§9.2 一元回归分析

本节讨论回归函数是一元线性函数或可线性化函数的情况.

一.一元线性回归模型

若回归函数是线性函数

$$\mu(x_1, x_2, \cdots, x_k) = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

其中b_0, b_1, …,b_k是未知常数，称为线性回归问题.

若Y关于X 的回归函数为

$$\mu(x)=E(Y\mid X=x)=ax+b$$

有一元线性回归模型：

$$Y=a+bx+\varepsilon,\qquad \varepsilon\sim N(0,\sigma^2)$$

其中a、b、σ^2为未知参数，且

a — 回归常数(又称截距)

b — 回归系数(又称斜率)

ε —— 随机误差（随机扰动项）

若随机误差$\varepsilon \sim N(0,\sigma^2)$, 称为**一元线性正态回归模型.**

取定自变量X 的一组值：$x_1, x_2, \cdots, x_n$,

对Y 做 n 次独立观察(试验), 试验结果记为

$$Y_1, \quad Y_2, \cdots, Y_n$$

由自变量X确定的成分

则有 $$Y_i = a + bx_i + \varepsilon_i, \qquad i = 1,\ldots,n$$

ε_i 是第 i 次观察时的随机误差，有

$$\begin{cases} 1)\ E(\varepsilon_i)=0,\quad D(\varepsilon_i)=\sigma^2,\quad i=1,2,\ \ldots,\ n; \\ 2)\ \varepsilon_1,\ \ldots,\ \varepsilon_n \text{ 相互独立.} \end{cases}$$

回归假定

二.一元线性回归模型的参数估计

需要对模型中的参数 a、b、σ^2 进行估计.

对自变量 X 的一组值 $x_1, x_2, \ldots, x_n$ 做 n 次独立试验，得独立观察值 $y_1, y_2, \ldots, y_n$.

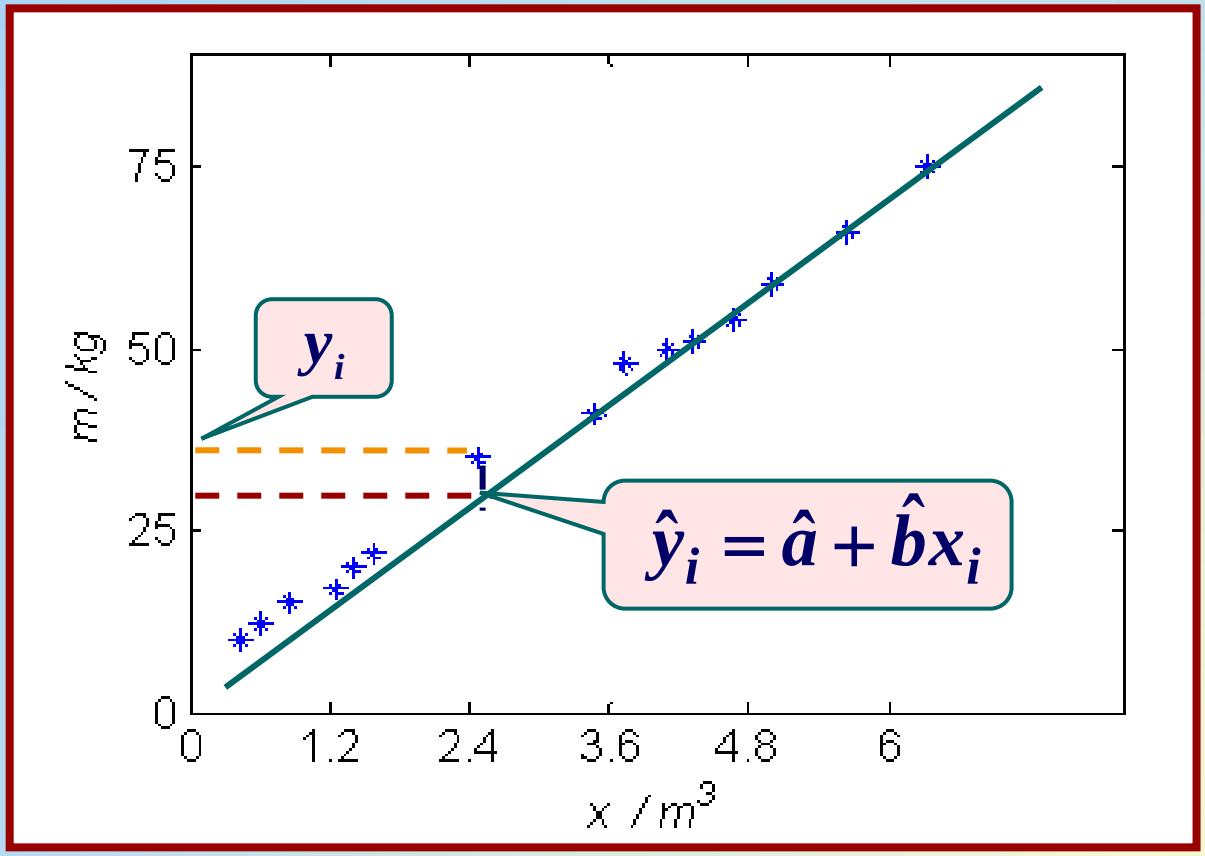
问题 如何依据观察值

$$(x_i,\ y_i),\quad i=1,\ 2,\ \ldots,\ n.$$

求a、b 的估计值 $\hat{a}, \hat{b}$?

记y_i 的估计值为 $\hat{y}_i = \hat{a} + \hat{b}x_i$

称为**回归值**.



对所有的i，应使偏差$y_i - \hat{y}_i$都尽可能小，有三种思路：

1) 使误差总和$\sum(y_i-\hat{y}_i)$最小；

缺点：可能正负误差抵消

2) 使误差绝对值之和$\sum|y_i-\hat{y}_i|$最小；

缺点：数学处理困难

3) 使误差的平方和$\sum(y_i-\hat{y}_i)^2$最小.

结论 应选a、b的估计使离差(误差)平方和：

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

达最小.

令 $Q(a,b)=\sum_{i=1}^{n}(y_i-a-bx_i)^2$

应用最小二乘法 分别对a,b求一阶偏导，并建立方程组：

$$\begin{cases}\sum_{i=1}^{n}(y_i-a-bx_i)=0\\ \sum_{i=1}^{n}(y_i-a-bx_i)x_i=0\end{cases}$$

或
$$\begin{cases} na+(\sum_{i=1}^{n}x_i)b=\sum_{i=1}^{n}y_i & (1) \\ (\sum_{i=1}^{n}x_i)a+(\sum_{i=1}^{n}x_i^2)b=\sum_{i=1}^{n}x_iy_i & (2) \end{cases}$$

称为**正规方程组**，由克莱姆法则，解得：

$$\begin{cases} \hat{b}=\dfrac{l_{xy}}{l_{xx}} \\ \hat{a}=\bar{y}-\hat{b}\bar{x} \end{cases}$$

其中

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\quad \bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$$

$$l_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{n}x_iy_i-n\bar{x}\,\bar{y}$$

$$l_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-n\bar{x}^2$$

由回归假定

$E(\varepsilon_i)=0$，　$D(\varepsilon_i)=\sigma^2$，$i=1,2$，…，$n$；

有$\sigma^2=D(\varepsilon)=E(\varepsilon^2)$，故

$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2$$

是σ^2的矩估计量.

可证明σ^2的无偏估计量为 $\hat{\sigma}^2=\dfrac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

代入 $\hat{y}_i = \hat{a} + \hat{b}x_i$，得

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

其中

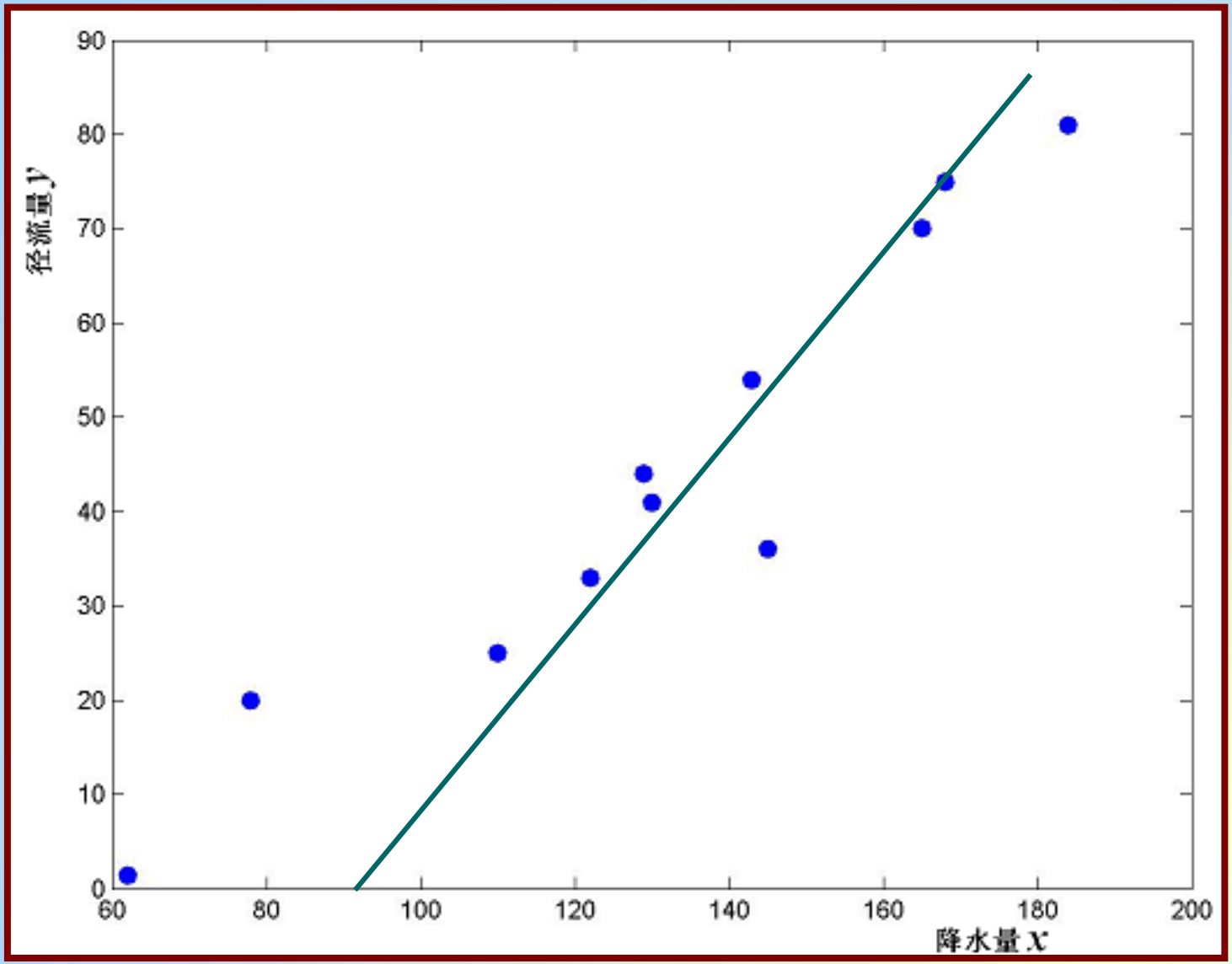
$$l_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

例9.2.1 流经某地区的降雨量X和该地河流的径流量Y的观察值如下表，

降雨量x_i:	110	184	145	122	165	143	78
径流量y_i:	25	81	36	33	70	54	20
	129	62	130	168	1436(∑)		
	44	1.41	41	75	480.4（∑）		

求Y关于X的(经验)线性回归方程,试估计降雨量为200时径流量为多少?

回归分析



解 $n=11,\ \bar{x}=130.5,\ \bar{y}=43.7$

$$l_{xx}=\sum_{i=1}^{11}(x_i-\bar{x})^2=13768.7$$

$$l_{xy}=\sum_{i=1}^{n}x_iy_i-n\bar{x}\,\bar{y}$$

$$=71424.8-62731.35=8693.45$$

$$\begin{cases}\hat{b}=\dfrac{l_{xy}}{l_{xx}}=\dfrac{8693.45}{13768.7}=0.6314\\ \hat{a}=\bar{y}-\hat{b}\bar{x}=43.7-0.6314\times130.5=-38.7\end{cases}$$

所求经验回归方程为

$$\hat{y} = \hat{a} + \hat{b}x = -38.7 + 0.6314\,x$$

降雨量为200时的径流量值为

$$\hat{y}(200) = -38.7 + 0.6314 \times 200 = 87.58$$

又 $$l_{yy} = \sum_{i=1}^{n}(y_i - 43.7)^2 = 6050.59$$

随机误差的方差σ^2的估计为

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

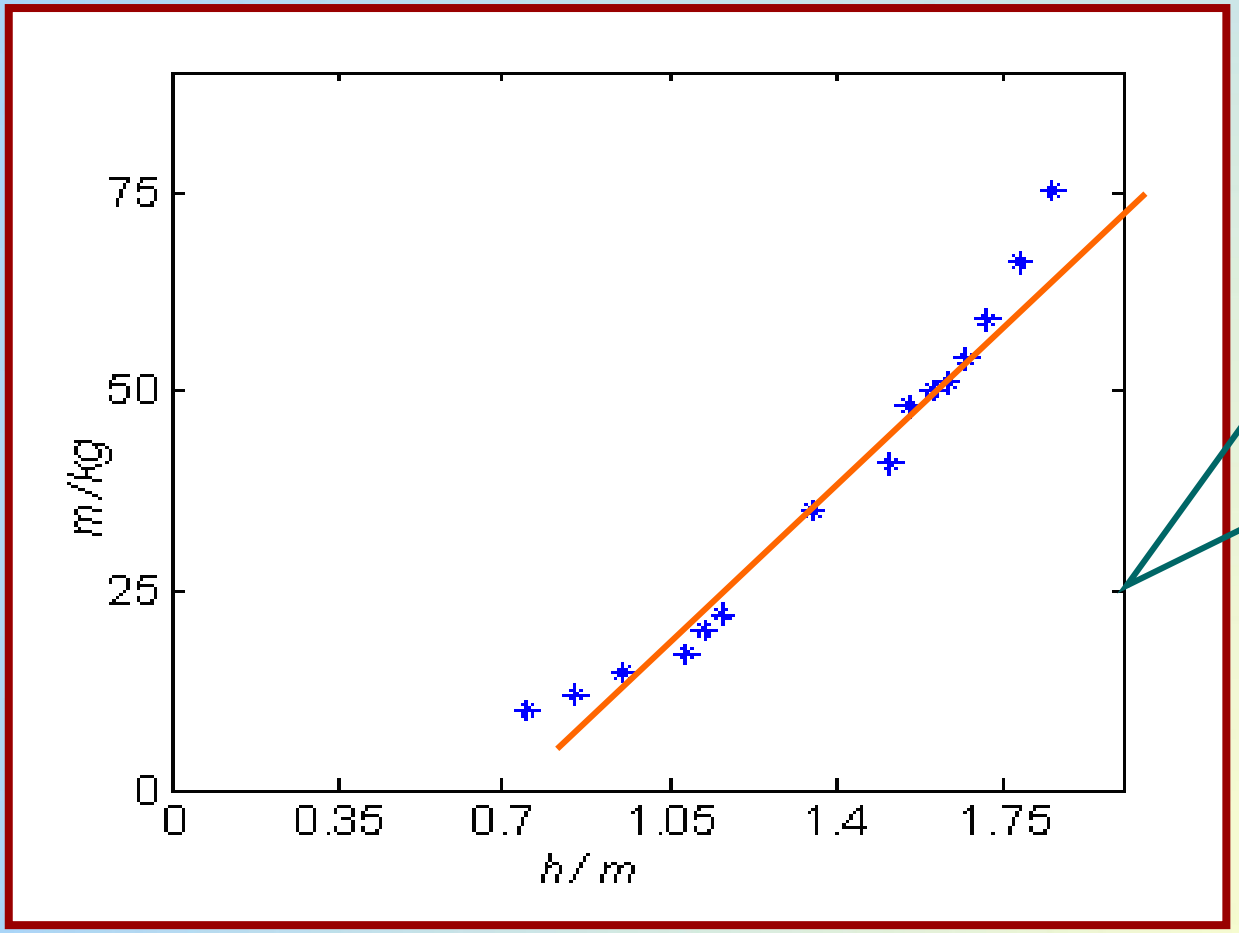
$$= (6050.59 - 0.6314^2 \times 13768.7)/9 = 62.3864$$

问题

随机变量Y与X间是否存在线性相关关系?

是否能由数据散布图完全确定回归函数?

身高体重关系



身高h 和体重m无明显的线性相关关系.

形式地估计回归系数和回归常数,并建立经验线性回归方程无实际意义.

2. 一元线性回归的假设检验（相关系数法）

相关系数法是基于试验数据检验随机变量间线性相关关系是否显著的一种方法.

相关系数

$$\rho_{XY}=\frac{E\{[X-E(Y)][Y-E(Y)]\}}{\sqrt{D(X)}\sqrt{D(Y)}}$$

是表征随机变量Y与X的**线性**相关程度的数字特征.

样本相关系数：

$$\hat{\rho}_{XY} = R = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$= \frac{l_{xy}}{\sqrt{l_{xx}}\sqrt{l_{yy}}}$$

作为ρ_{XY}的估计值.

有 $|R|\leq 1$，统计量R描述了X与Y间的线性相关关系的密切程度.

1) 当$R=0$ $\Rightarrow$ $l_{xy}=0$

$\Rightarrow$ $\hat{b}=\dfrac{l_{xy}}{l_{xx}}=0$

$\Rightarrow$ 经验回归方程形为 $\hat{y}=\hat{a}$

说明自变量X的变化不会引起因变量Y的变化(不相关).

2）$|R|=1 \Longrightarrow \dfrac{l_{xy}}{\sqrt{l_{xx}}\sqrt{l_{yy}}}=\pm 1$

$\Longrightarrow l_{xy}=\pm\sqrt{l_{xx}}\sqrt{l_{yy}}$

$\Longrightarrow \hat{b}=\dfrac{l_{xy}}{l_{xx}}=\dfrac{\pm\sqrt{l_{xx}}\sqrt{l_{yy}}}{l_{xx}}=\pm\sqrt{\dfrac{l_{yy}}{l_{xx}}}\neq 0,$

可认为X与Y间存在线性相关关系.

结论 1) $|R|$越接近于1，X与Y间的线性相关关系越显著；

2) $|R|$越靠近于0，X与Y间的线性相关关系越不显著.

根据附表6 相关系数临界值表，有

判别准则 给定显著性水平α（0.05，0.01）

当$|R| > R_{\alpha}(n-2)$,

认为X与Y之间的线性相关关系显著

当 $|R| \le R_\alpha(n-2)$,

认为X与Y之间的线性相关关系不显著

例9.2.2(续前例)利用相关系数显著性检验法，检验降雨量X和径流量Y的线性相关关系是否显著.

解 X与Y的样本相关系数为

$$R = \frac{l_{XY}}{\sqrt{l_{XX}}\sqrt{l_{YY}}}$$

$$= \frac{8693.45}{\sqrt{13768.7}\sqrt{6050.58}} = 0.952$$

查表得

$$R_{\alpha}(n-2) = R_{0.01}(9) = 0.735 < 0.952 = R$$

可认为X与Y的线性相关关系显著.

3. 非线性回归问题的线性化处理

在实际问题中,变量间的相关关系未必是线性关系，即其回归函数

$$y = \mu(x_1, x_2, \cdots, x_k)$$

往往是非线性函数.

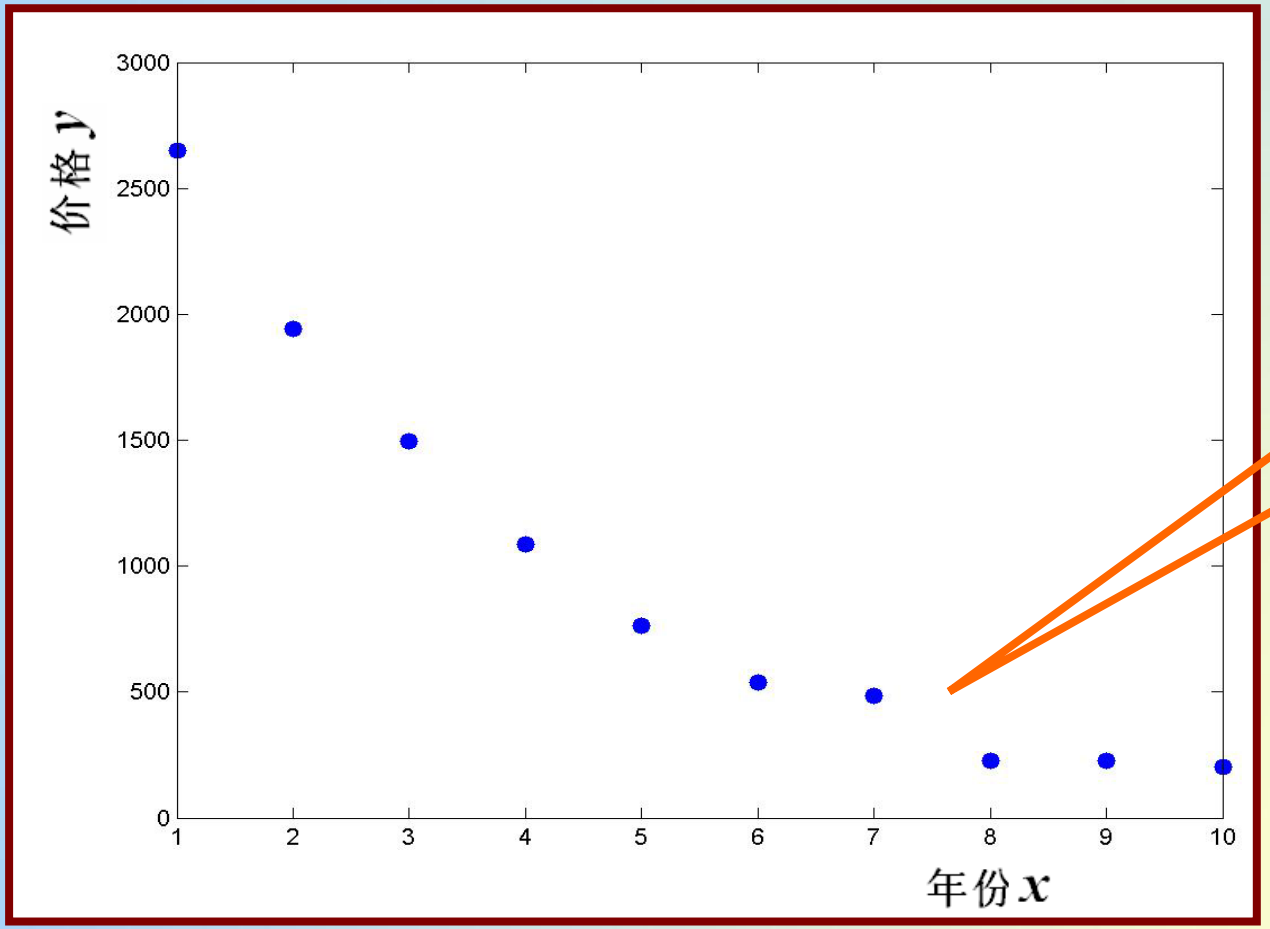
可通过适当的变换，将其转化为线性回归问题.

例9.2.3 下表是1957年美国旧轿车的调查数据表

使用年数x_i	1	2	3	4	5	6	7	8	9	10
平均价格y_i	2651	1943	1494	1087	765	538	484	226	226	204

求平均价格Y关于使用年数X的回归方程.

解 观察试验数据的散布图



y与 x呈指数关系

设经验回归方程为

$$y = ae^{bx}, \quad (a > 0, b < 0)$$

两边取对数，得 $\ln y = \ln a + bx$

令 $z = \ln y$, $x = x$, 记 $a' = \ln a$

经变换得回归方程为 $z = a' + bx$

记 $z_i = \ln y_i$, 将原数据转换为

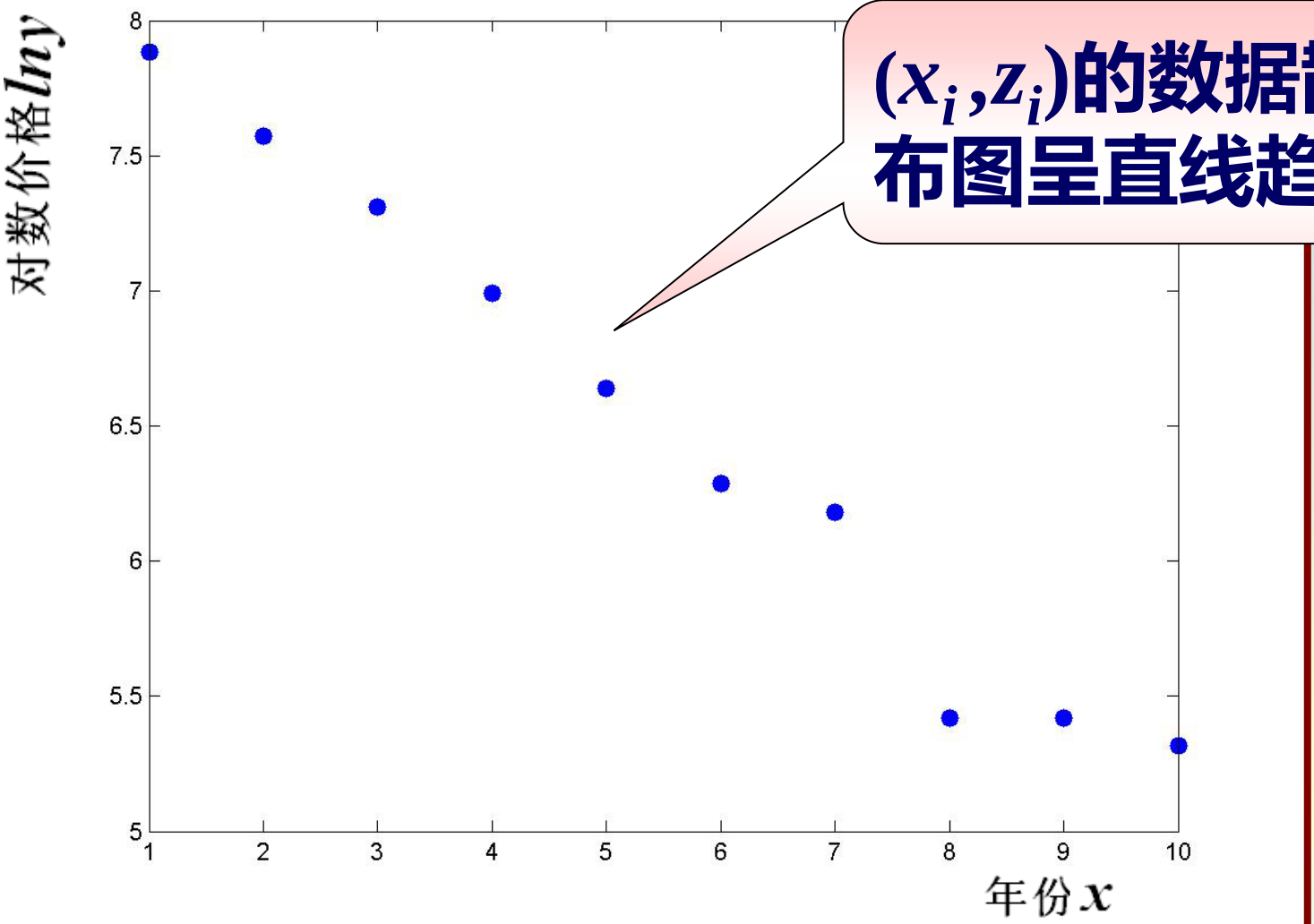
$$(x_i, z_i), \ i = 1, 2, \ldots, 10.$$

x_i	1	2	3	4	5	6	7
z_i	7.88	7.57	7.31	6.99	6.64	6.29	6.18

x_i	8	9	10
z_i	5.67	5.42	5.32

$z=\ln y$

回归分析



(x_i, z_i)的数据散布图呈直线趋势

$$\bar{x} = 5.5, \quad \bar{z} = 6.527,$$

$$l_{xx} = \sum_{i=1}^{10} x_i^2 - 10(\bar{x})^2 = 38.5 - 10 \times 5.5^2 = 82.5$$

$$l_{xz} = \sum_{i=1}^{10} x_i z_i - 10\bar{x}\bar{z} = -24.554$$

$$\hat{b} = \frac{l_{xz}}{l_{xx}} = -\frac{24.5538}{82.5} = -0.2976$$

$$\hat{a}' = \bar{z} - \hat{b}\bar{x} = 6.527 + 0.2976 \times 5.5 = 8.1642$$

从而　　$\hat{z} = 8.1642 - 0.2976x,$

代入原变量，得非线性经验回归方程为

$$\hat{y} = e^{\hat{a}'} e^{\hat{b}x} = 3512.91e^{-0.2976x}$$

检验X与Y是否存在显著的指数相关关系

检验X与$\ln Y$的线性相关关系是否显著

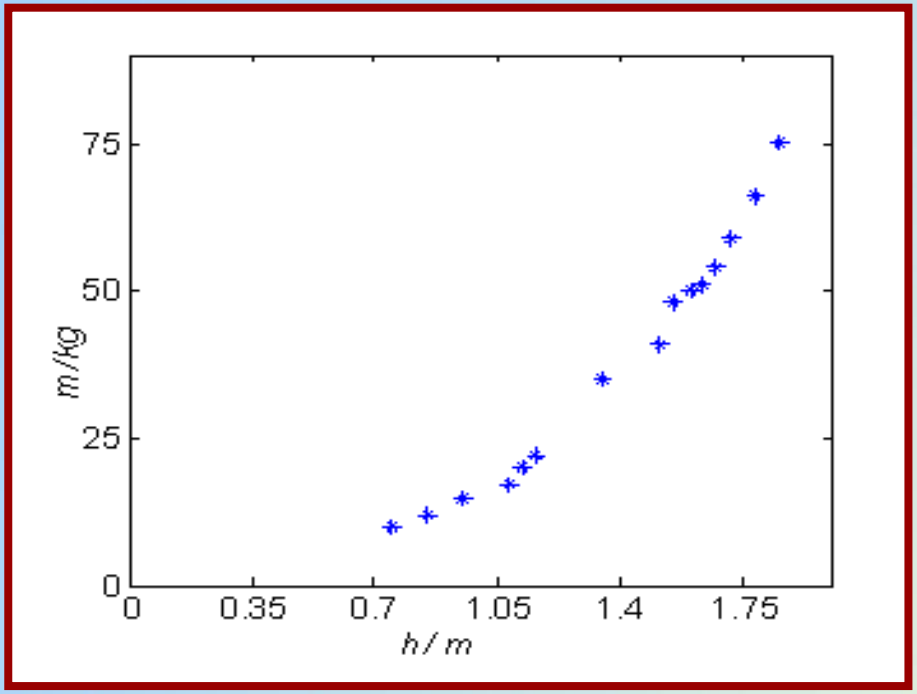
有 $R=\dfrac{l_{xz}}{\sqrt{l_{xx}}\sqrt{l_{zz}}}=-0.996,$

$|R|=0.996>0.765=R_{0.01}(8),$

可以认为X与Y存在显著的指数相关关系.

例9.2.4建模范例（身高体重关系）

现有15对某地区人的身高h 和体重数据m，希望用简洁的函数关系式描述该地区人的身高体重的对应关系.



选定回归函数的估计函数为

$$m = dh^a$$

两边取对数得

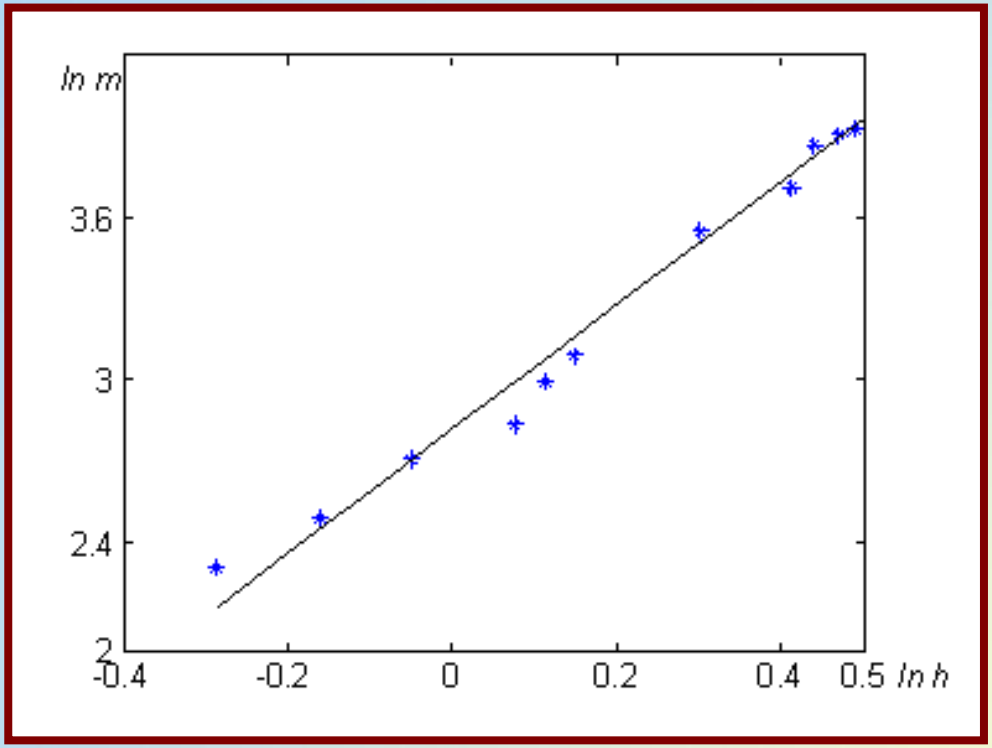
$$\ln m = a \ln h + \ln d$$

记 $y = \ln m, x = \ln h, b = \ln d$ 得线性方程

$$y = ax + b, b \neq 0$$

去掉m=0, h=0, 对原数据做相应变换.

变换数据的散布图



可得经验线性回归方程

$$y = 2.30x + 2.82$$

即 $\ln m = 2.30 \ln h + 2.28$

从而 $m = 16.78 h^{2.3}$

可用来描述该地区人的身高和体重的关系.

问题 请考虑还需做什么工作?