

第九章、回归分析

§9.1 引言

- 自变量/解释变量: x , 因变量/响应变量: y .
- 回归函数: f , 未知.
- 回归模型/方程/关系: $y = f(x) + e$. 其中, e 是误差.
- 例: $x =$ 路程(可设定), $y =$ 耗油量.
 $x =$ 父亲身高(不可设定, 只可测量), $y =$ 儿子身高.
关心 f , 不关心自变量如何变化.
将 x 视为已知参数, 将 y, e 视为随机变量或其取值.

- 一元线性回归(正态)模型:

$$y = b_0 + b_1x + e, \quad e \sim N(0, \sigma^2),$$

其中 b_0, b_1, σ^2 为未知参数.

- 数据 $(x_i, y_i), i = 1, \dots, n$.

$$y_i = b_0 + b_1x_i + e_i, \quad i = 1, \dots, n.$$

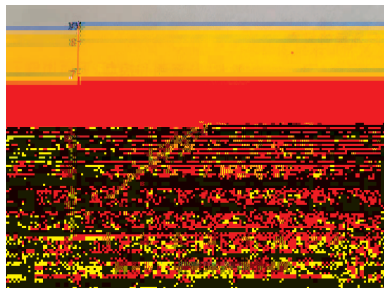
- x_i 是已知参数,
 y_i 是随机变量(或其取值), 可观测.
 e_1, \dots, e_n 是i.i.d. 随机变量(或其取值), 不可观测(因为 b_0, b_1 是未知参数).

例1.1. x 与 y 分别代表某个体的两个特征. 数据: (x_i, y_i) ,
 $i = 1, \dots, n = 50$.

问: x 与 y 之间什么依赖关系?

- 散点图:

-



回归方程的应用.

- 例1.2 (预测). $x =$ 水的沸点, $y =$ 大气压.

由 $n = 17$ 组数据得到预测公式:

$$y = -43.131 + 0.895x + e.$$


某地测得 $x = x_0$, 那么, 可预测 $Y_0 = \hat{b}_0 + \hat{b}_1 x_0 + e_0$.

- 例1.3 (预测与控制). $x =$ 某小区人口数, $y =$ 冬季用煤量, $z =$ 室温.

通过数据 (x_i, y_i, z_i) , $i = 1, \dots, n$ 得到回归关系:

$$y = a + bx + e, \quad z = d + fy + \varepsilon.$$

预测: 根据某小区人口数 x_0 , 预测用煤量 $Y_0 = \hat{a} + \hat{b}x_0 + e_0$.

控制: 为控制 $z \in [17, 18]$, 应该储备  煤(反求 y)?

回归模型与最大似然估计 (§9.2)

§9.2 ~ §9.4 一元线性回归及其参数检验

$$y = b_0 + bx + e, \quad e \sim N(0, \sigma^2),$$

其中, σ^2 未知. 数据: $(x_i, y_i), i = 1, \dots, n$.

- 回归模型: $y_i = b_0 + bx_i + e_i, i = 1, \dots, n$.

$$p_{Y_i}(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - (b_0 + bx_i))^2},$$

x_i : 已知参数; b_0, b : 待估参数; σ^2 : 讨厌参数.

- 似然函数: $L(b_0, b, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}^n} e^{-\frac{1}{2\sigma^2} Q(b_0, b)}$.

均方误差: $Q(b_0, b) = \sum_{i=1}^n [y_i - (b_0 + bx_i)]^2$.

- 定义2.1. $Q(b_0, b)$ 的最小值点 \hat{b}_0, \hat{b} 被称为最小二乘拟合系数, 或 b_0 与 b 的最小二乘估计.
- 最大似然估计: $\hat{b}_0, \hat{b}, \hat{\sigma}^2 = \frac{1}{n} Q(\hat{b}_0, \hat{b})$.

定理2.2 & 定理3.2. 若 x_i 不全相等, 则 \hat{b}_0, \hat{b} 是最优线性无偏估计.

$$\hat{b}_0 = \bar{y} - \hat{b}\bar{x}, \quad \hat{b} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{l_{xy}}{l_{xx}}.$$

• \hat{b}_0, \hat{b} 是 (y_1, \dots, y_n) 的线性函数.

• $y_i = b_0 + bx_i + e_i, \quad \bar{y} = b_0 + b\bar{x} + \bar{e},$

$$y_i - \bar{y} = b(x_i - \bar{x}) + (e_i - \bar{e}).$$

• $E\hat{b} = b: e_1, \dots, e_n$ i.i.d., 且 $e_1 \sim N(0, \sigma^2),$

$$\hat{b} = b + \frac{1}{l_{xx}} \sum_{i=1}^n (x_i - \bar{x})(e_i - \bar{e}) = b + \frac{1}{l_{xx}} \sum_{i=1}^n (x_i - \bar{x})e_i,$$

• $E\hat{b}_0 = b_0:$

$$\hat{b}_0 = \bar{y} - \hat{b}\bar{x} = (b_0 + b\bar{x} + \bar{e}) - \hat{b}\bar{x} = b_0 + (b - \hat{b})\bar{x} + \bar{e}.$$



$$H_0 : b = 0 \leftrightarrow H_1 : b \neq 0.$$

- $Q(b_0, b) = \sum_{i=1}^n [y_i - (b_0 + bx_i)]^2,$

$$L(\hat{\theta}) = \left(\sqrt{2\pi\hat{\sigma}^2}\right)^{-n/2} e^{-\frac{n}{2}}, \quad \hat{\sigma}^2 = \frac{1}{n}Q(\hat{b}_0, \hat{b})$$

$$L(\hat{\theta}_0) = \left(\sqrt{2\pi\check{\sigma}_0^2}\right)^{-n/2} e^{-\frac{n}{2}}, \quad \check{\sigma}_0^2 = \frac{1}{n}Q(\bar{y}, 0).$$

- $Q = Q(\hat{b}_0, \hat{b}) = \sum_{i=1}^n (y_i - \hat{y}_i)^2$, 残差平方和;

$$Q(\bar{y}, 0) = \sum_{i=1}^n (y_i - \bar{y})^2 = l_{yy},$$

- 广义似然比: $\lambda(\vec{y}) = L(\hat{\theta})/L(\hat{\theta}_0) = (l_{yy}/Q)^{n/2}.$

- 广义似然比否定域:

$$\mathcal{W} = \{\vec{y} : l_{yy}/Q > c_1\}.$$

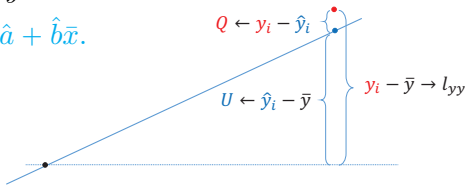
正交分解. $\hat{a} = \bar{y} - \hat{b}\bar{x}$, $\hat{b} = \frac{l_{xy}}{l_{xx}}$. 产生直线 \hat{f} : $\hat{y} = \hat{a} + \hat{b}x$,

- \hat{f} 过点 (\bar{x}, \bar{y}) , $\bar{y} = \hat{f}(\bar{x}) = \bar{\hat{y}}$.

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

- 残差平方和 Q :

$$\begin{aligned} Q &= \sum_{i=1}^n (y_i - (\hat{a} + \hat{b}x_i))^2 \\ &= \sum_{i=1}^n \underbrace{(y_i - \hat{y}_i)}^2. \end{aligned}$$



- 回归平方和 U :

$$U = \sum_{i=1}^n \underbrace{(\hat{y}_i - \bar{y})}^2 = \hat{b}^2 \sum_{i=1}^n (x_i - \bar{x})^2.$$

- 引理2.1. $l_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = U + Q$.

- $\sum_{i=1}^n \underbrace{y_i}_{**} \underbrace{(x_i - \bar{x})}_{**} = \sum_{i=1}^n \underbrace{(y_i - \bar{y})}_{**} \underbrace{\hat{b}(x_i - \bar{x})}_{**} - \sum_{i=1}^n \underbrace{(\hat{y}_i - \bar{y})}_{**} \underbrace{(\hat{y}_i - \bar{y})}_{**}$
 $= \hat{b}l_{xy} - \hat{b}^2 l_{xx} = 0$.

- 广义似然比否定域: $\mathcal{W} = \{\vec{y} : U/Q > c_2\}$.

命题4.1. 残差平方和 $Q = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ 与回归平方和 $U = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \hat{b}^2 \ell_{xx}$ 相互独立, 且

$$\frac{1}{\sigma^2} Q \sim \chi^2(n-2); \quad \text{若 } b=0, \text{ 则 } \frac{U}{\sigma^2} \sim \chi^2(1).$$

- $\hat{b}_0 = b_0 + (b - \hat{b})\bar{x} + \bar{e}, \quad \hat{b} = b + \frac{1}{\ell_{xx}} \sum_i (x_i - \bar{x})(e_i - \bar{e}).$
- $\begin{aligned} \hat{y}_i - y_i &= (\hat{b}_0 + \hat{b}x_i) - (b_0 + bx_i + e_i) \\ &= (b - \hat{b})\bar{x} + \bar{e} + (\hat{b} - b)x_i - e_i \\ &= (\hat{b} - b)(x_i - \bar{x}) - (e_i - \bar{e}). \end{aligned}$
- $\begin{aligned} (\hat{y}_i - y_i)^2 &= (\hat{b} - b)^2 (x_i - \bar{x})^2 + (e_i - \bar{e})^2 \\ &\quad - 2(\hat{b} - b)(x_i - \bar{x})(e_i - \bar{e}). \end{aligned}$
- $\begin{aligned} Q &= (\hat{b} - b)^2 \ell_{xx} + \sum_i (e_i - \bar{e})^2 - 2(\hat{b} - b) \ell_{xx} (\hat{b} - b), \\ &= \sum_i (e_i - \bar{e})^2 - \ell_{xx} (\hat{b} - b)^2. \end{aligned}$

命题4.1(续). $Q = \sum_{i=1}^n (y_i - \hat{y}_i)^2$. 则 $\frac{1}{\sigma^2}Q \sim \chi^2(n-2)$.

- $Q = \sum_i (e_i - \bar{e})^2 - l_{xx}(\hat{b} - b)^2$, $\hat{b} = b + \sum_i \frac{x_i - \bar{x}}{l_{xx}} e_i$.

- $\sum_i (e_i - \bar{e})^2$:

$$\begin{aligned} \sum_i (e_i - \bar{e})^2 &= \sum_i e_i^2 - n\bar{e}^2 \\ &= \sum_i e_i^2 - \left(\sum_i \frac{1}{\sqrt{n}} e_i \right)^2 = \sum_i e_i^2 - \left(\sum_i a_{1i} e_i \right)^2. \end{aligned}$$

- $l_{xx}(\hat{b} - b)^2$:

$$l_{xx}(\hat{b} - b)^2 = \left(\sum_i \frac{x_i - \bar{x}}{\sqrt{l_{xx}}} e_i \right)^2 = \left(\sum_i a_{2i} e_i \right)^2.$$

- 再补 $n-2$ 行, 可得正交矩阵 $A = (a_{ki})_{n \times n}$:

$$\sum_i a_{1i}^2 = \sum_i a_{2i}^2 = 1, \quad \sum_i a_{1i} a_{2i} = 0.$$

命题4.1(续). $Q = \sum_{i=1}^n (y_i - \hat{y}_i)^2$. 则 $\frac{1}{\sigma^2}Q \sim \chi^2(n-2)$.

- $Q = \sum_i e_i^2 - (\sum_i a_{1i}e_i)^2 - (\sum_i a_{2i}e_i)^2$.

$A = (a_{ki})_{n \times n}$ 为正交矩阵.

- $W_i = e_i/\sigma$: W_1, \dots, W_n i.i.d., $W_1 \sim N(0, 1)$, 则

$$\begin{aligned}EZ_k Z_\ell &= E \sum_i a_{ki} W_i \sum_j a_{lj} W_j \\ &= \sum_{i,j} a_{ki} a_{lj} E W_i W_j = \sum_i a_{ki} a_{li} = 1_{\{k=\ell\}}.\end{aligned}$$

$$(W_1, \dots, W_n)^T \stackrel{d}{=} (Z_1, \dots, Z_n)^T := A(W_1, \dots, W_n)^T.$$

- $Q = \sigma^2 (\sum_i Z_i^2 - Z_1^2 - Z_2^2) = \sigma^2 \sum_{i=3}^n Z_i^2$.

命题4.1(续). 回归平方和 $U = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \hat{b}^2 l_{xx}$ 与 Q 相互独立, 且若 $b = 0$, 则

$$\frac{U}{\sigma^2} \sim \chi^2(1).$$



- $Q = \sigma^2 \sum_{i=3}^n Z_i^2, \quad U = (\sqrt{\ell_{xx}b} + \sigma Z_2)^2.$



例2.1. $x =$ 注射后天数,
 $y =$ 金残留百分数.

- 根据散点图建立函数

$$\ln y = b_0 + bx + e.$$

- 求 \bar{x} , \bar{z} ; $\hat{z}_i = \hat{b}_0 + \hat{b}x_i$:

$$\hat{b} = l_{xz}/l_{xx}, \hat{b}_0 = \bar{z} - \hat{b}\bar{x};$$

- 求残差平方和: $Q = \sum_i (z_i - \hat{z}_i)^2$
 与回归平方和: $U = \sum_i (\hat{z}_i - \bar{z})^2$.

- $n = 10$, 根据 $\lambda = F_{0.95}(1, n - 2) = F_{0.95}(1, 8) = 5.32$.

$\frac{U}{Q/(n-2)} = 344.82 > \lambda$ 否定 H_0 , 强烈可 z 线性依赖于 x .

