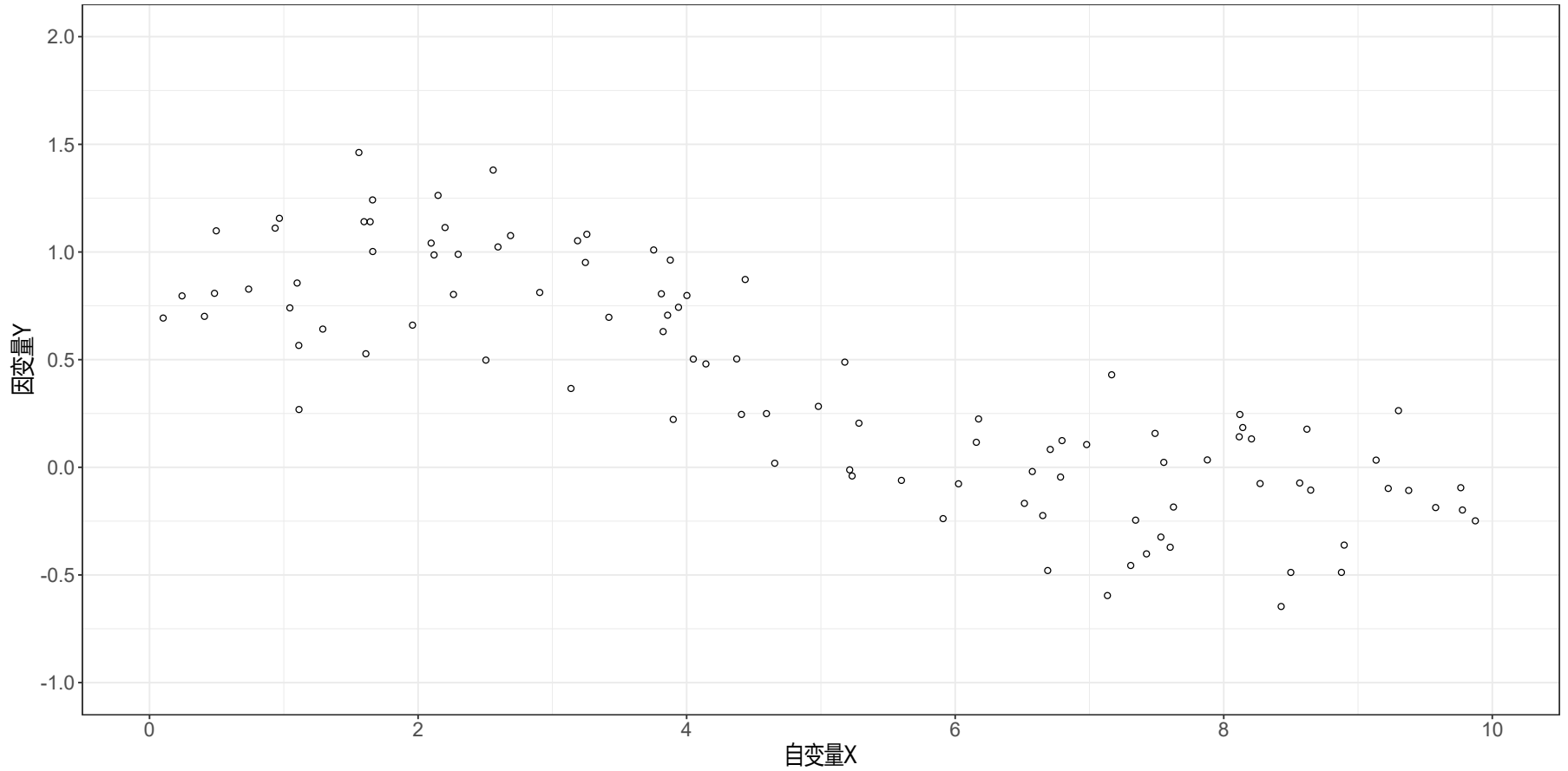
$$\hat{\sigma}^2 = \frac{\sum \hat{\epsilon}^2}{n-q-1}$$

- 步骤5：根据上述全部结果计算得到经验谱宽：

$$h_{\text{rot}} = 0.58\left(\frac{\widehat{\sigma}^2\left(\xi_2-\xi_1\right)}{n\widehat{B}}\right)^{1/5}$$

（示例）参考谱宽的计算：数据散点图

- 我们继续使用前述的蒙特卡洛模拟数据集：



（示例）参考谱宽的计算：多项式回归

下面我们按前述步骤来计算参考谱宽值 h_{rot}：

- **步骤1**：根据案例数据集，设定权重取值范围 $\{\xi_1 \leq x \leq \xi_2\} = \{0, 10\}$
- **步骤2**：构建多项式回归

$$Y_i = + \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3 + \beta_5 X_i^4 + u_i$$

> 直接使用OLS进行估计，得到估计方程：

$$\begin{array}{llllll} \widehat{Y} = & + 0.4943 & + 0.6986 X_i & - 0.2807 X_i^2 & + 0.0326 X_i^3 & - 0.0012 X_i^4 \\ (s) & (0.1521) & (0.2015) & (0.0802) & (0.0119) & (0.0006) \end{array}$$

> 进而得到拟合值 $\widehat{m}(x)$及其二阶导 $\widehat{m}''(x)$

$$\begin{aligned} \widehat{m}(x) &= +0.4943 + 0.6986 x_i - 0.2807 x_i^2 + 0.0326 x_i^3 - 0.0012 x_i^4 \\ \widehat{m}''(x) &= -2 \times 0.2807 + 6 \times 0.0326 x_i - 12 \times 0.0012 x_i^2 \end{aligned}$$

（示例）参考谱宽的计算：多项式回归

- 拟合值 $\widehat{m}(x)$及其二阶导 $\widehat{m}''(x)$以及残差 $\hat{\epsilon}$

多项式回归计算表

mx	mx_q2	epsilon
0.8671	-0.0580	-0.0551
-0.1482	0.0268	-0.3401
-0.0739	0.0489	0.0285
0.7813	-0.0393	0.3009
-0.1368	0.0162	-0.3513
0.9279	-0.1956	0.1830
0.4530	0.0093	-0.2074
0.6082	-0.0107	0.3543

Showing 1 to 8 of 100 entries

Previous 1 2 3 4 5 … 13 Next

（示例）参考谱宽的计算：结果

- 步骤3：利用上述估计结果计算

$$\widehat{B}=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}\hat{m}''\left(X_i\right)\right)^2 1\left\{\xi_1\leq X_i\leq\xi_2\right\}=0.0089$$

- 步骤4：多项式模型的回归误差方差

$$\hat{\sigma}^2=\frac{\sum\hat{\epsilon}^2}{n-q-1}=0.0687$$

- 步骤5：根据上述全部结果计算得到经验谱宽：

$$h_{\text{rot}}=0.58\left(\frac{\widehat{\sigma}^2\left(\xi_2-\xi_1\right)}{n\widehat{B}}\right)^{1/5}=0.58\times\left(\frac{0.0687\times(10-0)}{100\times0.0089}\right)^{1/5}=0.5508$$

3.4 谱宽选择：交叉验证谱宽（目标问题）

我们期望选择谱宽，以实现估计量 $\widehat{m}(x,h)$ 最小化**积分均方误**（Integrated mean-squared error, IMSE），也即：

$$\text{IMSE}_n(h) = \int_S \mathbb{E}\left[(\widehat{m}(x,h) - m(x))^2\right] f(x) w(x) dx$$

- 其中 $f(x)$ 为 X 的边际密度（marginal density）
- $w(x)$ 为可积分权重函数（integrable weight function）

3.4 谱宽选择：交叉验证谱宽（可行计算）

上述最小化问题中，偏差值 $(\widehat{m}(x,h)-m(x))$可以通过**留一法**下的预测误差来代替，也即：

$$\tilde{e}_i(h)=Y_i-\tilde{m}_{-i}\left(X_i,h\right)$$

因此，上述最小化问题 $IMSE_n(h)$的一个**可行计算方案**可以表达为如下的**交叉验证准则函数**（Cross-Validation Criterion）：

$$\mathrm{CV}(h)=\frac{1}{n}\sum_{i=1}^{n}\tilde{e}_i(h)^2w\left(X_i\right)$$

3.4 谱宽选择：交叉验证谱宽（准则函数）

对于**交叉验证准则函数**（Cross-Validation Criterion）：

$$\mathrm{CV}(h) = \frac{1}{n}\sum_{i=1}^{n}\tilde{e}_i(h)^2 w\left(X_i\right)$$

我们可以证明它是去掉一个样本的积分均方误$IMSE_(n-1)(h)$加上一个常数项的**无偏估计**，也即：

$$\mathbb{E}[\mathrm{CV}(h)] = \bar{\sigma}^2 + \mathrm{IMSE}_{n-1}(h)$$

- 其中：$\bar{\sigma}^2 = \mathbb{E}\left[e^2 w(X)\right]$，它是不依赖于谱宽 h的。
- 显然，最小化 $\mathbb{E}[\mathrm{CV}(h)]$和最小化 $\mathrm{IMSE}_{n-1}(h)$将是等价的。
- 而且，当 n比较大时，最小化 $\mathrm{IMSE}_{n-1}(h)$和最小化 $\mathrm{IMSE}_{n}(h)$也将是等价的，二者各自得到的谱宽也是无偏的。

3.4 谱宽选择：交叉验证谱宽（受约束）

另外，为了避免谱宽值选择过小，需要给定约束条件 $h \geq h_\ell$，此时：

$$h_{\mathrm{CV}} = \underset{h \geq h_\ell}{\operatorname{argmin}} \, \mathrm{CV}(h)$$

3.4 谱宽选择：交叉验证谱宽（主要过程）

因此，最优交叉验证谱宽选择的主要过程如下：

- 构建 h的序贯数值（grid value）$[h_1, h_2, \cdots, h_J]$，并评估交叉效用准则函数的最小化取值 $CV(h_j), j \in (1, 2, \cdots, J)$，从而得到最优谱宽：

$$h_{\text{CV}} = \underset{h \in [h_1, h_2, \ldots, h_J]}{\text{argmin}} \ \text{CV}(h)$$

- 需要注意的是，以上方法得到的谱宽理论上可以是无界化的。意味着，交叉验证准则函数 $CV(h)$是单调下降的，以至于最优谱宽 $h = \infty$。
- 此时，将表明**全样本**回归估计也可能是一个最优结果。也即，使用全样本数据，NW估计方法有 $\widehat{m}(x) = \overline{Y}$；而局部线性估计方法则有 $\widehat{m}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$

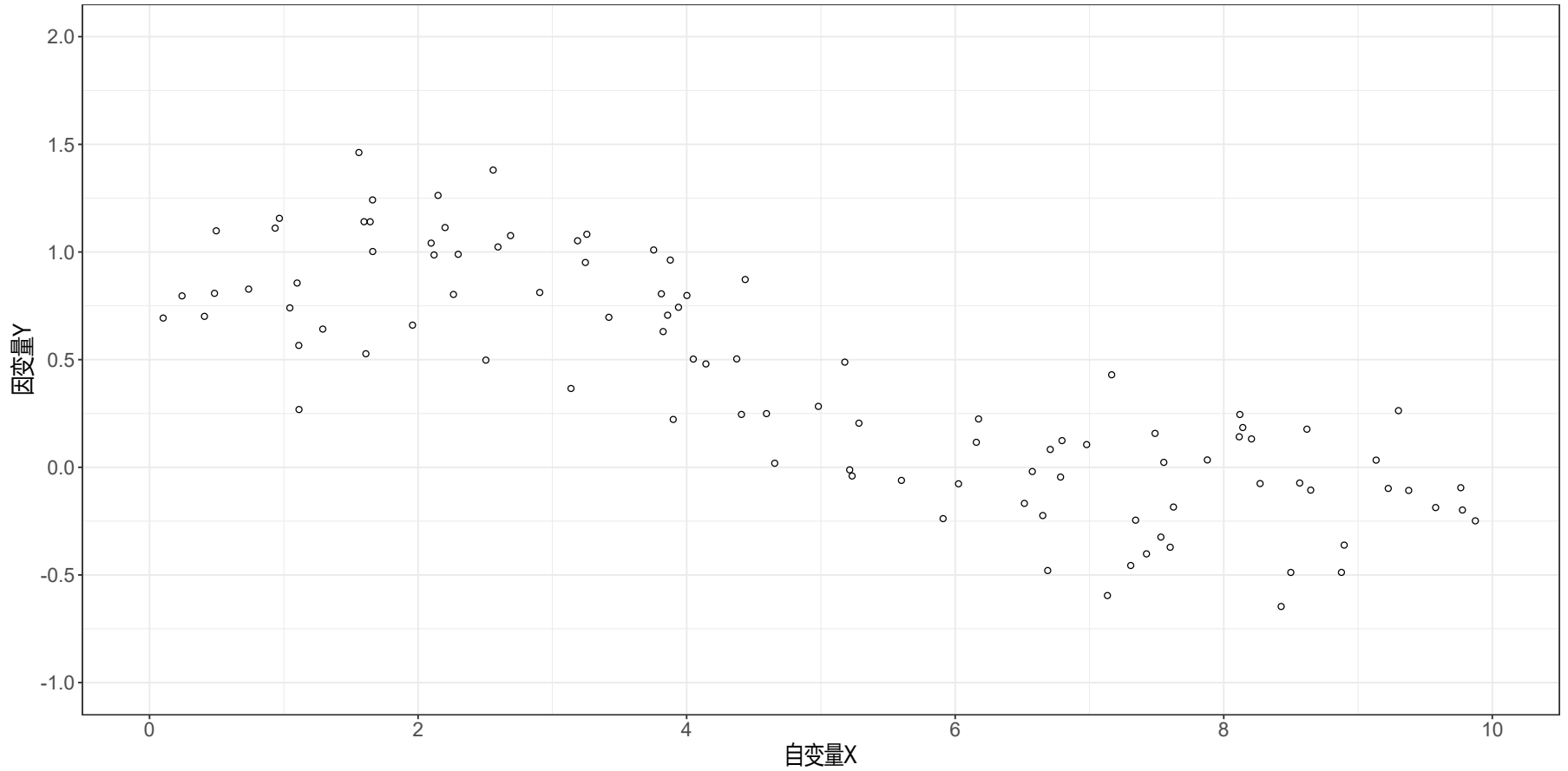
3.4 谱宽选择：交叉验证谱宽（计算步骤）

- 步骤1：设定**初始值**。也即选定一个**参考谱宽**作为初始值，例如前面提到的**经验谱宽** h_{rot}。
- 步骤2：设定**调参谱宽**（tuning bandwidth）。也即给定待评估的谱宽范围，及其待评估序贯值（grid value）。

> - 一个经验谱宽范围可供参考： $[h_{rot}/3, 3h_{rot}]$。
> - 范围内的待评估序贯值的个数 g也是需要做出尝试性选择。

（示例）交叉验证谱宽的计算：数据集

- 我们继续使用前述的蒙特卡洛模拟数据集：



（示例）交叉验证谱宽的计算：规则

- 步骤1：设定**经验谱宽** $h_{rot}=0.5508$作为**初始值**。。
- 步骤2：设定**调参谱宽**（tuning bandwidth）。
- 一个经验谱宽范围可供参考： $[h_{rot}/3, 3h_{rot}]=[0.1836, 1.6524]$。
- 给定范围内的搜寻总数为 $n=201$。则待评估序贯值为 $h \in (0.1836, 0.1909, 0.1983, 0.2056, 0.2130, \cdots, 1.6304, 1.6377, 1.6451, 1.6524)$ 。

（示例）交叉验证谱宽的计算：MW方法下的预测误差平方

MW方法LOO交叉验证的预测误差平方

id	h	loo	ei_sqr
1	0.1836	loo_001	0.0080
1	0.1836	loo_002	0.1278
1	0.1836	loo_003	0.0017
1	0.1836	loo_004	0.0841
1	0.1836	loo_005	0.1126
1	0.1836	loo_006	0.1153
1	0.1836	loo_007	0.0735
1	0.1836	loo_008	0.0902

Showing 1 to 8 of 20,100 entries

Previous 1 2 3 4 5 … 2,513 Next

[a] `loo`表示交叉验证**留一法**，例如`loo_001`表示第1个样本数据不进入局部回归估计。

（示例）交叉验证谱宽的计算：LL方法下的预测误差平方

LL方法LOO交叉验证的预测误差平方

id	h	loo	ei_sqr
1	0.1836	loo_001	0.0093
1	0.1836	loo_002	0.1267
1	0.1836	loo_003	0.0007
1	0.1836	loo_004	0.0820
1	0.1836	loo_005	0.1150
1	0.1836	loo_006	0.0816
1	0.1836	loo_007	0.0713
1	0.1836	loo_008	0.0815

Showing 1 to 8 of 20,100 entries

Previous 1 2 3 4 5 … 2,513 Next

[a] `loo`表示交叉验证**留一法**，例如`loo_001`表示第1个样本数据不进入局部回归估计。

（示例）交叉验证谱宽的计算：CV计算表

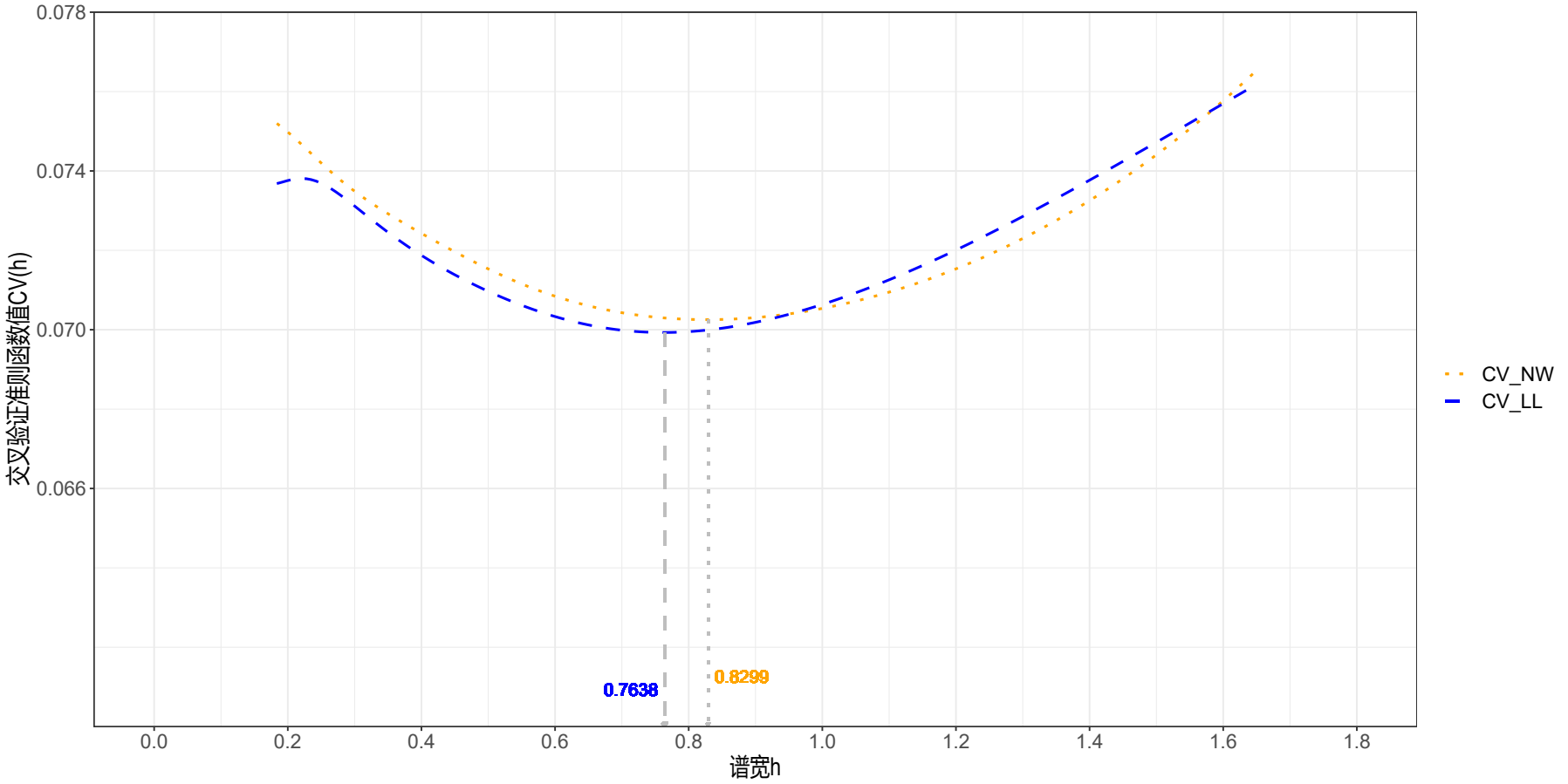
NW和LL方法下的CV计算表

id	h_tune	cv_NW	cv_LL
1	0.1836	0.0752	0.0737
2	0.1909	0.0751	0.0737
3	0.1983	0.0750	0.0738
4	0.2056	0.0749	0.0738
5	0.2130	0.0748	0.0738
6	0.2203	0.0747	0.0738
7	0.2277	0.0746	0.0738
8	0.2350	0.0744	0.0738

Showing 1 to 8 of 201 entries

Previous 1 2 3 4 5 … 26 Next

（示例）交叉验证谱宽的计算：谱宽与CV变化



（示例）最优谱宽下的估计表

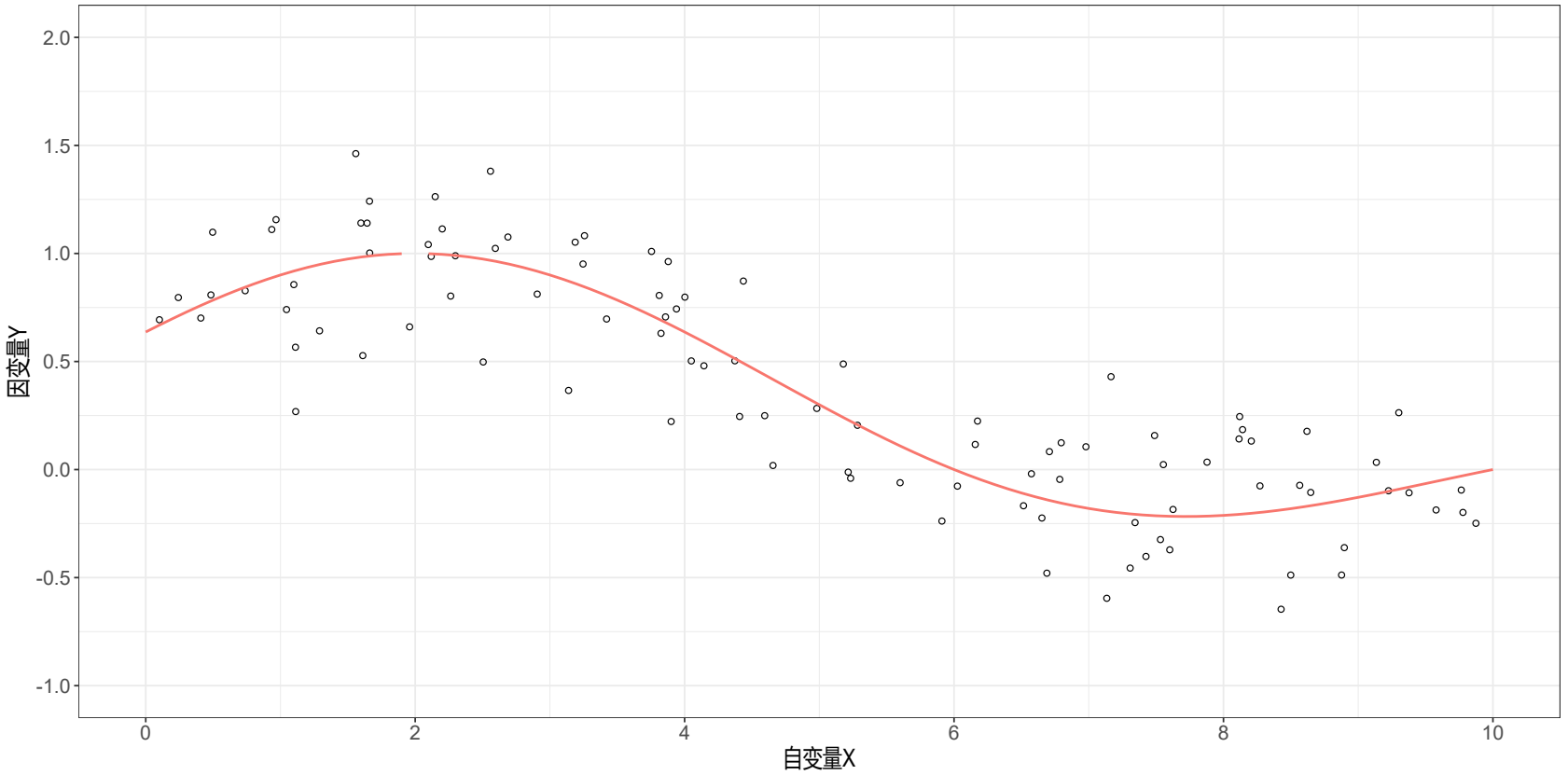
使用不同谱宽下LL方法对m(x)的估计结果

index	xg	mx_rot	mx_cv
1	0.00	0.7603	0.7648
2	0.05	0.7699	0.7696
3	0.10	0.7785	0.7743
4	0.15	0.7862	0.7789
5	0.20	0.7929	0.7833
6	0.25	0.7989	0.7877
7	0.30	0.8041	0.7921
8	0.35	0.8087	0.7965

Showing 1 to 8 of 201 entries

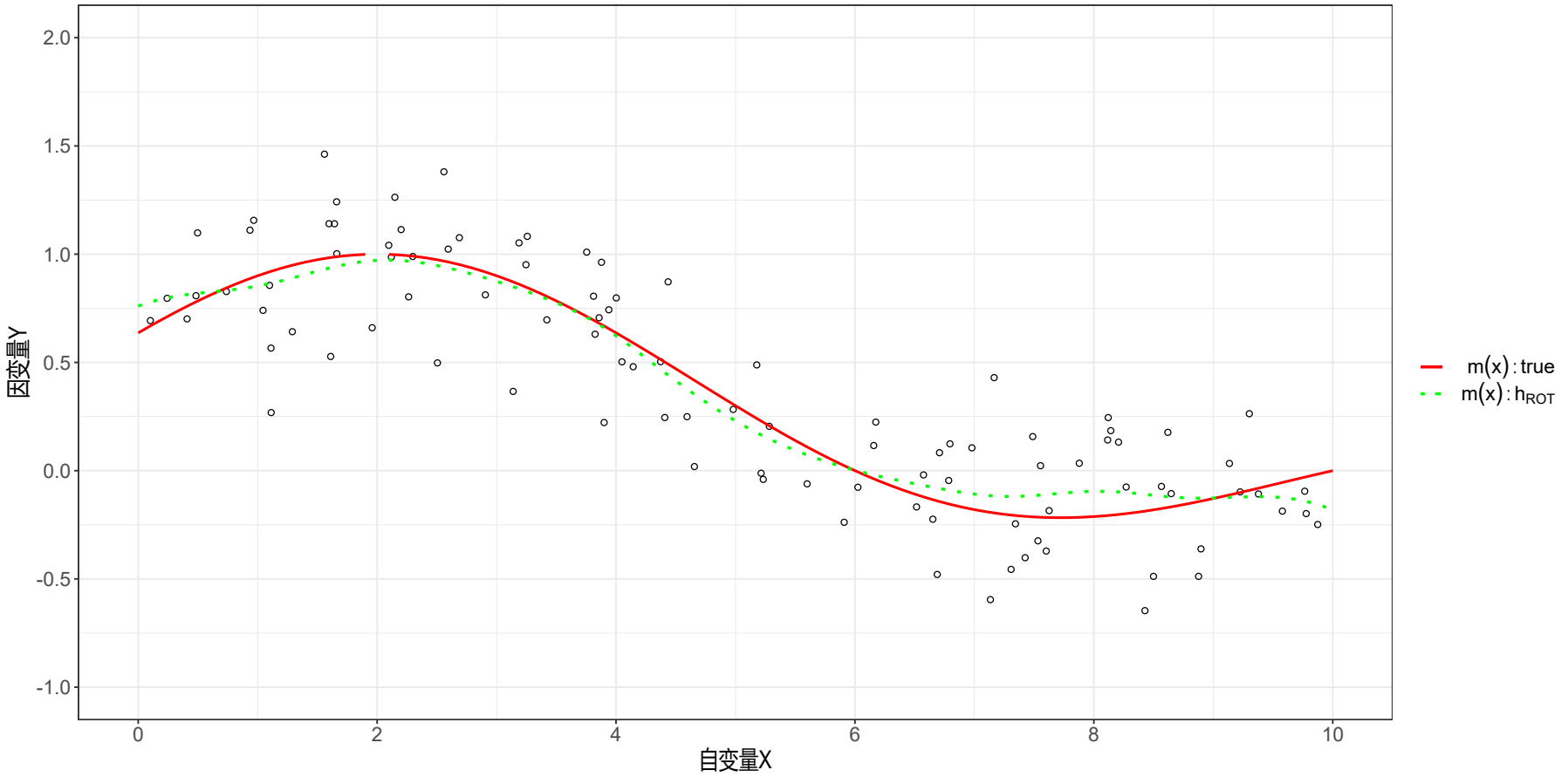
Previous 1 2 3 4 5 … 26 Next

（示例）真实的条件期望函数m(x)

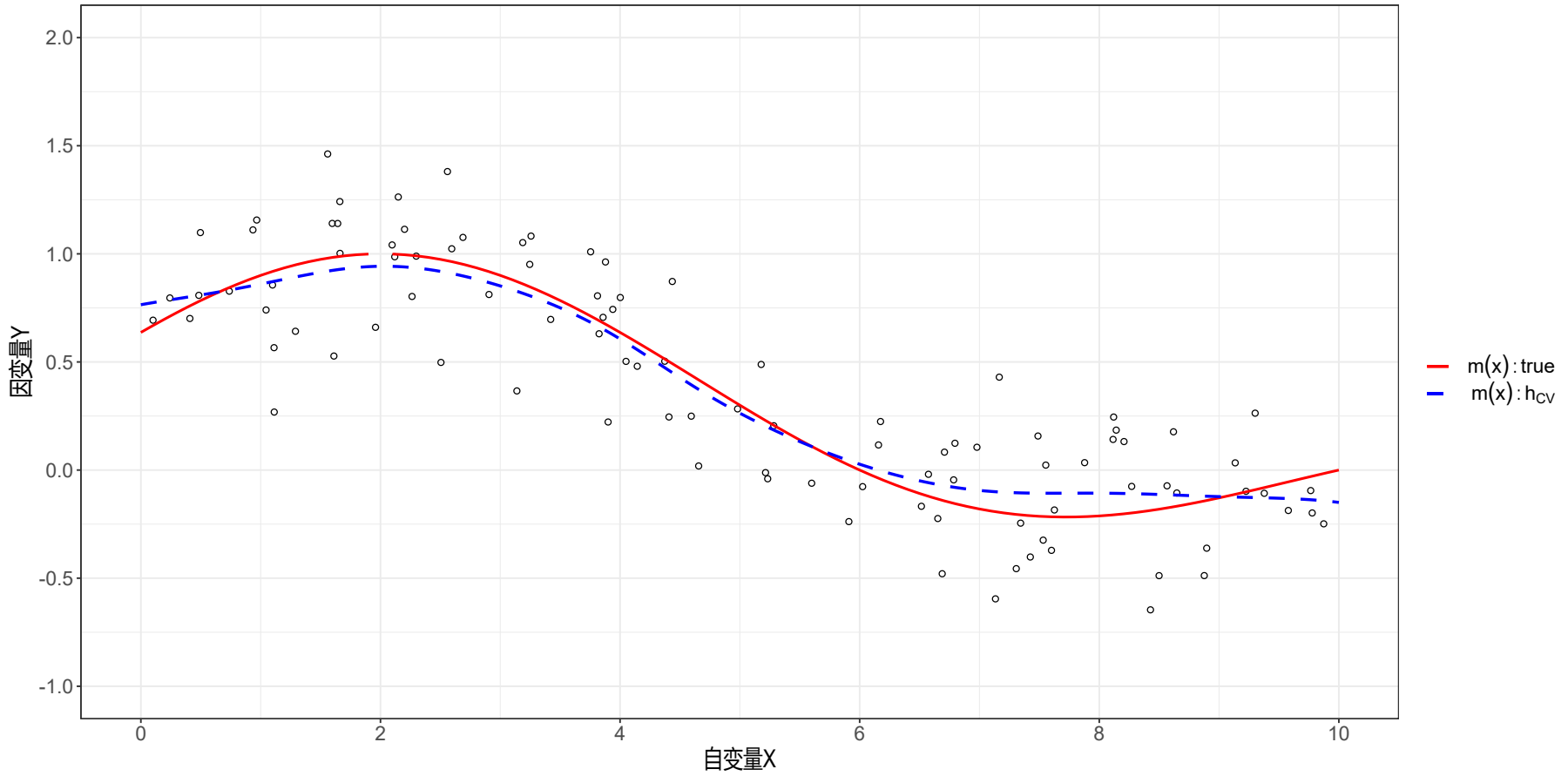


$$m(X) = \frac{sin(\frac{\pi}{4} \cdot (X_i - 2))}{\frac{\pi}{4} \cdot (X_i - 2)}$$

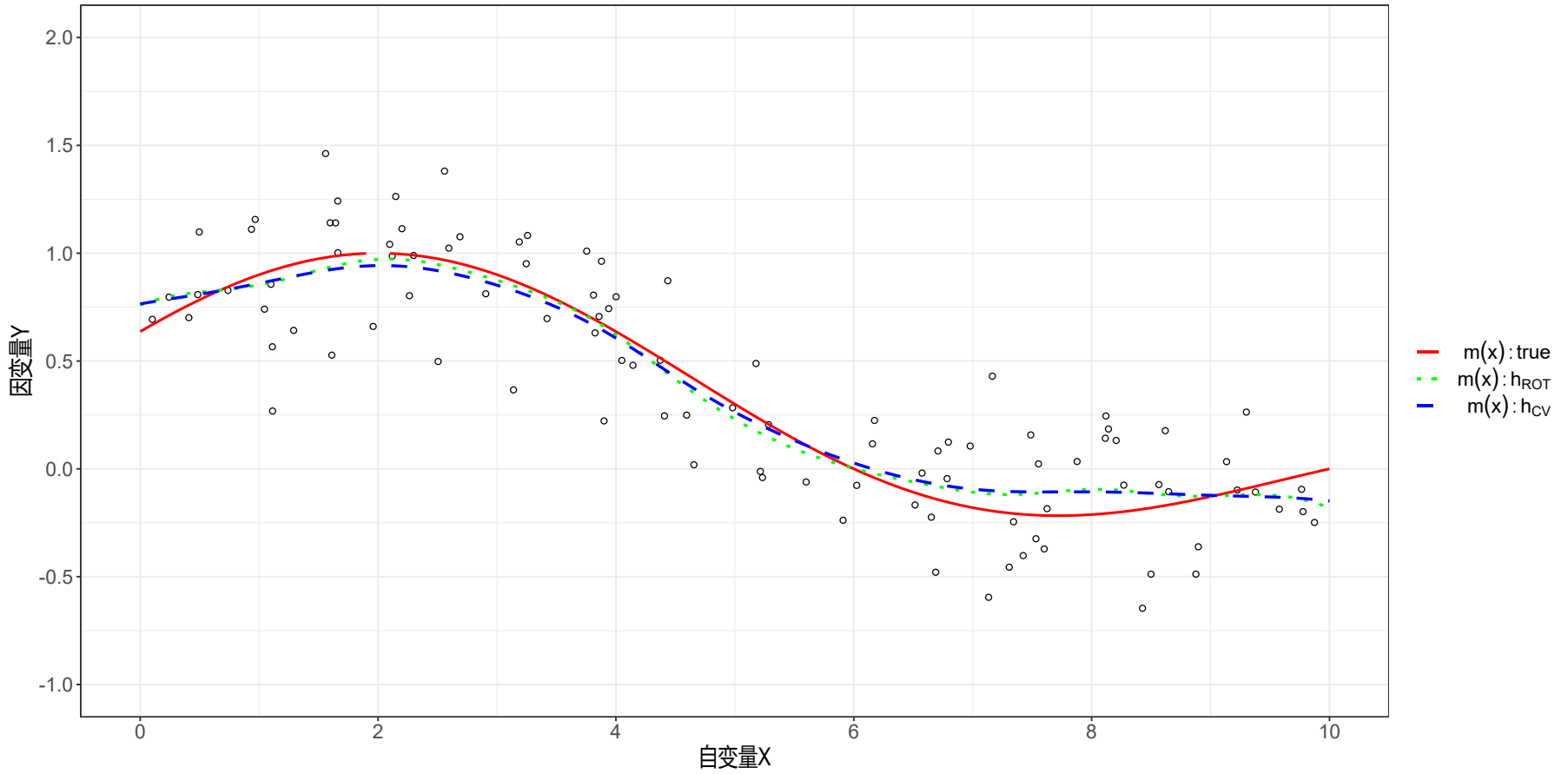
（示例）LL方法下使用ROT带宽估计得到的m(x)



（示例）LL方法下使用最优CV谱宽估计得到的m(x)



(示例)LL方法下使用不同谱宽估计得到的m(x):对比



（示例）模拟数据集

为了更好地进行数据验证，我们将根据如下规则生成蒙特卡洛模拟数据集：

$$m(X) = \frac{sin(\frac{\pi}{4} \cdot (X_i - 2))}{\frac{\pi}{4} \cdot (X_i - 2)}$$

- 此时，我们具有**上帝视角**，实际上已经知道**数据生成机制**（DGP）
- 此时，我们心里面已知真实模型为**非线性的**

3.5 渐近分布：渐近正态分布

渐近分布（Asymptotic Distribution）

- 对于局部NW估计：

$$\sqrt{nh}\left(\widehat{m}_{\mathrm{nw}}(x)-m(x)-h^2B_{\mathrm{nw}}(x)\right)\underset{d}{\longrightarrow}\mathrm{N}\left(0,\frac{R_K\sigma^2(x)}{f(x)}\right)$$

- 对于局部线性估计：

$$\sqrt{nh}\left(\widehat{m}_{\mathrm{LL}}(x)-m(x)-h^2B_{\mathrm{LL}}(x)\right)\underset{d}{\longrightarrow}\mathrm{N}\left(0,\frac{R_K\sigma^2(x)}{f(x)}\right)$$

3.5 渐近分布：超平滑情形

$$\sqrt{nh}\left(\hat{m}_{\mathrm{nw}}(x)-m(x)\right) \underset{d}{\longrightarrow} \mathrm{N}\left(0, \frac{R_K \sigma^2(x)}{f(x)}\right)$$

$$\sqrt{nh}\left(\widehat{m}_{\mathrm{LL}}(x)-m(x)\right) \underset{d}{\longrightarrow} \mathrm{N}\left(0, \frac{R_K \sigma^2(x)}{f(x)}\right)$$

3.6 方差估计：条件方差

$$\sigma^2(x) = \operatorname{var}[Y \mid X = x] = \mathbb{E}\left[e^2 \mid X = x\right]$$

- NW方法下条件方差的一个**理想**估计为：

$$\bar{\sigma}^2(x) = \frac{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) e_i^2}{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)}$$

- NW方法下条件方差的一个**可行**估计为：

$$\widehat{\sigma}^2(x) = \frac{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) \tilde{e}_i^2}{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)}.$$

> 其中 $\tilde{e}_i = Y_i - \hat{m}_{-i}(X_i)$为**留一法**下的预测误差（leave-one-out prediction error）

3.6 方差估计：直接公式

如前所属，NW方法、LL方法和LP方法的CEF估计可以统一表达为：

$$\begin{aligned}\widehat{\beta}(x) &= \left(\boldsymbol{Z}'\boldsymbol{K}\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{K}\boldsymbol{Y}\right) \\ &= \left(\boldsymbol{Z}'\boldsymbol{K}\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{K}\boldsymbol{m}\right) + \left(\boldsymbol{Z}'\boldsymbol{K}\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{K}\boldsymbol{e}\right)\end{aligned}$$

- 那么，我们可以直接使用下列条件方差公式：

$$\boldsymbol{V}_{\widehat{\beta}}(x) = \operatorname{var}[\widehat{\beta} \mid \boldsymbol{X}] = \left(\boldsymbol{Z}'\boldsymbol{K}\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{K}\boldsymbol{D}\boldsymbol{K}\boldsymbol{Z}\right)\left(\boldsymbol{Z}'\boldsymbol{K}\boldsymbol{Z}\right)^{-1}$$

- 其中，我们可以直接使用**平方误差** $\hat{e}_i^2$ 或者**平方预测误差** $\tilde{e}_i^2$：

$$\widehat{\boldsymbol{V}}_{\widehat{\beta}}(x) = \left(\boldsymbol{Z}'\boldsymbol{K}\boldsymbol{Z}\right)^{-1}\left(\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)^2 Z_i(x)Z_i(x)'\tilde{\boldsymbol{e}}_i^2\right)\left(\boldsymbol{Z}'\boldsymbol{K}\boldsymbol{Z}\right)^{-1}$$

3.6 方差估计：渐近公式

更为简洁地，我们也可以使用如下条件方差的渐近公式：

$$\widehat{V}_{\widehat{m}(x)} = \frac{R_K \widehat{\sigma}^2(x)}{nh\widehat{f}(x)}$$

- 其中：

$$\widehat{f}(x) = \frac{1}{nb}\sum_{i=1}^{n} K\left(\frac{X_i - x}{b}\right)$$

$$\widehat{\sigma}^2(x) = \frac{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)\tilde{e}_i^2}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)}.$$

逐点置信区间（Pointwise Confidence Interval）

$$\widehat{m}(x) \pm z_{1-\alpha/2}(n-1) \cdot \sqrt{\widehat{V}_{\widehat{m}(x)}}$$

$$\widehat{m}(x) \pm 1.96 \sqrt{\widehat{V}_{\widehat{m}(x)}}$$

3.7 工资案例：背景说明

工资案例：

- 案例基于CPS数据集，重点分析其中的**子样本数据**（黑人、男性、拥有12年受教育程度——高中毕业），样本数为n=762。
- 关注的问题：**时均工资**的对数（`Y =log(wage)`）对**职业经历**（`X=experience`）的非参数回归估计。
- 后面的分析中，我们会重点划定观测窗口为：职业经历（年数）范围为$[0, 40]$。因为样本中90%以上的观测对象都落在这个范围之内。

（工资案例）样本数据集

CPS数据集(n=762)

obs	X	Y
1	18	2.7176
2	-2	1.7525
3	30	2.7940
4	30	2.7940
5	18	2.3587
6	28	2.5998
7	14	2.9565
8	22	3.0053

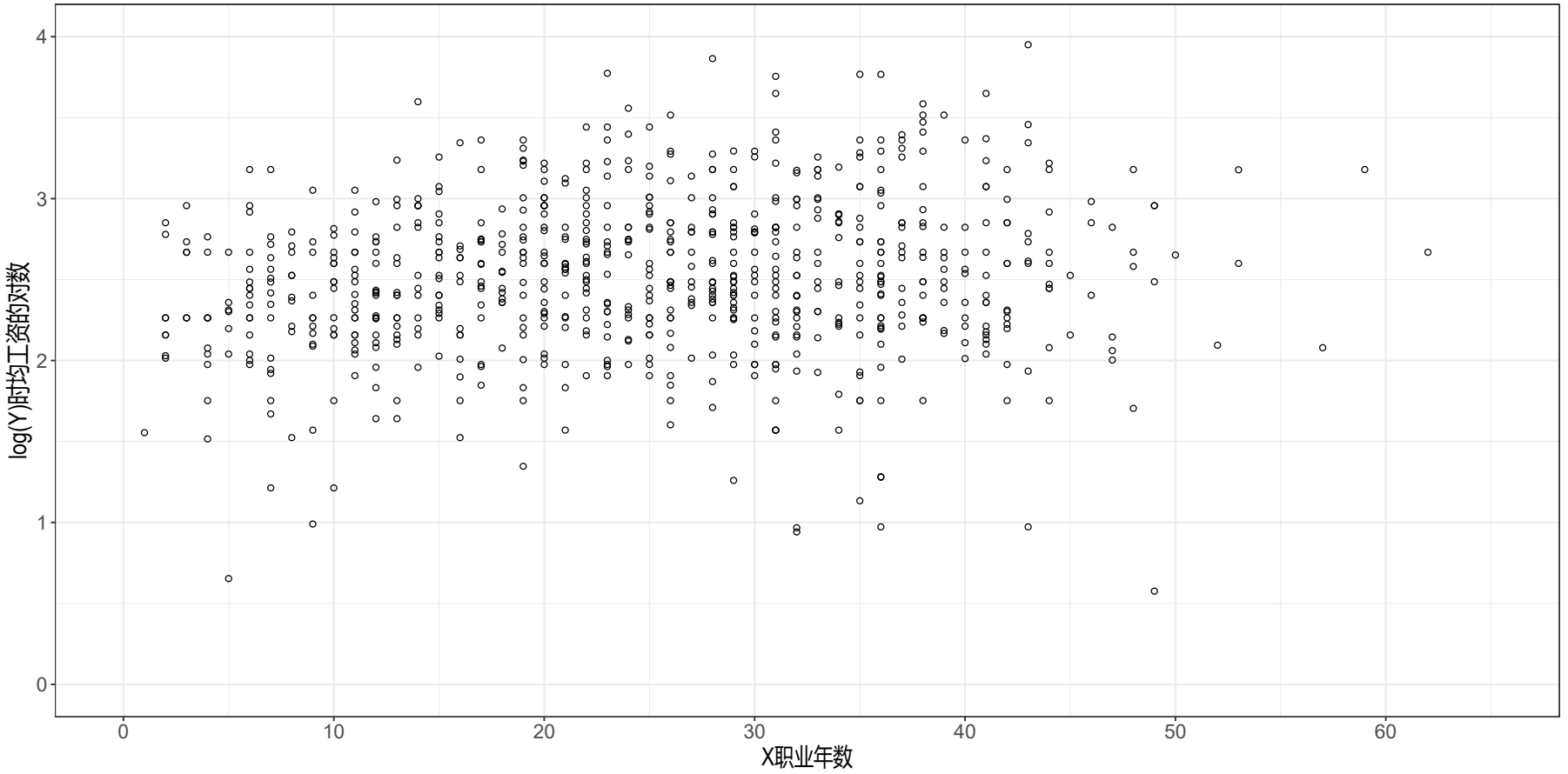
Showing 1 to 8 of 762 entries

Previous 1 2 3 4 5 … 96 Next

- 样本数据的描述性统计如下：

```
    X.age          Y.earnings
 Min.   :-2     Min.   :-3.1
 1st Qu.:15     1st Qu.: 2.2
 Median :25     Median : 2.5
 Mean   :25     Mean   : 2.5
 3rd Qu.:34     3rd Qu.: 2.8
 Max.   :62     Max.   : 4.0
```

（工资案例）样本数据散点图



（工资案例）参考谱宽的计算：多项式回归

下面我们按前述步骤来计算参考谱宽值 h_{rot}：

- **步骤1**：根据案例数据集，设定权重取值范围 $\{\xi_1 \leq x \leq \xi_2\} = \{0, 40\}$
- **步骤2**：构建多项式回归

$$Y_i = + \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3 + \beta_5 X_i^4 + u_i$$

> 直接使用OLS进行估计，得到估计方程：

$$\begin{array}{lllll} \widehat{Y} = + 2.094395 + 0.030951 X_i - 0.000103 X_i^2 - 0.000021 X_i^3 + 0.000000 X_i^4 \\ (s) \quad (0.1371) \quad (0.0284) \quad (0.0019) \quad (0.0000) \quad (0.0000) \end{array}$$

> 进而得到拟合值 $\widehat{m}(x)$及其二阶导 $\widehat{m}''(x)$及残差 $\hat{\epsilon}$

$$\begin{aligned} \widehat{m}(x) &= 2.094395 + 0.030951 x_i - 0.000103 x_i^2 - 0.000021 x_i^3 + 0.000000 x_i^4 \\ \widehat{m}''(x) &= -2 \times 0.000103 - 6 \times 0.000021 x_i + 12 \times 0.000000 x_i^2 \end{aligned}$$

（工资案例）参考谱宽的计算：结果

- 步骤3：利用上述估计结果计算

$$\widehat{B}=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}\hat{m}''\left(X_i\right)\right)^2 1\left\{\xi_1\leq X_i\leq\xi_2\right\}=0.00000025$$

- 步骤4：多项式模型的回归误差方差

$$\hat{\sigma}^2=\frac{\sum\hat{\epsilon}^2}{n-q-1}=0.2592$$

- 步骤5：根据上述全部结果计算得到经验谱宽：

$$h_{\text{rot}}=0.58\left(\frac{\widehat{\sigma}^2\left(\xi_2-\xi_1\right)}{n\widehat{B}}\right)^{1/5}=0.58\times\left(\frac{0.2592\times(40-0)}{762\times0.000000248}\right)^{1/5}=5.1442$$

（工资案例）交叉验证谱宽的计算：规则

- 步骤1：设定**经验谱宽** $h_{rot}=5.1442$作为**初始值**。。
- 步骤2：设定**调参谱宽**（tuning bandwidth）。
- 一个经验谱宽范围可供参考： $[h_{rot}/3, 3h_{rot}]=[1.7147, 15.4326]$。
- 给定范围内的搜寻总数为 $n=201$。则待评估序贯值为 $h \in (1.7147, 1.7833, 1.8519, 1.9205, 1.9891, \cdots, 15.2268, 15.2954, 15.3640, 15.4326)$ 。

（工资案例）交叉验证谱宽的计算：CV计算表

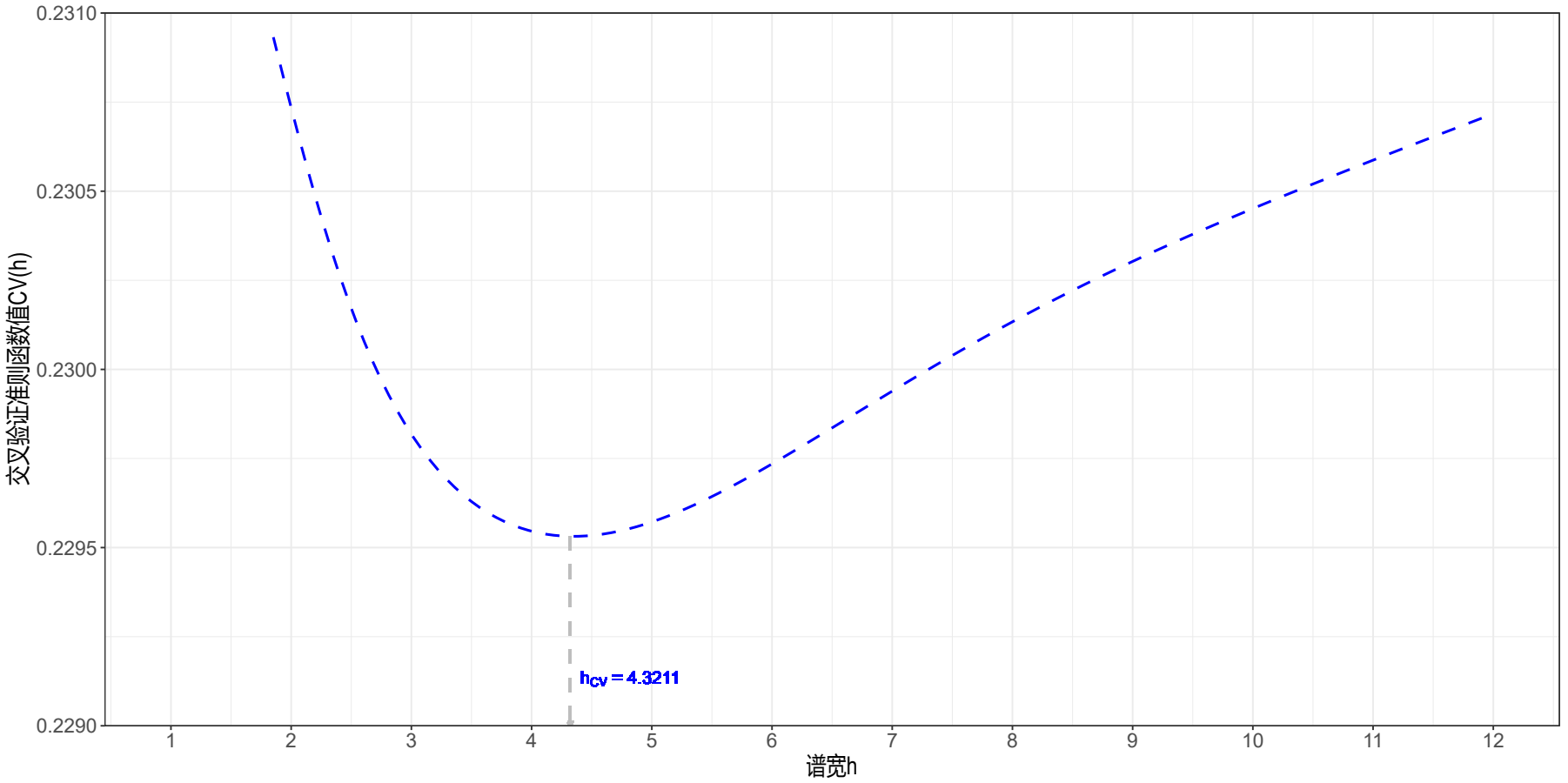
LL方法下的CV计算表

id	h_tune	cv_LL
1	1.7147	0.2311
2	1.7833	0.2310
3	1.8519	0.2309
4	1.9205	0.2308
5	1.9891	0.2307
6	2.0577	0.2307
7	2.1263	0.2306
8	2.1949	0.2305

Showing 1 to 8 of 201 entries

Previous 1 2 3 4 5 … 26 Next

（工资案例）交叉验证带宽的计算：带宽与CV变化



（工资案例）最优谱宽选择下的估计表

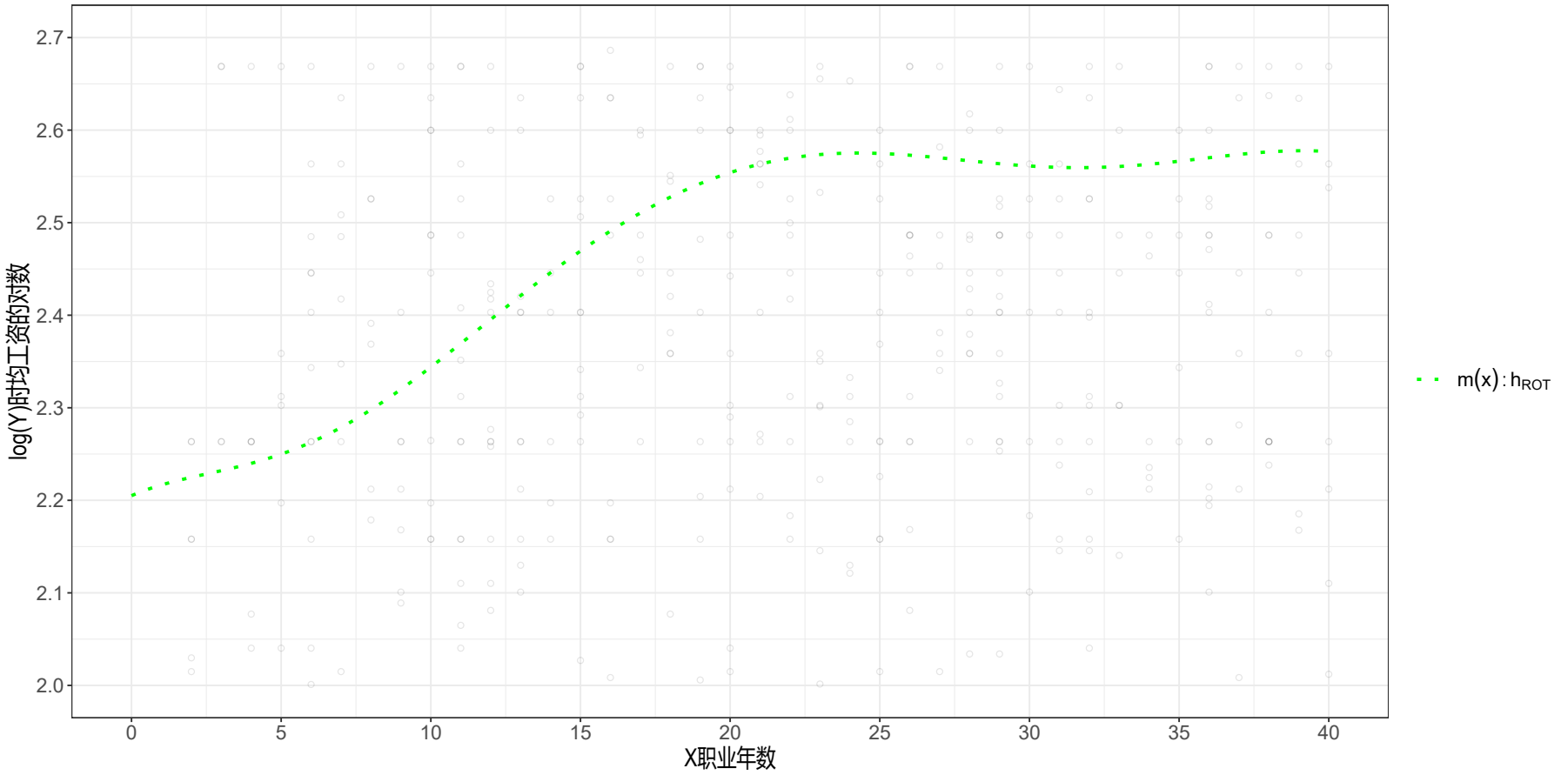
使用不同谱宽下LL方法对m(x)的估计结果

index	xg	mx_rot	mx_cv
1	0.00	2.2050	2.2153
2	0.20	2.2077	2.2205
3	0.40	2.2102	2.2251
4	0.60	2.2125	2.2292
5	0.80	2.2146	2.2328
6	1.00	2.2166	2.2359
7	1.20	2.2184	2.2386
8	1.40	2.2202	2.2410

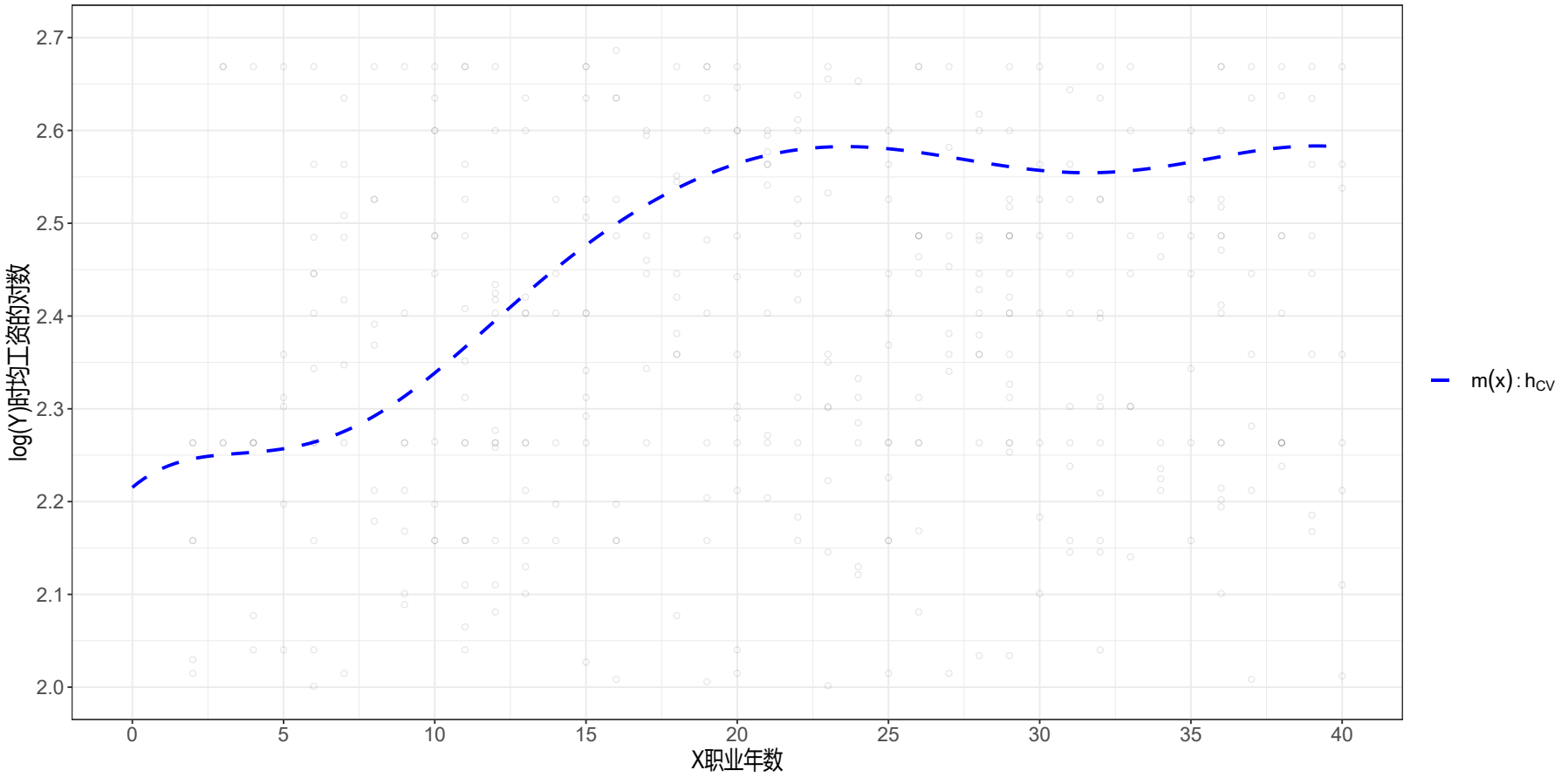
Showing 1 to 8 of 201 entries

Previous 1 2 3 4 5 … 26 Next

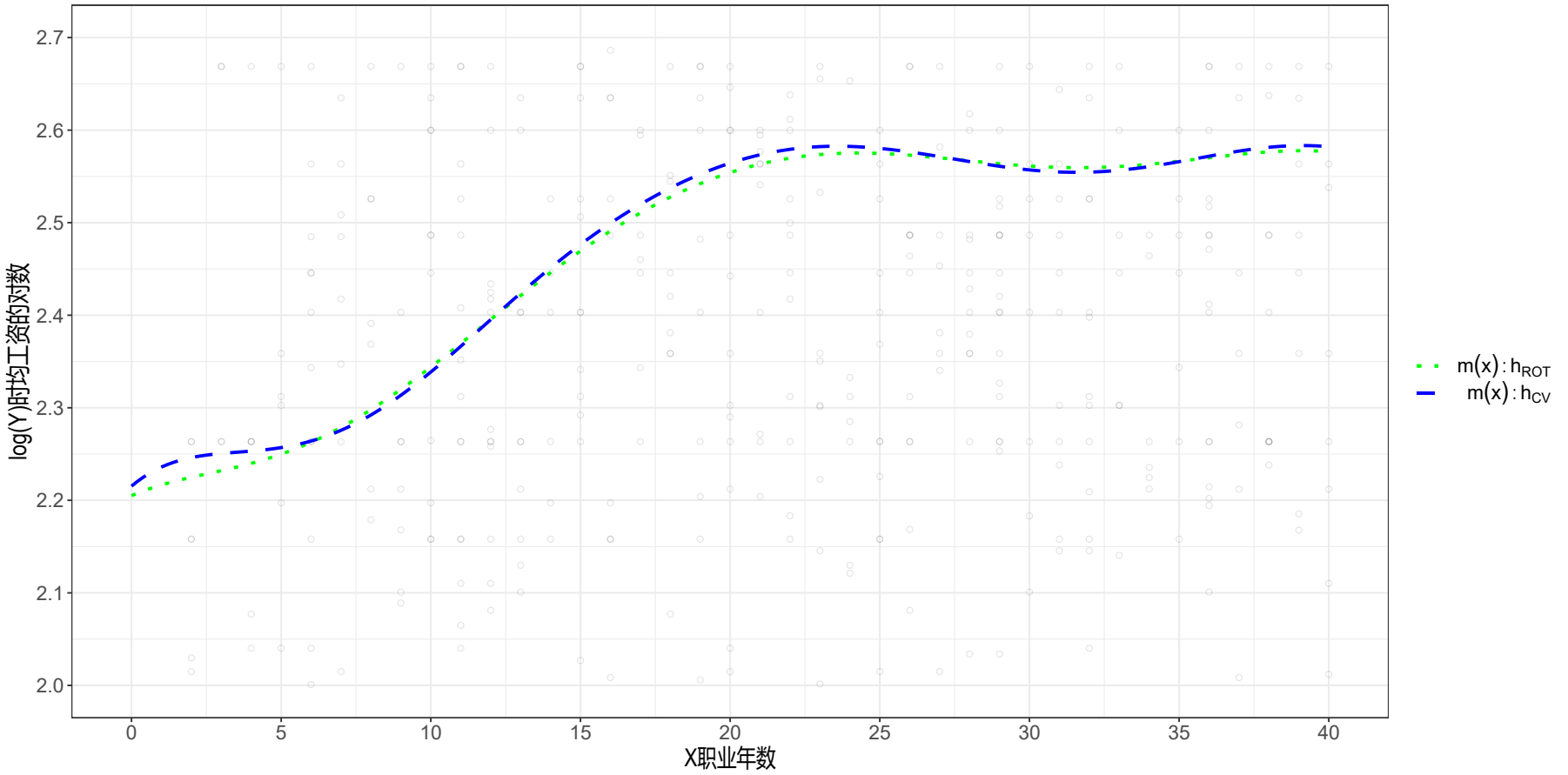
（工资案例）LL方法下使用ROT带宽估计得到的m(x)



（工资案例）LL方法下使用最优CV谱宽估计得到的m(x)



(工资案例)LL方法下使用不同谱宽估计得到的m(x):对比



（工资案例）方差估计：获得预测残差

（1）我们首先可以计算得到LL估计下的预测残差的平方 $\tilde{e}_i^2$

- 这一步可以直接采用前述的参考谱宽 $h_{rot} = 5.1442$

（工资案例）方差估计：再次获得参考谱宽

（2）然后开始计算方差估计下的最优谱宽。这里我们再进行一次参考谱宽的计算流程。

- 构建残差平方的多项式回归模型

$$\tilde{e}_i^2 = + \gamma_0 + \gamma_1 X_i + \gamma_2 X_i^2 + \gamma_3 X_i^3 + \gamma_4 X_i^4 + v_i$$

- 利用ROT公式流程，再次获得参考谱宽

$$hv_{\text{rot}} = 0.58 \left(\frac{\widehat{\sigma}^2 \left(\xi_2 - \xi_1 \right)}{n \widehat{B}} \right)^{1/5} = 0.58 \times \left(\frac{1.2384 \times (40 - 0)}{762 \times 0.000000300} \right)^{1/5} = 6.7708$$

（工资案例）方差估计：再次获得最优交叉验证谱宽

（3）然后开始计算方差估计下的最优谱宽。这里我们再进行一次参考谱宽的计算流程。

- 步骤1：设定**经验谱宽** $hv_{rot} = 6.7708$作为**初始值**。。
- 步骤2：设定**调参谱宽**（tuning bandwidth）。
- 一个经验谱宽范围可供参考： $[2.0, 40.0]$。
- 给定范围内的搜寻总数为 $n = 202$。则待评估序贯值为 $h \in (2.0000, 2.1891, 2.3781, 2.5672, 2.7562, \cdots, 39.4328, 39.6219, 39.8109, 40.0000)$。
- 步骤3：采用交叉验证**留一法**，分别遍历计算**NW估计**和**LL估计**下的全部CV值（见后面计算表）
- 步骤4：最小CV值对应的谱宽评估值，则为最优交叉验证谱宽。当然，我们最终发现**NW估计**和**LL估计**下的结果是一样的，都选择了最大边界值 $hv_{CV}(NW) = hv_{CV}(LL) = 40$。（见后面的CV比较图）

（工资案例）方差估计：CV值计算表（附表）

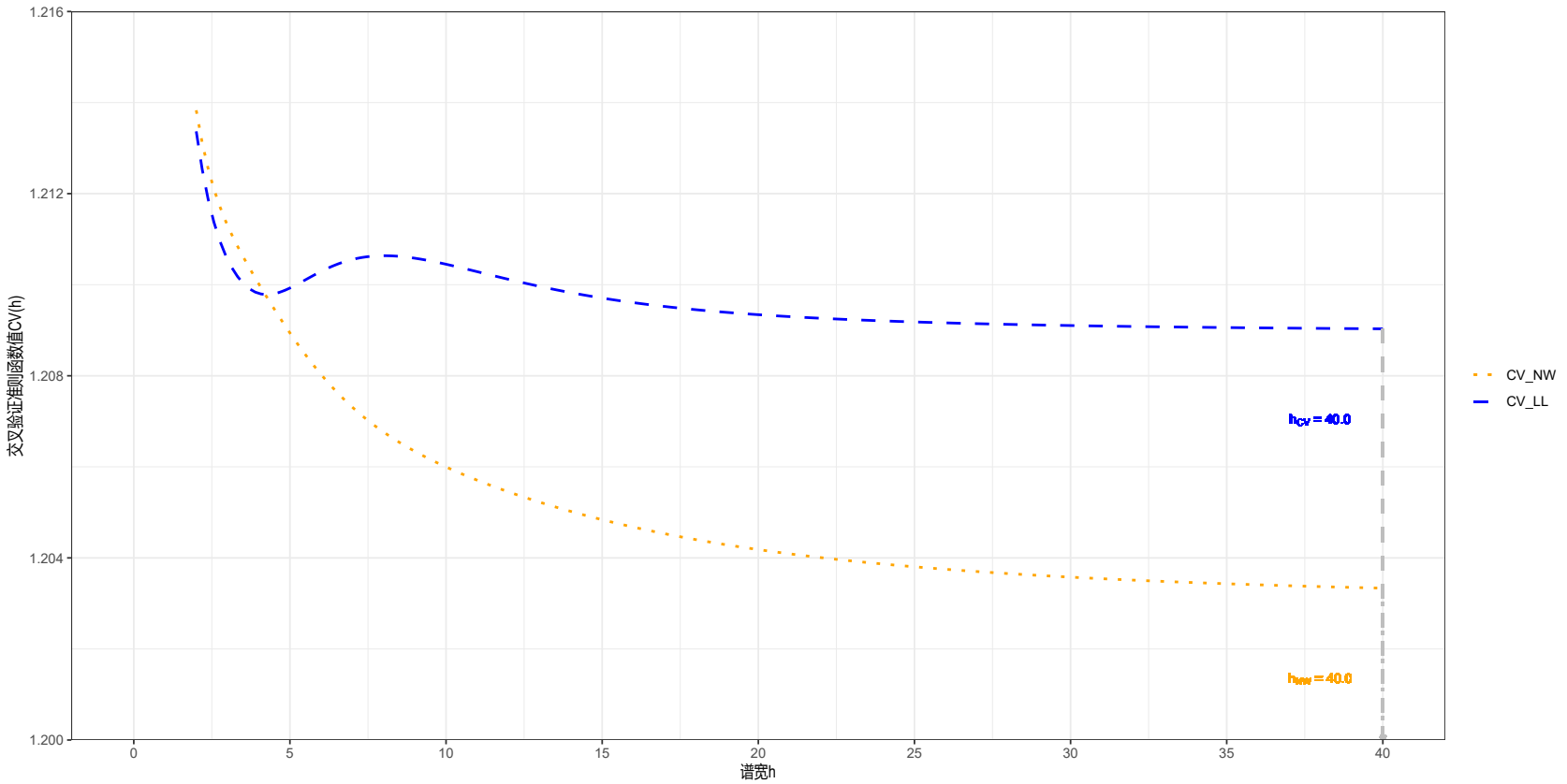
NW和LL方法下的CV计算表

id	h_tune	cv_NW	cv_LL
1	2.0000	1.2138	1.2134
2	2.1891	1.2131	1.2125
3	2.3781	1.2126	1.2119
4	2.5672	1.2121	1.2114
5	2.7562	1.2117	1.2110
6	2.9453	1.2114	1.2106
7	3.1343	1.2111	1.2104
8	3.3234	1.2108	1.2102

Showing 1 to 8 of 202 entries

Previous 1 2 3 4 5 … 26 Next

（工资案例）方差估计：CV值与谱宽（附图）



- LL估计的CV值在局部上具有最小值，也即局部最优谱宽约为 $hv_{cv}(LL) \simeq 5$
- 但是从全局来看，无论是NW估计，还是LL估计，CV函数值都表现为下降趋势。因此它们都选择了最大边界值 $hv_{CV}(NW) = hv_{CV}(LL) = 40$。

（工资案例）方差估计：计算方差、标准差

（4）利用前面的平方预测误差，并使用谱宽 $h=5.1442$进行LL估计，最终得到方差和标准差估计值（见后面附表）。

$$\widehat{\boldsymbol{V}}_{\hat{\beta}}(x)=\left(\boldsymbol{Z}'\boldsymbol{K}\boldsymbol{Z}\right)^{-1}\left(\sum_{i=1}^{n}K\left(\frac{X_i-x}{h}\right)^2 Z_i(x)Z_i(x)'\tilde{\boldsymbol{e}}_i^2\right)\left(\boldsymbol{Z}'\boldsymbol{K}\boldsymbol{Z}\right)^{-1}$$

（工资案例）方差估计：计算方差估计值（附表）

LL方法下的方差和标准差、置信区间

id	xg	mx	s	s2	lwr	upr
1	0	2.2050	0.1016	0.0103	2.0058	2.4042
2	0.2	2.2077	0.0979	0.0096	2.0159	2.3995
3	0.4	2.2102	0.0944	0.0089	2.0252	2.3952
4	0.6	2.2125	0.0911	0.0083	2.0340	2.3910
5	0.8	2.2146	0.0880	0.0078	2.0420	2.3872
6	1	2.2166	0.0852	0.0073	2.0496	2.3836
7	1.2	2.2184	0.0826	0.0068	2.0565	2.3803
8	1.4	2.2202	0.0802	0.0064	2.0629	2.3774

Showing 1 to 8 of 201 entries

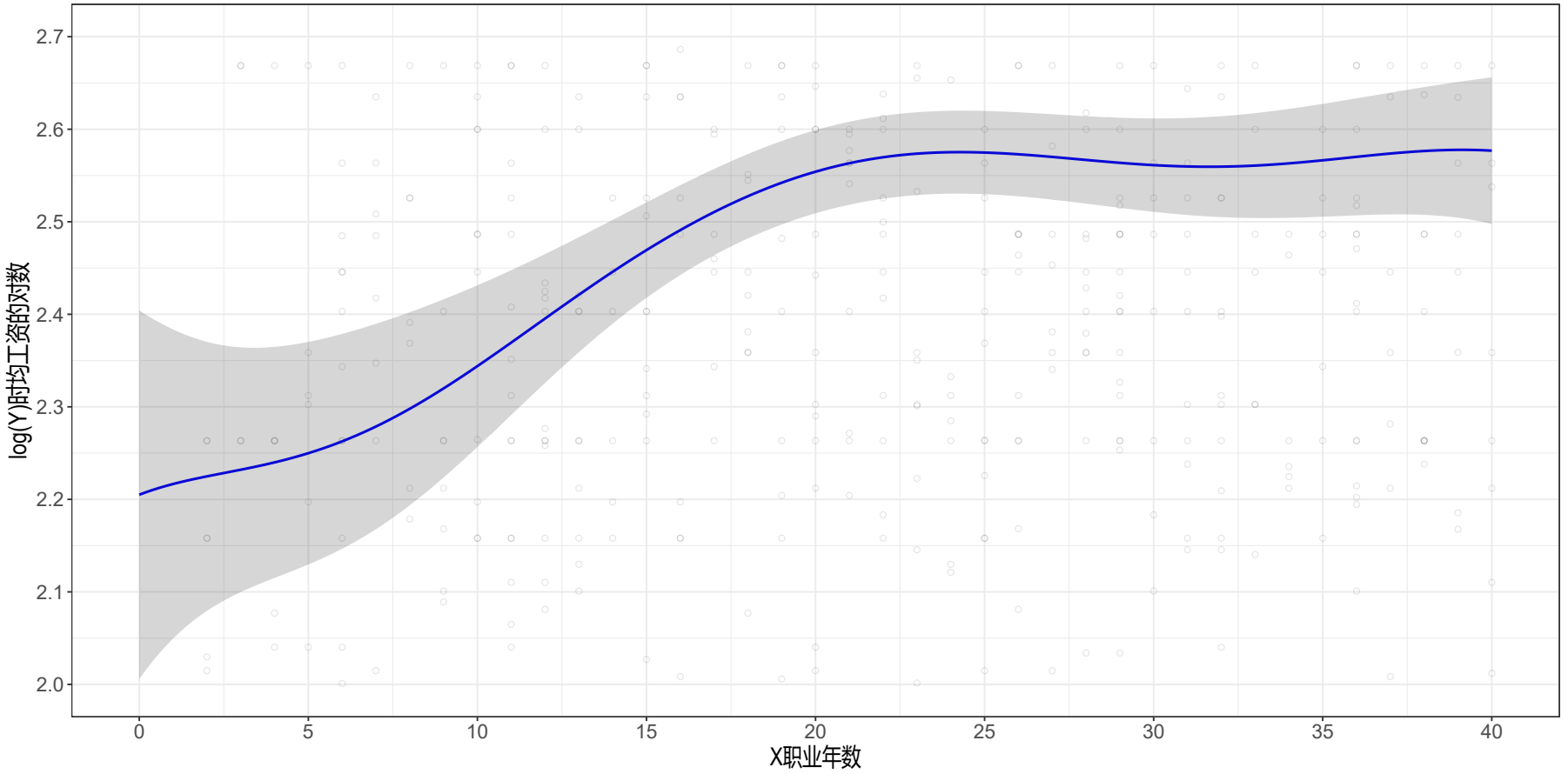
Previous 1 2 3 4 5 … 26 Next

（5）进一步计算**逐点置信区间**（Pointwise Confidence Interval）（见前面附表），并得到置信带（见后面附图）。

$$\widehat{m}(x) \pm z_{1-\alpha/2}(n-1) \cdot \sqrt{\widehat{V}_{\widehat{m}(x)}}$$

$$\widehat{m}(x) \pm 1.96 \sqrt{\widehat{V}_{\widehat{m}(x)}}$$

（工资案例）方差估计：置信区间和置信带（附图）



9.群组分析(Cluster observations)

9.1 基本原理

9.2 成绩案例

群组观测(Clustered observations)定义为：观测个体 $i=1,2,\cdots,n_g$ 分别处在群组 $g=1,2,\cdots,G$ 内，并可观测到变量对 $\left(Y_{ig},X_{ig}\right)$。

- 此时，可以定义模型为：

$$
\begin{aligned}
Y_{ig} &= m\left(X_{ig}\right)+e_{ig} \\
\mathbb{E}\left[e_{ig}\mid \boldsymbol{X}_g\right] &= 0
\end{aligned}
$$

> 其中：
>
> - $\boldsymbol{X}_g$ 是 X_{ig} 的堆栈形式（stacked）。
> - 并假定群组之间是相互独立的。

进一步地，我们定义如下：

- $\boldsymbol{X}_g$是 X_{ig}的堆栈形式（stacked）。
- $\boldsymbol{Y}_g$是 Y_{ig}的堆栈形式（stacked）。
- $\boldsymbol{Z}_g(x)$是 z_{ig}的堆栈形式（stacked），其中：

$$Z_{ig}(x) = \begin{pmatrix} 1 \\ X_{ig} - x \end{pmatrix}$$

- $\boldsymbol{K}_g(x) = \text{diag}\left\{K\left(\frac{X_{ig}-x}{h}\right)\right\}$

9.1 基本原理：LL估计

$$\begin{aligned}\widehat{\beta}(x) &= \left(\sum_{g=1}^{G}\sum_{i=1}^{n_g} K\left(\frac{X_{ig}-x}{h}\right) Z_{ig}(x)Z_{ig}(x)'\right)^{-1}\left(\sum_{g=1}^{G}\sum_{i=1}^{n_g} K\left(\frac{X_{ig}-x}{h}\right) Z_{ig}(x)Y_{ig}\right)\\ &= \left(\sum_{g=1}^{G} \boldsymbol{Z}_g(x)'\boldsymbol{K}_g(x)\boldsymbol{Z}_g(x)\right)^{-1}\left(\sum_{g=1}^{G} \boldsymbol{Z}_g(x)'\boldsymbol{K}_g(x)\boldsymbol{Y}_g\right).\end{aligned}$$

- 其中LL的估计量 $\hat{m}(x)=\widehat{\beta}_1(x)$等于上述估计的**截距项**。

4.1基本原理：删组回归法及其预测误差

为了得到**预测误差**（prediction error），我们可以采用**删组回归法**（delete cluster regression）进行遍历估计：

$$\tilde{\beta}_{(-g)}(x)=\left(\sum_{j\neq g}\boldsymbol{Z}_j(x)'\boldsymbol{K}_j(x)\boldsymbol{Z}_j(x)\right)^{-1}\left(\sum_{j\neq g}\boldsymbol{Z}_j(x)'\boldsymbol{K}_j(x)\boldsymbol{Y}_j\right)$$

然后得到 $m(x)$的估计量 $\tilde{m}_1(x)=\tilde{\beta}_{1(-g)}(x)$，并进一步得到观测个体 ig的**删组预测误差**（delete-cluster prediction error）：

$$\tilde{e}_{ig}=Y_{ig}-\tilde{\beta}_{1(-g)}\left(X_{ig}\right)$$

4.1 基本原理：条件方差

与之前类似，我们可以得到条件方差：

$$\boldsymbol{V}_{\widehat{\beta}}(x)=\left(\sum_{g=1}^{G}\boldsymbol{Z}_g(x)'\boldsymbol{K}_g(x)\boldsymbol{Z}_g(x)\right)^{-1}\left(\sum_{g=1}^{G}\boldsymbol{Z}_g(x)'\boldsymbol{K}_g(x)\boldsymbol{S}_g(x)\boldsymbol{K}_g(x)\boldsymbol{Z}_g(x)\right)\left(\sum_{g=1}^{G}\boldsymbol{Z}_g(x)'\boldsymbol{K}_g(x)\boldsymbol{Z}_g(x)\right)$$

- 其中：$\boldsymbol{S}_g=\mathbb{E}\left[\boldsymbol{e}_g\boldsymbol{e}_g'\mid\boldsymbol{X}_g\right]$，

因此，上述理论协方差矩阵可以通过$\boldsymbol{e}_g\boldsymbol{e}_g'$的估计量$\tilde{\boldsymbol{e}}_g\tilde{\boldsymbol{e}}_g'$计算得出，也即：

$$\widehat{\boldsymbol{V}}_{\widehat{\beta}}(x)=\left(\sum_{g=1}^{G}\boldsymbol{Z}_g(x)'\boldsymbol{K}_g(x)\boldsymbol{Z}_g(x)\right)^{-1}\left(\sum_{g=1}^{G}\boldsymbol{Z}_g(x)\boldsymbol{K}_g(x)\tilde{\boldsymbol{e}}_g\tilde{\boldsymbol{e}}_g'\boldsymbol{K}_g(x)\boldsymbol{Z}_g(x)\right)\left(\sum_{g=1}^{G}\boldsymbol{Z}_g(x)\boldsymbol{K}_g(x)\boldsymbol{Z}_g(x)\right)$$

$\widehat{m}(x)$的方差也就是上述估计得到的协方差的第1个对角线元素。

4.1 基本原理：交叉验证准则函数

与前面标准的**留一法**交叉验证略有不同，群组交叉验证准则函数可以表达为：

$$\mathrm{CV}(h) = \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\tilde{e}_{ig}^2$$

此时，最优CV谱宽将出现在上述函数的最小值处：

$$h_{\mathrm{CV}} = \underset{h \geq h_\ell}{\operatorname{argmin}}\, \mathrm{CV}(h)$$

4.2（成绩案例）：背景说明

成绩案例：

- 案例基于(Duflo, Dupas, and Kremer, 2011)学生排位追踪（percentile tracking）对成绩（testscore）影响的数据集。其中**学生排位追踪**为分位数变量（1-100之间）。我们这里重点分析其中的**子样本数据**（女生、实施了学生排位追踪），样本数为n=1487。
- 关注的问题：**学生排位跟踪**（`X=percentile`）对**学生成绩**（`Y=testscore`）的非参数回归估计，并且我们对根据**学生所在学校**（school ID）对样本进行了分组。
- 后面的分析中，我们会重点划定观测窗口为：**成绩**范围 $[0, 40]$。非参数估计中我们会采用基于**高斯核函数**（Gaussian Kernel）的局部线性回归LL。

4.2（成绩案例）样本数据集

数据集(n=1487)

obs	schoolid	Y	X
1	430	2.9000	3.2051
2	430	9.7429	12.1795
3	430	13.2071	16.0256
4	430	14.2500	17.3077
5	430	13.8286	18.5897
6	430	8.5071	21.1538
7	430	9.2286	22.4359
8	430	13.5714	26.2820

Showing 1 to 8 of 1,487 entries

Previous 1 2 3 4 5 … 186 Next

- 样本数据的描述性统计（未分群组）如下：

```
  schoolid                Y
Length:1487        Min.   : 0   Min.
Class :character   1st Qu.: 7   1st Qu
Mode  :character   Median :13   Median
                   Mean   :14   Mean
                   3rd Qu.:21   3rd Qu
                   Max.   :40   Max.
```

4.2（成绩案例）样本数据集：群组描述性统计

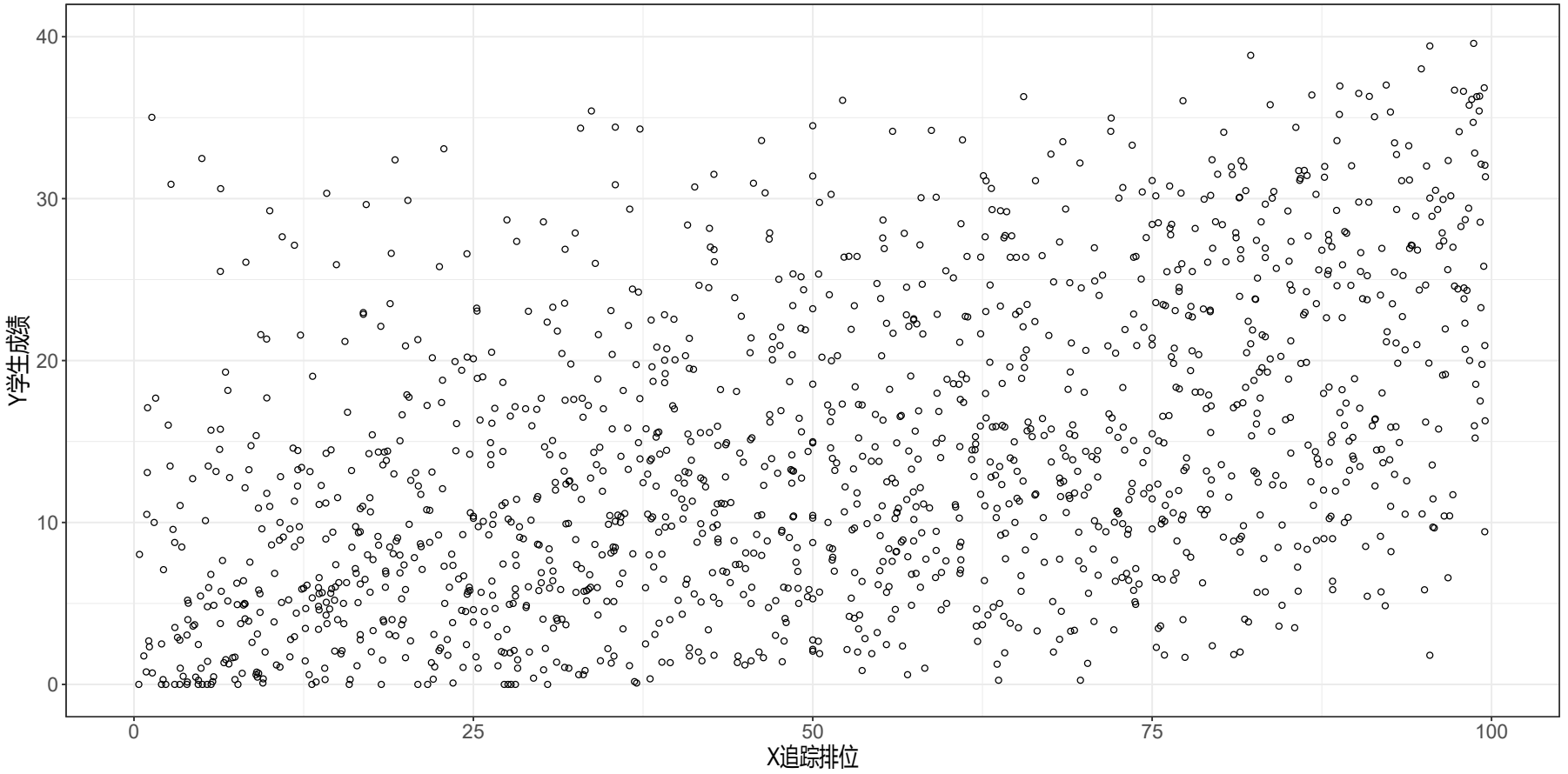
学生所在学校的群组描述性统计(q=60)

obs	schoolid	n	x_mean	x_min	x_max	x_sd	y_mean	y_min	y_max
1	1006	28	55.38	4.94	95.68	28.49	8.61	1.00	22.16
2	1012	21	52.43	7.61	98.91	27.19	13.51	4.03	36.50
3	1014	26	52.76	3.52	99.30	30.63	10.42	1.00	19.76
4	1015	21	53.25	9.17	99.17	27.10	17.87	3.69	31.96
5	1020	22	49.34	8.18	99.09	30.73	17.32	0.86	35.41
6	430	21	43.59	3.21	96.79	28.93	11.48	2.83	30.05
7	432	15	59.52	3.97	96.03	23.88	13.58	4.00	29.33
8	436	21	60.83	14.79	96.48	25.99	15.00	0.26	28.16

Showing 1 to 8 of 60 entries

Previous 1 2 3 4 5 … 8 Next

4.2（成绩案例）样本数据散点图



4.2（成绩案例）参考谱宽的计算：多项式回归

下面我们按前述步骤来计算参考谱宽值 h_{rot}：

- **步骤1**：根据案例数据集，设定权重取值范围 $\{\xi_1 \leq x \leq \xi_2\} = \{0, 100\}$
- **步骤2**：构建多项式回归

$$Y_i = + \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3 + \beta_5 X_i^4 + u_i$$

> 直接使用OLS进行估计，得到估计方程：

$$\begin{array}{lllll} \widehat{Y} = + 6.815543 + 0.078095 X_i + 0.004483 X_i^2 - 0.000100 X_i^3 + 0.000001 X_i^4 \\ (s) \quad (1.2764) \quad (0.1656) \quad (0.0065) \quad (0.0001) \quad (0.0000) \end{array}$$

> 进而得到拟合值 $\widehat{m}(x)$及其二阶导 $\widehat{m}''(x)$及残差 $\hat{\epsilon}$

$$\begin{aligned} \widehat{m}(x) &= 6.815543 + 0.078095 x_i + 0.004483 x_i^2 - 0.000100 x_i^3 + 0.000001 x_i^4 \\ \widehat{m}''(x) &= 2 \times 0.004483 - 6 \times 0.000100 x_i^3 + 12 \times 0.000001 x_i^2 \end{aligned}$$

4.2（成绩案例）参考谱宽的计算：结果

- 步骤3：利用上述估计结果计算

$$\widehat{B}=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}\hat{m}''\left(X_i\right)\right)^2 1\left\{\xi_1\leq X_i\leq\xi_2\right\}=2098.61641815$$

- 步骤4：多项式模型的回归误差方差

$$\hat{\sigma}^2=\frac{\sum\hat{\epsilon}^2}{n-q-1}=66.4390$$

- 步骤5：根据上述全部结果计算得到经验谱宽：

$$h_{\text{rot}}=0.58\left(\frac{\widehat{\sigma}^2\left(\xi_2-\xi_1\right)}{n\widehat{B}}\right)^{1/5}=0.58\times\left(\frac{66.4390\times(1-0)}{1487\times2098.616418148}\right)^{1/5}=6.7463$$

4.2（成绩案例）交叉验证谱宽的计算：规则

- 步骤1：设定**经验谱宽** $h_{rot}=6.7463$作为**初始值**。。
- 步骤2：设定**调参谱宽**（tuning bandwidth）。
- 人为选定谱宽范围：$[4,20]$。
- 给定范围内的搜寻总数为 $n=202$。则待评估序贯值为
  $h \in (4.0000, 4.0796, 4.1592, 4.2388, 4.3184, \cdots, 19.7612, 19.8408, 19.9204, 20.0000)$
  。

4.2（成绩案例）交叉验证谱宽的计算：CV计算表

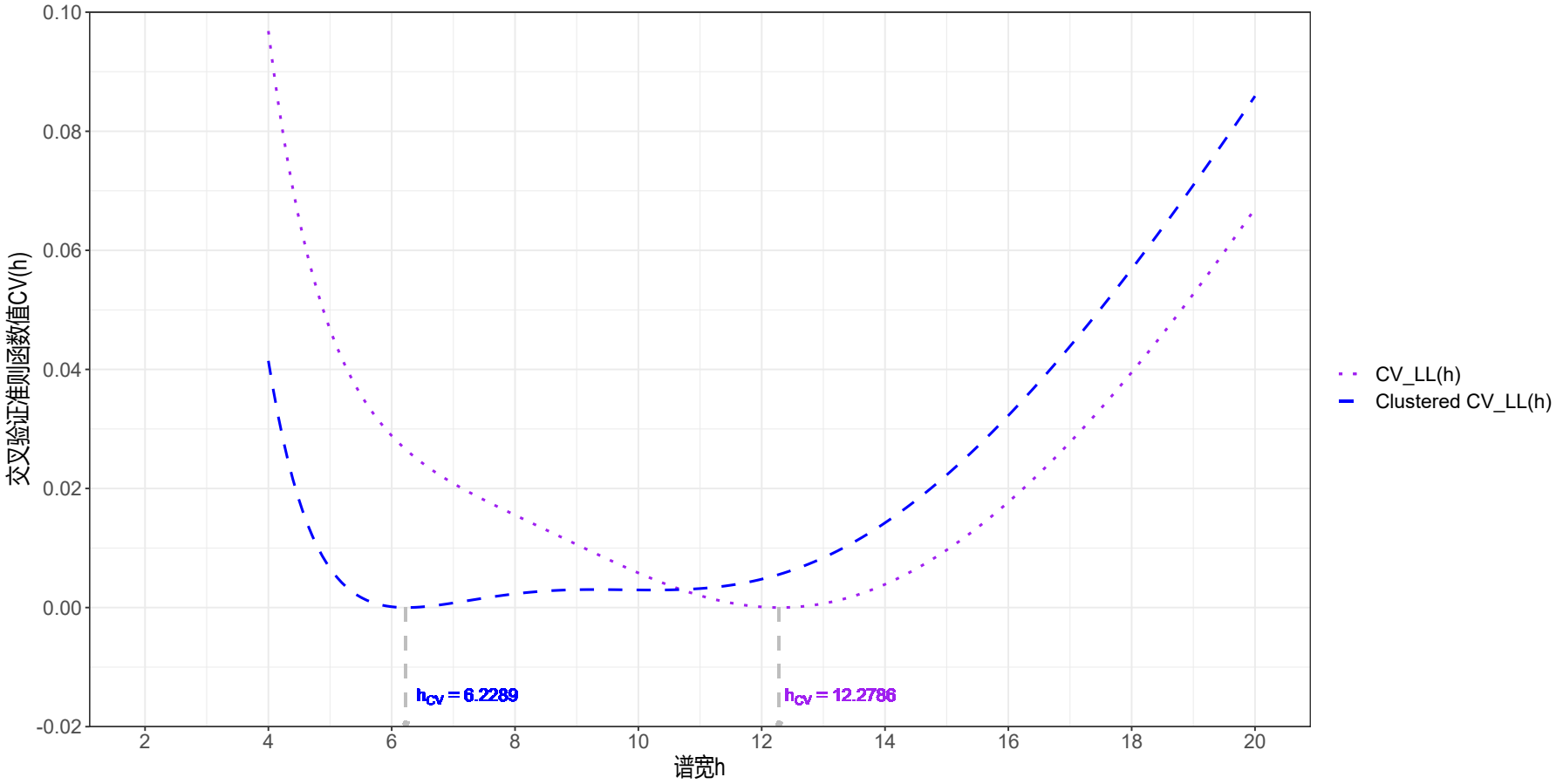
两种CV方法下的计算表

id	h_tune	cv_LL	cv_LLC
1	4.0000	0.0968	0.0414
2	4.0796	0.0906	0.0366
3	4.1592	0.0848	0.0322
4	4.2388	0.0795	0.0282
5	4.3184	0.0747	0.0247
6	4.3980	0.0703	0.0215
7	4.4776	0.0662	0.0187
8	4.5572	0.0625	0.0162

Showing 1 to 8 of 202 entries

Previous 1 2 3 4 5 … 26 Next

4.2（成绩案例）交叉验证谱宽的计算：谱宽与CV变化（对比）



4.2（成绩案例）交叉验证谱宽的计算：谱宽与CV变化（对比）

- 为了使得两种方法具有可比较性，此处对CV值做了去最小化的尺度变换，也即$cv^* = cv - min(cv)$。
- 常规的（未分组）局部线性LL交叉验证最优谱宽结果为12.2786，且CV函数相对比较陡峭；而群组化的（按学校分组）局部线性LL交叉验证最优谱宽结果为6.2289，并且在$[5, 11]$之间CV函数值表现得比较平稳。

4.2（成绩案例）CEF m(x)估计：基于群组CV最优谱宽（计算表）

使用不同谱宽下LL方法对m(x)的估计结果

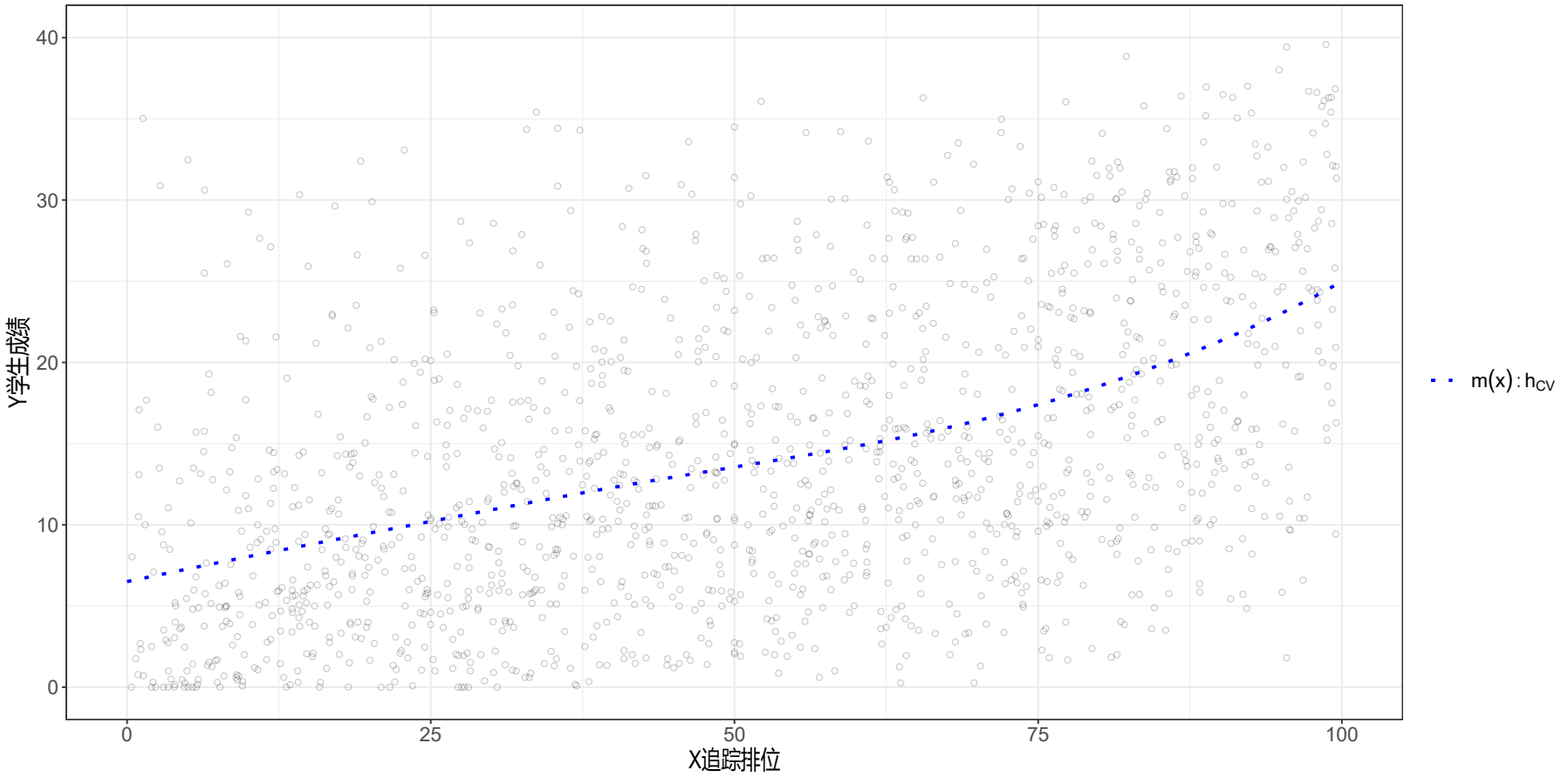
index	xg	mx_rot	mx_cv
1	0.00	6.5023	6.6031
2	0.50	6.5801	6.6335
3	1.00	6.6580	6.6711
4	1.50	6.7360	6.7155
5	2.00	6.8140	6.7659
6	2.50	6.8920	6.8219

Showing 1 to 6 of 201 entries

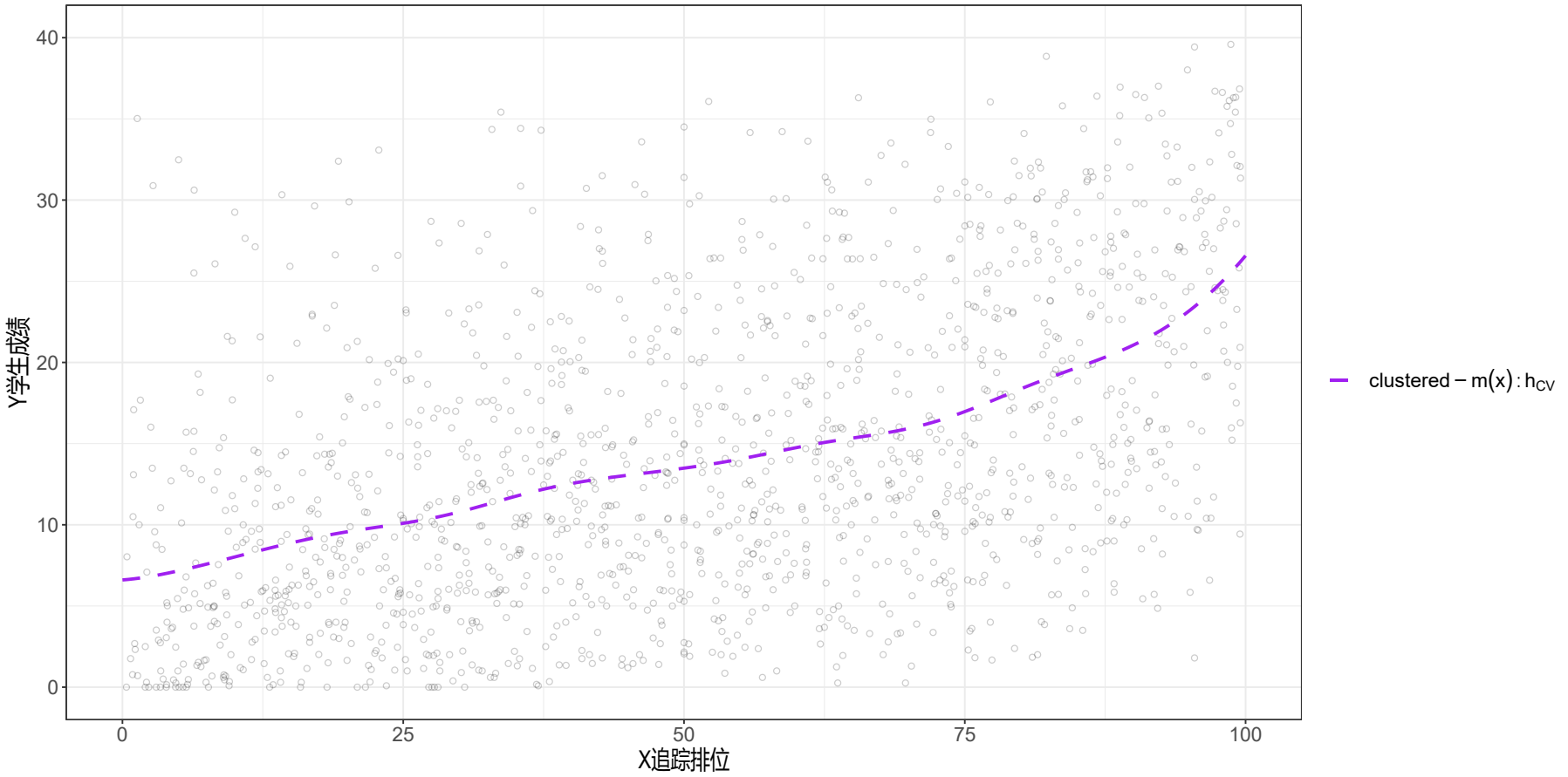
Previous 1 2 3 4 5 … 34 Next

- 我们采用群组化的（按学校分组）局部线性LL交叉验证最优谱宽结果为6.2289进行CEF估算 $\widehat{m}(x)$

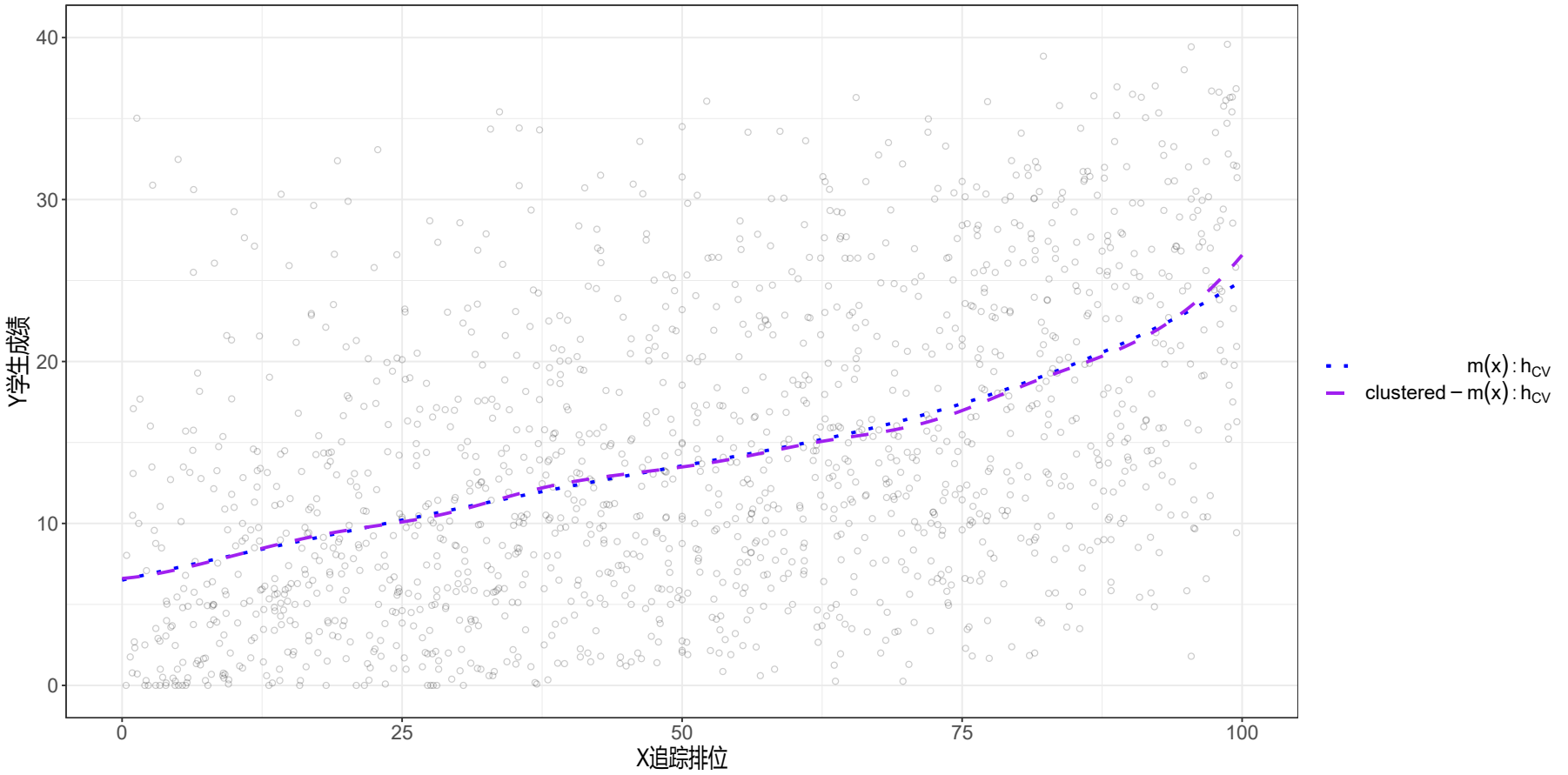
4.2（成绩案例）CEF m(x)估计：基于常规局部线性LL方法



4.2（成绩案例）CEF m(x)估计：基于群组局部线性LL方法



4.2（成绩案例）CEF m(x)估计：两种局部线性LL方法对比



4.2（成绩案例）方差估计：计算方差、标准差

（4）直接使用CEF估计中**群组LL交叉验证**最优谱宽[a] $h = 6.2289$进行局部线性LL估计，并利用删组法计算得到预测误差 $\tilde{\boldsymbol{e}}_g$，并最终分别得到未分组的和群组化的协方差矩阵（见下式），从而得到CEF估计值的方差和标准差（见后面附表）。

$$\widehat{\boldsymbol{V}}_{\hat{\beta}}(x) = \left(\sum_{g=1}^{G} \boldsymbol{Z}_g(x)'\boldsymbol{K}_g(x)\boldsymbol{Z}_g(x)\right)^{-1} \left(\sum_{g=1}^{G} \boldsymbol{Z}_g(x)\boldsymbol{K}_g(x)\tilde{\boldsymbol{e}}_g\tilde{\boldsymbol{e}}_g'\boldsymbol{K}_g(x)\boldsymbol{Z}_g(x)\right) \left(\sum_{g=1}^{G} \boldsymbol{Z}_g(x)\boldsymbol{K}_g(x)\boldsymbol{Z}_g(x)\right)$$

[a] 这里我们没有再次评估条件方差估计中的最优谱宽，而是简单直接地使用了CEF估计的交叉验证最优谱宽。但是我们还是要注意，二者的最优谱宽可以完全不相同！

4.2（成绩案例）方差估计：计算方差估计值（附表）

LL方法下的方差和标准差、置信区间

id	xg	mx	mxl	s_uc	s2_uc	lwr_uc	upr_uc	s	s2	lwr
1	0	6.6031	5.9789	1.5141	2.2924	3.6355	9.5707	1.6551	2.7392	3.3592
2	0.5	6.6335	6.0585	1.4074	1.9807	3.8751	9.3919	1.5694	2.4631	3.5574
3	1	6.6711	6.1381	1.3087	1.7127	4.1061	9.2362	1.4909	2.2229	3.7489
4	1.5	6.7155	6.2178	1.2178	1.4831	4.3285	9.1024	1.4191	2.0139	3.9340
5	2	6.7659	6.2974	1.1343	1.2867	4.5426	8.9892	1.3536	1.8322	4.1129
6	2.5	6.8219	6.3770	1.0579	1.1191	4.7485	8.8953	1.2939	1.6740	4.2860
7	3	6.8830	6.4567	0.9880	0.9762	4.9464	8.8195	1.2394	1.5362	4.4537
8	3.5	6.9485	6.5363	0.9245	0.8547	5.1365	8.7605	1.1899	1.4158	4.6163

Showing 1 to 8 of 201 entries

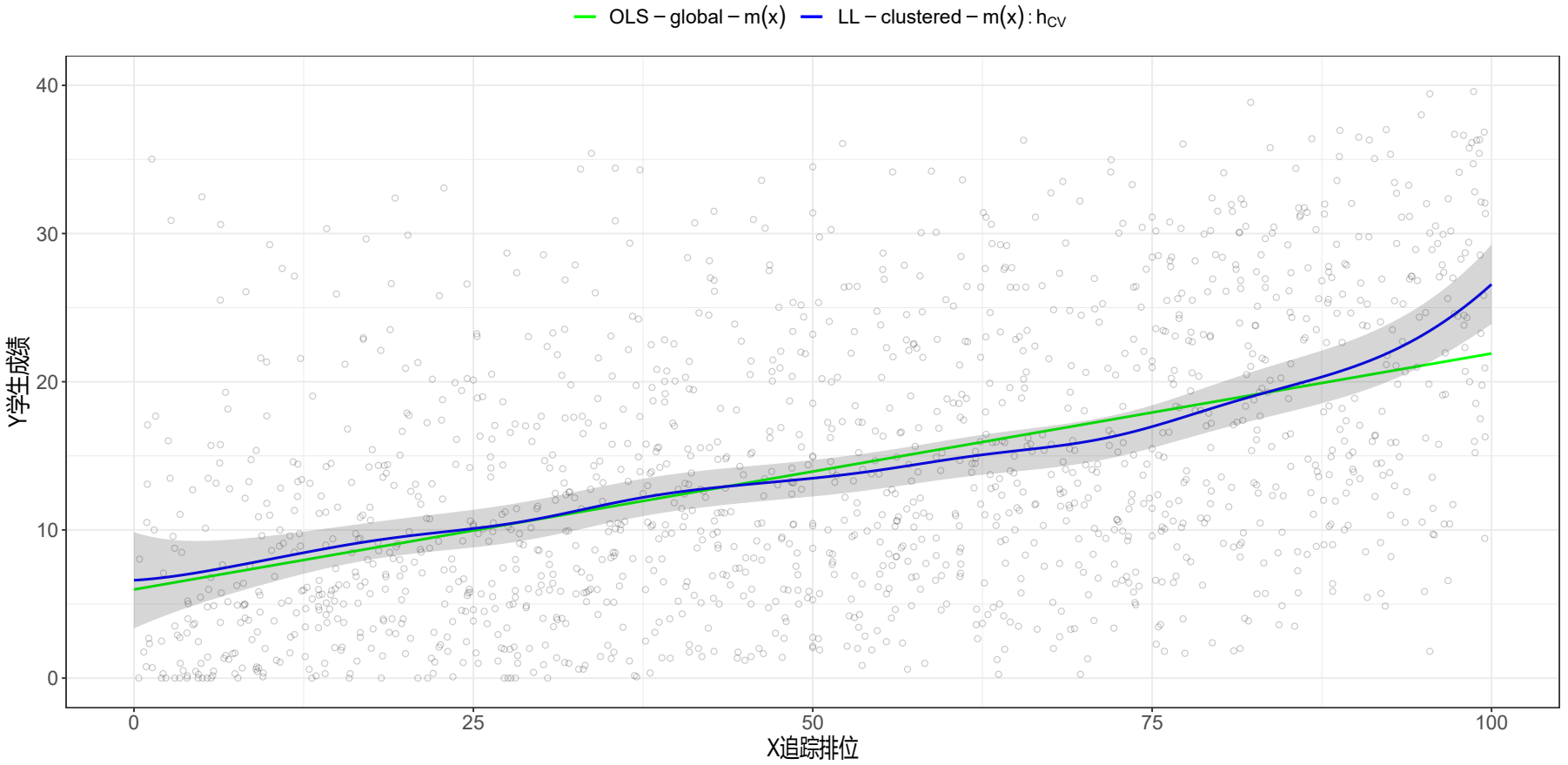
Previous 1 2 3 4 5 … 26 Next

（5）进一步计算**群组局部线性估计**下的**逐点置信区间**（Pointwise Confidence Interval）（见前面附表），并得到置信带（见后面附图）。

$$\widehat{m}(x) \pm z_{1-\alpha/2}(n-1) \cdot \sqrt{\widehat{V}_{\widehat{m}(x)}}$$

$$\widehat{m}(x) \pm 1.96 \sqrt{\widehat{V}_{\widehat{m}(x)}}$$

4.2（成绩案例）置信区间和置信带（附图）



4.2（成绩案例）置信区间和置信带：OLS回归结果（附录）

作为比较，我们还简单使用了全样本OLS估计得到估计值（绿色）

- 全局OLS回归模型为

$$Y_i = + \beta_1 + \beta_2 X_i + u_i$$

- 直接使用OLS进行估计，得到估计结果：

$$\begin{array}{lll} \widehat{Y} = & +5.9789 & +0.1593 X_i \\ (s) & (0.4472) & (0.0077) \\ (t) & (+13.37) & (+20.74) \\ (over) & n = 1487 & \hat{\sigma} = 8.2000 \\ (fit) & R^2 = 0.2247 \bar{R}^2 = 0.2242 & \\ (Ftest) & F^* = 430.35 p = 0.0000 & \end{array}$$

本章參考文献

参考文献（References）：1/2

Duflo, E., P. Dupas, and M. Kremer (2011). "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya". In: *American economic review* 101.5, pp. 1739-74.

本章结束