直线相关与回归

第二军医大学卫生统计学教研室 贺 佳

复习：以前所学的为一个变量，如三种药物治疗三组病人，观察其血压值。而今天我们学习两个变量之间的相关、回归。

医学中有很多说明两者关系的实例：年龄与血压、血脂的关系；身高、体重与胸围、AFP 与肝癌、T_3、T_4与甲亢的关系等。

第一节 直线相关(Linear Correlation)

一、直线相关的概念

相关：表示事物数量相互关系的密切程度，如身高与体重的关系。

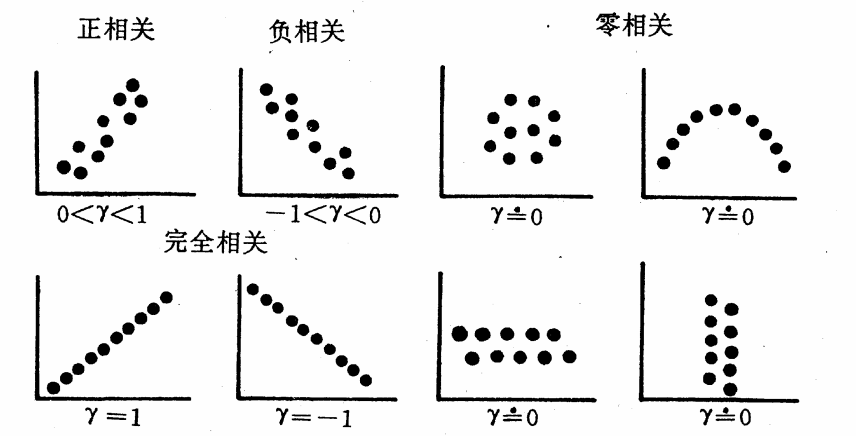

图 11.1 相关系数示意

英文专业词汇：

positive correlation

negative correlation

zero correlation

perfect positive correlation

perfect negative correlation

二、相关系数的意义

直线相关系数简称相关系数(correlation coefficient)。**由英国统计学者 F.Y. Edgeworth 于 1892 年首创，以符号 r 表示。它说明了两变量间相关的密切程度与相关方向。计算公式为：**

$$r=\frac{\sum(X_1-\bar{X}_1)(X_2-\bar{X}_2)}{\sqrt{\sum(X_1-\bar{X}_1)^2\sum(X_2-\bar{X}_2)^2}}=\frac{l_{12}}{\sqrt{l_{11}l_{22}}}$$

$$l_{11}=\sum(X_1-\bar{X}_1)^2=\sum X_1^2-\frac{(\sum X_2)^2}{n}$$

$$l_{22}=\sum(X_2-\bar{X}_2)^2=\sum X_2^2-\frac{(\sum X_2)^2}{n}$$

$$l_{12}=\sum(X_1-\bar{X}_1)(X_2-\bar{X}_2)=\sum X_1X_2-\frac{(\sum X_1)(\sum X_2)}{n}$$

$\sum(X_1-\bar{X}_1)^2$ **为** X_1 **的离均差平方和，** $\sum(X_2-\bar{X}_2)^2$ **为** X_2 **的离均差平方和，** $\sum(X_1-\bar{X}_1)(X_2-\bar{X}_2)$**为** X_1 **与** X_2 **的离均差积和。**

意义：

1.相关系数 r 表示两个变量相互间的直线关系，并可据此判断其密切程度。

2. r 在-1～+1 之间变动，没有单位。

3.其绝对值越大，越接近 1，两变量间的直线关系越密切，越接近 0，相关越不密切。

4.r>0**，说明一变量随另一变量增减而增减，如年龄越大，血压越高，方向相同；**r<0**，表示一变量增加，另一变量减少，方向相反。如年龄越大，体力越差。r 的符号由** l_{12} **决定。**

**三、相关系数的计算

$$r=\frac{\sum(X_1-\bar{X}_1)(X_2-\bar{X}_2)}{\sqrt{\sum(X_1-\bar{X}_1)^2\sum(X_2-\bar{X}_2)^2}}=\frac{l_{12}}{\sqrt{l_{11}l_{22}}}$$

例 11.1　测定 8 名健康成人血清胆固醇与低密度脂蛋白含量，试计算相关系数。P199。

代入公式计算：

l_{11}＝281286－$1478^2/8$=8225.5

l_{22}=92304-$822^2/8$=7843.5

l_{12}=159687-1478×822/8=7822.5

$$r = 7822.5/\sqrt{8225.5\times 7843.5} = 0.9739$$

r 为正值，表示两者呈正相关的关系，其绝对值为 0.97，表明两者相关较密切。

介绍应用计算器计算。

四、直线相关系数的假设检验

1. t 检验法。

复习样本、总体、抽样误差。

看 r 是否来自 ρ =0 的总体，因有抽样误差，r≠0。相关系数呈对称分布，故用 t 检验来进行。

⑴假设：H_0:ρ =0, H_1:ρ ≠0，α =0.01

⑵计算:

$$t = \frac{r-r}{s_r} = \frac{r-0}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.97}{\sqrt{\frac{1-0.97^2}{8-2}}} = 10.47$$

s_r 为相关系数标准误。

⑶查 t 值表作结论：

P432，ν =n-2=8-2=6

单、双侧皆可，以单侧较好，$t_{0.001,6}$=5.959，P<0.001,。

在α =0.001 处，拒绝 H_0，接受 H_1，认为两者有正相关关系。

2.查表法：P456。结论相同。

根据 t 检验的公式，可算出 r 值，统计学家已将各种自由度各种 α 的 r 值求出，列成相关系数表。

根据 ν 值查表，分析界值表：不能只看 r 值，还需看 ν 值，ν 值越大，越容易显著。

第二节　直线回归(Linear Regression)

一、直线回归的概念

回归：即了解两变量中一个变量依另一变量而变化的规律。如肝炎患者预后与并发症的关系。

回归一词源于生物学，由 Galton 于 1887 年提出，研究 1078 对父子身高，父高子更高的可能性小，向中轴力，即回归；生活中，父母很强，子女一般；回归大自然。

**二、直线回归方程的计算

若某一变量（y）随另一个变量（x）的变动而变动，则称 x 为自变量（independent variable），y 为应变量（dependent variable）。直线回归的任务是拟合直线方程，$\hat{y}=a+bx$（$\hat{y}$ hat, cap, peak），a 为截距或常数项（intercept, constant），是 X=0 时的 $\hat{y}$ 值即回归直线与纵轴的交点；b 为回归系数（regression coefficient），是直线的斜率（slope），即 X 每增加一个单位时，$\hat{y}$ 相应增（或减）b 个单位；$\hat{y}$ 为由 X 代入直线方程计算所得，为 y 的估计值。拟合的原理为最小二乘法（least square method），即各点到直线的纵向距离的平方和为最小，即 $\sum(y-\hat{y})^2$ 为最小。Y 可称因变量或应变量。

$$b=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}=\frac{\sum xy-\frac{\sum x\sum y}{n}}{\sum x^2-\frac{(\sum x)^2}{n}}=\frac{l_{xy}}{l_{xx}}$$

$$a=\bar{y}-b\bar{x}$$

例 11.2 表 11.2 资料，P201，年龄(x)与平均收缩压(y)的回归关系。

1、计算 Σx、Σy、Σx^2、Σy^2 及 Σxy。

2、计算 x、y 的均数 $\bar{x}$、$\bar{y}$，离均差平方和 l_{xx}、l_{yy} 及离均差积和 l_{xy}。

$$\bar{x}=\sum x/n=240/4=60$$

$$\bar{y}=\sum y/n=70.2/4=17.55$$

$$l_{xx}=\sum(x-\bar{x})^2=\sum x^2-(\sum x)^2/n=14900-240^2/4=500$$

$$l_{yy}=\sum(y-\bar{y})^2=\sum y^2-(\sum y)^2/n=1233.18-70.2^2/4=1.17$$

$$l_{xy}=\sum(x-\bar{x})(y-\bar{y})=\sum xy-(\sum x)(\sum y)/n$$

$$=4236.0-240\times70.2/4=24$$

3、计算回归系数 b **与截距** a，**列出直线回归方程。**

b=24/500=0.048

a=17.55-0.048×60=14.67

$$\hat{y}=14.67+0.048x$$

三、直线回归方程的图示

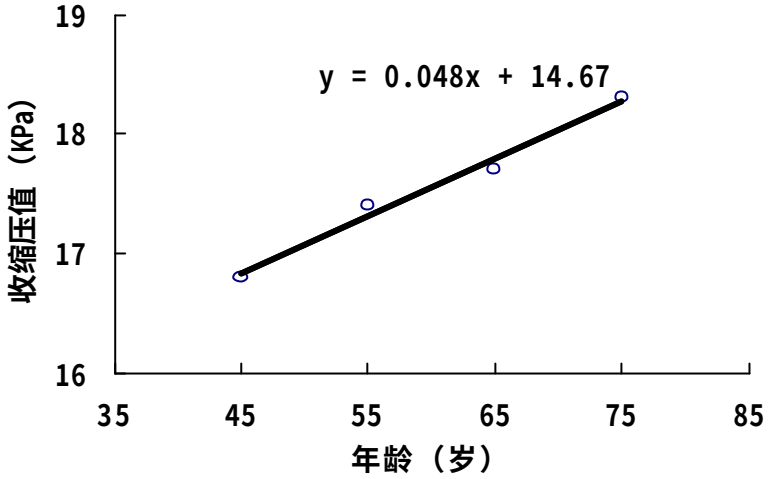

图11.2 收缩压值与年龄的散点图及回归直线

在自变量 x 的实测全距范围内取相距较远且易读的两个 x 值，代入所求的回归方程中。如本例取 x_1=45，$\hat{y}_1$=16.83；x_2=75，$\hat{y}_2$=18.27。在图上确定（45，16.83）和（75，18.27）两个点，连接这两点，划直线即得直线方程 $\hat{y}=14.67+0.048x$。该直线通过（$\bar{x}$，$\bar{y}$）点，$\hat{y}$ 在线上，y 散在线的两边及线上。

四、直线回归系数的假设检验

由于抽样误差，即使样本回归系数 b 来自总体回归系数 β 为 0 的总体，其值不一定为 0，因此需作 β 是否为 0 的假设检验。

1、 直线回归的方差分析（F 检验）

y 平方和划分示意图

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

则各点相加，划简整理有：

则：$SS_{总}=SS_{回}+SS_{剩}$

$SS_{总}$：y 的总变异；

$SS_{回}$：回归平方和，总平方和中 x 解释的部分；

$SS_{剩}$：剩余平方和，其越小估计误差越小。

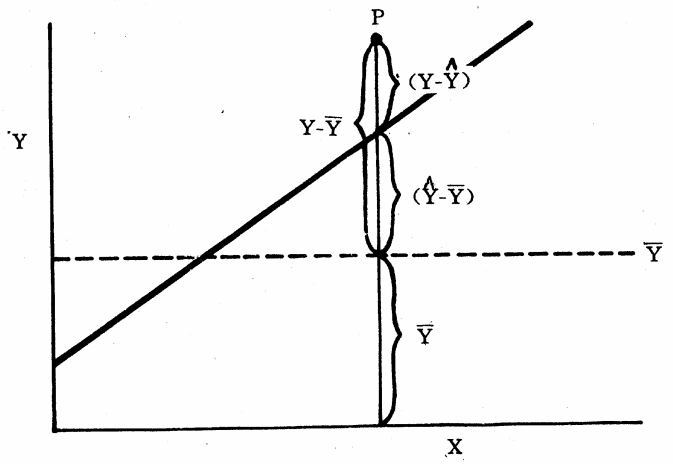

图 11.3　应变量 Y 的平方和划分示意

具体计算：

$SS_{总}=l_{yy}$　　$\nu_{总}=n-1$

$SS_{回}=bl_{xy}=l_{xy}^2/l_{xx}$　　$\nu_{回}=1$

$SS_{剩}=SS_{总}-SS_{回}$　　$\nu_{剩}=n-2$

分析表 11.2 资料，计算结果列于方差分析表中：

变异来源	SS	ν	MS	F
总变异	1.170	3		
回归	1.152	1	1.152	128.00
剩余	0.018	2	0.009	

查 F 值表，$F_{0.01,(1,2)}=98.49$**，本例** $F>F_{0.01,(1,2)}$**，故** $P<0.01$**。在**$\alpha=0.01$**水平处拒绝** H_0**，接受** H_1**，认为平均收缩压高低与年龄大小之间有线性关系。**

2、t 检验法

$$t=\frac{b-0}{s_b}=\frac{b}{s_{y.x}/\sqrt{l_{xx}}},\quad \nu=n-2,$$

$$s_{y.x}=\sqrt{\frac{\sum(y-\hat{y})^2}{n-2}}$$

$s_{y.x}$ **称剩余标准差；**s_b **为样本回归系数** b **的标准误。**

本例：$\sum(y-\hat{y})^2=SS_{剩}=0.018$

$$s_{y.x}=\sqrt{\frac{0.018}{4-2}}=0.0949$$

$$t=\frac{0.048}{0.0949/\sqrt{500}}=11.3099$$

ν=2，查 t 值表，结论与 F 检验完全相同。

两者关系：$F=t^2$。

3、查 r 界值表法

根据数理统计学的理论：$t_r=t_b$，故可通过查 r 值表得出检验结果。

查 r 界值表，r=0.9923，结论与 t、F 检验完全一致。

五、直线回归的应用

1、描述两变量间的依存关系。

2、根据一个较易测得的变量值，推算另一个不易测得的变量值。如由体重推算体表面积。

3、利用回归方程进行预测（由已知到未知）。

4、估计正常值范围。

**第三节　　直线相关与回归的区别与联系

一、区别

1、相关分析要求两个变量均服从正态分布，而回归分析则有两种不同的模型:Ⅰ型回归:定 x 后对 y 进行测量，y 须服从正态分布；Ⅱ型回归:x、y 均须服从正态分布，如体重依身高的变动关系。

2、对于同一资料，只能计算一个相关系数，而Ⅱ型回归可以计算由 x 推 y 和由 y 推 x 的两个回归方程，但两者不是反函数的关系。

3、回归反映两变量间的依存关系，相关反映两变量间的相互关系。有相关联系不一定是因果联系。

二、联系

1、同一资料 r 与 b 符号相同。

2、同一资料 r 与 b 的假设检验结果是等价的。

3、r 与 b 可以互相换算。

Ⅰ型回归：$r = b\sqrt{\frac{l_{xx}}{l_{yy}}}$，$b = r\sqrt{\frac{l_{yy}}{l_{xx}}}$

Ⅱ型回归：$b_{y\cdot x} = r\sqrt{\frac{l_{yy}}{l_{xx}}} = r\frac{s_y}{s_x}$，$b_{x\cdot y} = r\sqrt{\frac{l_{xx}}{l_{yy}}} = r\frac{s_x}{s_y}$

$r = \sqrt{b_{y\cdot x}b_{x\cdot y}}$

4、相关系数的平方：r^2 称为<u>决定系数</u>(coefficient of determination)。

$$r^2 = \frac{l_{xy}^2}{l_{xx}l_{yy}} = \frac{l_{xy}^2/l_{xx}}{l_{yy}} = \frac{SS_{回}}{SS_{总}}$$

指回归平方和在总平方和中所占的比重，即 x 对 y 的影响，其值越接近 1，说明回归效果越好。

5、归纳：

相关：<u>相互,双方向</u>,-1≤r≤+1,<u>无单位,有相关不一定有回归</u>。

回归：<u>依存,单方向，　无限，　有单位,有回归一定有相关</u>。

**第四节　　应用直线相关与回归的注意事项

1.作相关回归要有<u>实际意义</u>，不要把无关的两个事物用来作相关与回归。如美国某年离婚率高，香烟销售量大，两者相关非常显著，只是数字游戏。生个小孩，种棵小树，两者一起长，都与时间有关。

2.首先绘制散点图。

3.在回归分析中，由 X 推算 Y 与由 Y 推算 X 的回归方程不同，不可混淆。

	回归系数	截距
<u>由 X 推算 Y</u>	$b=l_{xy}/l_{xx}$	$a=\bar{y}-b\bar{x}$
<u>由 Y 推算 X</u>	$b=l_{xy}/l_{yy}$	$a=\bar{x}-b\bar{y}$

<u>一般选原因或个体变异小的变量作自变量</u>。

4.相关与回归仅适用与原数值的范围，<u>不可任意外推</u>。

5.如果有两个不同质的子群，可能产生实际上不存在的相关与回归，也可能忽视了确实存在的相关与回归。如下图。

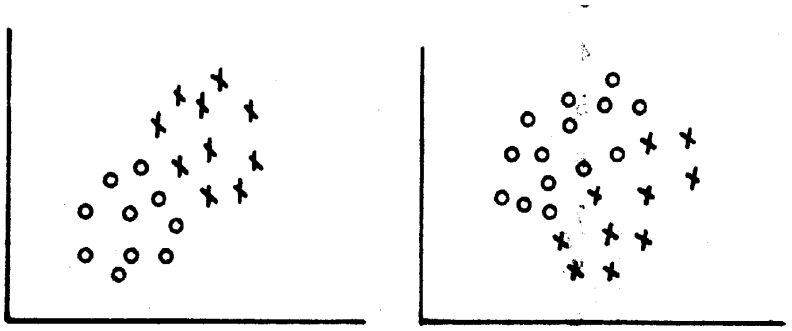

图 12-4　存在两个不同质子群对相关的影响

（a）误为有相关　　（b）相关被忽视

小结：1.**相关系数的计算**。

2.**回归方程的拟合**。

3.**相关、回归的区别与联系**。

Summary:

Today we study the new statistics methods of linear correlation and regression. We should be able to calculate of correlation and regression coefficient. Should be able to understand the distinguish and relation between the correlation and regression.