AI 中的数学

第二十四讲

方聪，概率统计部分参考章复熹和张原老师课件

2024 年秋季

实际问题中，两个变量之间往往有某种依赖关系

$$y = f(x) + e$$

其中 e 是误差项，为一个随机变量，该方程称为回归模型或回归方程，y 和 x 的这种关系称为回归关系 (或者相关关系)，称 x 为自变量/解释变量，称 y 为因变量/响应变量，f 为回归函数.

实际问题中，两个变量之间往往有某种依赖关系

$$y = f(x) + e$$

其中 e 是误差项，为一个随机变量，该方程称为回归模型或回归方程，y 和 x 的这种关系称为回归关系 (或者相关关系)，称 x 为自变量/解释变量，称 y 为因变量/响应变量，f 为回归函数.

- 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:

$$y = b_0 + b_1 x + e, \quad e \sim N\left(0, \sigma^2\right),$$

其中 b_0, b_1, σ^2 为未知参数.

- 数据 $(x_i, y_i), i = 1, \cdots, n.$

$$y_i = b_0 + b_1 x_i + e_i, \quad i = 1, \cdots, n.$$

- x_i 是已知参数, y_i 是随机变量 (或其取值), 可观测.
$e_1, \cdots, e_n$ 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: $(x_i, y_i), i = 1, \cdots, n = 50.$
问: x 与 y 之间什么依赖关系?

- 散点图:

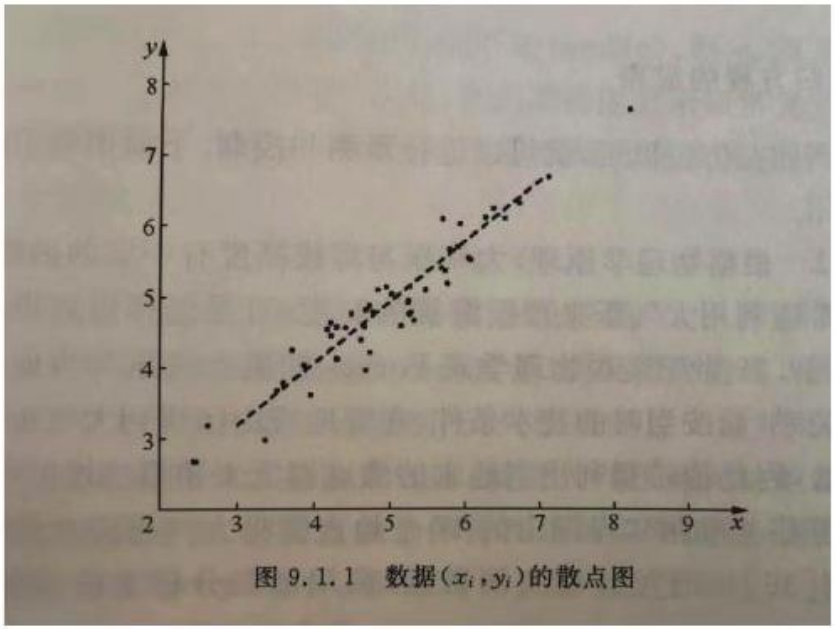

图 9.1.1 数据(x_i, y_i)的散点图

- 初步判断:

$$y_i = b_0 + b_1 x_i + e_i,$$
$$i = 1, \cdots, n.$$

回归方程的应用.

- $x =$ 父亲身高 (不可设定, 只可测量), $y =$ 儿子身高.
- 例: $x =$ 路程 (可设定), $y =$ 耗油量
- 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 $(x_i, y_i, z_i), i = 1, \cdots, n$ 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 $z \in [17, 18]$, 应该储备多少煤 (反求 y)?

考虑一元线性回归问题

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

其中，σ^2 未知. 数据：$\left(x_i, y_i\right), i = 1, \cdots, n.$
回归模型：$y_i = b_0 + bx_i + e_i, i = 1, \cdots, n.$

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

x_i：已知参数；b_0, b：待估参数；σ^2：讨厌参数.
似然函数：$L\left(b_0, b, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}^n} e^{-\frac{1}{2\sigma^2}Q(b_0, b)}$，其中
$Q\left(b_0, b\right) = \sum_{i=1}^{n}\left[y_i - \left(b_0 + bx_i\right)\right]^2$ 称为均方误差。

定义： $Q(b_0, b)$ 的最小值点 $\hat{b}_0, \hat{b}$ 被称为最小二乘拟合系数，或 b_0 b 的最小二乘估计.

最大似然估计: $\hat{b}_0, \hat{b}, \hat{\sigma}^2 = \frac{1}{n} Q\left(\hat{b}_0, \hat{b}\right)$

定理： 假设 $x_1, \cdots x_n$ 不完全相同，则

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

其中, $\ell_{uv} = \sum_{i=1}^{n}(u_i - \bar{u})(v_i - \bar{v})$.

证明：只需找 $Q(a,b)=\sum_{i=1}^{n}\left[y_i-(b_0+bx_i)\right]^2$ 的最小值点，注意到对 $w_i=y_i-(b_0+bx_i)$ 有

$$\sum_{i=1}^{n}w_i^2=\sum_{i=1}^{n}(w_i-\bar{w})^2+n\bar{w}^2.$$

因此

$$Q(a,b)=\sum_{i=1}^{n}\left[(y_i-\bar{y})-b\,(x_i-\bar{x})\right]^2+n\left(\bar{y}-(b_0+b\bar{x})\right)^2.$$

最小值点有 $\hat{b}_0=\bar{y}-b\bar{x}$，代入 $Q(a,b)$ 得

$$Q(a,b)=\ell_{yy}-2b\ell_{xy}+b^2\ell_{xx},$$

最小值点为 $\hat{b}=\frac{\ell_{xy}}{\ell_{xx}}$

定理：若 x_i 不全相等，则 $\hat{b}_0, \hat{b}$ 是 (最优) 线性无偏估计.

证明：$\hat{b}_0, \hat{b}$ 是 $(y_1, \cdots, y_n)$ 的线性函数.

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

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

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

由于 $e_1, \cdots, e_n$ i.i.d., 且 $e_1 \sim N(0, \sigma^2)$,

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

故 $E(\hat{b}) = b$。

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

故 $E(\hat{b}_0) = b_0$.

对于参数: $\theta = \left(b_0, b, \sigma^2\right)$，考虑假设检验问题（通常称为相关性检验问题）:

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

否定 H_0, 则表明 y 与 x 之间有线性依赖关系，因此

$$\Theta = \left\{\theta : b_0, b \in \mathbb{R}, \sigma^2 > 0\right\}, \Theta_0 = \left\{\theta \in \Theta : b = 0\right\}$$

对于参数: $\theta = \left(b_0, b, \sigma^2\right)$，考虑假设检验问题（通常称为相关性检验问题）:

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

否定 H_0, 则表明 y 与 x 之间有线性依赖关系，因此

$$\Theta = \left\{\theta : b_0, b \in \mathbb{R}, \sigma^2 > 0\right\}, \Theta_0 = \left\{\theta \in \Theta : b = 0\right\}$$

似然函数: $L(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\pi\sigma^2} Q(b_0, b)}$，其中
$Q\left(b_0, b\right) = \sum_{i=1}^{n} \left[y_i - \left(b_0 + b x_i\right)\right]^2$..
Θ 上的最大似然估计: $\hat{\theta} = \left(\hat{b}_0, \hat{b}, \hat{\sigma}^2\right), \hat{b}_0 = \bar{y} - \hat{b}\bar{x}, \hat{b} = \frac{\ell_{xy}}{\ell_{xx}}$,

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

Θ_0 上的最大似然估计: $\check{\theta}_0 = \left(\check{b}_0, \check{b}, \check{\sigma}^2\right), \check{b}_0 = \bar{y}, \check{b} = 0$

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

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

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

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

- $Q = Q\left(\hat{b}_0, \hat{b}\right) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$, 残差平方和;
  $Q(\bar{y}, 0) = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \ell_{yy}$,
- 广义似然比: $\lambda(\vec{y}) = L(\hat{\theta}) / L\left(\hat{\theta}_0\right) = (\ell_{yy}/Q)^{n/2}$.
- 广义似然比否定域:

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

- 残差平方和 Q：
$$Q = \sum_{i=1}^{n} \left(y_i - \left(\hat{a} + \hat{b} x_i \right) \right)^2$$
$$= \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 .$$

- 回归平方和 U：$U = \sum_{i=1}^{n} \underline{(\hat{y}_i - \bar{y})^2} = \hat{b}^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$.
- 引理 2.1. $\ell_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = U + Q$.
- $\ell_{yy} - (U + Q) = 2 \sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$

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