

第九章 回归分析

§ 9.1 相关关系与回归分析

§ 9.2 一元回归分析

§ 1. 回归分析模型

事物之间的关系 $\xleftrightarrow{\text{量化}}$ 变量之间的关系

确定性关系：比如 $S=L^2$


相关关系：比如农作物亩产量 Y 与播种量 X_1 、施肥量 X_2

回归分析：找出相关关系中变量之间的近似关系

我们把要考察的目标作为因变量(记为 Y),
而把影响它的因素称为自变量(记为 X_1 ,
 $X_2, \dots X_k$)。

设自变量(X_1, X_2, \dots, X_k)的取值为: (x_1, x_2, \dots, x_k)

多元回归模型:

$$Y = \mu(x_1, \dots, x_k) + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$


Y 对(X_1, X_2, \dots, X_k)的回归函数

$$y = \mu(x_1, \dots, x_k) \longrightarrow \text{回归方程}$$

在回归分析中: 因变量被看作随机变量

自变量则是可控制的!

分为: 多元回归分析
一元回归分析

回归分析涉及三个问题：

(1) 建立模型(找出自变量与因变量)

(2) 确定回归函数 $\mu(\mathbf{x})$ 的类型

(3) 估计参数

在回归模型中最简单的为一元回归模型：

$$Y = \mu(x) + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

相应的一元回归方程为： $y = \mu(x)$

一元回归最简单的情形：线性回归

§ 9.2一元线性回归分析

一元正态回归模型:

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

其中, a — 回归常数(又称截距)

b — 回归系数(又称斜率)

ε — 随机扰动项

对于任意一组样本, 有 $Y_i = a + bx_i + \varepsilon_i, \quad i=1, \dots, n$

其中: (1) 各次试验相互独立

(2) $E(\varepsilon_i)=0$

(3) $D(\varepsilon_i)=\sigma^2$

(4) $\varepsilon_1, \dots, \varepsilon_n$ 相互独立

从而, $Y_i \sim N(a+bx_i, \sigma^2)$, 且相互独立, 其形状如图:

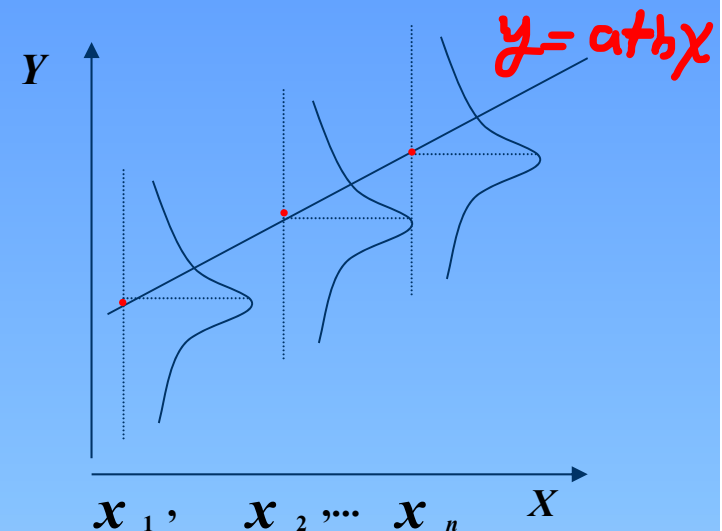
记 \hat{Y}_i 为 Y_i 的估计值, 则

$$Y_i = a + b x_i + \varepsilon_i = \hat{Y}_i + \varepsilon_i$$

这可写成 :

$$\varepsilon_i = Y_i - \hat{Y}_i = Y_i - (a + b x_i)$$

这表明 ε_i 是 Y 的实际观测值与估计值之差, 即拟合误差。



为确定合适的回归方程, 使 Y 的估计值尽可能接近真实值, 为此可选 a, b 的估计值, 使 $\sum \varepsilon_i^2 = \sum (Y_i - \hat{Y}_i)^2$ 达到最小

使误差平方和达到最小以寻求估计值的方法,
叫最小二乘法, 用最小二乘法得到的估计叫
最小二乘估计

由误差 (离差)平方和:

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

要使 Q 达到最小, 应使 Q 对应于 a 、 b 的一阶偏导为0, 即

$$\begin{cases} \frac{\partial Q}{\partial a} = -2 \sum (y_i - a - b x_i) = 0 \\ \frac{\partial Q}{\partial b} = -2 \sum (y_i - a - b x_i) x_i = 0 \end{cases}$$

$$\text{可化为 : } \begin{cases} na + b \sum x_i = \sum y_i & (1) \end{cases}$$

$$\begin{cases} a \sum x_i + b \sum x_i^2 = \sum x_i y_i & (2) \end{cases}$$

(1) $\times \sum x_i$ - (2) $\times n$ 得 :

$$b[n \sum x_i^2 - (\sum x_i)^2] = n \sum x_i y_i - \sum x_i \sum y_i$$

故

$$\begin{aligned}\hat{b} &= \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2} = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2} \\ &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \\ &= \frac{l_{xy}}{l_{xx}}, \text{ 其中, } l_{xx} = \sum (x_i - \bar{x})^2, l_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})\end{aligned}$$

由(1): $\hat{a} = \frac{\sum y_i}{n} - \hat{b} \frac{\sum x_i}{n} = \bar{y} - \hat{b} \bar{x}$

同时, 由于 $\sigma^2 = D(\varepsilon) = E(\varepsilon^2)$

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

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

$$l_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x} \bar{y}$$

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

$$\hat{b} = \frac{l_{xy}}{l_{xx}}$$

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

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

例题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，求Y关于X的线性回归方程。

降雨量 x_i : 110 184 145 122 165 143 78 129

径流量 y_i : 25 81 36 33 70 54 20 44

60 130 168 **1434 (Σ)**

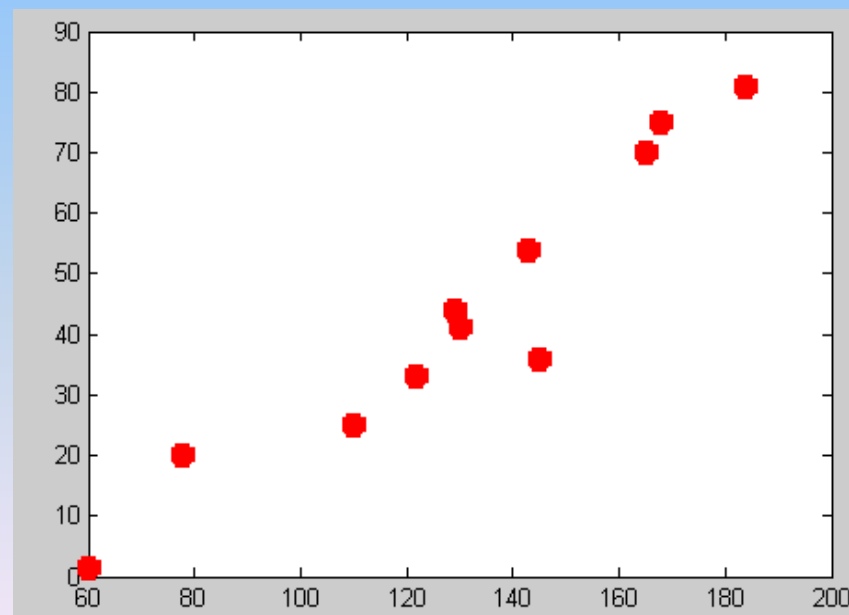
1.4 41 75 **480.4 (Σ)**

解： $n=11$

$$\bar{x} = 130.4 \quad \bar{y} = 43.7$$

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

$$= 14047$$



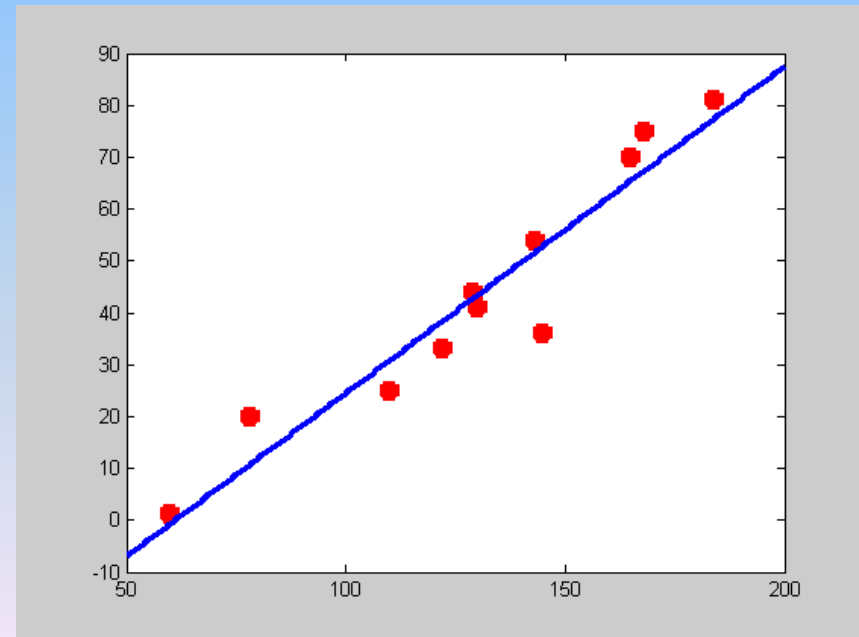
$$l_{xy} = \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} = 71424.8 - 62731.35 = 8795.3$$

$$\begin{cases} \hat{b} = \frac{l_{xy}}{l_{xx}} = \frac{8795.3}{14047} = 0.63 \end{cases}$$

$$\begin{cases} \hat{a} = \bar{y} - \hat{b} \bar{x} = 43.7 - 0.63 \times 130.4 = -38.0 \end{cases}$$

所求经验回归方程为

$$\hat{y} = \hat{a} + \hat{b}x = 0.63x - 38.0$$



随机误差的方差 σ^2 的估计为

$$l_{yy} = \sum_{i=1}^n (y_i - 43.7)^2 = 6050.6$$

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

$$= (6050.6 - 0.63^2 \times 14047) / 9 = 60.37$$

2. 一元线性回归的假设检验（相关系数法）

问题：变量Y与X间是否存在线性相关关系？

相关系数法：是基于试验数据检验变量间线性相关关系是否显著的一种方法。

相关系数

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

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:

$$\begin{aligned}\hat{\rho}_{XY} = R &= \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}} \\ &= \frac{l_{xy}}{\sqrt{l_{xx}} \sqrt{l_{yy}}}\end{aligned}$$

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。

判别准则: 给定显著性水平 α

当 $|R| > R_{\alpha}(n-2)$ 时

认为 X 与 Y 之间的线性相关关系显著。

当 $|R| \leq R_{\alpha}(n-2)$ 时

认为 X 与 Y 之间的线性相关关系不显著。

EX.（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

$$\begin{aligned} R &= \frac{l_{xy}}{\sqrt{l_{xx}} \sqrt{l_{yy}}} \\ &= \frac{8795.3}{\sqrt{14047} \sqrt{6050.6}} = 0.954 \end{aligned}$$

查表得

$$R_{\alpha}(n-2) = R_{0.01}(9) = 0.735 < 0.954 = R$$

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

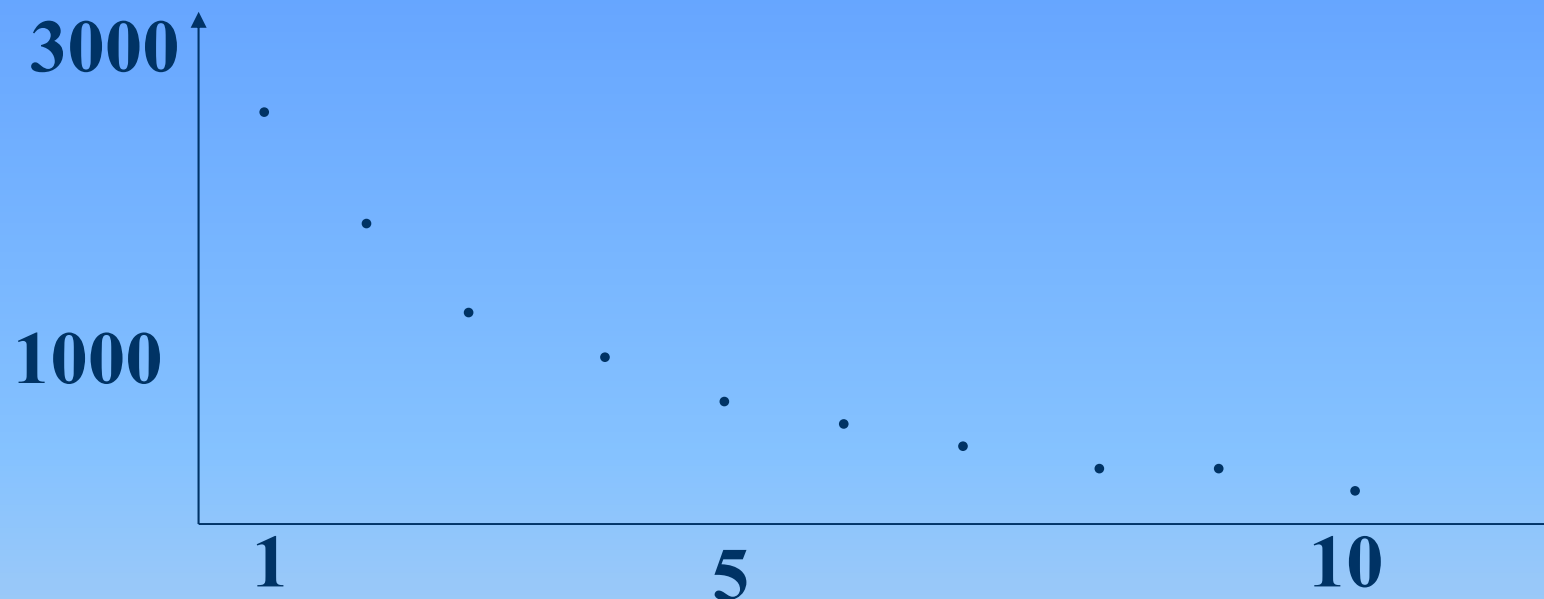
$$y = \mu(x_1, x_2, \dots, x_k)$$

往往是非线性函数。有时可通过适当的变换，将其转化为线性回归问题。

例题2：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
平均价格 y_i	2651	1943	1494	1087	765	538	484
	8	9	10				
	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

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

两边取对数，得 $\ln y = \ln a + bx$

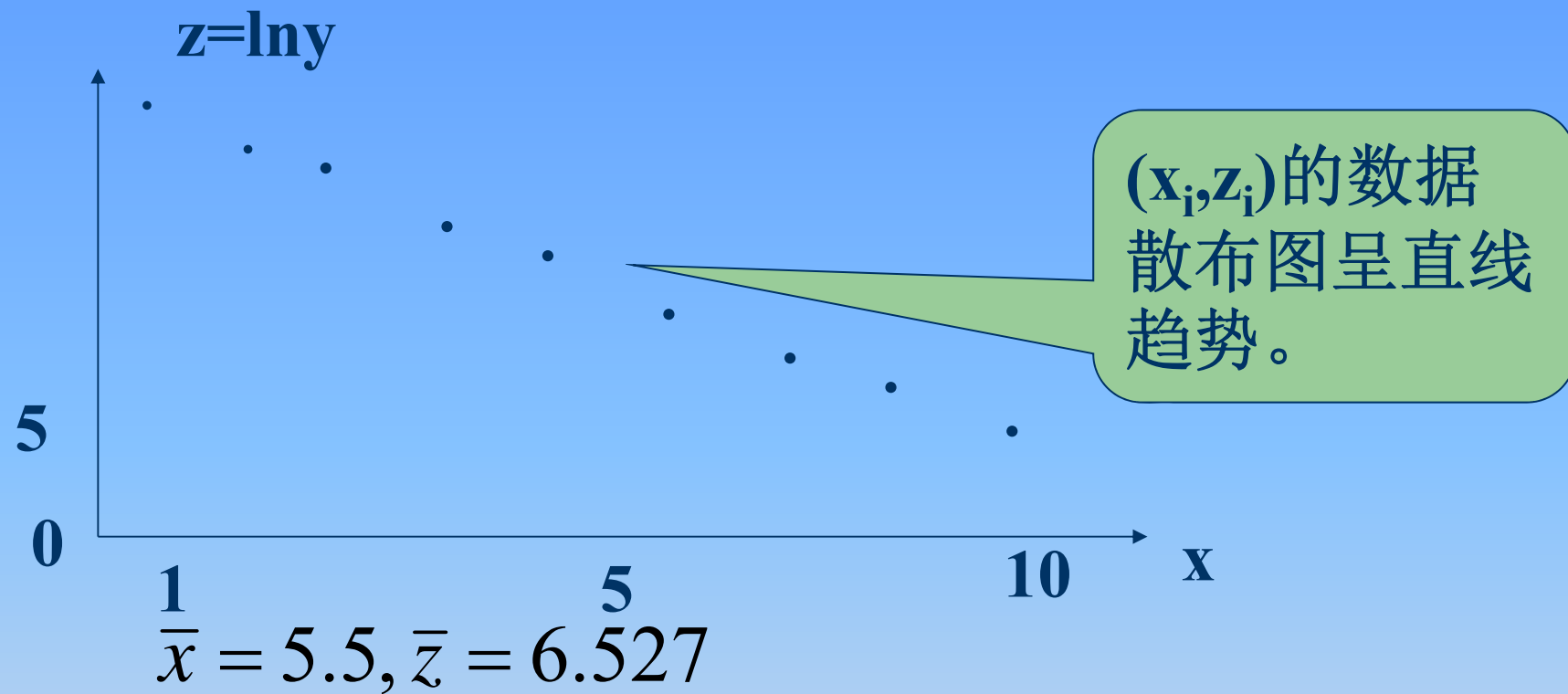
令 $z = \ln y$, $x = x$, 记 $a' = \ln a$

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) , $i=1, 2, \dots, 10$.

x_i	1	2	3	4	5	6	7	8
z_i	7.88	7.57	7.31	6.99	6.64	6.29	6.18	5.67

x_i	9	10
z_i	5.42	5.32



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

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

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

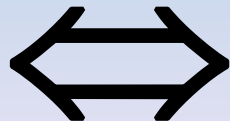
$$\hat{a}' = \bar{z} - \hat{b}\bar{x} = 6.527 + 0.2976 \times 5.5 = 8.1642$$

从而 $\hat{z} = 8.1642 - 0.2976x$

代入原变量，得非线性经验回归方程为

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

检验X与Y是否存在显著的指数相关关系



检验X与lnY的线性相关关系是否显著

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

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

可以认为X与Y存在显著的指数相关关系。