8.1 成对数据的统计相关性

FIG · E3-8-1

01 · 核心定理、公式与阶段边界

样本相关系数主公式：

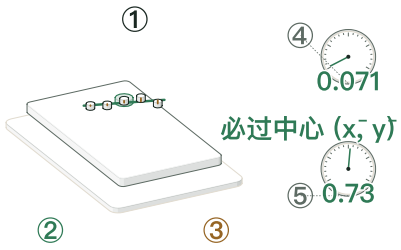
$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$(-1\le r\le 1)$$

- 强弱比较看 $|r|$，但不存在适用于一切情境的固定强、中、弱阈值；
- 本页用0.95、0.70、0.20三组同尺度散点比较线性紧密程度；
- 不能只由r判断预测误差的绝对大小或因果关系；
- $r=0$：**无线性相关**（注意：不等于无任何关系，可能存在完美的非线性抛物线等关系！）。

02 · 动手实验台

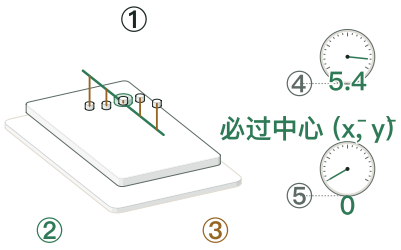
姿态 A · 基准状态



① 最小二乘残差弹簧平衡杆台 · 斜率 b=0.73, 中心偏离 Δa=0.00
② 最优回归斜率 b̂ = 0.73 (过样本中心)
③ 残差平方和 RSS = ∑e_i² = 0.071
④ 残差平方和 RSS · 量程[0, 6]
⑤ 回归斜率 b · 量程[0, 1.4]

斜率 b 0.73 · 截距偏离 Δa 0.0

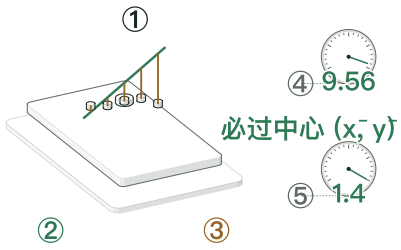
姿态 B · 下界/初始端



① 最小二乘残差弹簧平衡杆台 · 斜率 b=0.00, 中心偏离 Δa=0.00
② 最优回归斜率 b̂ = 0.73 (过样本中心)
③ 残差平方和 RSS = ∑e_i² = 5.400
④ 残差平方和 RSS · 量程[0, 6]
⑤ 回归斜率 b · 量程[0, 1.4]

斜率 b 0.00 · 截距偏离 Δa 0.0

姿态 C · 上界/满量程端



① 最小二乘残差弹簧平衡杆台 · 斜率 b=1.40, 中心偏离 Δa=1.00
② 最优回归斜率 b̂ = 0.73 (过样本中心)
③ 残差平方和 RSS = ∑e_i² = 9.560
④ 残差平方和 RSS · 自动量程[0, 10]
⑤ 回归斜率 b · 量程[0, 1.4]

斜率 b 1.40 · 截距偏离 Δa 1.0

03 · 诊断与易错辨析自测（先独立作答，再核对答案页）

1. x：$1,2,3$；y：$6,4,2$。样本相关系数 r 是?

☐ A -2　☐ B -1　☐ C 1　☐ D 0

2. 四组成对数据的相关系数：0.3、-0.85、0.6、-0.1。线性相关最强的是哪一组?

☐ A 0.6　☐ B -0.85　☐ C -0.1　☐ D 0.3

3. 一组成对数据算出 $r=0$。下面哪句话对?

☐ A x 和 y 之间没有任何关系　☐ B y 一定不是 x 的函数

☐ C这组数据没有线性相关，但可能有曲线关系

4. 某城市逐月统计，冰淇淋销量和溺水人数的相关系数接近 1。下面哪个结论合理?

☐ A多吃冰淇淋会导致溺水

☐ B两者正相关，但可能都受气温影响，不能推出因果

☐ C禁止卖冰淇淋就能减少溺水

参考答案与防坑要点：

(1) [B]

B。$\bar{x}=2$，$\bar{y}=4$；偏离 $-1,0,1$ 和 $2,0,-2$；乘积和 $-2+0-2=-4$；平方和 2 和 8。$r=\dfrac{-4}{\sqrt{16}}=-1$。三个点正好在直线 $y=-2x+8$ 上。选 A 的同学把斜率 -2 当成了 r：r 不会超出 $[-1,1]$。

(2) [B]

B。比 $|r|$：0.85 最大。选 A 的同学只在正数里挑了最大的；选 C 的同学以为数小就“最负、最强”，可 $|-0.1|$ 是最小的。

(3) [C]

C。r 只量直线关系。“讲清楚”里 $y=x^2$ 取 $x=-2,\dots,2$ 的 5 个点，$r=0$，可 y 是 x 的函数，A、B 都错。“动手”③里 $h=5$ 就是这种情形。

(4) [B]

B。相关不等于因果。天热时买冰淇淋的人多，去游泳的人也多，气温同时推高两者。A、C 都把“一起变”当成了“一个导致另一个”。