人民卫生出版社
PEOPLE´S MEDICAL PUBLISHING HOUSE

目录

人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE

人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE

第一节

直线回归

人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE

一、直线回归的概念

在定量描述健康成年人血清总胆固醇含量与年龄数量上的依存关系时，将年龄称为自变量（independent variable），用X 表示；血清总胆固醇含量称为应变量（dependent variable），也称因变量或反应变量，用Y表示。

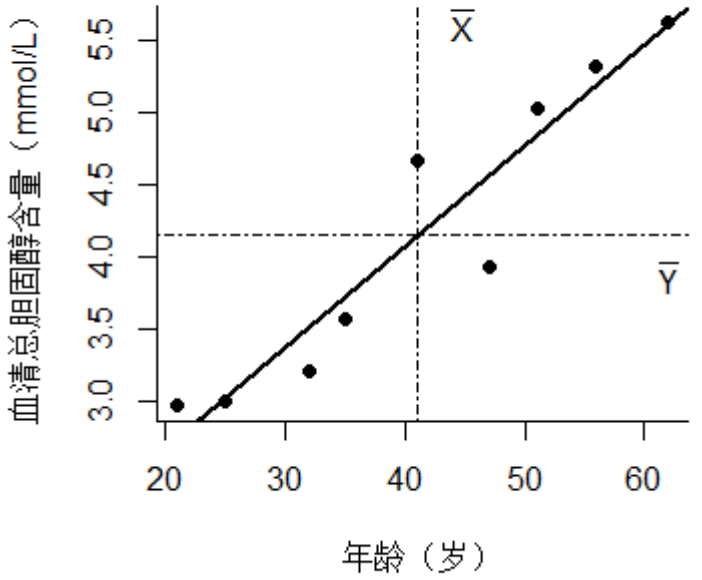

图9-1 10名健康成年人的年龄与血清总胆固醇含量散点图

图注或图片来源（限人卫社图片）

人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE

一、直线回归的概念

直线回归方程（linear regression equation）

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

总体间线性关系

$$\mu_{Y|X} = \alpha + \beta X$$

人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE

一、直线回归的概念

$\hat{Y}$ 回归方程的预测值（predicted value）

α 常数项（constant term），截距，*X*取0时 *Y*的总体均数

β 回归系数（coefficient of regression），斜率，*X*变化一个单位时 *Y*的平均变化量

β=0时，*Y*与*X*无直线关系

人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE

二、直线回归方程的求法

1. 残差（residual）或剩余值

Y的实测值与回归线上估计值的纵向距离 $Y-\hat{Y}$

2. 最小二乘法（least squares method）

使各点残差平方和最小的回归系数所对应的直线

人民卫生出版社 PEOPLE'S MEDICAL PUBLISHING HOUSE

二、直线回归方程的求法

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

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

$$l_{XY}=\sum(X-\bar{X})(Y-\bar{Y})=\sum XY-\frac{(\sum X)(\sum Y)}{n}$$

$$\hat{Y}=1.2917+0.0695X$$

9名健康成年人的年龄X（岁）与血清总胆固醇含量Y（mmol/L）

编号	年龄X	血清总胆固醇含量Y
1	56	5.32
2	32	3.21
3	41	4.67
4	51	5.03
5	25	3.01
6	35	3.57
7	21	2.98
8	47	3.93
9	62	5.62

人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE

三、直线回归的统计推断

（一）回归方程的假设检验

（二）总体回归系数的置信区间

（三）利用回归方程进行估计和预测

人民卫生出版社 PEOPLE'S MEDICAL PUBLISHING HOUSE

（一）回归方程的假设检验

1.方差分析

（1）平方和分解 $SS_{总} = SS_{回} + SS_{残}$

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

（2）自由度分解 $\nu_{总} = \nu_{回} + \nu_{残}$

$$\nu_{总} = n - 1 \qquad \nu_{回} = 1 \qquad \nu_{残} = n - 2$$

（3）F检验

$$F = \frac{SS_{回}/\nu_{回}}{SS_{残}/\nu_{残}} = \frac{MS_{回}}{MS_{残}}$$

$$SS_{回} = bl_{XY} = l_{XY}^2/l_{XX} = b^2 l_{XX}$$

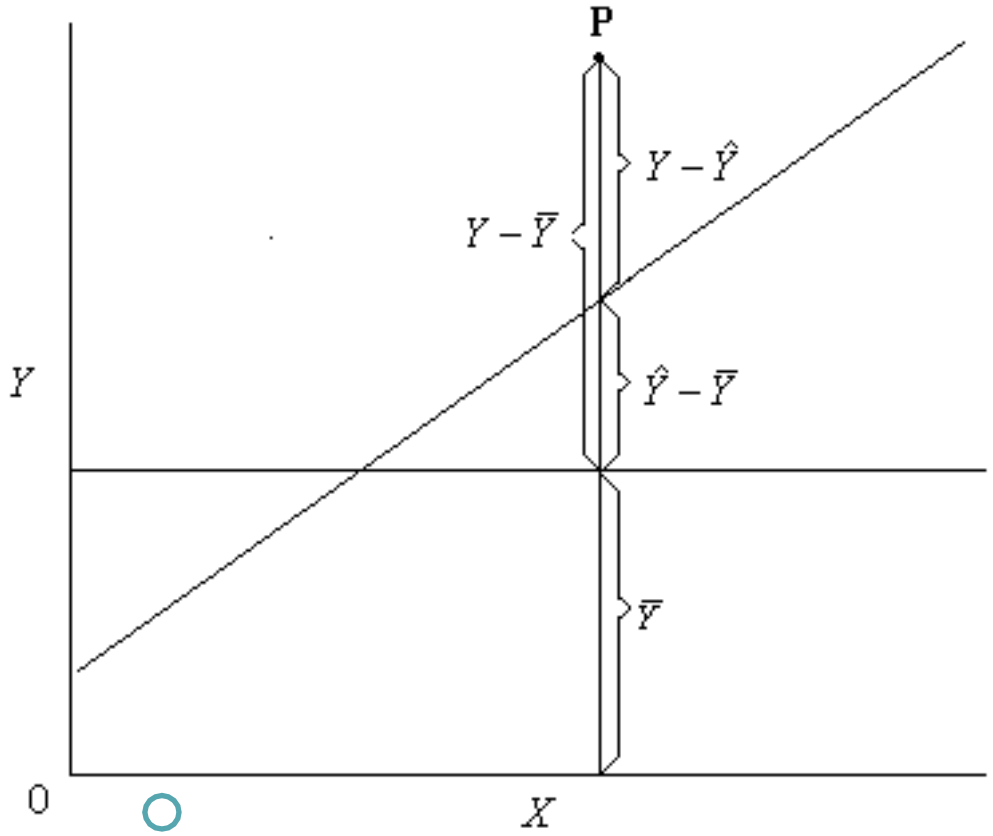

图9-2 平方和划分示意图

人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE

2. t检验

$$t = \frac{b - 0}{S_b} \qquad \nu = n - 2$$

$$S_b = \frac{S_{Y \cdot X}}{\sqrt{l_{XX}}}$$ 回归系数标准误

$$S_{Y \cdot X} = \sqrt{\frac{SS_{残}}{n-2}}$$ 剩余数标准差

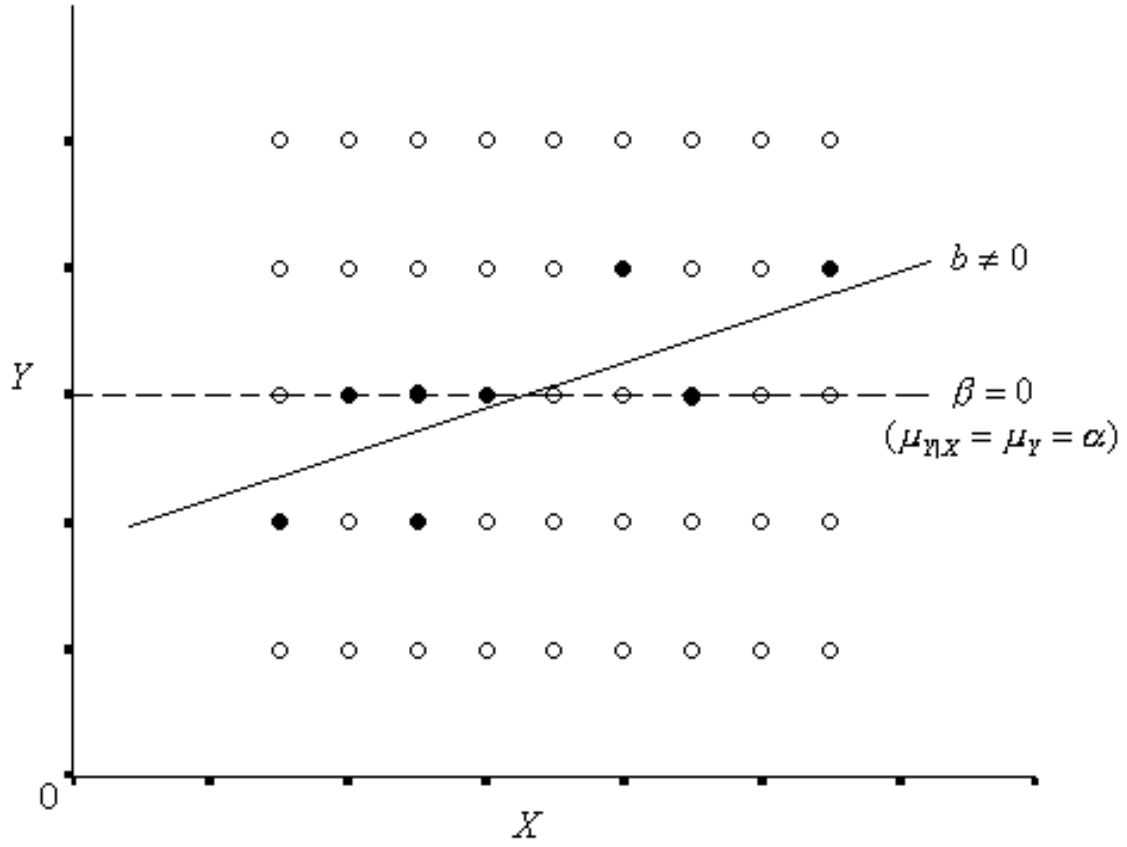

图9-3 总体回归系数与样本回归系数的示意图

人民卫生出版社 PEOPLE'S MEDICAL PUBLISHING HOUSE

3. 实例

H_0 $\beta=0$ 即血清总胆固醇含量与年龄之间无直线关系

H_1 $\beta\neq0$ 即血清总胆固醇含量与年龄之间有直线关系

$\alpha=0.05$

$F_{0.05(1,7)}=5.591, F=59.461>5.591, P<0.05$

$$t=\frac{0.0695}{0.009}=7.722$$

$\sqrt{F}=\sqrt{59.461}=7.711\approx t$

方差分析表

变异来源	自由度	*SS*	*MS*	*F*	*P*
总变异	8	8.511			
回归	1	7.611	7.611	59.461	<0.05
残差	7	0.900	0.128		

按 $\alpha=0.05$ 水准，拒绝 H_0 ， 可认为血清总胆固醇含量与年龄之间有直线关系

人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE

（二）总体回归系数的置信区间

$$b \pm t_{\alpha/2,\nu} s_b$$

本例中已计算得到：

$b=0.0695 \qquad s_b=0.009$

按自由度 $\nu=7$ 查t界值表，得到 $t_{0.05/2,7}=2.365$

代入上式计算得到区间：

$$(0.0695-2.365\times0.009,0.0695+2.365\times0.009)=(0.0482,0.0908)$$

人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE

（三）利用回归方程进行估计和预测

1.总体条件均数的置信区间

$$S_{\hat{Y}_0} = S_{Y\cdot X}\sqrt{\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum(X-\bar{X})^2}}$$

条件均数的标准误估计值

$$\hat{Y}_0 \pm t_{\alpha/2,\nu} S_{\hat{Y}_0}$$

条件均数的置信区间

人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE

（三）利用回归方程进行估计和预测

2.个体值的预测区间

$$S_{Y_0}=S_{Y\cdot X}\sqrt{1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum(X-\bar{X})^2}}$$

个体Y值的标准差估计值

$$\hat{Y}_0 \pm t_{\alpha/2,\nu}S_{Y_0}$$

个体值的预测区间

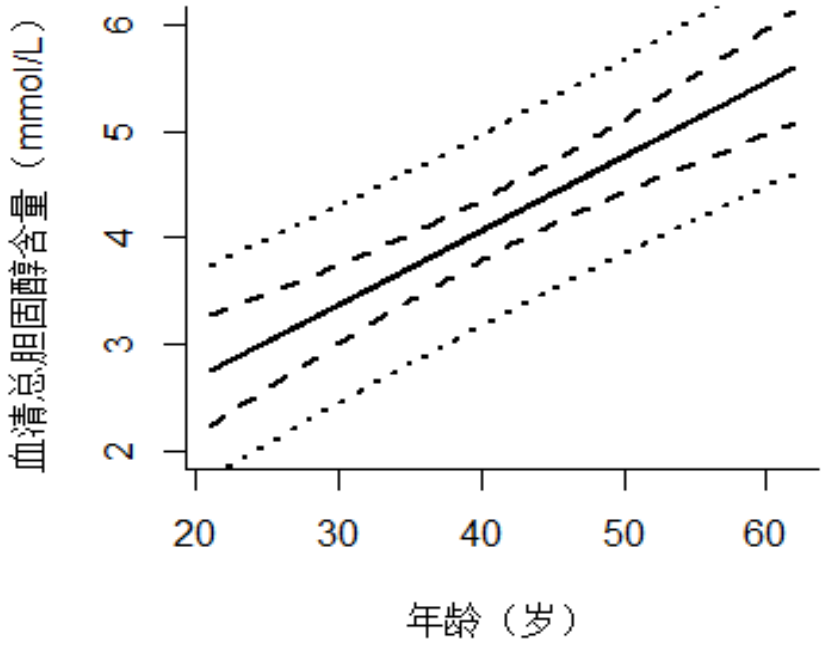

图9-4 $\mu_{Y|X}$ 的置信区间和个体值预测区间示意图

人民卫生出版社
PEOPLE´S MEDICAL PUBLISHING HOUSE

第二节

直线相关

人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE

一、直线相关的概念

直线相关（linear correlation）又称简单相关（simple correlation），用于双变量正态分布（bivariate normal distribution）资料。

- 正相关
- 负相关
- 零相关
- 完全相关

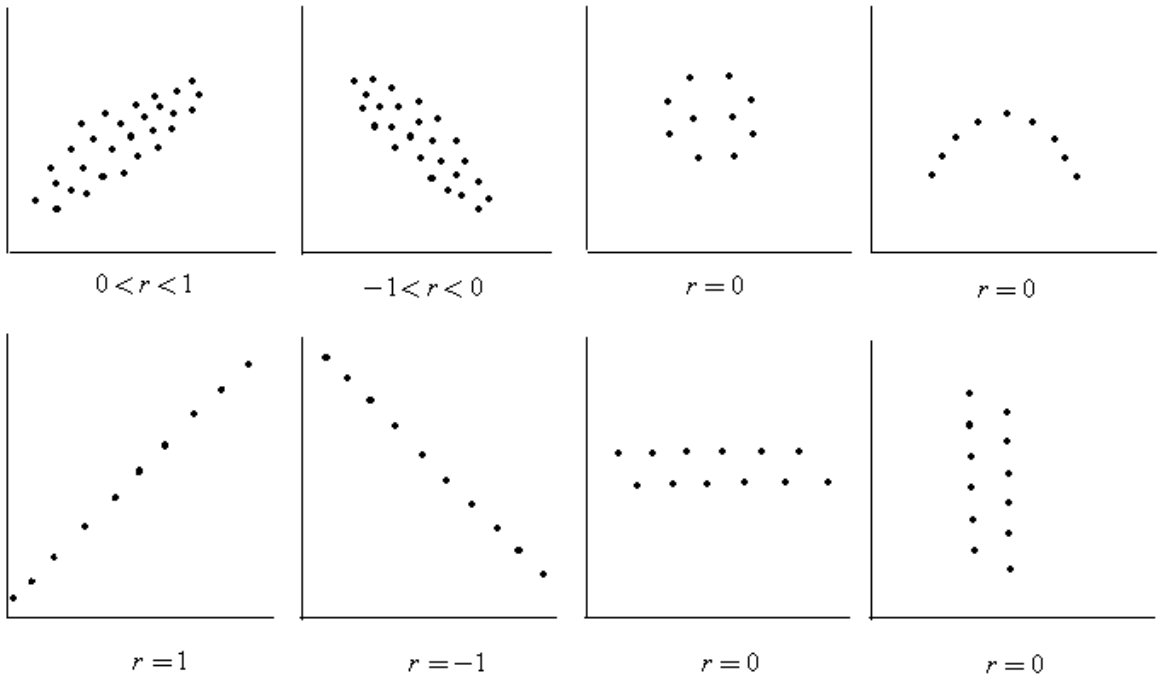

图9-5 直线相关示意图