

第十一章 线性回归与相关

第一节 线性回归

一、线性回归的概念

一句话抓住本质

线性回归不是在“画一条最好看的直线”，
而是在问：

👉 当 X 变化 1 个单位时， Y 的平均水平会如何变化？

1. 线性回归在解决什么问题？ (Purpose)

当你面对的是：

- 两个定量变量 X, Y
- 观察到：

X 变 $\rightarrow Y$ 也跟着变

- 你想要的不是“相关强不强”，
而是：

数量关系如何？能不能预测？

👉 线性回归 = 用一个数学模型刻画这种依存关系

2. 模型长什么样？ (Model)

$$\hat{Y} = a + bX$$

这是整章里最重要的一行式子。

每一项的真实含义（一定要说清）

- \hat{Y} :
 - 👉 在给定 X 时, Y 的总体均数的估计值
(不是某个个体的值)
- a (截距, intercept):
 - 👉 当 $X = 0$ 时, Y 的理论平均水平
 - 👉 有时有实际意义, 有时只是数学量
- b (回归系数, regression coefficient):
 - 👉 X 每增加 1 个单位, Y 的平均变化量

🔴 真正有解释力的是 b 。

3. 回归系数 b 在“数学上”干了什么? (直觉解释)

$$b = \frac{l_{XY}}{l_{XX}} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2}$$

你可以把它拆成一句话:

$b =$
 X 和 Y 的“共同变化程度”
 \div
 X 自己的“变化程度”

进一步直觉化

- 分子 $\sum(X - \bar{X})(Y - \bar{Y})$:
 - 👉 X 偏大时, Y 是不是也偏大?
(同向 \rightarrow 正; 反向 \rightarrow 负)
- 分母 $\sum(X - \bar{X})^2$:
 - 👉 X 本身到底波动得有多厉害?

👉 所以:

b 本质上是：
把 Y 的变化，
归因到 X 的变化上。

4.截距 a 在干什么？（别神化）

$$a = \bar{Y} - b\bar{X}$$

直觉解释：

既然直线要经过
 (\bar{X}, \bar{Y}) ，
那剩下的位置自然就被确定了。

📌 **重要结论（常考）：**

最小二乘回归直线一定经过样本均值点
 (\bar{X}, \bar{Y}) 。

5.最小二乘法到底在“最小”什么？（Very important）

最小二乘法的目标函数是：

$$Q = \sum (Y - \hat{Y})^2$$

注意关键词：**纵向距离**

- 不是点到直线的最短距离
 - 而是：
👉 **Y 方向的残差**
-

直觉翻译

在所有可能的直线中,
选那一条,
使“预测值和真实值之间的偏差平方和”最小。

✦ 这意味着:

- 回归是 **以 Y 为被解释对象**
- X 被当作“已知、无误差”的变量
(这是回归的一个重要隐含假设)

6.线性回归在统计上的角色 (别混)

相关分析:

问“有没有关系? 强不强?”

回归分析:

问“关系怎么写? 能不能预测?”

👉 **回归是有方向性的 ($X \rightarrow Y$)**

👉 **不是对称的**

7.什么时候“可以”用线性回归? (先有直觉)

你暂时只要记住这三个关键词:

- **线性趋势明显**
- **X 是解释变量**
- **Y 是连续型变量**

其他严格条件 (误差正态、同方差、独立性),
会在后面章节系统展开。

8.考试级一句话总结（直接可写）

线性回归是通过最小二乘法建立因变量与自变量之间数量关系的统计方法，用回归方程 $\hat{Y} = a + bX$ 描述当自变量变化时因变量总体均数的变化趋势，其中回归系数 b 表示自变量每增加一个单位时因变量平均变化的大小。

给你一个“不会跑偏”的心智锚点

相关：对称、描述

回归：有方向、可预测

只要你在脑中始终记住

“我是在解释 Y，不是在解释 X”

线性回归这一章就不会被你学成“公式堆”。

二、回归方程的估计

【例 11-1】研究成人体重指数 (BMI) (kg/m^2) 与肝脏硬度值 (LSM) (kPa) 间的关系,得到了表 11-1 中所示的资料,试进行线性回归分析。

(1) 根据表 11-1 的数据绘制散点图(图 11-1)。从绘制的散点图中可以看出,成人 BMI 与肝脏硬度值 (LSM) 之间存在着明显的直线趋势,因此可进一步考虑建立二者之间的线性回归方程。

(2) 计算回归系数与常数项

本例: $\sum X = 499.29, \sum X^2 = 12\,731.621, \bar{X} = 24.965$

(李康,贺佳主编, 2024, p. 110) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

表 11-1 成人 BMI(kg/m^2)与 LSM(kPa)回归分析数据

调查对象	BMI(X)	LSM(Y)	XY	X^2	Y^2
1	32.06	8.37	268.342	1 027.844	70.057
2	31.20	8.47	264.264	973.440	71.741
3	30.04	7.37	221.395	902.402	54.317
4	28.93	7.90	228.547	836.945	62.410
5	28.41	8.10	230.121	807.128	65.610
6	27.43	7.00	192.010	752.405	49.000
7	26.35	5.80	152.830	694.323	33.640
8	25.88	5.33	137.940	669.774	28.409
9	25.20	7.00	176.400	635.040	49.000
10	24.84	5.80	144.072	617.026	33.640
11	24.11	7.43	179.137	581.292	55.205
12	23.89	4.67	111.566	570.732	21.809
13	23.23	5.50	127.765	539.633	30.250
14	22.43	6.30	141.309	503.105	39.690
15	22.03	6.20	136.586	485.321	38.440
16	21.72	4.31	93.613	471.758	18.576
17	21.11	5.00	105.550	445.632	25.000
18	20.72	3.90	80.808	429.318	15.210
19	20.03	5.13	102.754	401.201	26.317
20	19.68	6.00	118.080	387.302	36.000
合计	499.29	125.58	3 213.090	12 731.621	824.320

(李康,贺佳主编, 2024, p. 111) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

$$\sum Y = 125.58, \sum Y^2 = 824.320, \bar{Y} = 6.279$$

$$\sum XY = 3213.090$$

代入公式 (11-2) 和公式 (11-3) 得

$$b = \frac{l_{XY}}{l_{XX}} = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}}$$

$$= \frac{3213.090 - \frac{499.29 \times 125.58}{20}}{12731.621 - \frac{499.29^2}{20}}$$

$$= \frac{78.048}{267.096} = 0.292$$

$$a = \bar{Y} - b\bar{X} = 6.279 - 0.292 \times 24.965 = -1.011$$

则回归方程为

$$\hat{Y} = -1.011 + 0.292X$$

(3) 作回归直线

按上述回归方程, 在 X 实测值的范围内, 任取两个相距较远的点 $A(X_1, \hat{Y}_1)$ 和 $B(X_2, \hat{Y}_2)$, 连接 A

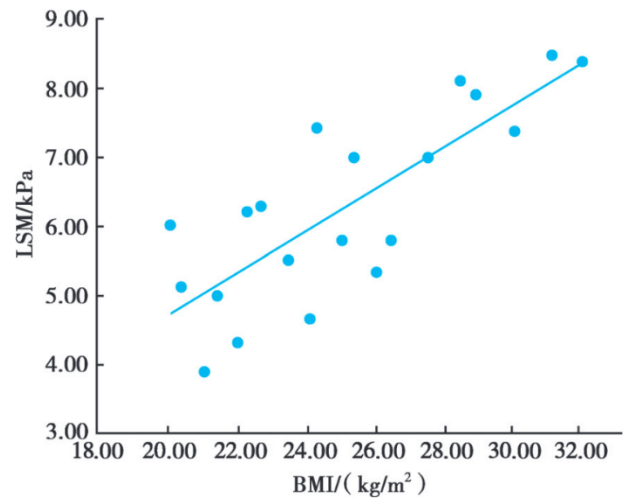


图 11-1 成人 BMI 与肝脏硬度值间关系图

(李康, 贺佳主编, 2024, p. 111) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

和 B 两点即得到回归直线。本例可取 $X_1 = 20$, 计算出 $\hat{Y}_1 = 4.829$; $X_2 = 32$, 计算出 $\hat{Y}_2 = 8.333$; 两点的连线即为所求回归直线 (图 11 - 1)。

由图 11 - 1 可见, 成人 BMI 升高, 肝脏硬度值也趋向升高, 肝脏硬度值 (Y) 随成人 BMI (X) 的变化呈直线趋势, 但并不完全在一条直线上, 说明除了受 X 的影响外, 还有其他因素对 Y 起作用。

三、线性回归的假设检验

一句话抓住本质

线性回归的假设检验, 本质只检验一件事:

👉 回归系数 β 是否显著不为 0。

F 检验和 t 检验, 只是两种“说法”。

一、为什么“看起来 $b \neq 0$ ”还不够？(Why)

即使真实世界里：

$$\beta = 0$$

由于**抽样误差**：

- 样本回归系数 b 几乎不可能正好等于 0

👉 问题从来不是： b 是不是 0

👉 而是：

b 大到，能不能用“随机波动”来解释？

二、回归里的“方差分析思想”（核心世界观）

你可以把回归完全当作一种 **特殊的 ANOVA**。

任意观测值的分解：

$$Y - \bar{Y} = (\hat{Y} - \bar{Y}) + (Y - \hat{Y})$$

直觉翻译：

- $Y - \bar{Y}$ ：总变异
 - $\hat{Y} - \bar{Y}$ ：被 X 的线性关系解释的部分
 - $Y - \hat{Y}$ ：解释不了的随机误差
-

三、三个平方和分别在说什么？（一定要会讲）

1 总平方和

$$SS_{\text{总}} = \sum (Y - \bar{Y})^2$$

👉 **Y 的全部变异**

2 回归平方和

$$SS_{\text{回归}} = \sum(\hat{Y} - \bar{Y})^2$$

👉 **X→Y 的线性关系能解释的变异（信号）**

3 残差平方和

$$SS_{\text{残差}} = \sum(Y - \hat{Y})^2$$

👉 **随机因素 + 模型解释不了的部分（噪声）**

恒等式（必须刻进脑子）

$$SS_{\text{总}} = SS_{\text{回归}} + SS_{\text{残差}}$$

这和单因素 ANOVA 的“组间 + 组内”完全同构。

四、F 检验到底在比什么？（ANOVA 语言）

自由度

- 回归: $v_{\text{回归}} = 1$
 - 残差: $v_{\text{残差}} = n - 2$
-

均方

$$MS_{\text{回归}} = \frac{SS_{\text{回归}}}{1}, MS_{\text{残差}} = \frac{SS_{\text{残差}}}{n - 2}$$

F 统计量

$$F = \frac{MS_{\text{回归}}}{MS_{\text{残差}}}$$

一句话直觉

👉 被回归解释的变异

和

👉 随机误差的变异

相比, 是否 “明显更大” ?

原假设在 F 语言里等价于:

$$H_0: \beta = 0 \Leftrightarrow MS_{\text{回归}} \approx MS_{\text{残差}}$$

五、t 检验在干什么? (参数语言)

t 检验直接盯着 **回归系数本身**:

$$t_b = \frac{|b - 0|}{S_b}, v = n - 2$$

每一项的真实含义

- b : 样本回归系数
- S_b : **b 的标准误** (抽样波动尺度)
- $S_{Y|X} = \sqrt{MS_{\text{残差}}}$:
👉 点到回归线的平均离散程度

t 检验的直觉一句话

- 👉 b 离 0 有多远?
- 👉 这点距离, 是否超过了它应有的随机波动?

六、最关键的一点: F 和 t 根本不是两套结论

在一元线性回归中, 存在严格关系:

$$F = t^2$$

所以:

- F 显著 \Leftrightarrow t 显著
- P 值完全一致
- 结论必然相同

✦ 你在例 11-1 中看到的:

$$\sqrt{31.589} \approx 5.62 \approx t_b$$

不是巧合, 而是**数学必然**。

七、那到底什么时候用 F? 什么时候用 t? (现实决策)

一元线性回归:

- **二者等价**
- 软件通常两者都会给

多元线性回归 (提前给你一个前瞻):

- **F 检验:**
 - 👉 检验 “模型整体是否有意义”
- **t 检验:**
 - 👉 检验 “某一个自变量是否有独立贡献”

八、把这一节压缩成 5 句“考场级结论”

- 1 回归检验本质是检验 $\beta = 0$
 - 2 回归变异 + 残差变异 = 总变异
 - 3 F 比的是“信号 / 噪声”
 - 4 t 比的是“系数 / 标准误”
 - 5 一元回归中: $F = t^2$
-

九、考试级一句话总结 (可直接写)

线性回归的假设检验通过方差分析将因变量总变异分解为回归变异和残差变异, 构造 F 统计量检验回归方程的整体显著性; 同时也可通过对回归系数进行 t 检验判断其是否显著不为零。在一元线性回归中, 两种检验方法在统计上是等价的。

给你一个不会再混的心智锚点

F 问的是: 这条线值不值得信?

t 问的是: 这个斜率是不是噪声?

在一元回归里, 它们问的是同一件事。

下面对例 11-1 数据建立的回归方程进行假设检验:

(1) 建立检验假设, 确定检验水准

$H_0: \beta = 0$, 即 BMI 和 LSM 间无线性回归关系

$H_1: \beta \neq 0$, 即 BMI 和 LSM 间有线性回归关系

(李康, 贺佳主编, 2024, p. 112) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

$\alpha = 0.05$

(2) 计算统计量

$$SS_{\text{总}} = \sum Y^2 - \frac{(\sum Y)^2}{n} = 824.32 - \frac{125.58^2}{20} = 35.803$$

$$SS_{\text{回归}} = \frac{l_{XY}^2}{l_{XX}} = \frac{78.05^2}{267.10} = 22.807$$

$$SS_{\text{残差}} = SS_{\text{总}} - SS_{\text{回归}} = 12.996$$

$$F = \frac{MS_{\text{回归}}}{MS_{\text{残差}}} = \frac{SS_{\text{回归}}/\nu_{\text{回归}}}{SS_{\text{残差}}/\nu_{\text{残差}}} = \frac{22.807/1}{12.996/18} = 31.589$$

(3) 确定 P 值,得出统计结论

查 F 界值表, $\nu_{\text{回归}} = 1, \nu_{\text{残差}} = 18, F_{0.01(1,18)} = 8.28, F > 8.28$, 故 $P < 0.01$, 按 $\alpha = 0.05$ 检验水准, 拒绝 H_0 , 可以认为成人 BMI 与 LSM 之间存在线性回归关系。上面结果可以归纳成表 11-2 的形式。

表 11-2 方差分析表

变异来源	SS	ν	MS	F	P
总变异	35.803	19			
回归	22.807	1	22.807	31.589	< 0.001
残差	12.996	18	0.722		

(李康,贺佳主编, 2024, p. 113) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

例 11-1 数据建立回归方程后,进行 t 检验,过程如下:

(1) 建立检验假设,确定检验水准

$H_0: \beta = 0$, 即 BMI 和 LSM 间无线性回归关系

$H_1: \beta \neq 0$, 即 BMI 和 LSM 间有线性回归关系

$\alpha = 0.05$

(2) 计算统计量

$$S_{Y|X} = \sqrt{\frac{12.996}{18}} = 0.850, \quad S_b = \frac{0.850}{\sqrt{267.10}} = 0.052$$

$$t_b = \frac{|0.292 - 0|}{0.052} = 5.615, \quad \nu = 20 - 2 = 18$$

(3) 确定 P 值,作出结论

根据 $\nu = 18$, 查 t 界值表(附表 2), $t_{0.01/2, 18} = 2.878, t_b > 2.878$, 故 $P < 0.01$, 按 $\alpha = 0.05$ 检验水准, 拒绝 H_0 , 结论与方差分析相同。实际上, 统计量 F 与 t 之间存在确定的数量关系, 即 $\sqrt{F} = t$, 本例 $\sqrt{31.589} = 5.620$, 与 5.615 略有差异, 是由于计算过程中的四舍五入。

(李康,贺佳主编, 2024, p. 113) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

第二节 | 线性相关 (Linear correlation)

一句话抓住本质

线性相关不问“谁解释谁”，

只问：

👉 两个变量是否一起变、变得有多一致。

一、线性相关在研究什么？ (Purpose)

当研究目标是：

- 不关心用 X 去预测 Y
- 不设定因果方向
- 只想知道：

两个变量之间有没有直线关系？关系强不强？方向如何？

👉 用线性相关，而不是回归。

🔴 这是“对称”的分析：

- X 和 Y 地位完全相同

二、线性相关的适用前提 (Scope)

教材给了一个很重要但容易被忽略的前提：

线性相关用于分析双变量正态分布资料
(bivariate normal distribution) 。

直觉理解为：

- 每个变量本身近似正态
- 两者的联合分布呈“椭圆云团”

三、散点图是第一步，不是装饰 (Very important)

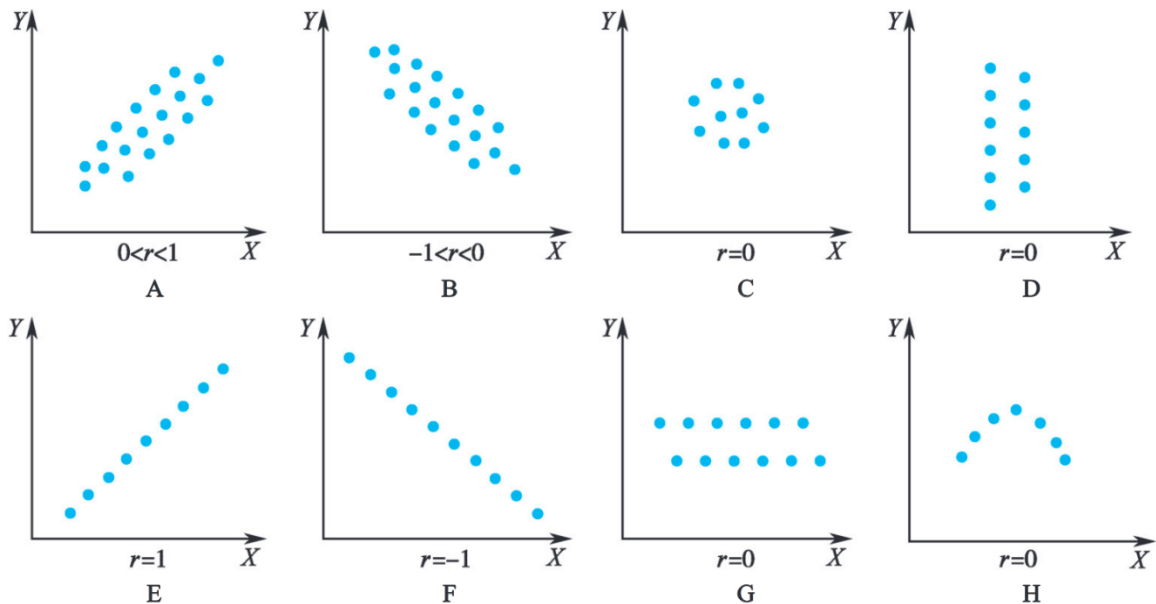


图 11-4 直线相关示意图

(李康,贺佳主编, 2024, p. 114) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

在任何相关分析之前，**第一反应永远是画散点图。**

散点图能直接告诉你四件事：

1 正相关

- $X \uparrow \rightarrow Y \uparrow$
- 越集中，相关越强
- 完全在直线上 \rightarrow 完全正相关

2 负相关

- $X \uparrow \rightarrow Y \downarrow$
- 完全在直线上 \rightarrow 完全负相关

3 无线性相关

- 点云杂乱
- 或直线平行于坐标轴
- $r \approx 0$

4 非线性相关

- 有关系，但不是直线
- 🖐️ r 会误判

📌 一句铁律：

r 只刻画“直线相关”，
看不见曲线关系。

四、相关系数 r 在数学上做了什么？（Core idea）

$$r = \frac{l_{XY}}{\sqrt{l_{XX}l_{YY}}} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}}$$

直觉拆解（非常重要）

- 分子 l_{XY} ：
🖐️ X 偏大时， Y 是否也偏大？
- 分母：
🖐️ 用 X 、 Y 各自的变异做**标准化**

🖐️ 所以：

$r =$
共同变化程度

÷

各自变化的几何平均

五、为什么 r 没有单位？（理解点）

- 分子、分母单位相互抵消
- r 是**纯比例量**

✦ 这带来两个直接后果：

- 可以跨变量比较强弱
 - 但 $r \neq$ **变化幅度大小**
-

六、 r 的取值与含义（一定要说清）

$$-1 \leq r \leq 1$$

- $r > 0$ ：正相关
- $r < 0$ ：负相关
- $r = 0$ ：**无线性相关**（ \neq 完全无关系）
- $|r| = 1$ ：完全线性相关（医学中几乎不存在）

✦ **强弱判断只看 $|r|$ ，方向看正负。**

七、相关 \neq 回归（这是分水岭）

相关分析：

- 对称
- 描述关系强弱
- 不涉及预测

回归分析:

- 有方向 ($X \rightarrow Y$)
- 定量解释
- 可用于预测

👉 **r 大, 不代表能预测好;**

回归显著, 也不等于 r 很大。

【例 11-2】从男青年总体中随机抽取 11 名男青年组成样本, 分别测量每个男青年的身高和前臂长, 测量结果如表 11-3 所示, 试计算身高与前臂长之间的相关系数。

表 11-3 11 名男青年身高与前臂长的测量结果

单位: cm

编号	身高 (X)	前臂长 (Y)	XY	X^2	Y^2
1	170	47	7 990	28 900	2 209
2	173	42	7 266	29 929	1 764
3	160	44	7 040	25 600	1 936
4	155	41	6 355	24 025	1 681
5	173	47	8 131	29 929	2 209
6	188	50	9 400	35 344	2 500
7	178	47	8 366	31 684	2 209
8	183	46	8 418	33 489	2 116
9	180	49	8 820	32 400	2 401
10	165	43	7 095	27 225	1 849
11	166	44	7 304	27 556	1 936
合计	1 891	500	8 6185	326 081	22 810

经计算:

$$\sum X = 1\,891, \quad \sum X^2 = 326\,081$$

$$\sum Y = 500, \quad \sum Y^2 = 22\,810$$

$$\sum XY = 86\,185, \quad n = 11$$

(李康, 贺佳主编, 2024, p. 115) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

代入计算公式得

$$l_{XX} = \sum X^2 - \frac{(\sum X)^2}{n} = 326\,081 - \frac{1\,891^2}{11} = 1\,000.909$$
$$l_{YY} = \sum Y^2 - \frac{(\sum Y)^2}{n} = 22\,810 - \frac{500^2}{11} = 82.727$$
$$l_{XY} = \sum XY - \frac{(\sum X)(\sum Y)}{n} = 86\,185 - \frac{1\,891 \times 500}{11} = 230.455$$

按公式 (11-14) 计算相关系数为

$$r = \frac{230.455}{\sqrt{1\,000.909 \times 82.727}} = 0.801$$

由计算结果可见, 本例 r 为正值, 表示前臂长与身高之间呈正相关关系, 且有较高的相关性。

(李康, 贺佳主编, 2024, p. 116) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

八、相关系数的假设检验在检验什么? (Hypothesis)

$$H_0: \rho = 0$$

直觉翻译:

总体中, X 与 Y 之间
没有线性相关关系

九、t 检验在干嘛? (别死背公式)

$$t = \frac{|r|}{\sqrt{\frac{1-r^2}{n-2}}}, v = n - 2$$

直觉一句话:

r 离 0 有多远?

这点距离, 是否超过了随机波动?

- 分子: 观测到的相关强度
- 分母: r 的抽样误差

👉 若 t 足够大 → 拒绝 H_0

下面对例 11-2 计算得到的 r 值进行假设检验:

(1) 建立检验假设, 确定检验水准

$H_0: \rho = 0$, 即身高与前臂长之间不存在线性相关关系

$H_1: \rho \neq 0$, 即身高与前臂长之间存在线性相关关系

$\alpha = 0.05$

(2) 计算统计量

$$t = \frac{|0.801 - 0|}{\sqrt{\frac{1 - 0.801^2}{11 - 2}}} = 4.013, \quad \nu = 11 - 2 = 9$$

(3) 确定 P 值, 作出结论

查 t 界值表(附表 2), 得 $t_{0.005/2, 9} = 3.690, t > t_{0.005/2, 9}, P < 0.005$, 故按 $\alpha = 0.05$ 检验水准, 拒绝 H_0 , 接受 H_1 , 可以认为男青年身高与前臂长之间存在正相关关系。或直接查 r 界值表(附表 11), $t_{0.005/2, 9} = 0.766, r > r_{0.005/2, 9}, P < 0.005$, 结论相同。

(李康, 贺佳主编, 2024, p. 116) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

十、r 界值表 vs t 检验 (怎么选)

- 原理 完全等价
- r 界值表只是把:

$$t_{\alpha/2, n-2}$$

换算成:

$$r_{\alpha/2, n-2}$$

✦ 考试写哪种都对，
但本质是同一件事。

十一、考试级一句话总结（直接可写）

线性相关用于分析两个数值变量之间是否存在线性关系及其方向和密切程度，常用 Pearson 积差相关系数 r 进行定量描述，并通过 t 检验或 r 界值表对总体相关系数是否为零进行假设检验。

给你一个“永不混淆”的终极锚点

相关：有没有关系、关系强不强（对称）

回归：关系怎么写、能不能预测（有方向）

只要你在做题时先问一句

“我到底关不关心预测？”

相关和回归就再也不会混。

第三节 线性回归与相关应用的注意事项

一句话抓住本质

回归和相关本身不难，

难的是：

👉 什么时候该用、怎么用不越界、结论敢不敢说。

一、线性回归的应用边界（什么时候“该用”）

1 线性回归能干的三件事（按重要性）

① 判断是否存在线性数量关系（最常用）

X 变, Y 的平均水平是否系统性改变?

② 预测 (Prediction)

用 X 估计 Y (点估计 / 区间估计)

③ 统计控制 (Control)

反过来:

通过控制 X, 把 Y 限制在目标区间

✦ 现实中, 第一点占 90% 的使用场景。

2 X 和 Y 怎么放, 不能随意 (非常重要)

- 有明确因果关系
 - 👉 原因 = X
 - 👉 结果 = Y
- 因果不清 / 难以判断
 - 👉 把更稳定、测量误差更小的变量当 X

✦ 这是统计建模的责任问题, 不是数学问题。

3 X 是随机的, 还是给定的? (模型意识)

- I 型回归模型
 - X 给定
 - Y 随机
 - 最常见 (实验、设计型研究)
- II 型回归模型
 - X、Y 都是随机变量
 - 服从双变量正态
 - 更接近相关分析的世界

✦ 但一元线性回归的估计形式不变。

4 非线性怎么办？（别硬上）

- Y 不正态
- X-Y 呈曲线关系

👉 两条路：

- 变量变换（对数、平方根等）
- 直接拟合曲线模型

✦ 强行线性回归 = 自欺欺人。

5 外推是“高风险操作”（考试爱考）

不要把回归方程用到样本 X 取值范围之外。

- 样本支持的，只是局部线性
- 区间外是否仍线性，无从判断

✦ 预测 ≠ 占卜

外推要非常谨慎。

二、线性相关的应用边界（什么时候“只能用它”）

1 r 的前提比你想象得苛刻

- 双变量正态分布
- 线性关系

若不满足：

- 先尝试变量变换
- 仍不满足 / 有序资料
 - 👉 Spearman / Kendall

✦ Pearson r 不是万能相关指标。

2 r 多大 “算有意义”？（经验尺度）

这是经验判断，不是数学定理：

- $r \leq 0.3$ ：弱相关
- $0.3 < r \leq 0.6$ ：中度相关
- $0.6 < r \leq 0.8$ ：较强相关
- $r > 0.8$ ：高度相关

✦ 是否 “重要”，必须结合具体问题。

3 最重要的一条：相关 \neq 因果（一定要会说）

r 显著，只说明 “一起变”，
不说明 “谁导致谁”。

- 因果判断：
 - 👉 靠专业知识、机制、实验设计
 - 👉 不是靠 P 值

这是统计伦理问题。

三、回归与相关的 “深层联系”（一次打通）

1 方向一致性

- 对同一资料：

- r 与 b 符号必然相同
- 正正、负负

👉 方向一致，但含义不同。

2 r 与 b 可以互相转化 (非常重要)

$$b = r \frac{S_Y}{S_X}$$

直觉解释：

**回归系数 =
相关强度 × 尺度变换**

📌 所以：

- r 是 “无量纲关系”
- b 是 “有单位变化率”

3 决定系数 R^2 ：两者的桥梁

$$R^2 = r^2 = \frac{SS_{\text{回归}}}{SS_{\text{总}}}$$

真正含义（一定要会说）：

**Y 的变异中，有多少比例
可以被 X 解释？**

例如：

- $r = 0.5$
- $R^2 = 0.25$

👉 只有 25% 的变异被解释

👉 剩下 75% 仍是未知因素

📌 这是对“相关强但解释力有限”的最好提醒。

四、相关 vs 回归：终极对照（别再混）

维度	相关	回归
方向性	无（对称）	有 ($X \rightarrow Y$)
核心问题	有没有关系、强不强	怎么变、能解释多少
因果适配	不适合	更适合
预测能力	✗	✓
关键指标	r	b, R^2

五、考试级一句话总结（可直接写）

线性回归主要用于分析变量间的数量依存关系及预测因变量，而线性相关用于描述两个变量之间线性关系的方向和密切程度。两者在统计上密切相关，但研究目的、模型假设和解释侧重点不同，应用时应根据研究问题合理选择，并避免将相关性误解释为因果关系。

给你一个终极心智锚点

相关是“描述世界”，
回归是“试图解释世界”。

统计只能给证据，
因果必须你来负责。

[BIBLIOGRAPHY] Please click Zotero - Refresh in Word/LibreOffice to update all fields