

第七章、回归分析方法

§7.1 一元线性回归

- 回归分析方法是数理统计的重要工具，是处理多个变量之间**相关关系**的一种数学方法。
- **函数**：确定性关系. $h = \frac{1}{2}gt^2$.
- **相关关系**：给定 x 后，不能确定 y 的值. 如，存在测量误差。
- **回归分析**：建立关系，判断公式的有效性，预测、控制。
- **一元线性回归**：**随机变量** Y 与**普通变量** x 之间的线性关系。
- **数据成对观测**：

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n).$$

- **重点**： x_1, x_2, \dots, x_n 不全相等。

例1.1. 某纤维的强度 Y 与拉伸倍数 x 有关. $n = 24$ 数据:

(1.9, 1.4), (2.0, 1.3), (2.1, 1.8), \dots , (9.5, 8.1), (10.0, 8.1).

目标: 找出 x 和 Y 的关系式.

- 散点图:

散点围绕在
一条直线周围.

- 建立回归方程:

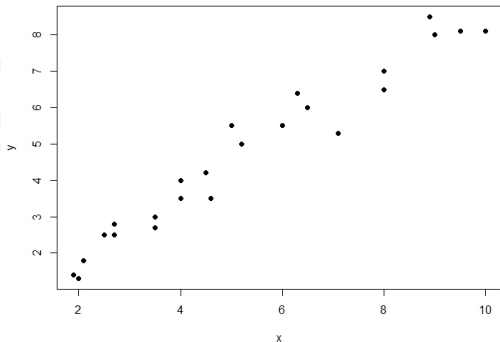
$$\hat{y} = a + bx,$$

回归系数: b .

- 线性:

x 每增加1, y 的变化量是恒定的.

- 非线性: x 每增加1, y 的变化量不是恒定的.



点估计 \hat{a} , \hat{b} .

最小二乘法: 求均方误差 Q 的最小值点 \hat{a} , \hat{b} .

$$Q = Q(a, b) = \sum_{i=1}^n \left(y_i - (a + bx_i) \right)^2. \quad (1.2)$$

- 方法一(微分法)、求解方程组:

$$\frac{\partial Q}{\partial a} = -2 \sum_{i=1}^n (y_i - (a + bx_i)) = 0, \quad (1.3)$$

$$\frac{\partial Q}{\partial b} = -2 \sum_{i=1}^n (y_i - (a + bx_i)) \cdot x_i = 0. \quad (1.4)$$

- 由(1.3) 解得 $a = \bar{y} - b\bar{x}$. 代入(1.4) 以求解 b .

- 由(1.3) 解得 $a = \bar{y} - b\bar{x}$. 代入

$$\frac{\partial Q}{\partial b} = -2 \sum_{i=1}^n (y_i - (a + bx_i)) \cdot x_i = 0. \quad (1.4)$$

- 解: $\star \Rightarrow \sum_{i=1}^n ((y_i - \bar{y}) - b(x_i - \bar{x}))(x_i - \bar{x}) = 0$. 记

$$\ell_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2, \quad \ell_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

则 $\ell_{xy} = b \cdot \ell_{xx}$. 解得 $b = \ell_{xy}/\ell_{xx}$. (前提: x_i 's 不全相等.)

- $\hat{a} = \bar{y} - \hat{b}\bar{x}$, $\hat{b} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) / \sum_{i=1}^n (x_i - \bar{x})^2$.

可以证明 (\hat{a}, \hat{b}) 是 $Q(a, b)$ 的最小值点. (理由: 二阶导数矩阵/海色阵 H 正定.)

$$H = \begin{pmatrix} 2n & 2n\bar{x} \\ 2n\bar{x} & 2(\sum_{i=1}^n (x_i - \bar{x})^2 + n\bar{x}^2) \end{pmatrix}.$$

最小二乘法: 求均方误差 Q 的最小值点 \hat{a}, \hat{b} .

$$Q = Q(a, b) = \sum_{i=1}^n \left(y_i - (a + bx_i) \right)^2. \quad (1.2)$$

- 方法二(配方法)、 (注: $\sum_{i=1}^n \star \cdot \star_i = 0$, $l_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$.)

$$\begin{aligned} Q(a, b) &= \sum_{i=1}^n \left((y_i - \bar{y}) + (\bar{y} - (a + b\bar{x})) - b(x_i - \bar{x}) \right)^2 \\ &= l_{yy} + n \cdot \star^2 + \underbrace{b^2 l_{xx} - 2bl_{xy}} \\ &= l_{yy} + n \cdot \star^2 + \underbrace{l_{xx} \left(b - \frac{l_{xy}}{l_{xx}} \right)^2}_{\text{完全平方项}} - \frac{l_{xy}^2}{l_{xx}}. \end{aligned}$$

- 最小值点: $\star = \star = 0$ 的解, 即, $\hat{a} = \bar{y} - \hat{b}\bar{x}$, $\hat{b} = l_{xy}/l_{xx}$.
最小值: $l_{yy} - l_{xy}^2/l_{xx}$.

总结:

- 均方误差 Q 的最小值点为 $\hat{a} = \bar{y} - \hat{b}\bar{x}$, $\hat{b} = \frac{l_{xy}}{l_{xx}}$.

$$Q = Q(a, b) = \sum_{i=1}^n \left(y_i - (a + bx_i) \right)^2. \quad (1.2)$$

- 回归方程/回归直线/经验公式:

$$y = \hat{a} + \hat{b}x.$$

- $\bar{y} = \hat{a} + \hat{b}\bar{x}$. 因此,

(1) 点 (\bar{x}, \bar{y}) 落在回归直线上, (2) \hat{y}_i 's 的平均值也为 \bar{y} .

$$\hat{y}_i := \hat{a} + \hat{b}x_i.$$

- 例1.1. 由24 个数据计算得 $\hat{y} = \hat{a} + \hat{b}x = 0.15 + 0.859x$.
回归系数 \hat{b} 的含义: 拉伸倍数 x 每增加1, 强度 Y 平均增加0.859.

非线性关系之线性化

例1.2. 彩色显影中, 染料光学密度 Y 与析出银的光学密度 x 关系如下, $A, B > 0$:

$$Y \approx Ae^{-B/x},$$

- 这不是线性关系. 两边取对数得

$$Y^* \approx \ln A - Bx^*, \quad \text{其中, } Y^* = \ln Y, \quad x^* = \frac{1}{x}.$$

- 数据: $(x_1, y_1), \dots, (x_n, y_n) \rightarrow (x_1^*, y_1^*), \dots, (x_n^*, y_n^*)$.
- 建立线性回归方程: $\hat{y}^* = \hat{a} + \hat{b}x^*$.
- 反变换: $\hat{A} = e^{\hat{a}}, \hat{B} = -\hat{b}$.
- 回归方程: $\hat{Y} = \hat{A}e^{-\hat{B}/x}$.

例1.3. 炼钢钢包容积 Y 随使用次数 x 增大. $n = 13$, 数据:

$(2, 106.42), (3, 108.20), (4, 109.58), \dots, (19, 111.20)$

- 画散点图. 用双曲线 $\frac{1}{y} \approx a + b \cdot \frac{1}{x}$.
- 线性回归方程: $\hat{y}^* = \hat{a} + \hat{b}x^*$, $x^* = 1/x$, $y^* = 1/y$,
其中 $\hat{a} = 0.008967$, $\hat{b} = 0.0008292$.
- 经验公式:

$$\frac{1}{\hat{y}} = 0.008967 + 0.0008292 \cdot \frac{1}{x}$$

平方和分解公式

- 前提: x_1, x_2, \dots, x_n 不全相等.

结论: 均方误差 Q 的(唯一的)最小值点为 $\hat{a} = \bar{y} - \hat{b}\bar{x}$, $\hat{b} = \frac{l_{xy}}{l_{xx}}$.

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

估计: $\hat{y} = \hat{a} + \hat{b}x$, $\bar{y} = \hat{a} + \hat{b}\bar{x}$, $\hat{y}_i = \hat{a} + \hat{b}x_i$.

- 注: 经验公式并不都能反映实际情况. 需要判别 x 与 Y 之间是否真的具有线性相关关系: Y 是否随着 x 增大而线性地增大(或者线性地减小).

平方和分解公式:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2. \quad (1.7)$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}, \hat{b} = \frac{\ell_{xy}}{\ell_{xx}}, \hat{y} = \hat{a} + \hat{b}x, \bar{y} = \hat{a} + \hat{b}\bar{x}, \hat{y}_i = \hat{a} + \hat{b}x_i, \bar{\hat{y}} = \bar{y}.$$

- 证: $\star = \star + \star$. 即, $\star = \star - \star$.
- 交叉项: $\sum_{i=1}^n \star \cdot \star = \sum_{i=1}^n \star \cdot \star - \sum_{i=1}^n \star^2$.
- $\star = \hat{b}(x_i - \bar{x})$. 故 $\sum_{i=1}^n \star \cdot \star = \hat{b}\ell_{xy}$, $\sum_{i=1}^n \star^2 = \hat{b}^2\ell_{xx}$.
- $\hat{b} = \frac{\ell_{xy}}{\ell_{xx}}$, 即 $\ell_{xy} = \hat{b}\ell_{xx}$. 故交叉项为0, 从而(1.7) 成立.

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2.$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}, \hat{b} = \frac{l_{xy}}{l_{xx}}, \hat{y} = \hat{a} + \hat{b}x, \bar{y} = \hat{a} + \hat{b}\bar{x}, \hat{y}_i = \hat{a} + \hat{b}x_i, \bar{\hat{y}} = \bar{y}.$$

- **离差平方和**: $l_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$.

(注: $l_{\alpha\alpha}$ 称为 $\alpha = \alpha_i$'s 的离差平方和.)

- **残差平方和**: $Q = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \hat{\varepsilon}_i^2$.

模型拟合的误差的度量, 是自变量和模型不能解释的部分.

越小越好, 说明模型与实际数据越相符. $\hat{\varepsilon}_i$'s 称为残差.)

- **回归平方和**: $U = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \hat{b}^2 l_{xx}$.

源于自变量 x 的分散程度, 通过线性关系影响了因变量 Y , 造成 Y 的变差, 是模型可以解释的部分.

- 一般地, 计算流程如下:

(1) 计算 l_{xx}, l_{yy}, l_{xy} .

(2) 估计 $\hat{b} = l_{xy}/l_{xx}$. 计算 U : $U = \hat{b}^2 l_{xx}$.

(3) 计算 Q : $Q = l_{yy} - U$.

正态模型的方差估计与相关性检验

- 正态模型:

$$Y_i = a + bx_i + \varepsilon_i, \quad i = 1, \dots, n \quad (1.10)$$

其中 $\varepsilon_1, \dots, \varepsilon_n$ 独立同分布, 都服从 $\sim N(0, \sigma^2)$, σ^2 未知.

- 注: 易得 (Y_1, \dots, Y_n) 的联合密度.
- 可以证明: 残差平方和 $Q = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \hat{\varepsilon}_i^2$ 满足

$$\frac{1}{\sigma^2} Q \sim \chi^2(n-2),$$

- $\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \bar{y} - \hat{b}(x_i - \bar{x})$ 满足两个约束方程:

$$\sum_{i=1}^n \hat{\varepsilon}_i = 0, \quad \sum_{i=1}^n \hat{\varepsilon}_i (x_i - \bar{x}) = 0.$$

- σ^2 的无偏估计: $\hat{\sigma}^2 = \frac{1}{n-2} Q$. (因为 $E \frac{1}{\sigma^2} Q = n - 2$.)

- 相关性检验. $H_0 : b = 0$.

(注: H_0 不成立, 则 Y 与 x 有线性相关关系.)

- 相对而言, $U = \hat{b}^2 \ell_{xx}$ 越大 & $Q = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ 越小, 则模型越准确描述 x 和 Y 之间的线性相关关系. 反之, 则 x 和 Y 之间没有线性相关关系.

- 检验统计量:

$$F = F(\vec{x}, \vec{Y}) = \frac{U}{Q/(n-2)}. \quad (1.9)$$

若 F 相当大, 则表明 x 对 Y 的线性影响强, 两者有线性相关性. 否则, 没有线性相关性.

- 可证明: 在 H_0 下, $F \sim F(1, n-2)$.
- 否定域: $\mathcal{W} = \{(\vec{x}, \vec{y}) : F(\vec{x}, \vec{y}) > \lambda\}$, 其中 $\lambda = F_{1-\alpha}(1, n-2)$.
- 结论: $F(\vec{x}, \vec{y}) > \lambda$, 则拒绝 H_0 , 认为 Y 显著地线性依赖于 x .

否则, 接受 H_0 , 认为 x 与 Y 之间没有显著的线性相关性.

- 有些书的检验统计量取样本相关系数:

$$R = R(\vec{x}, \vec{y}) = l_{xy} / \sqrt{l_{xx}l_{yy}}.$$

(回顾, 随机变量 ξ 与 η 的(线性)相关系数: $\rho_{\xi\eta} = \frac{\text{Cov}(\xi, \eta)}{\sqrt{D(\xi)D(\eta)}}$.)

- 当 $|R|$ 很大时, 拒绝 $H_0 : b = 0$.
- 复相关系数平方:

$$R^2 = \frac{l_{xy}^2}{l_{xx}^2} \cdot \frac{l_{xx}}{l_{yy}} = \frac{\hat{b}^2 l_{xx}}{l_{yy}} = \frac{U}{l_{yy}} = 1 - \frac{Q}{l_{yy}}.$$

注: $0 \leq R^2 \leq 1$. R^2 越接近1, 回归模型对数据拟合得越好.

- F 与 R^2 的关系: 一一对应, 严格增,

$$\begin{aligned} F &= \frac{U}{Q/(n-2)} = (n-2) \frac{U}{l_{yy} - U} \\ &= (n-2) \frac{R^2}{1 - R^2} = \frac{n-2}{1/R^2 - 1}. \end{aligned}$$

例1.4. 炼钢, 碳含量 x 越高, 冶炼时间 Y 越长. $n = 34$, 数据:

$(180, 200), (104, 100), \dots, (143, 160)$

- 画散点图, 观察数据, 建线形回归模型.
- 计算结果: $\hat{a} = -23.20, \hat{b} = 1.270, F = 145.0, R^2 = 0.8192$.
- 查 $F(1, n - 2) = F(1, 32)$ 表: $\alpha = 0.01$, 得 $\lambda = 4.15$.
- 下结论: $F > \lambda$, 否定 H_0 , 认为 x, Y 存在线性相关关系. 或: 直线回归是显著的.

应用1. 预报.

- 对新的自变量值 x_0 , 预报

$$Y_0 = a + bx_0 + \varepsilon_0.$$

- 点估计: 用 $\hat{y}_0 = \hat{a} + \hat{b}x_0$ 预报 Y_0 .

- 还需要衡量预报精度.

区间估计: 用 $\hat{\sigma}^2 = \frac{1}{n-2}Q$ 代替 σ^2 .

$$T := \frac{Y_0 - \hat{y}_0}{s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}}} \sim t(n-2).$$

- 查 $t(n-2)$ 表: $\lambda = t_{1-\alpha/2}(n-2)$.

- 区间两端点: $\hat{y}_{\pm} = \hat{y}_0 \pm \lambda s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}}$.

- 区间估计: $\lambda = t_{1-\alpha/2}(n-2)$,

$$\left[\hat{y}_0 - \lambda s^*, \hat{y}_0 + \lambda s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}} \right].$$

- x_0 离 \bar{x} 越远, 预报区间长度越长.
误差标准差的估计 s 越小, 预报区间越短, 预报越精确.
- 注: x_0 不能超出原数据的范围.
- 当 n 较大且 $x_0 - \bar{x}$ 较小时, $\star \approx 1$. 于是预测区间近似为

$$[\hat{y}_0 - \lambda s, \hat{y}_0 + \lambda s].$$

- 当 n 较大时, λ 可用标准正态分布的临界值 $z_{1-\alpha/2}$.

应用2: 控制.

- 要求控制 Y 在区间 $[A, B]$ 内, 如何选取 x ?
- 办法: 让 $\hat{y}_{\pm} \in [A, B]$, 反解出 x 的区间.

回归诊断和残差分析

- 即使否定了 $H_0 : b = 0$, 也并不说明模型就是合适的.
- 常见问题包括:
 - (1) 缺少重要自变量;
 - (2) 有非线性相关;
 - (3) 误差项方差非恒定;
 - (4) 误差项存在序列相关;
 - (5) 自变量严重共线 (多元回归中);
 - (6) 数据有异常值或强影响点.
 - (7) 可以用残差散点图等进行回归诊断.

- 残差: $\hat{\varepsilon}_i = y_i - \hat{y}_i$.
- 令

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{l_{xx}}, \quad s = \sqrt{\frac{Q}{n-2}}, \quad r_i = \frac{\hat{\varepsilon}_i}{s\sqrt{1-h_i}}.$$

- 近似地, r_1, \dots, r_n 相互独立, 且都服从 $N(0, 1)$. 故

$$P(|r_i| > 2) \approx 0.05.$$

- 当 n 比较大时, r_i 's 应该只有约 $[0.05n]$ 个的绝对值大于 2.

注: 可用来检验模型关于误差项的假设是否成立, 以及发现异常值点.