



基础数理统计

(研究生公共课)

肖柳青 教授 博士主讲



上海交通大学

SHANGHAI JIAO TONG UNIVERSITY



第4章 回归分析 (Regress Analysis)



一元
线性
回归

参数估计

假设检验

预 测

回归检验

系数检验

一元非线性回归

多元
线性
回归

参数估计

假设检验

预 测

回归检验

系数检验

线性岭回归与LASSO回归

本章
内容

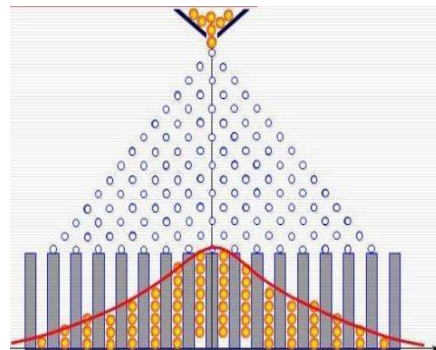
第四章 回归分析



数学上“回归”一词第一个被英国生物统计学家高尔顿 (Galton) 用于研究人类身高的遗传问题上.



他研究的结论是：很高(或矮)的双亲的儿子们一般高(或低)于平均值, 但不像他们的双亲那么高(或矮). 因此儿子们的高度将“回归”到平均值, 而不是更趋极端, 这也是“回归”一词的最初含义.





变量关系

确定性关系

由函数关系描述（如圆的面积与圆的半径的关系。）

不确定性关系

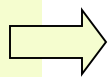
身高 X 与体重 Y 之间关系；
农作物的总产量与种植面积,农作物的单位面积产量与施肥量等等

回归关系

相关关系

对于**相关关系**,虽然不能找出变量之间确定的函数表达式,但通过大量的观测数据,可以发现它们之间存在一定的统计规律性.回归分析就是研究**相关关系**的一种有效方法.

回归分析



自变量可控时变量间关系的分析

一元回归分析

商场某商品的利润 Y 与进货量 x 间的关系

随机因变量

可控变量

某农作物的亩产量 W 与施肥量 氮磷钾 x, y, z 间的关系

多元回归分析

非回归分析问题

细纱的
强力

Y

与原棉的纤维

长度
细度
强力

x
 y
 z

的关系

随机因变量

不可控变量
(随机自变量)

§ 4.1 一元线性回归中的参数估计

1. 一元线性回归模型

设 x 是可控变量, Y 是依赖 x 的随机变量, 两者关系式为

$$Y = a + bx + \varepsilon, \varepsilon \sim N(0, \sigma^2), a, b, \sigma = \text{const.}$$

当 x 取定值时, $Y \sim N(a + bx, \sigma^2)$

记 $\tilde{y}(x) = E(Y|x)$, 则有 $\tilde{y}(x) = a + bx$

Y 对 x 的回归方程

回归系数

当 $\tilde{y}(x)$ 为 x 的线性函数时, 称为线性回归, 否则称为非线性回归.



一元线性回归的主要任务

- (1) 用 n 对试验值 $(x_i, y_i) i = 1 \sim n$, 对未知参数 a, b, σ^2 进行点估计;
- (2) 对回归系数 b 作假设检验;
- (3) 在 $x = x_0$ 处对 Y 作预测, 即对 Y 作区间估计.

2. 未知参数 a, b, σ^2 的估计

选择参数 a, b , 使离差平方和 Q 最小, 即

$$Q = \sum_{i=1}^n (y_i - \tilde{y}_i)^2 = \sum_{i=1}^n (y_i - a - bx_i)^2 = \min$$

$$\text{令} \begin{cases} Q'_a = -2 \sum_{i=1}^n (y_i - a - bx_i) = 0 \\ Q'_b = -2 \sum_{i=1}^n (y_i - a - bx_i)x_i = 0 \end{cases} \Rightarrow$$

$$\text{正规方程组} \begin{cases} na + b \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \\ a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i \end{cases}$$

用 \hat{a}, \hat{b} 取代 a, b

$$n\hat{a} + \hat{b} \sum_{i=1}^n x_i = \sum_{i=1}^n y_i$$

$$\hat{a} \sum_{i=1}^n x_i + \hat{b} \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i$$

方程两边都除以 n

$$\begin{cases} \hat{a} + \hat{b}\bar{x} = \bar{y} \\ \hat{a}\bar{x} + \hat{b}\overline{x^2} = \overline{xy} \end{cases}$$

解得

$$\begin{cases} \hat{a} = \bar{y} - \hat{b}\bar{x} \\ \hat{b} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{cases}$$

(本方法称最小二乘法—由高斯(Gauss)建立)

把 \hat{a}, \hat{b} 代入回归直线方程得

$$\hat{y} = \hat{a} + \hat{b}x = \bar{y} + \hat{b}(x - \bar{x})$$

称为 Y 对 x 的经验回归直线方程, 它表明

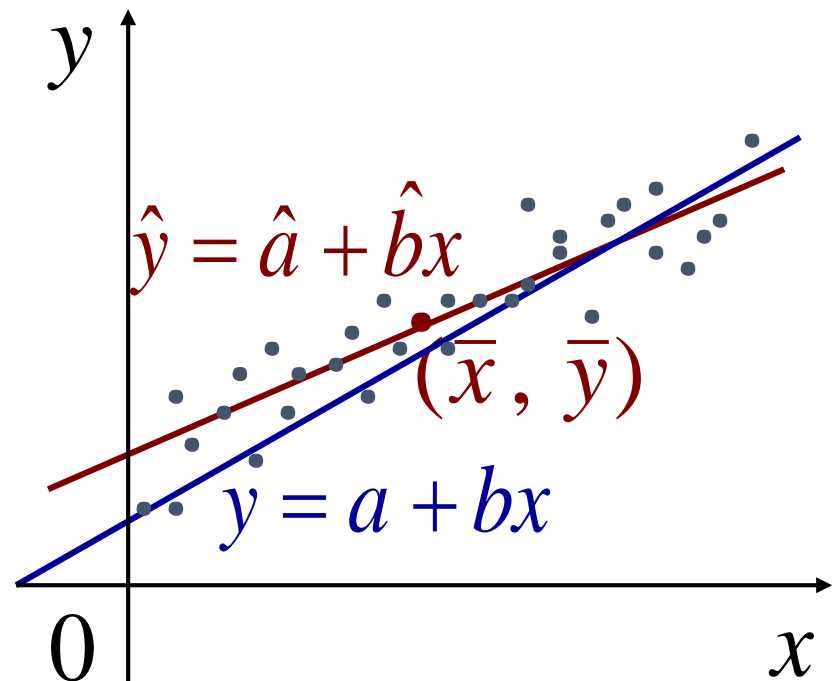
对于子样观察值

$(x_1, y_1), (x_2, y_2), \dots,$

(x_n, y_n) , 经验回归

直线通过散点图的

几何中心 (\bar{x}, \bar{y}) .



为计算方便引入下述记号:

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2$$

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2$$

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \frac{1}{n} \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}, \quad \hat{b} = \frac{S_{xy}}{S_{xx}}.$$


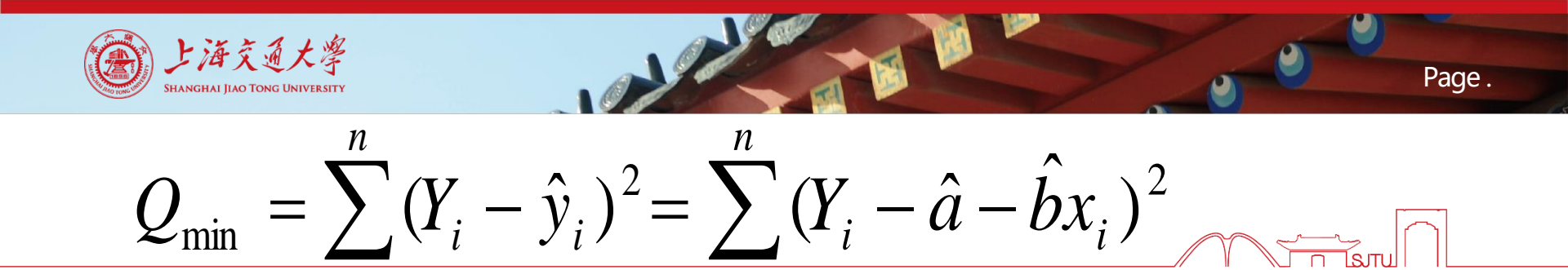
上面用最小二乘法估计 a, b , 下面用矩法估计 σ^2 .

由于 $\sigma^2 = D\varepsilon = E\varepsilon^2$, 故用 $\frac{1}{n} \sum_{i=1}^n \varepsilon_i^2$ 作估计.

其中 $\varepsilon_i = Y_i - a - bx_i$, 于是 σ^2 的估计量为

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{a} - \hat{b}x_i)^2$$

下面推导 $\hat{\sigma}^2$ 的简化计算公式



$$\begin{aligned} Q_{\min} &= \sum_{i=1}^n (Y_i - \hat{y}_i)^2 = \sum_{i=1}^n (Y_i - \hat{a} - \hat{b}x_i)^2 \\ &= \sum_{i=1}^n (Y_i - \bar{Y} + \hat{b}\bar{x} - \hat{b}x_i)^2 = \sum_{i=1}^n [(Y_i - \bar{Y}) - \hat{b}(x_i - \bar{x})]^2 \\ &= \sum_{i=1}^n (Y_i - \bar{Y})^2 - 2\hat{b} \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}) + \hat{b}^2 \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \sum_{i=1}^n (Y_i - \bar{Y})^2 - \hat{b}^2 \sum_{i=1}^n (x_i - \bar{x})^2 = S_{yy} - \hat{b}S_{xy} \\ \hat{\sigma}^2 &= \frac{Q_{\min}}{n} = \frac{1}{n}(S_{yy} - \hat{b}S_{xy}) \end{aligned}$$

例1 为研究某一化学反应过程中, 温度 x ($^{\circ}\text{C}$) 对产品收率 y (%) 的影响, 测得数据如下

温度 x	100	110	120	130	140	150	160	170	180	190
收率 y	45	51	54	61	66	70	74	78	85	89

求 Y 对 x 的经验回归直线方程, 并计算 σ^2 的估计值 $\hat{\sigma}^2$.

解 列表计算



x	y	x^2	y^2	xy
100	45	10000	2025	4500
110	51	12100	2601	5610
120	54	14400	2916	6480
130	61	16900	3721	7930
140	66	19600	4356	9240
150	70	22500	4900	10500
160	74	25600	5476	11840
170	78	28900	6084	13260
180	85	32400	7225	15300
190	89	36100	7921	16910
$\Sigma 1450$	673	218500	47225	101570

$$S_{xx} = 218500 - \frac{1}{10} \times 1450^2 = 8250$$

$$S_{xy} = 101570 - \frac{1}{10} \times 1450 \times 673 = 3985$$

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = 0.483,$$

$$\begin{aligned}\hat{a} &= \bar{y} - \hat{b}\bar{x} = \frac{1}{10} \times 673 - 0.483 \times \frac{1}{10} \times 1450 \\ &= -2.735 ,\end{aligned}$$

$$\therefore \hat{y} = -2.735 + 0.483x.$$



$$S_{yy} = 47225 - \frac{1}{10} \times 673^2 = 1932.1$$

$$\hat{\sigma}^2 = \frac{1}{n} (S_{yy} - \hat{b}S_{xy})$$

$$= \frac{1}{10} (1932.1 - 0.483 \times 3985)$$

$$= 0.7345.$$



解二 令 $u = x - 150$, $v = y - 45$

则原数据变为

温度 u	-50	-40	-30	-20	-10	0	10	20	30	40
收率 v	0	6	9	16	21	25	29	33	40	44

解 列表计算

x	y	x^2	y^2	xy
-50	0	2500	0	0
-40	6	1600	36	-240
-30	9	900	81	-270
-20	16	400	256	-320
-10	21	100	441	-210
0	25	0	625	0
10	29	100	841	290
20	33	400	1089	660
30	40	900	1600	1200
40	44	1600	1936	1760
$\Sigma -50$	223	8500	6905	2870

$$S_{uu} = 8500 - \frac{1}{10} \times (-50)^2 = 8250$$

$$S_{uv} = 2870 - \frac{1}{10} \times (-50 \times 223) = 3985$$

$$\hat{b}' = \frac{S_{uv}}{S_{uu}} = 0.483 ,$$

$$\hat{a}' = \bar{v} - \hat{b}'\bar{u} = \frac{1}{10} \times 223 - 0.483 \times \frac{1}{10} \times (-50) = 24.715,$$

$$\therefore \hat{v} = 24.715 + 0.483u.$$

还原 $\hat{y} = -2.735 + 0.483x.$



$$S_{vv} = 6905 - \frac{1}{10} \times 223^2 = 1932.1$$

$$\hat{\sigma}^2 = \frac{1}{n} (S_{vv} - \hat{b}' S_{uv})$$

$$= \frac{1}{10} (1932.1 - 0.483 \times 3985)$$

$$= 0.7345.$$

3. 估计量的分布

经验回归系数 \hat{b} 的分布

由于独立正态变量 Y_1, Y_2, \dots, Y_n 的线性组合仍是正态变量, 所以

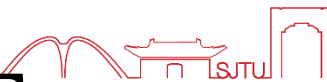
$$\hat{b} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \sim N(\hat{Eb}, \hat{Db})$$

其中

$$Eb = b, \hat{Db} = \sigma^2 / \sum_{i=1}^n (x_i - \bar{x})^2.$$

$$\bar{Y} \sum_{i=1}^n (x_i - \bar{x}) = 0$$

事实上


$$E\hat{b} = \frac{\sum_{i=1}^n (x_i - \bar{x}) EY_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(a + bx_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{b \sum_{i=1}^n (x_i - \bar{x})x_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = b$$

可见 \hat{b} 是 b 的无偏估计.

$$D\hat{b} = \frac{\sum_{i=1}^n (x_i - \bar{x}) DY_i}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

考察 $\hat{\sigma}^2 = \frac{Q_{\min}}{n} = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2 - \hat{b}^2 \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \dots \dots \dots (1)$

是不是 σ^2 的无偏估计量.

$$\begin{aligned}
 E\left[\sum_{i=1}^n (Y_i - \bar{Y})^2\right] &= E\left[\sum_{i=1}^n (a + bx_i + \varepsilon_i - a - b\bar{x} - \bar{\varepsilon})^2\right] \\
 &= E\left\{\sum_{i=1}^n [b(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon})]^2\right\} \\
 &= b^2 \sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n D(\varepsilon_i - \bar{\varepsilon}) \\
 &= b^2 \sum_{i=1}^n (x_i - \bar{x})^2 + (n-1)\sigma^2 \dots \dots \dots (2)
 \end{aligned}$$

$$\begin{aligned}
 E[\hat{b}^2 \sum_{i=1}^n (x_i - \bar{x})^2] &= [D\hat{b} + (E\hat{b})^2] \sum_{i=1}^n (x_i - \bar{x})^2 \\
 &= [\sigma^2 / \sum_{i=1}^n (x_i - \bar{x})^2 + b^2] \sum_{i=1}^n (x_i - \bar{x})^2 \\
 &= \sigma^2 + b^2 \sum_{i=1}^n (x_i - \bar{x})^2 \quad \dots\dots\dots (3)
 \end{aligned}$$

综合 (1) (2) (3) 得

$$E(\hat{\sigma}^2) = E \frac{Q_{\min}}{n} = \frac{n-2}{n} \sigma^2 \neq \sigma^2$$

$\hat{\sigma}^2$ 不是 σ^2 的无偏估计量.



记 $\hat{\sigma}_{*2}^2 = \frac{Q_{\min}}{n-2}$ 则 $E(\hat{\sigma}_{*2}^2) = E\left(\frac{Q_{\min}}{n-2}\right) = \sigma^2$

显然
$$\begin{aligned}\hat{\sigma}_{*2}^2 &= \frac{n}{n-2} \hat{\sigma}^2 \\ &= \frac{1}{n-2} \sum_{i=1}^n (Y_i - \bar{Y})^2 - \hat{b}^2 \frac{1}{n-2} \sum_{i=1}^n (x_i - \bar{x})^2\end{aligned}$$

例2 求例1中 σ^2 的无偏估计量的值.

解

$$\hat{\sigma}^{*2} = \frac{n}{n-2} \hat{\sigma}^2 = \frac{10}{10-2} \times 0.7345$$
$$= 0.9181.$$

定理 对一元线性回归, 有

$$(1) (n-2) \hat{\sigma}^{*2} / \sigma^2 \sim \chi^2(n-2)$$

(2) $\hat{\sigma}^{*2}$ 分别与 \hat{a}, \hat{b} 独立.

证明仿前面 S^{*2} 的分布的证明 (用到线性代数中的正交变换), 这里从略.

两个正规方程 是二次型 $Q_{\min} = \sum_{i=1}^n (Y_i - \hat{a} - \hat{b}x_i)^2$ 的约束条件, 所以 $\hat{\sigma}^{*2} = Q_{\min} / (n-2)$ 的自由度为 $n-2$.

2. 一元线性回归中的检验与预测

1. 一元线性回归中的假设检验

要检验一元线性回归模型

$$\begin{cases} y_i = a + bx_i + \varepsilon_i, & i = 1, 2, \dots, n, \\ \varepsilon_1, \varepsilon_2, \dots, \varepsilon_n \stackrel{i.i.d.}{\sim} N(0, \sigma^2) \end{cases}$$

是否成立是一件比较复杂的事情.

检验一元线性回归模型“三步曲”

1. 检验 x 取定时, Y 是否服从同方差的正态分布.
2. 检验 x 取定时, Y 是不是 x 的线性函数.
3. 检验 x 取定时, 相应 Y 值是否相互独立.

可见要严格地检验线性回归这一假设, 其计算麻烦且冗长.

1. 回归系数假设检验

上节求得的线性回归方程是否具有实用价值, 需要经过假设检验才能确定, 若假设符合实际, 则 $b \neq 0$, 否则 y 不依赖于 x .

因此需要检验假设 $H_0: b=0; H_1: b \neq 0$.

(1) t检验

由 § 5.1 知 $\hat{b} \sim N(b, \sigma^2 / S_{xx})$

$$\Rightarrow U = \frac{\hat{b} - b}{\sigma / \sqrt{S_{xx}}} \sim N(0, 1), \text{ 其中 } S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2.$$

$$\chi^2 = (n-2) \hat{\sigma}^{*2} / \sigma^2 \sim \chi^2(n-2)$$



且 \hat{b} 与 $\hat{\sigma}^{*2}$ 相互独立, 由 t 分布定义得

$$T = \frac{\hat{b} - b}{\hat{\sigma}^*} \sqrt{S_{xx}} \sim t(n-2) \quad \dots\dots\dots (1)$$

给定显著性水平 α , 取检验统计量 $T = \frac{\hat{b}}{\hat{\sigma}^*} \sqrt{S_{xx}}$

当 $|T| \geq t_{\alpha/2}(n-2)$ 时, 拒绝 H_0 , 认为线性回归显著, 否则不显著.

例1 为研究某一化学反应过程中, 温度 x ($^{\circ}\text{C}$) 对产品收率 y (%) 的影响, 测得数据如下

温度 x	100	110	120	130	140	150	160	170	180	190
收率 y	45	51	54	61	66	70	74	78	85	89

上节已求出经验回归直线方程

$$\hat{y} = a + bx$$

现检验线性回归是否显著, 显著水平取5%.

解 作假设 $H_0 : b = 0 ; H_1 : b \neq 0$.

取统计量 $T = \frac{\hat{b}}{\hat{\sigma}^*} \sqrt{S_{xx}} \sim t(n-2)$

拒绝域 $W: |T| \geq t_{0.025}(8) = 2.306$

$$\hat{\sigma}^{*2} = \frac{n}{n-2} \hat{\sigma}^2 = \frac{10}{8} \times 0.7345 = 0.9181$$

$$|T| = \frac{0.483}{\sqrt{0.9181}} \sqrt{8250} = 45.7844 > 2.306 \in W$$

拒绝 H_0 , 认为线性回归效果显著.

注

当回归效果显著时,常需要对回归系数 b 作区间估计. 事实上,由 (1) 式得到 b 的置信度为 $1-\alpha$ 的置信区间为

$$\left(\hat{b} - t_{\alpha/2}(n-2) \frac{\hat{\sigma}^*}{\sqrt{S_{xx}}}, \hat{b} + t_{\alpha/2}(n-2) \frac{\hat{\sigma}^*}{\sqrt{S_{xx}}} \right)$$

如例1中回归系数 b 的置信区间是

$$\left(0.483 \pm 2.306 \frac{0.9181}{\sqrt{8250}} \right) = (0.460, 0.506)$$

推广



若检验假设 $H_0: b = b_0$; $H_1: b \neq b_0$.

取检验统计量 $T = \frac{\hat{b} - b_0}{\hat{\sigma}^*} \sqrt{S_{xx}}$

当 $|T| \geq t_{\alpha/2}(n-2)$ 时, 拒绝 H_0 , 认为回归系数与 b_0 有显著差异, 否则无显著差异.

2) F检验

平方和分解式 $S_T = S_R + S_E$, $f_T = f_R + f_E$,

总平方和 $S_T = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2 = S_{yy}$

自由度 $f_T = n - 1$;

回归平方和 $S_R = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^n (\hat{a} + \hat{b}x_i - \bar{y})^2$
 $= \hat{b}S_{xy} = \hat{b}^2 S_{xx}$

自由度 $f_R = 1$;

残差
平方和

$$S_E = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{a} - \hat{b}x_i)^2$$

自由度

$$f_E = n - 2.$$

在原假设 $H_0: b=0$ 成立的条件下, 检验

统计量
$$F = \frac{S_R}{S_E / (n - 2)} \sim F(1, n - 2)$$

拒绝域 $W: F \geq F_{1-\alpha}(1, n - 2).$

注

上面介绍的t检验与F检验等价



因为两种检验所用统计量

$$T = \frac{\hat{b}}{\hat{\sigma}^*} \sqrt{S_{xx}} \sim t(n-2)$$

$$F = \frac{S_R}{S_E / (n-2)} \sim F(1, n-2)$$

存在关系 $T^2 = F$.

线性回归效果不显著的可能原因



1. 影响 Y 取值的除 x 外还有其它不可忽略的因素.
2. Y 与 x 的存在非线性关系.
3. Y 与 x 无关.

若要对以上三种情形配线性回归模型，都有 $b = 0$ ，即

$$Y = a + \varepsilon$$

2. 预测


回归方程的一个重要应用是预测. 即当 $x = x_0$ 时, 以一定的置信度对 Y 作区间估计. 气象台的气温预报便是典型的预测.

当 $x = x_0$ 时, Y 的值为 $Y_0 = a + bx_0 + \varepsilon_0$, $\varepsilon_0 \sim N(0, \sigma^2)$

取 x_0 处的回归值 $\hat{y}_0 = \hat{a} + \hat{b}x_0$ 作为 Y_0 的预测值, 则两者之差仍服从正态分布, 即

$$Y_0 - \hat{y}_0 = Y_0 - \hat{a} - \hat{b}x_0 \sim N\left(0, \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\right)$$

事实上,


$$\begin{aligned} E(Y_0 - \hat{y}_0) &= E(a + bx_0 + \varepsilon_0 - \hat{a} - \hat{b}x_0) \\ &= a + bx_0 - E\hat{a} - bx_0 = a - E(\bar{Y} - \hat{b}\bar{x}) \\ &= a - E\bar{Y} + b\bar{x} = a - (a + b\bar{x}) + b\bar{x} = 0, \end{aligned}$$

$$\begin{aligned} D(Y_0 - \hat{y}_0) &= D(a + bx_0 + \varepsilon_0) + D(\hat{a} + \hat{b}x_0) \\ &= \sigma^2 + D[\bar{Y} + \hat{b}(x_0 - \bar{x})] = \sigma^2 + D\bar{Y} \\ &\quad + D[\hat{b}(x_0 - \bar{x})] + 2\text{cov}(\bar{Y}, \hat{b}(x_0 - \bar{x})) \end{aligned}$$

$$= \sigma^2 + \frac{\sigma^2}{n} + \frac{(x_0 - \bar{x})^2 \sigma^2}{S_{xx}} + 2(x_0 - \bar{x}) E[(\bar{Y} - E\bar{Y})(\hat{b} - E\hat{b})]$$

$$E[(\bar{Y} - E\bar{Y})(\hat{b} - E\hat{b})]$$

$$= E\left\{ \frac{1}{n} \sum_{i=1}^n (Y_i - EY_i) \left[\sum_{i=1}^n (x_i - \bar{x})(Y_i - EY_i) \right] / S_{xx} \right\}$$

$$= \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n (x_i - \bar{x}) E[(Y_i - EY_i)(Y_j - EY_j)] / S_{xx}$$

$$= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) \sigma^2 / S_{xx} = 0.$$

$$D(Y_0 - \hat{y}_0) = \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right] \sigma^2$$

由 $U = \frac{Y_0 - \hat{a} - \hat{b}x_0}{\sqrt{1 + \frac{1}{n} + \sigma^2(x_0 - \bar{x})^2 / S_{xx}}} \sim N(0, 1)$

又 $\chi^2 = (n-2) \hat{\sigma}^{*2} / \sigma^2 \sim \chi^2(n-2)$

得 $T = \frac{Y_0 - \hat{a} - \hat{b}x_0}{\sqrt{1 + \frac{1}{n} + (x_0 - \bar{x})^2 / S_{xx}} \hat{\sigma}^*} \sim t(n-2)$

给定置信概率 $1-\alpha$, 由

$$P(|T| < t_{\alpha/2}(n-2)) = 1 - \alpha$$

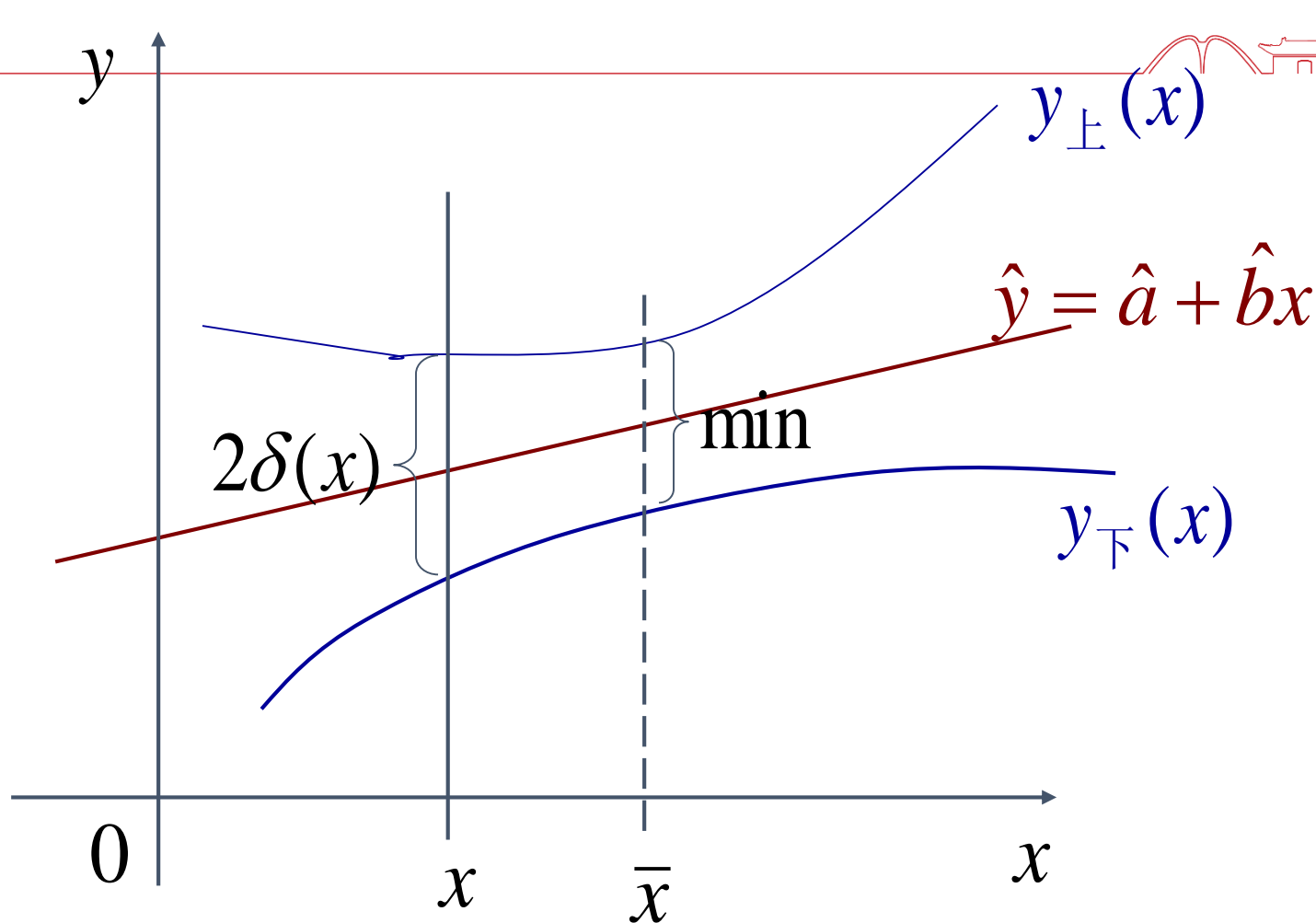
得 Y_0 的置信区间

$$(y_{\text{下}}(x), y_{\text{上}}(x)) = (\hat{y} - \delta(x), \hat{y} + \delta(x)) \quad (*)$$

其中 $\hat{y} = \hat{a} + \hat{b}x_0$

$$\delta(x) = t_{\alpha/2}(n-2) \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \hat{\sigma}^*$$

置信区间的几何意义



x 离 \bar{x} 愈近, 置信区间愈短, 估计精度愈高.

例2 求例1中温度 $x_0 = 135$ ($^{\circ}\text{C}$) 时产品收率 $y(\%)$ 的预测区间, 置信度取95%.

解 $\hat{y} = \hat{a} + \hat{b}x_0 = -2.735 + 0.483 \times 135 = 62.470$

$$\begin{aligned}\delta(x) &= t_{\alpha/2}(n-2) \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \hat{\sigma}^* \\ &= 2.306 \sqrt{1 + \frac{1}{10} + \frac{(135 - 145)^2}{8250}} \times 0.9582 = 2.330\end{aligned}$$

由(*)式得预测区间为

$$(\hat{y} - \delta(x), \hat{y} + \delta(x)) = (60.14, 64.80)$$

注 在实际回归问题中, 子样容量常很大.
此时对于在 \bar{x} 附近的 x , 不仅能得到较短
的预测区间, 还可简化(*)式

$$n > 45 \Rightarrow t_{\alpha/2}(n-2) \approx u_{\alpha/2}, \quad \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \approx 1,$$

于是(*)式变为

$$(\hat{y} - u_{\alpha/2} \hat{\sigma}^*, \hat{y} + u_{\alpha/2} \hat{\sigma}^*)$$

若取 $u_{0.025} = 1.96 \approx 2$, 则 $(\hat{y} - 2\hat{\sigma}^*, \hat{y} + 2\hat{\sigma}^*)$.

3. 可线性化的一元非线性回归



1. 非线性函数形式

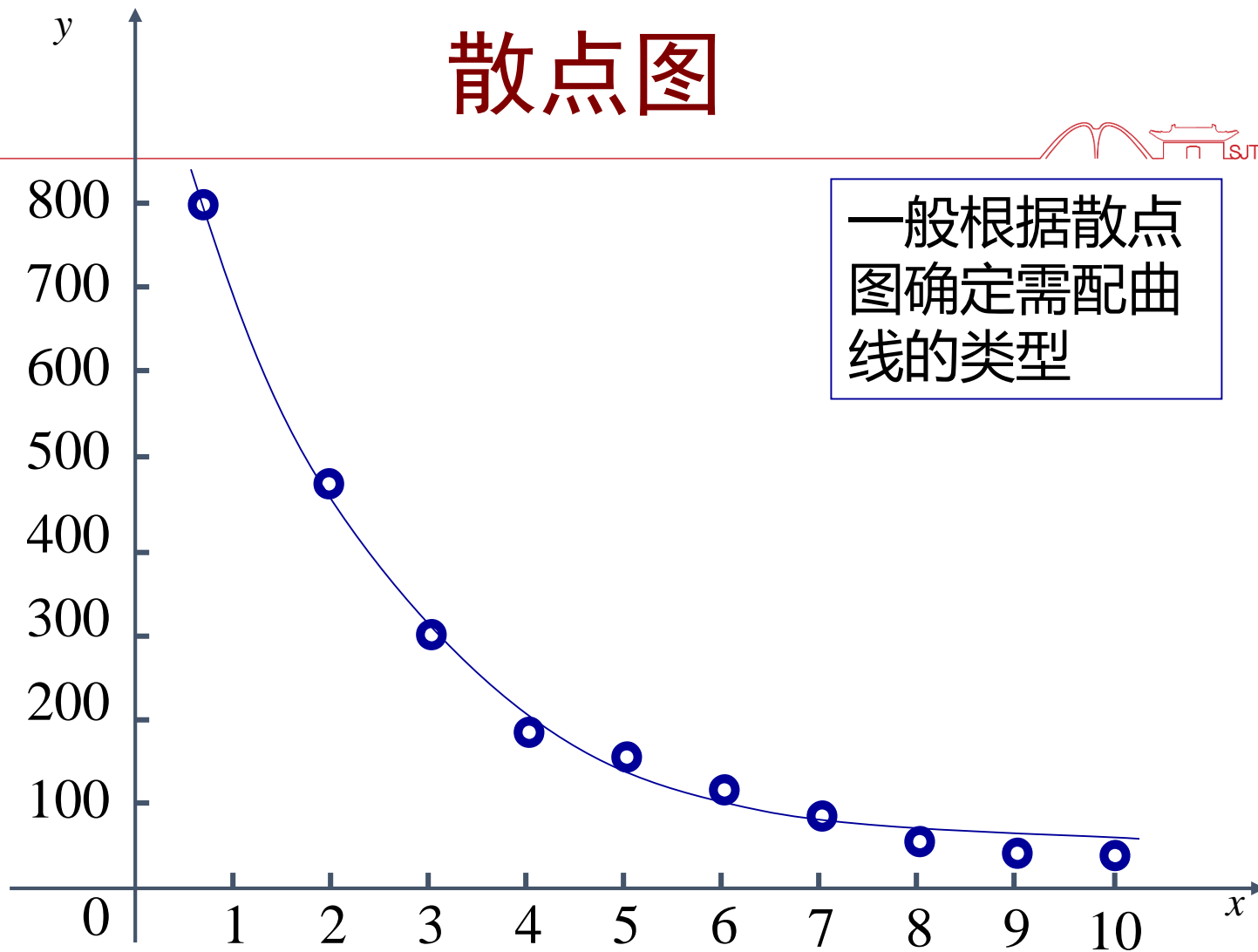
在工程技术中, 两个变量之间的关系可以不是直线(即线性)的相关关系, 而是某种曲线(即非线性)的相关关系.

一般, 可根据二维子样的散点图来确定可能的非线性函数形式, 也可利用专业知识确定曲线类型.

例1 为了检验X射线的杀菌作用,用200kv的X射线照射杀菌,每次照射 6 min ,照射次数为 X , 照射后所剩细菌数为 y ,试验结果如下

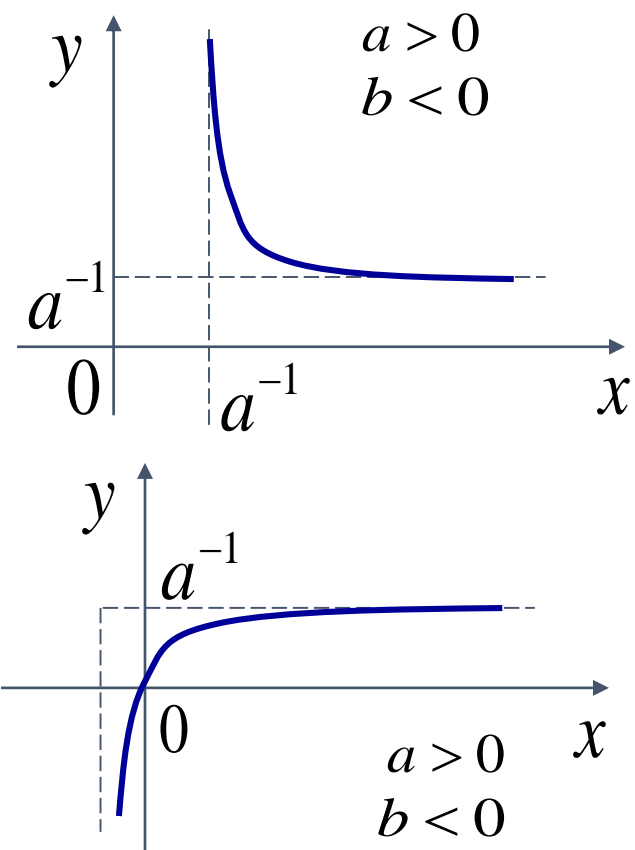
x	y	x	y
1	783	6	72
2	433	7	43
3	287	8	28
4	175	9	16
5	129	10	9

散点图



非线性回归(曲线回归)

常用五类曲线配置方法

名 称	表达式	图 像	方 法
双曲线	$\frac{1}{y} = a + \frac{b}{x}$	 <p> $a > 0$ $b < 0$ </p>	$u = 1/x$ $v = 1/y$

名 称

表达式

图 像

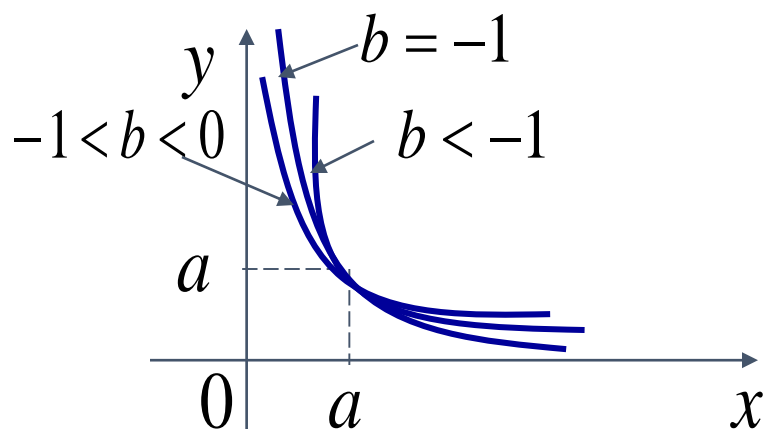
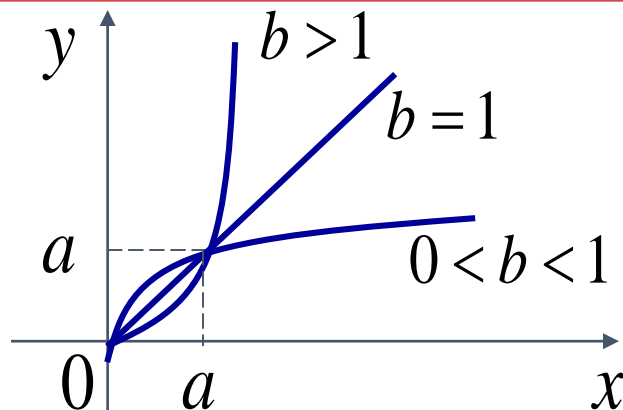
方 法

幂函数
曲线

$$y = ax^b$$

$$x > 0$$

$$a > 0$$



$$u = \ln x$$

$$v = \ln y$$

名 称

表达式

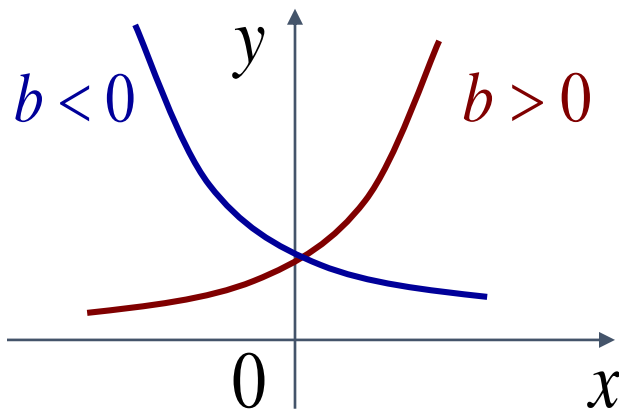
图 像

方 法

指 数
曲 线

$$y = ae^{bx}$$

$$a > 0$$

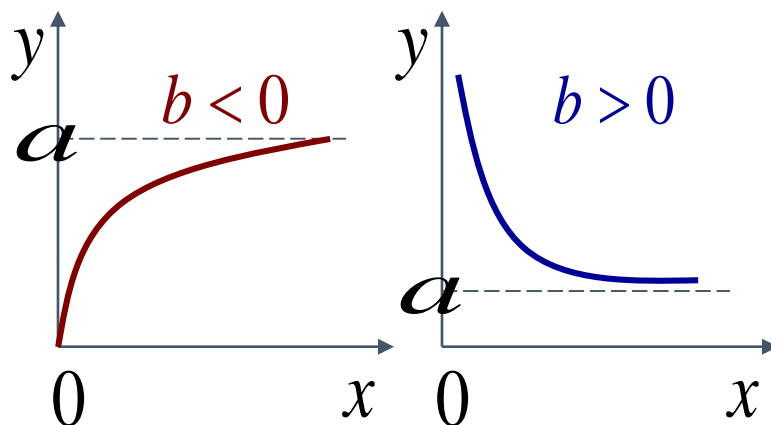


$$u = x$$

$$v = \ln y$$

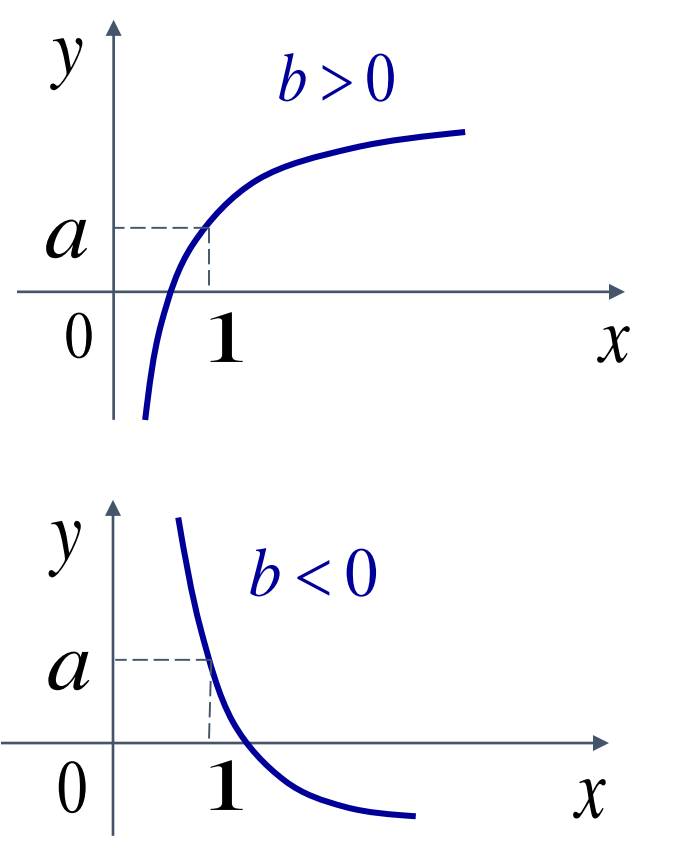
$$y = ae^{b/x}$$

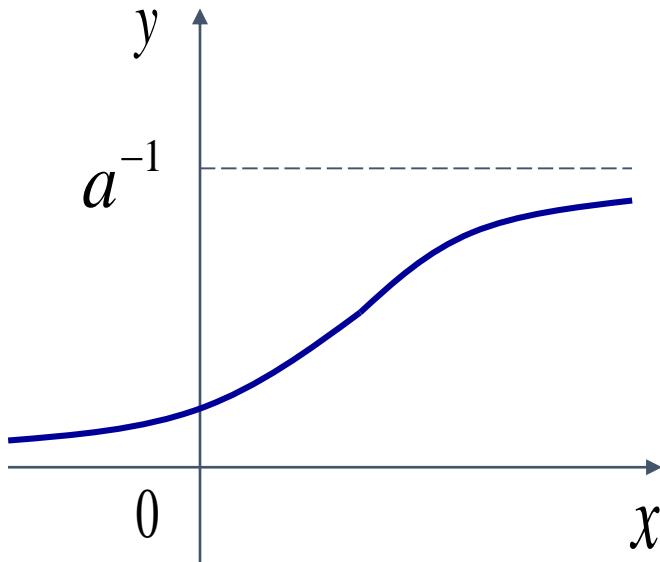
$$a > 0$$



$$u = 1/x$$

$$v = \ln y$$

名 称	表达式	图 像	方 法
对数 曲线	$y = a + b \ln x$ $x > 0$	 <p>Top graph: $b > 0$</p> <p>Bottom graph: $b < 0$</p>	$u = \ln x$ $v = y$

名 称	表达式	图 像	方 法
S型 曲线	$y = \frac{1}{a + be^{-x}}$		$u = e^{-x}$ $v = 1/y$

2. 参数估计



非线性函数

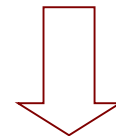
$$y = f(x)$$

变量代换

$$\begin{aligned} u &= g(x) \\ v &= h(x) \end{aligned}$$

线性函数

$$v = a + bu$$



求参数 a, b 的最小二乘估计.

3. 常用曲线回归方程好坏的评价标准

① 决定系数

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \quad \text{愈大愈好}$$

② 剩余标准差

$$S = \left[\frac{\sum (y_i - \hat{y}_i)^2}{n - 2} \right]^{1/2} \quad \text{愈小愈好}$$

注 这两个评价标准是一致的, 只是从两个不同侧面作出评价.

配曲线步骤

配曲线“三部曲”



1. 由试验数据 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

作出散点图;

2. 确定需配曲线的类型; 注

3. 根据试验数据计算所配曲线的未知
参数 a 和 b .

注 若有两个或两个以上非线性函数可用,
则分别拟合非线性回归并根据评价标
准进行选择.

解例1

给出例1具体的回归方程, 并求其对应的决定系数和剩余标准差.

解 根据经验知道, y 关于 x 的曲线回归方程形式可能有 $\hat{y} = Ae^{bx}$ (或 $\hat{y} = a + b \ln x$ 等)

令 $v = \ln y$, $a = \ln A$, 则回归方程

$$\hat{y} = Ae^{bx} \quad \triangleright \quad \hat{v} = a + bx$$

列表计算

x	y	$v = \ln y$	x^2	xv	y^2	\hat{y}	$(y - \hat{y})^2$
1	783	6.663	1	6.663	613089	772.01	120.78
2	433	6.071	4	12.142	187489	476.28	1873.16
3	287	5.660	9	16.980	82369	293.83	46.65
4	175	5.165	16	20.660	30625	181.27	39.31
5	129	4.860	25	24.300	16641	111.83	294.81
6	72	4.277	36	25.662	5184	68.99	9.06
7	43	3.761	49	26.327	1849	42.56	0.19
8	28	3.332	64	26.656	784	26.26	3.03
9	16	2.773	81	24.957	256	16.20	0.04
10	9	2.197	100	21.970	81	9.99	0.98
55	1975	44.759	385	206.317	938367		2388.01

$$S_{xx} = 385 - \frac{1}{10} \times 55^2 = 82.5$$

$$S_{xv} = 206.317 - \frac{1}{10} \times 55 \times 44.759 = -39.858$$

$$\hat{b} = \frac{S_{xv}}{S_{xx}} = -0.483 ,$$

$$\hat{a} = \bar{v} - \hat{b}\bar{x} = \frac{1}{10} \times 44.759 + 0.483 \times \frac{1}{10} \times 55 = 7.132$$

$\ln y$ 关于 x 的线性回归方程

$$\ln \hat{y} = 7.132 - 0.483x.$$

从而 y 关于 x 的曲线回归方程为

$$\hat{y} = 1251.371e^{-0.483x}.$$

决定系数

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{2388.01}{938367 - 1975^2 / 10} = 0.9956$$

表明曲线拟合程度相当好！

剩余标准差

$$S = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{2388.01}{8}} = 17.2772.$$

注 用其它曲线方程来拟合,可类似计算

$$(1) \quad \hat{y} = a + b \ln x \quad R^2 = 0.97, S = 38;$$

$$(2) \quad \hat{y} = (a + b\sqrt{x})^2 \quad R^2 = 0.98, S = 29.$$

例2 设曲线函数形式为 $y = 10 + Ae^{-x/B}$ ($B > 0$)
试给出一个变换, 将其化为一元线性回归的形式.

解 考虑如下变换: $u = x, v = \ln(y - 10)$

变换后的线性函数为: $v = \ln A - u / B,$

再令 $a = \ln A, b = -1 / B,$

则最后的线性回归函数为

$$v = a + bu.$$



第10次作业:

- 孙 p.115
- 习题四
- 1. 3. 5



谢谢!



上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

