

# 第 4 章实验建模

韩建伟

信息学院

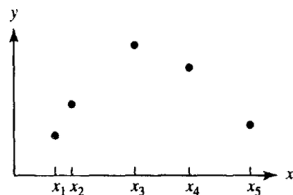
hanjianwei@zjgsu.edu.cn

2018/10/26

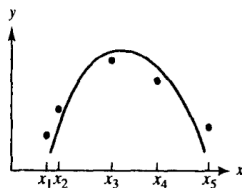
# 拟合和插值的选择

**拟合** 建模者利用一些假定来选择—个特定的模型类型，以解释观测值反应的状况.

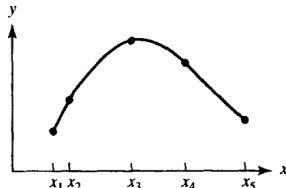
**插值** 建模者不可能构造一个满意地解释已知状况的易于处理的模型.



(a) 数据



(b) 拟合

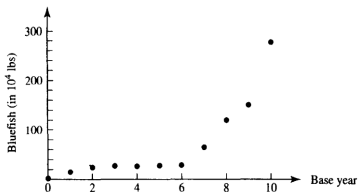


(c) 插值

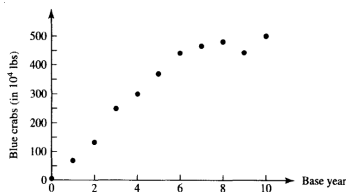
# Chesapeake 海湾的收成

| Year | Bluefish (lbs) | Blue crabs (lbs) |
|------|----------------|------------------|
| 1940 | 15,000         | 100,000          |
| 1945 | 150,000        | 850,000          |
| 1950 | 250,000        | 1,330,000        |
| 1955 | 275,000        | 2,500,000        |
| 1960 | 270,000        | 3,000,000        |
| 1965 | 280,000        | 3,700,000        |
| 1970 | 290,000        | 4,400,000        |
| 1975 | 650,000        | 4,660,000        |
| 1980 | 1,200,000      | 4,800,000        |
| 1985 | 1,500,000      | 4,420,000        |
| 1990 | 2,750,000      | 5,000,000        |

(a) 收成数据



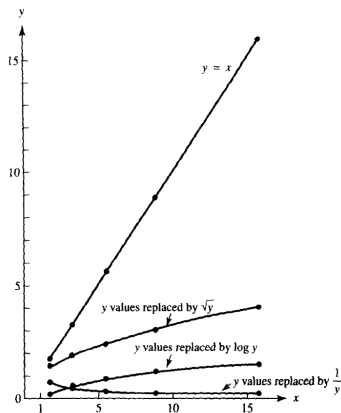
(b) 蓝鱼散点图



(c) 蓝蟹散点图

图: 1992 年《每日评论》收集的过去 50 年中 Chesapeake 海湾的海产收成。

# 如何对数据进行变换



(a) 三种变换

$$\begin{array}{c}
 \vdots \\
 z^3 \\
 z^2 \\
 z \text{ (no change)} \\
 * \left\{ \begin{array}{l} \sqrt{z} \\ \log z \\ -\frac{1}{\sqrt{z}} \\ -\frac{1}{z} \\ -\frac{1}{z^2} \end{array} \right. \\
 \vdots
 \end{array}$$

\*Most often used transformations

(b) 变换阶梯

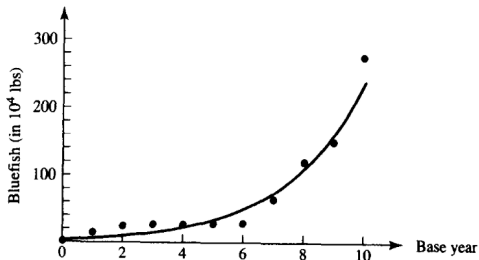
# 收获蓝鱼

- “向下压”曲线的右端，用最小二乘法拟合下列模型：

$$\log y = mx + b$$

| Year | Base year | Bluefish (lbs) |
|------|-----------|----------------|
|      | $x$       | $y$            |
| 1940 | 0         | 15,000         |
| 1945 | 1         | 150,000        |
| 1950 | 2         | 250,000        |
| 1955 | 3         | 275,000        |
| 1960 | 4         | 270,000        |
| 1965 | 5         | 280,000        |
| 1970 | 6         | 290,000        |
| 1975 | 7         | 650,000        |
| 1980 | 8         | 1,200,000      |
| 1985 | 9         | 1,550,000      |
| 1990 | 10        | 2,750,000      |

(a) 蓝鱼数据



(b) 蓝鱼拟合  $y = 5.457(1.4635)^x$

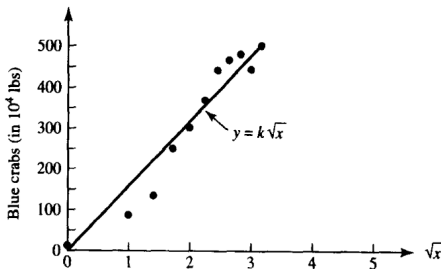
# 收获蓝蟹

- “向上扳”曲线的右端，用最小二乘法拟合下列模型：

$$y = k\sqrt{x}$$

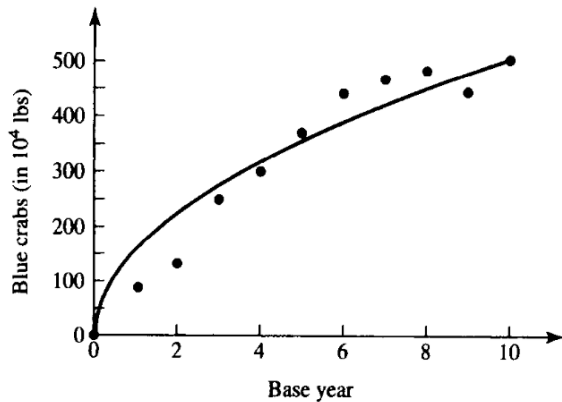
| Year | Base year | Blue crabs (lbs) |
|------|-----------|------------------|
|      | $x$       | $y$              |
| 1940 | 0         | 100,000          |
| 1945 | 1         | 850,000          |
| 1950 | 2         | 1,330,000        |
| 1955 | 3         | 2,500,000        |
| 1960 | 4         | 3,000,000        |
| 1965 | 5         | 3,700,000        |
| 1970 | 6         | 4,400,000        |
| 1975 | 7         | 4,660,000        |
| 1980 | 8         | 4,800,000        |
| 1985 | 9         | 4,420,000        |
| 1990 | 10        | 5,000,000        |

(a) 蓝蟹数据



(b) 蓝蟹拟合直线  $y = 158.344\sqrt{x}$

## 蓝蟹最终拟合数据



# 验证模型

- 针对每一对数据，计算残差和相对误差.
- 对未来进行预测或者外推，这些模型是否依然适用.
- 这些简单的单项模型应该用于插值而不是外推.



# 本节的想法

构造一个模型时，从细心分析收集到的数据开始：

- ① 看数据存在什么样的倾向
- ② 是否有明显处于倾向以外的点？如果有这样的异常值，想想是否抛弃它们？如果是实验观测到的，重复该实验做一个数据收集错误的检查.
- ③ 当倾向确实清楚存在时，下一步是找到一个将数据变换成一（近似）直线的函数.
- ④ 警惕变换的使用如何带来欺骗性，特别是当数据点集中在一起时.
- ⑤ 最后，用第 3 章讨论的指示量分析拟合优度，记住对原始的而不是变换后的数据画出所提供的模型.
- ⑥ 如果对这一拟合不满意，可以研究其它的单项模型.

# 高阶多项式模型

- 单项模型易于进行模型分析，包括敏感性分析、优化、变化率以及曲线下面积的估计。
- 单项模型的可用性有限。
- 解决方案：多项模型。

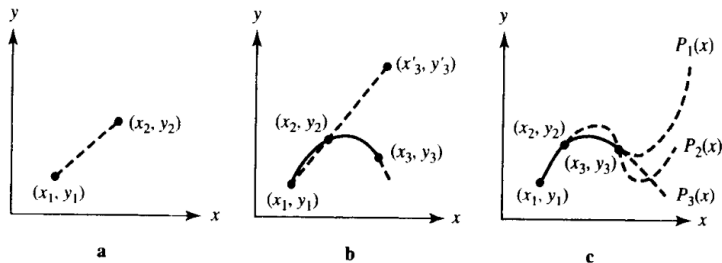


图: 通过三个数据点且最高阶为 2 的多项式只有唯一一个，而阶高于 2 的多项式有无穷多个

## 录音机的播放时间

| $c_i$             | 100 | 200 | 300 | 400 | 500  | 600  | 700  | 800  |
|-------------------|-----|-----|-----|-----|------|------|------|------|
| $t_i(\text{sec})$ | 205 | 430 | 677 | 945 | 1233 | 1542 | 1872 | 2224 |

图: 录音机计数器和播放时间关系表

我们有 8 个数据点, 应期望一个最高阶为 7 的唯一多项式, 记为

$$P_7(c) = a_0 + a_1 c + a_2 c^2 + a_3 c^3 + a_4 c^4 + a_5 c^5 + a_6 c^6 + a_7 c^7$$

把 8 个数据点代入

$$205 = a_0 + a_1 1 + a_2 1^2 + a_3 1^3 + a_4 1^4 + a_5 1^5 + a_6 1^6 + a_7 1^7$$

$$430 = a_0 + a_1 2 + a_2 2^2 + a_3 2^3 + a_4 2^4 + a_5 2^5 + a_6 2^6 + a_7 2^7$$

$$\vdots$$

$$2224 = a_0 + a_1 8 + a_2 8^2 + a_3 8^3 + a_4 8^4 + a_5 8^5 + a_6 8^6 + a_7 8^7$$

# 系数求解

$$\begin{aligned}a_0 &= -13.9999923 & a_4 &= -5.354166491 \\a_1 &= 232.9119031 & a_5 &= 0.8013888621 \\a_2 &= -29.08333188 & a_6 &= -0.0624999978 \\a_3 &= 19.78472156 & a_7 &= 0.0019841269\end{aligned}$$

| $c_i$      | 100 | 200 | 300 | 400 | 500  | 600  | 700  | 800  |
|------------|-----|-----|-----|-----|------|------|------|------|
| $t_i$      | 205 | 430 | 677 | 945 | 1233 | 1542 | 1872 | 2224 |
| $P_7(c_i)$ | 205 | 430 | 677 | 945 | 1233 | 1542 | 1872 | 2224 |

考虑第 3 章提出的评判标准... 该模型真的是最好的吗?

# 多项式的拉格朗日形式

## 定理 1

如果  $x_0, x_1, \dots, x_n$  是  $n+1$  个不同的点, 而  $y_0, y_1, \dots, y_n$  是这些点上对应的观测值, 那么, 存在一个唯一的最高阶为  $n$  的多项式  $P(x)$ , 具有性质

$$y_k = P(x_k), k = 0, 1, \dots, n$$

这一多项式由下式给定

$$P(x) = y_0 L_0(x) + \dots + y_n L_n(x)$$

其中

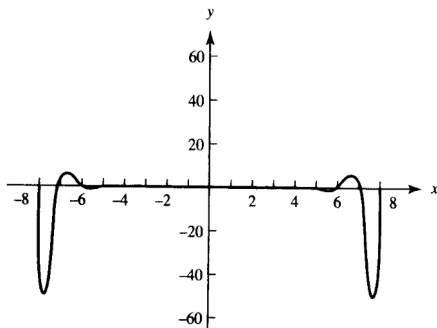
$$L_k(x) = \frac{(x - x_0)(x - x_1) \dots (x - x_{k-1})(x - x_{k+1}) \dots (x - x_n)}{(x_k - x_0)(x_k - x_1) \dots (x_k - x_{k-1})(x_k - x_{k+1}) \dots (x_k - x_n)}$$

# 高阶多项式的优缺点

## 缺点 1

在端点附近多项式有严重摆动, 进行数据预测时有很大问题

|       |    |    |    |    |    |    |    |    |   |   |   |   |   |   |   |   |   |
|-------|----|----|----|----|----|----|----|----|---|---|---|---|---|---|---|---|---|
| $x_i$ | -8 | -7 | -6 | -5 | -4 | -3 | -2 | -1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| $y_i$ | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

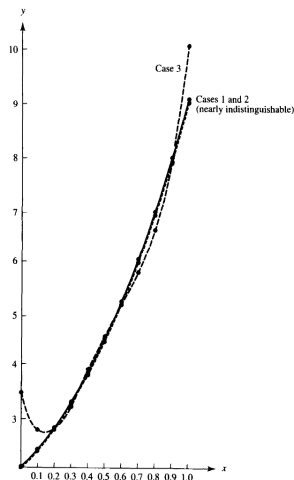


# 高阶多项式的优缺点

## 缺点 2

对数据微小变换非常敏感

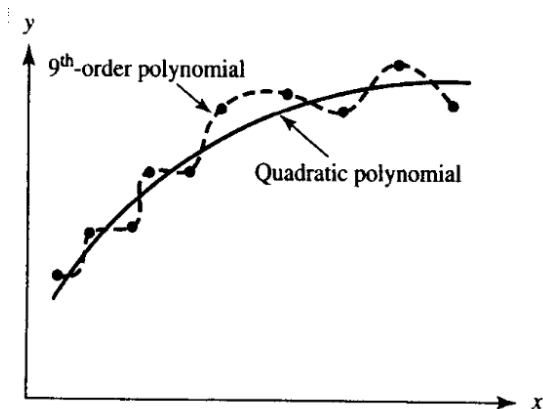
| $x_i$         | 0.2    | 0.3    | 0.4    | 0.6    | 0.9    |
|---------------|--------|--------|--------|--------|--------|
| Case 1: $y_i$ | 2.7536 | 3.2411 | 3.8016 | 5.1536 | 7.8671 |
| Case 2: $y_i$ | 2.754  | 3.241  | 3.802  | 5.154  | 7.867  |
| Case 3: $y_i$ | 2.7536 | 3.2411 | 3.8916 | 5.1536 | 7.8671 |



# 光滑化：低阶多项式模型

## 光滑化

不管数据的个数, 选取一个低阶多项式, 通过第 3 章的准则拟合给定的数据.





## 再论录音机的播放时间

|                   |     |     |     |     |      |      |      |      |
|-------------------|-----|-----|-----|-----|------|------|------|------|
| $c_i$             | 100 | 200 | 300 | 400 | 500  | 600  | 700  | 800  |
| $t_i(\text{sec})$ | 205 | 430 | 677 | 945 | 1233 | 1542 | 1872 | 2224 |

图: 录音机计数器和播放时间关系表

拟合一个下列形式的二阶多项式

$$P_2(c) = a + bc + dc^2$$

用最小二乘法确定  $a$ ,  $b$ ,  $c$ .

# 最小二乘法拟合

$$\text{Minimize } S = \sum_{i=1}^m [t_i - (a + bc_i + dc_i^2)]^2$$

存在极小点的必要条件

$$\frac{\partial S}{\partial a} = \frac{\partial S}{\partial b} = \frac{\partial S}{\partial d} = 0$$

故

$$ma + (\sum c_i)b + (\sum c_i^2)d = \sum t_i$$

$$(\sum c_i)a + (\sum c_i^2)b + (\sum c_i^3)d = \sum c_i t_i$$

$$(\sum c_i^2)a + (\sum c_i^3)b + (\sum c_i^4)d = \sum c_i^2 t_i$$

# 拟合结果

$$P_2(c) = 0.14286 + 1.94226c + 0.00105c^2$$

|                  |       |        |       |       |       |        |        |       |
|------------------|-------|--------|-------|-------|-------|--------|--------|-------|
| $c_i$            | 100   | 200    | 300   | 400   | 500   | 600    | 700    | 800   |
| $t_i$            | 205   | 430    | 677   | 945   | 1233  | 1542   | 1872   | 2224  |
| $t_i - P_2(c_i)$ | 0.167 | -0.452 | 0.000 | 0.524 | 0.119 | -0.214 | -0.476 | 0.333 |

- ① 应该用一个多项式吗？
- ② 如果应该，几阶的多项式合适？

# 多项式阶数的确定

$$P(x) = a + bx + cx^2$$

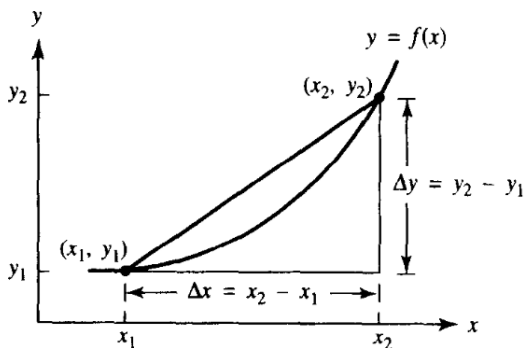
$$P'(x) = b + 2cx$$

$$P''(x) = 2c$$

$$P'''(x) = 0$$

# 导数的离散形式

$$\frac{dy}{dx} = \lim_{\Delta x \rightarrow 0} \frac{\Delta y}{\Delta x}$$



# 离散求导示例

| Data  |       | Differences |            |            |            |
|-------|-------|-------------|------------|------------|------------|
| $x_i$ | $y_i$ | $\Delta$    | $\Delta^2$ | $\Delta^3$ | $\Delta^4$ |
| 0     | 0     |             |            |            |            |
| 2     | 4     | 4           |            |            |            |
| 4     | 16    | 12          | 8          | 0          |            |
| 6     | 36    | 20          | 8          | 0          | 0          |
| 8     | 64    | 28          | 8          |            |            |

| Data  |       | First<br>divided difference   | Second<br>divided difference  |
|-------|-------|-------------------------------|---|
| $x_1$ | $y_1$ | $\frac{y_2 - y_1}{x_2 - x_1}$ |   |
| $x_2$ | $y_2$ |                               | $\frac{\frac{y_3 - y_2}{x_3 - x_2} - \frac{y_2 - y_1}{x_2 - x_1}}{x_3 - x_1}$ |
| $x_3$ | $y_3$ |                               |   |

图：一阶、二阶导数的离散计算

## 再论录音机的播放时间

| Data  |       | Divided differences |            |            |            |
|-------|-------|---------------------|------------|------------|------------|
| $x_i$ | $y_i$ | $\Delta$            | $\Delta^2$ | $\Delta^3$ | $\Delta^4$ |
| 100   | 205   | 2.2500              |            |            |            |
| 200   | 430   | 2.4700              | 0.0011     |            |            |
| 300   | 677   | 2.6800              | 0.0011     | 0.0000     |            |
| 400   | 945   | 2.8800              | 0.0010     | 0.0000     | 0.0000     |
| 500   | 1233  | 2.8800              | 0.0011     | 0.0000     | 0.0000     |
| 600   | 1542  | 3.0900              | 0.0011     | 0.0000     | 0.0000     |
| 700   | 1872  | 3.3000              | 0.0011     | 0.0000     | 0.0000     |
| 800   | 2224  | 3.5200              |            |            |            |

图: 录音机数据的均差表

可以看到, 二阶均为常数, 三阶均差到小数点后 4 位均为 0, 可以看出数据基本是二次的, 可以用二次多项式作为经验模型.

# 差分表中的观测值

- 对“小”的评判是相对的、定性的.
- 必须灵敏地感觉数据中出现的误差和不规则变化.



# 车辆的停止距离

|                      |    |    |      |      |     |       |     |       |     |       |     |     |     |
|----------------------|----|----|------|------|-----|-------|-----|-------|-----|-------|-----|-----|-----|
| Speed<br>$v$ (mph)   | 20 | 25 | 30   | 35   | 40  | 45    | 50  | 55    | 60  | 65    | 70  | 75  | 80  |
| Distance<br>$d$ (ft) | 42 | 56 | 73.5 | 91.5 | 116 | 142.5 | 173 | 209.5 | 248 | 292.5 | 343 | 401 | 464 |

| Data  |       | Divided differences |            |            |            |
|-------|-------|---------------------|------------|------------|------------|
| $v_i$ | $d_i$ | $\Delta$            | $\Delta^2$ | $\Delta^3$ | $\Delta^4$ |
| 20    | 42    |                     |            |            |            |
| 25    | 56    | 2.2800              |            |            |            |
| 30    | 73.5  | 3.5000              | 0.0700     |            |            |
| 35    | 91.5  | 3.6000              | 0.0100     | -0.0040    | 0.0006     |
| 40    | 116   | 4.9000              | 0.1300     | 0.0080     | -0.0007    |
| 45    | 142.5 | 4.9000              | 0.0400     | -0.0060    | 0.0004     |
| 50    | 173   | 5.3000              | 0.0800     | 0.0027     | 0.0000     |
| 55    | 209.5 | 6.1000              | 0.1200     | 0.0027     | -0.0004    |
| 60    | 248   | 7.3000              | 0.0400     | -0.0053    | 0.0005     |
| 65    | 292.5 | 7.7000              | 0.1200     | 0.0053     | -0.0003    |
| 70    | 343   | 8.9000              | 0.1200     | 0.0000     | 0.0001     |
| 75    | 401   | 10.1000             | 0.1500     | 0.0020     | -0.0003    |
| 80    | 464   | 11.6000             | 0.1000     | -0.0033    |            |
|       |       | 12.6000             |            |            |            |

## 用二阶模型进行拟合

$$P(v) = a + bv + cv^2$$

用最小二乘法求解:

$$\text{Minimize } S = \sum_{i=1}^m [d_i - (a + bv_i + cv_i^2)]^2$$

存在极小点的必要条件  $\frac{\partial S}{\partial a} = \frac{\partial S}{\partial b} = \frac{\partial S}{\partial c} = 0$ , 故

$$ma + (\sum v_i)b + (\sum v_i^2)c = \sum d_i$$

$$(\sum v_i)a + (\sum v_i^2)b + (\sum v_i^3)c = \sum v_i d_i$$

$$(\sum v_i^2)a + (\sum v_i^3)b + (\sum v_i^4)c = \sum v_i^2 d_i$$

# 模型验证

$$P(v) = 50.0594 - 1.9701v + 0.0886v^2$$

|                |        |        |        |        |       |       |        |
|----------------|--------|--------|--------|--------|-------|-------|--------|
| $v_i$          | 20     | 25     | 30     | 35     | 40    | 45    | 50     |
| $d_i$          | 42     | 56     | 73.5   | 91.5   | 116   | 142.5 | 173    |
| $d_i - P(v_i)$ | -4.097 | -0.182 | 2.804  | 1.859  | 2.985 | 1.680 | -0.054 |
| $v_i$          | 55     | 60     | 65     | 70     | 75    | 80    |        |
| $d_i$          | 209.5  | 248    | 292.5  | 343    | 401   | 464   |        |
| $d_i - P(v_i)$ | -0.719 | -2.813 | -3.838 | -3.292 | 0.323 | 4.509 |        |

图: 用二次多项式光滑化停止距离

当  $v = 0$  时呢?

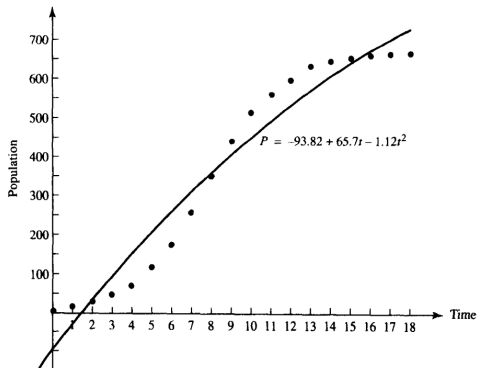
# 酵母培养物的增长

| Data  |        | Divided differences |            |            |            |
|-------|--------|---------------------|------------|------------|------------|
| $t_i$ | $P_i$  | $\Delta$            | $\Delta^2$ | $\Delta^3$ | $\Delta^4$ |
| 0     | 9.60   |                     |            |            |            |
| 1     | 18.30  | 8.70                |            |            |            |
| 2     | 29.00  | 10.70               | 1.00       |            |            |
| 3     | 47.20  | 18.20               | 3.75       | 0.92       | -0.31      |
| 4     | 71.10  | 23.90               | 2.85       | -0.30      | 0.84       |
| 5     | 119.10 | 48.00               | 12.05      | 3.07       | -1.46      |
| 6     | 174.60 | 55.50               | 3.75       | -2.77      | 1.51       |
| 7     | 257.30 | 82.70               | 13.60      | 3.28       | -1.51      |
| 8     | 350.70 | 93.40               | 5.35       | -2.75      | 0.11       |
| 9     | 441.00 | 90.30               | -1.55      | -2.30      | -0.05      |
| 10    | 513.30 | 72.30               | -9.00      | -2.48      | 0.29       |
| 11    | 559.70 | 46.40               | -12.95     | -1.32      | 0.94       |
| 12    | 594.80 | 35.10               | -5.65      | 2.43       | -0.16      |
| 13    | 629.40 | 34.60               | -0.25      | 1.80       | -1.40      |
| 14    | 640.80 | 11.40               | -11.60     | -3.78      | 1.87       |
| 15    | 651.10 | 10.30               | -0.55      | 3.68       | -1.10      |
| 16    | 655.90 | 4.80                | -2.75      | -0.73      | 0.37       |
| 17    | 659.60 | 3.70                | -0.55      | 0.73       | -0.20      |
| 18    | 661.80 | 2.20                | -0.75      | -0.07      |            |

图: 差分表

# 拟合结果

$$P = -93.82 + 65.70t - 1.12t^2$$



拟合失败!

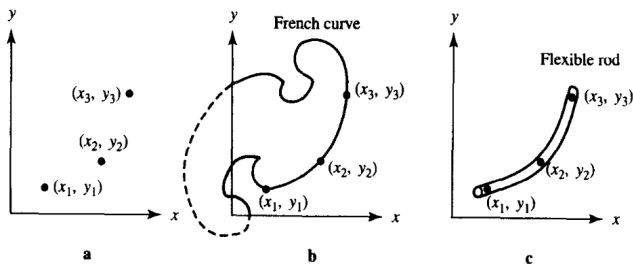
# 三阶样条模型

在连续的数据点对间使用不同的三阶多项式，追踪数据的趋势。

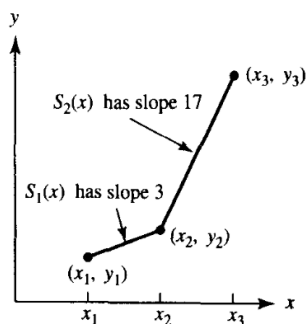
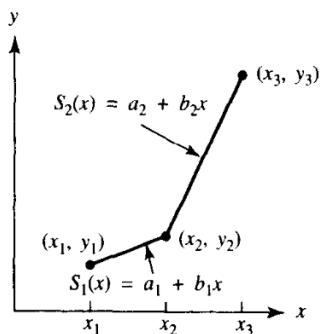
**高阶多项式** 端点处有摆动倾向，对数据中小的变化太敏感

**低阶多项式** 拟合效果不好（如速度很快时的刹车问题模拟）

**三阶样条插值** 既保证基本关系的特征，同时减少摆动的倾向和数据变换的敏感性

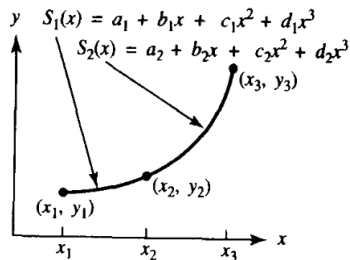


# 线性样条



- (a) 线性样条是含直线段的连续函数
- (b) 由于一阶导数不连续，线性样条不光滑

# 三阶样条



$$S_1(x) = a_1 + b_1x + c_1x^2 + d_1x^3, x \in [x_1, x_2)$$

$$S_2(x) = a_2 + b_2x + c_2x^2 + d_2x^3, x \in [x_2, x_3)$$



# 三阶样条求解

$$S_1'(x) = b_1 + 2c_1x + 3d_1x^2, x \in [x_1, x_2)$$

$$S_1''(x) = 2c_1 + 6d_1x, x \in [x_1, x_2)$$

$$S_2'(x) = b_2 + 2c_2x + 3d_2x^2, x \in [x_2, x_3)$$

$$S_2''(x) = 2c_2 + 6d_2x, x \in [x_2, x_3)$$

求解:

$$S_1(x_1) = y_1$$

$$S_1(x_2) = y_2$$

$$S_2(x_2) = y_2$$

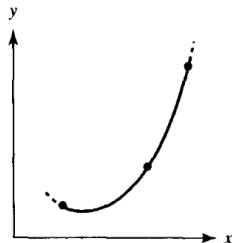
$$S_2(x_3) = y_3$$

# 约束条件

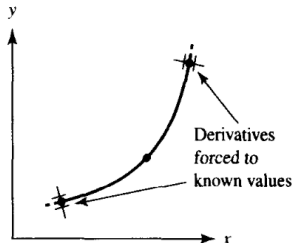
连接处导数约束:  $S_1'(x_2) = S_2'(x_2)$ ,  $S_1''(x_2) = S_2''(x_2)$

端点约束 (自然样条):  $S_1''(x_1) = 0$ ,  $S_2''(x_3) = 0$

端点约束 (强制样条):  $S_1'(x_1) = f'(x_1)$ ,  $S_2'(x_3) = f'(x_3)$



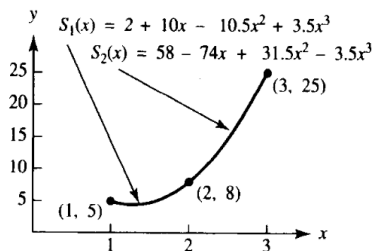
a. Natural spline



b. Clamped spline

# 拟合结果

| Interval          | Model                                  |
|-------------------|--|
| $1 \leq x < 2$    | $S_1(x) = 2 + 10x - 10.5x^2 + 3.5x^3$  |
| $2 \leq x \leq 3$ | $S_2(x) = 58 - 74x + 31.5x^2 - 3.5x^3$ |



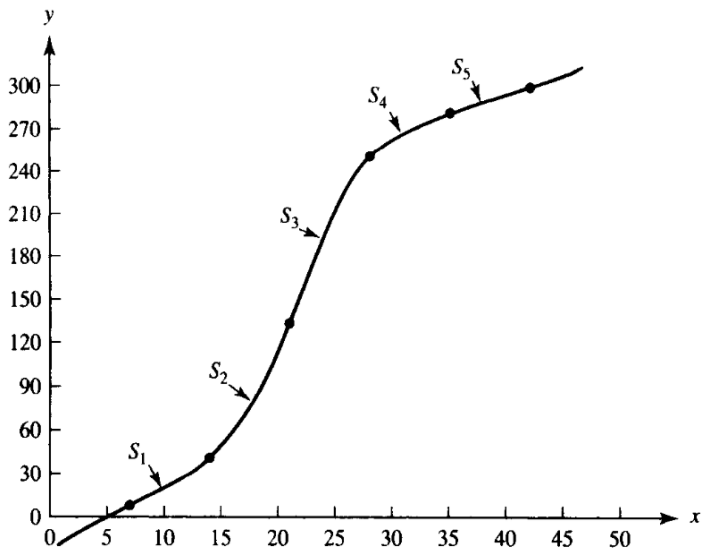
# 再论车辆的停止距离

|                      |    |    |      |      |     |       |     |       |     |       |     |     |     |
|----------------------|----|----|------|------|-----|-------|-----|-------|-----|-------|-----|-----|-----|
| Speed<br>$v$ (mph)   | 20 | 25 | 30   | 35   | 40  | 45    | 50  | 55    | 60  | 65    | 70  | 75  | 80  |
| Distance<br>$d$ (ft) | 42 | 56 | 73.5 | 91.5 | 116 | 142.5 | 173 | 209.5 | 248 | 292.5 | 343 | 401 | 464 |

| Interval         | Model   |
|------------------|---|
| $20 \leq v < 25$ | $S_1(v) = 42 + 2.596(v - 20) + 0.008(v - 20)^3$                         |
| $25 \leq v < 30$ | $S_2(v) = 56 + 3.208(v - 25) + 0.122(v - 25)^2 - 0.013(v - 25)^3$       |
| $30 \leq v < 35$ | $S_3(v) = 73.5 + 3.472(v - 30) - 0.070(v - 30)^2 + 0.019(v - 30)^3$     |
| $35 \leq v < 40$ | $S_4(v) = 91.5 + 4.204(v - 35) + 0.216(v - 35)^2 - 0.015(v - 35)^3$     |
| $40 \leq v < 45$ | $S_5(v) = 116 + 5.211(v - 40) - 0.015(v - 40)^2 + 0.006(v - 40)^3$      |
| $45 \leq v < 50$ | $S_6(v) = 142.5 + 5.550(v - 45) + 0.082(v - 45)^2 + 0.005(v - 45)^3$    |
| $50 \leq v < 55$ | $S_7(v) = 173 + 6.787(v - 50) + 0.165(v - 50)^2 - 0.012(v - 50)^3$      |
| $55 \leq v < 60$ | $S_8(v) = 209.5 + 7.503(v - 55) - 0.022(v - 55)^2 + 0.012(v - 55)^3$    |
| $60 \leq v < 65$ | $S_9(v) = 248 + 8.202(v - 60) + 0.161(v - 60)^2 - 0.004(v - 60)^3$      |
| $65 \leq v < 70$ | $S_{10}(v) = 292.5 + 9.489(v - 65) + 0.096(v - 65)^2 + 0.005(v - 65)^3$ |
| $70 \leq v < 75$ | $S_{11}(v) = 343 + 10.841(v - 70) + 0.174(v - 70)^2 - 0.005(v - 70)^3$  |
| $75 \leq v < 80$ | $S_{12}(v) = 401 + 12.245(v - 75) + 0.106(v - 75)^2 - 0.007(v - 75)^3$  |

图: 三阶样条模型

# 拟合结果



# 构造经验模型

