流行病与卫生统计学
双变量回归与相关

本章讲课内容

第一节　直线回归

第二节　直线有关

第三节　秩有关

第一节　直线回归

一、直线回归的概念

目的： 研究应变量Y对自变量X的数量依存关系。

资料：双变量计量资料，即每个个体有两个变量值。

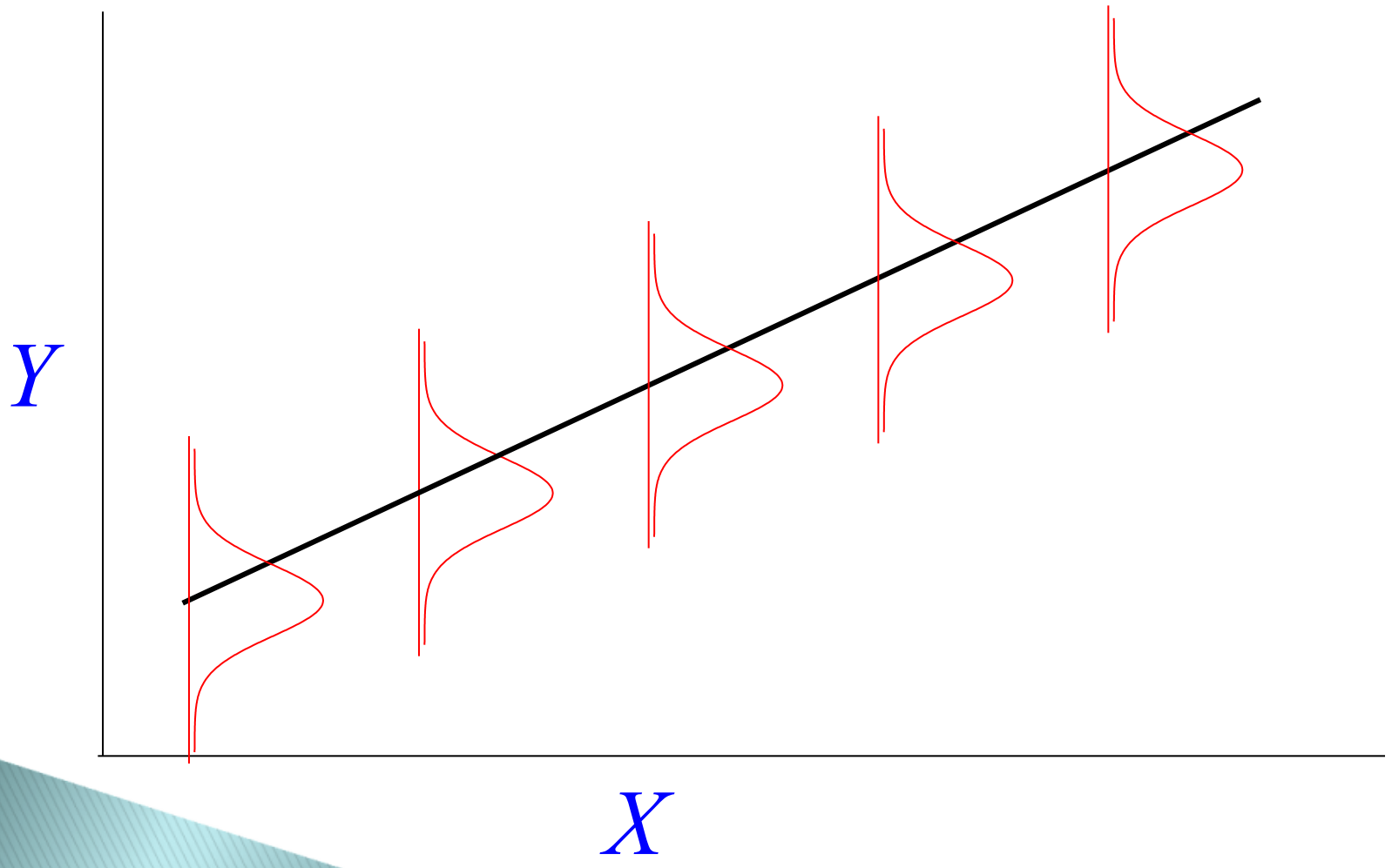
特点： 统计关系。 X值和Y的均数的关系，不同于一般数学上的X 和Y的函数关系。

回归模型的前提假设

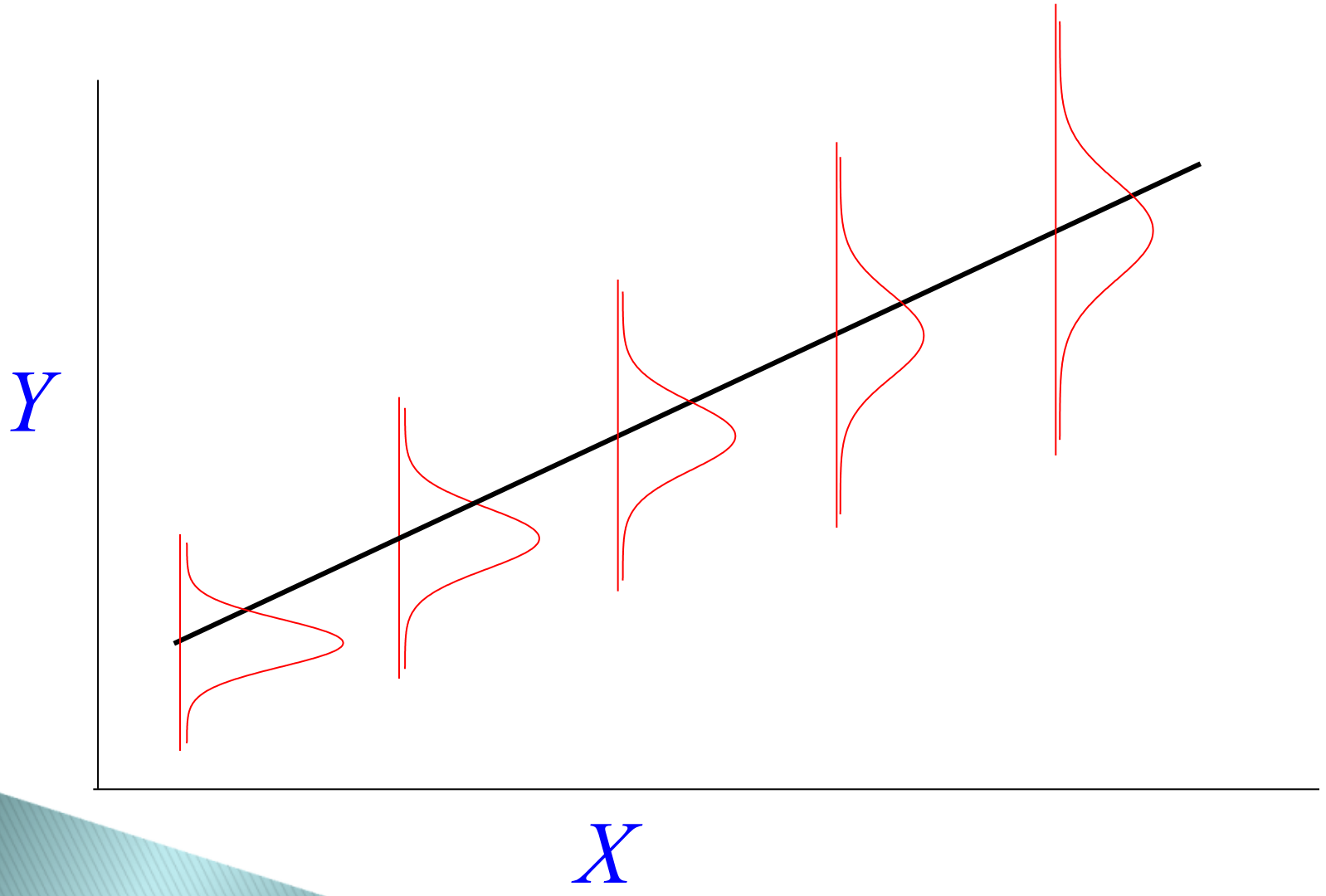
- 线性(linear)
- 独立(independent)
- 正态(normal)
- 等方差(equal variance)

恰好为“**LINE**”。

给定X时，Y是正态分布、等方差示意图



给定X时，Y是正态分布、不等方差示意图



例9-1 某地方病研究所调查了8名正常小朋友的尿肌酐含量（mmol/24h）如表9-1。估计尿肌酐含量（*Y*）对其年龄（*x*）的回归方程。

表9-1 8名正常小朋友的年龄*X*（岁）与尿肌酐含量*Y*（mmol/24h）

编号	1	2	3	4	5	6	7	8
年龄*X*	13	11	9	6	8	10	12	7
尿肌酐含量*Y*	3.54	3.01	3.09	2.48	2.56	3.36	3.18	2.65

自变量

反应变量

简朴回归

尿肌酐含量 *Y* 随年龄 *X* 增长而增大且呈直线趋势，但8个点并非恰好全都在一直线上，此与两变量间严格的直线函数关系不同，称为直线回归,其方程叫直线回归方程，以区别严格意义的直线方程。

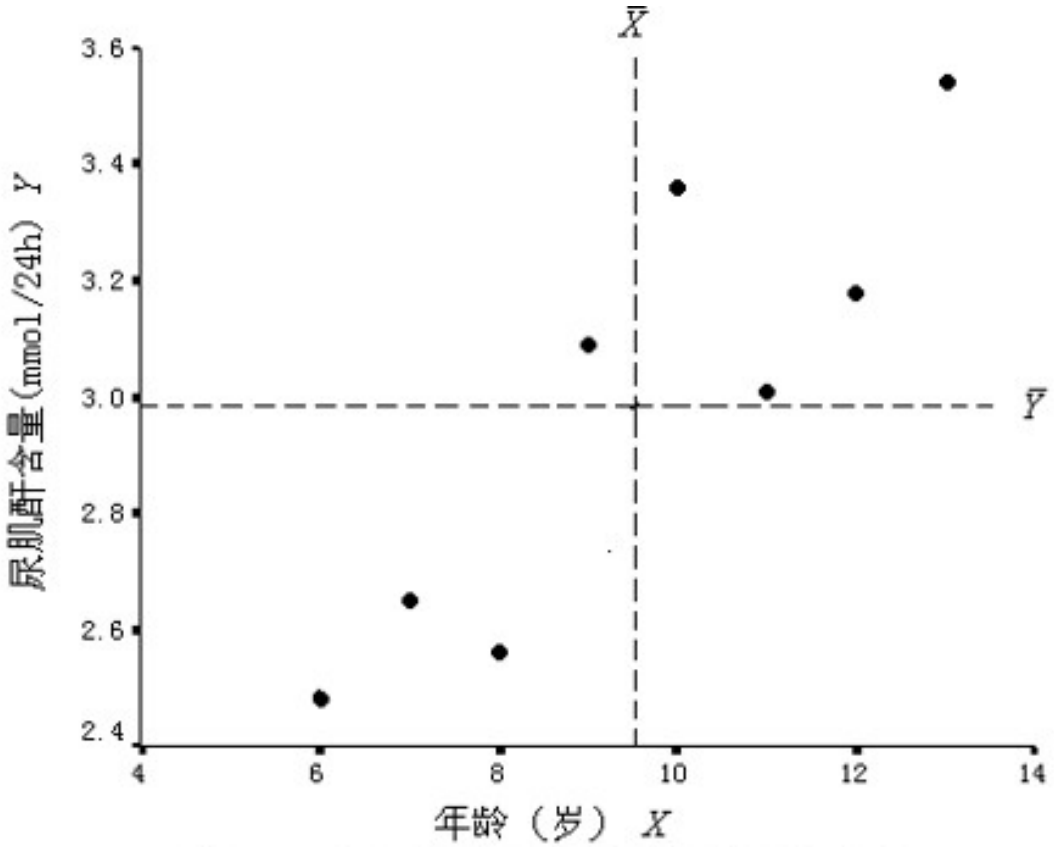


图9-1　8名儿童的年龄与其尿肌酐含量散点图

直线回归方程的一般体现式为

$$\hat{Y} = a + bX \qquad (9-1)$$

$\hat{Y}$为各X处Y的总体均数的估计。

称为样本回归方程，它是对两变量总体间线性关系的一个估计。根据散点图我们可以假定，对于 X 各个取值，相应 Y 的总体均数 $\mu_{Y|X}$ 在一条直线上（图 9-2），表示为

$$\mu_{Y|X} = \alpha + \beta X \qquad (9-2)$$

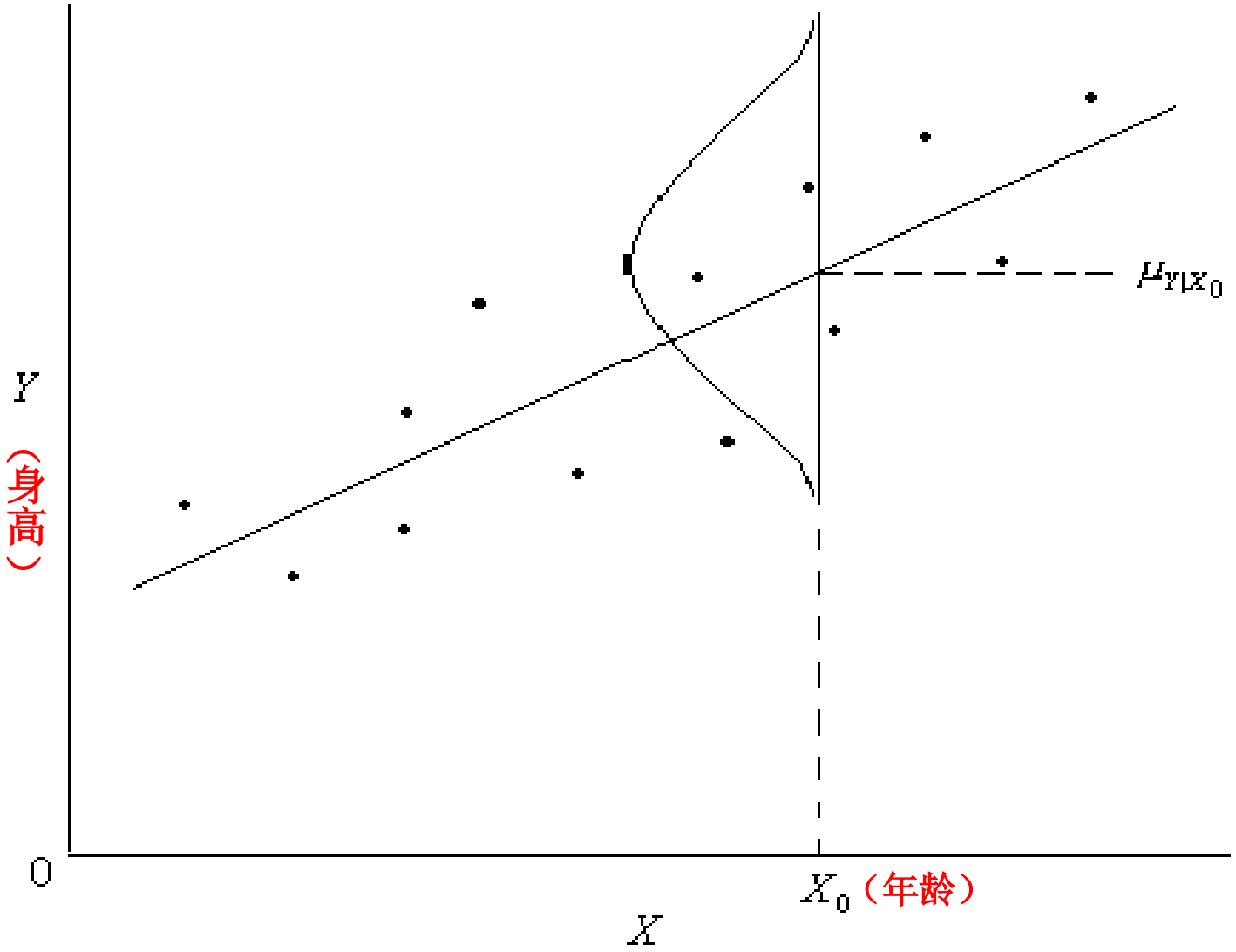


图9-2　直线回归的概念示意图

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

a 的意义

- ***a*** **截距或常数项(intercept, constant)**
- ***X*=0 时，*Y*的估计值**
- ***a*的单位与*Y*值相同**

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

b 的意义

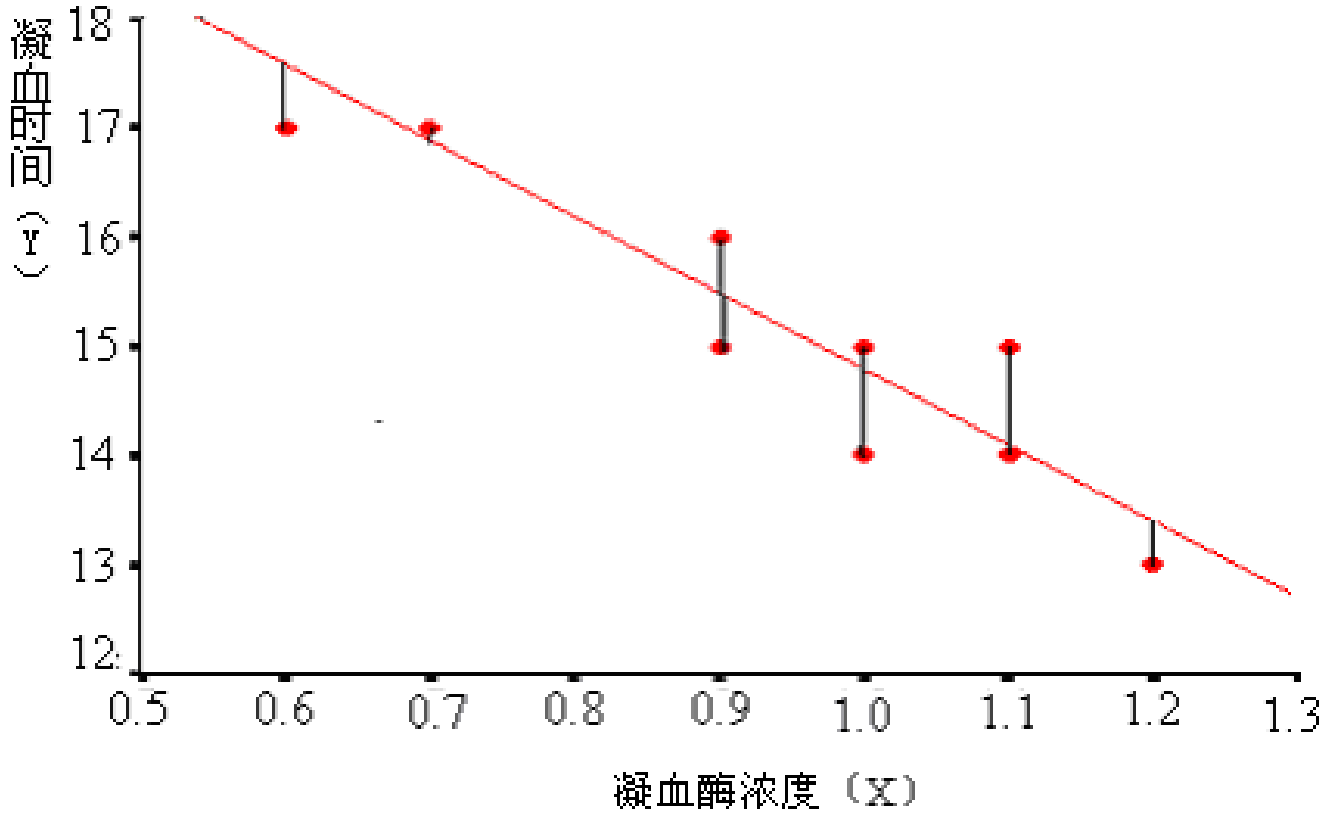
- 回归系数*b*称为斜率(slope)，其统计学意义是：*X* 每增长(减)一种单位，*Y* 平均变化*b*个单位。
- *b* 的单位为 (Y的单位/X的单位)

二、直线回归方程的求法

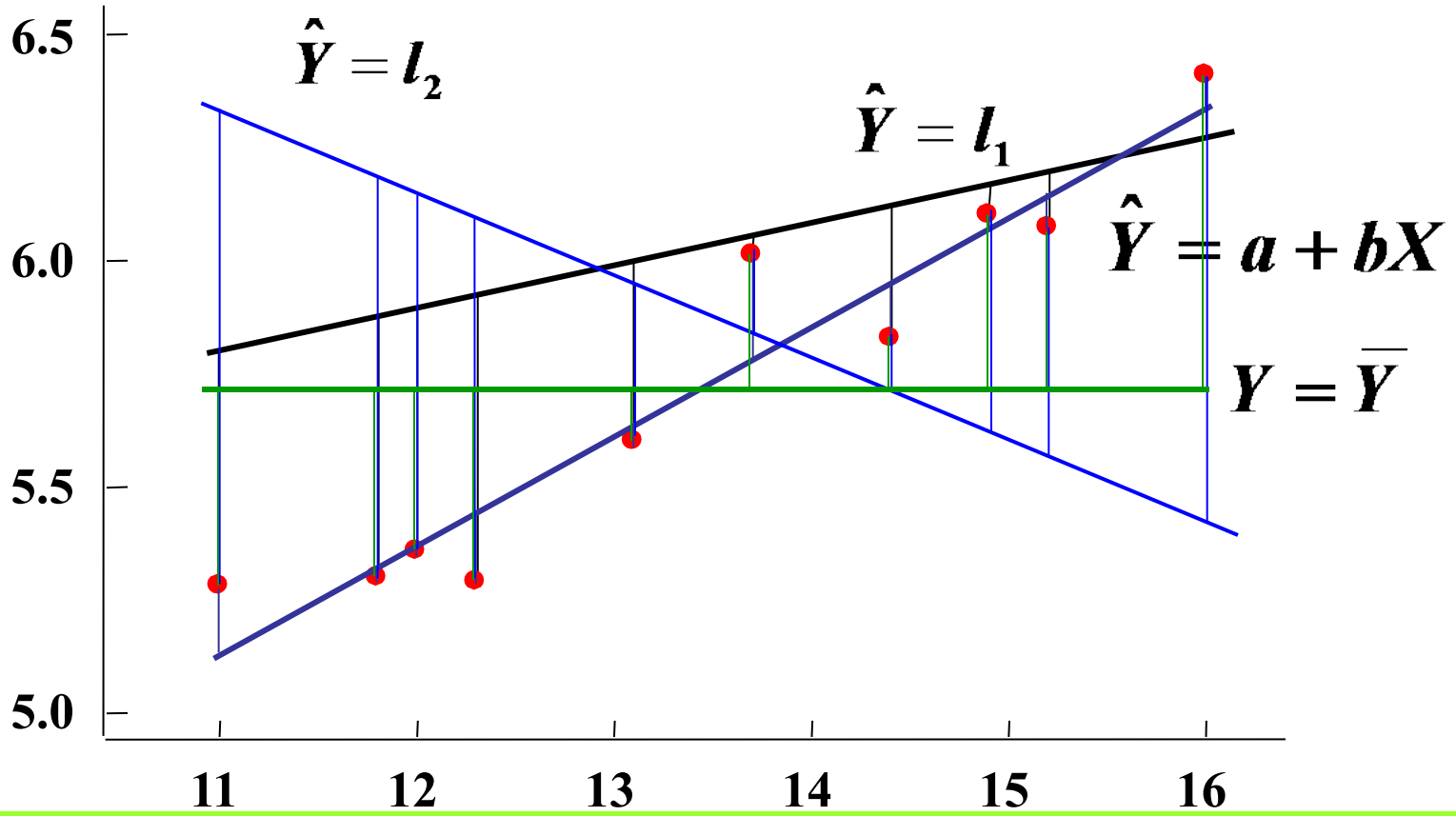
- 残差(residual)或剩余值，即实测值Y_i与假定回归线上的估计值 $\hat{Y}_i$ 的纵向距离 。
- 求解a、b实际上就是“合理地”找到一条能最佳地代表数据点分布趋势的直线。

$Y-\hat{Y}$的意义

为残差：点到直线的纵向距离。



点到直线的距离



原则：最小二乘法(least sum of squares)，即可确保各实测点至直线的纵向距离的平方和最小

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

$$b = \frac{l_{XY}}{l_{XX}} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2} \qquad (9\text{-}3)$$

$$a = \bar{Y} - b\bar{X} \qquad (9\text{-}4)$$

式中 l_{XY} 为 X 与 Y 的离均差乘积和：

$$l_{XY} = \sum (X - \bar{X})(Y - \bar{Y}) = \sum XY - \frac{(\sum X)(\sum Y)}{n} \qquad (9\text{–}5)$$

①先作散点图，以判断两变量间是否呈线性趋势

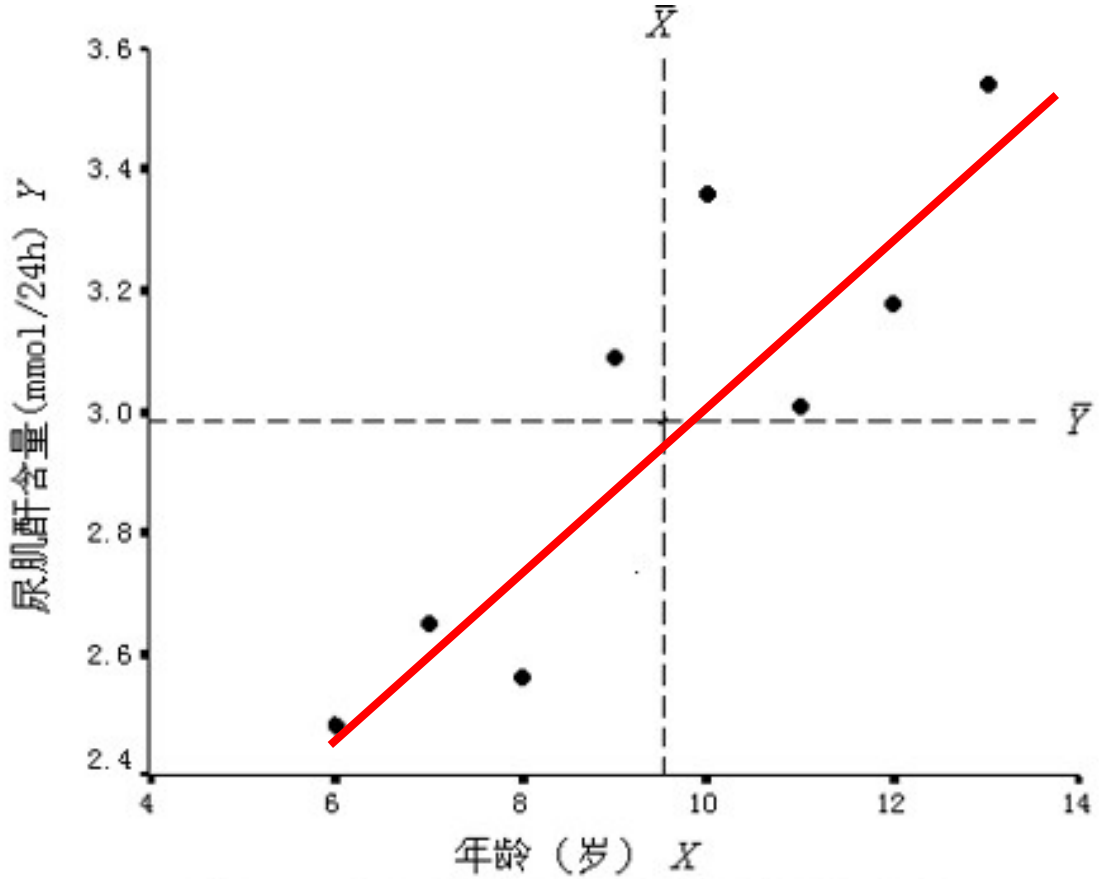


图9-1　8名儿童的年龄与其尿肌酐含量散点图

②求直线回归方程

$$\bar{X} = \frac{\sum X}{n} = \frac{76}{8} = 9.5$$

$$\bar{Y} = \frac{\sum Y}{n} = \frac{23.87}{8} = 2.9838$$

$$l_{XX} = \sum X^2 - \frac{(\sum X)^2}{n} = 764 - \frac{(76)^2}{8} = 42$$

$$l_{YY} = \sum Y^2 - \frac{(\sum Y)^2}{n} = 72.2683 - \frac{(23.87)^2}{8} = 1.0462$$

$$l_{XY} = \sum XY - \frac{(\sum X)(\sum Y)}{n} = 232.61 - \frac{(76)(23.87)}{8} = 5.8450$$

$$b = \frac{l_{XY}}{l_{XX}} = \frac{5.8450}{42} = 0.1392$$

$$a = \bar{Y} - b\bar{X} = 2.9838 - (0.1392)(9.5) = 1.6617$$

$$\hat{Y} = 1.6617 + 0.1392X$$

③绘制回归直线

在自变量实测范围内远端取易于读数的X值代入回归方程得到一种点的坐标，连接此点与点$(\bar{X},\bar{Y})$也可绘出回归直线。

此直线必然经过点$(\bar{X},\bar{Y})$且与纵坐标轴相交于截距a。

三、直线回归中的统计推断

（一）回归方程的假设检验

建立样本直线回归方程，只是完毕了统计分析中两变量关系的统计描述，研究者还须回答它所来自的总体的直线回归关系是否确实存在，即是否对总体有 $\beta \neq 0$ ？

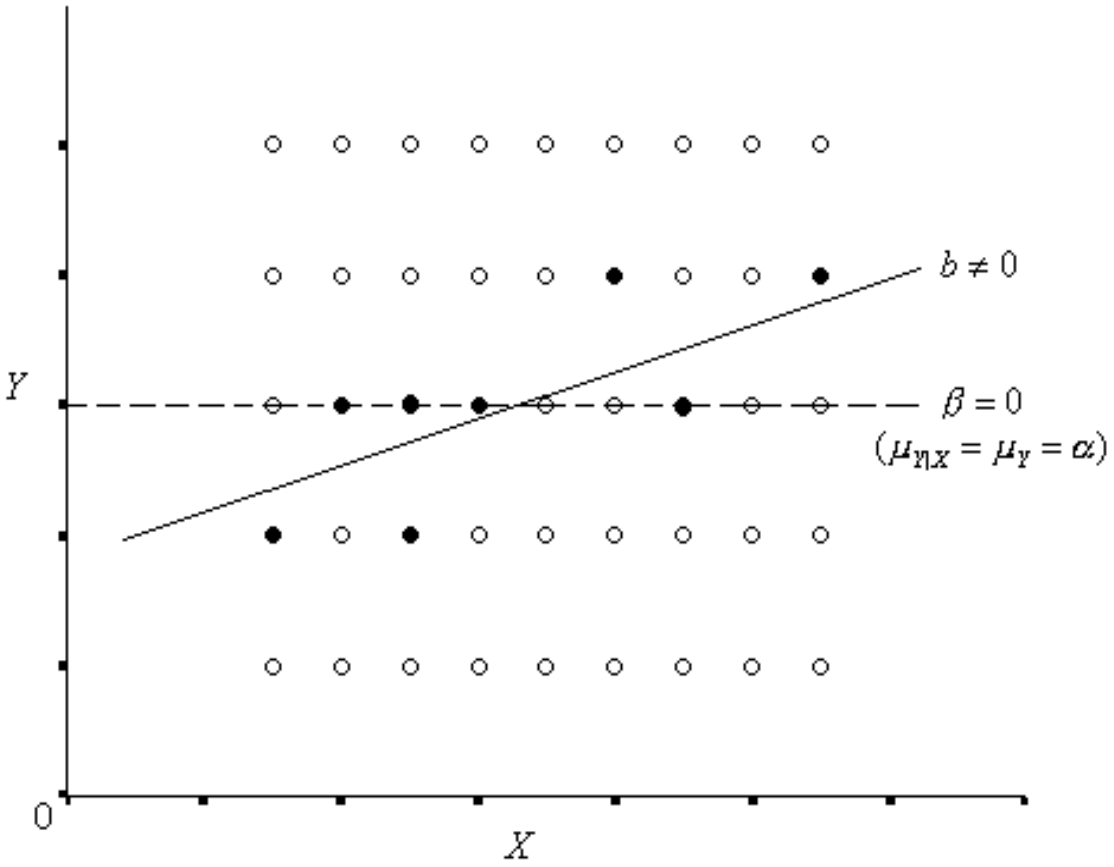


图9-3　总体回归系数与样本回归系数示意图

无论 X

$\mu_{Y|X}$

$\beta = 0$

Y　X

$\mu_{Y|X} = \mu_Y$

b　b

1．方差分析

理解回归中方差分析的基本思想，

需要对应变量 Y 的离均差平方和 l_{YY}

因变量总变异的分解

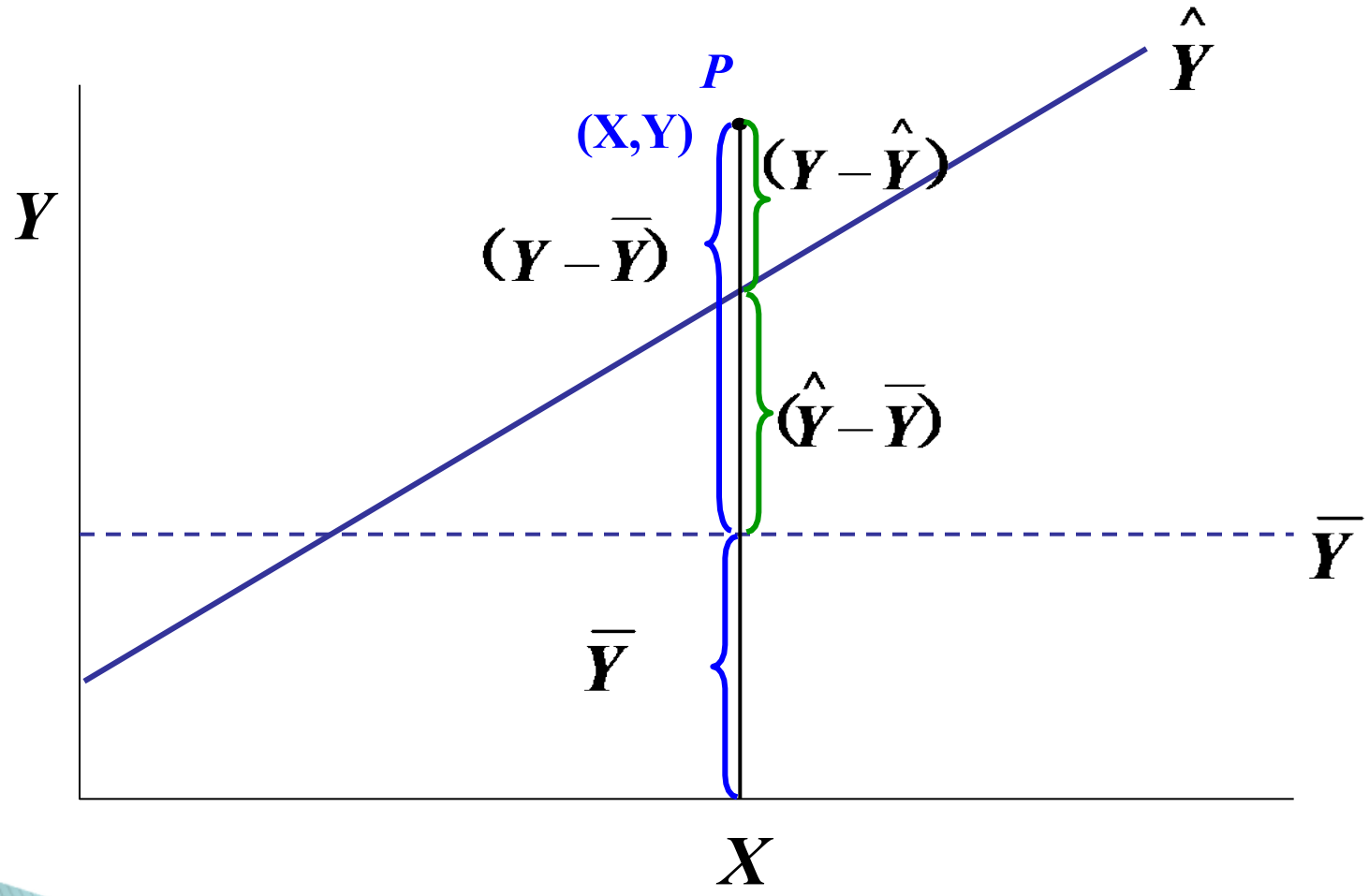


图5-4　平方和划分示意图

Y的总变异分解

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

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

总变异 $SS_{总}$

回归平方和 $SS_{回}$

剩余平方和 $SS_{剩}$

数理统计可证明： $\sum(\hat{Y}-\bar{Y})(Y-\hat{Y})=0$

$$SS_{总} = SS_{回} + SS_{残} \qquad (9\text{-}6)$$

$$\nu_{总} = \nu_{回} + \nu_{残}，\ \nu_{总} = n-1，\ \nu_{回} = 1，\ \nu_{残} = n-2 \qquad (9\text{-}7)$$

即 $SS_{回}$ $\sum(\hat{Y}-\overline{Y})^2$ ，为回归平方和。由于特定样本的均数

$\overline{Y}$ 是固定的，所以这部分变异由 $\hat{Y}_i$ 的大小不同引起。

当 X 被引入回归以后，

X_i 取值不同导致了 $\hat{Y}_i = a + bX_i$ 不同，故 $SS_{回}$

反映了在 Y 的总变异中可以用 X 与 Y 的直线关系解释的那部分变异。

b 离 0 越远，X 对 Y 的影响越大，

$SS_{回}$ 就越大，说明回归效果越好。

即 $SS_{残}$ $\sum(Y-\hat{Y})^2$ ，为残差平方和。它反应除了 X 对 Y 的线性影响之外的一切因素对 Y 的变异的作用，也就是在总平方和中无法用 X 解释的部分，表示考虑回归之后 Y

父母身高与子女身高：遗传+其他原因

- 政治
- 经济
- 环境
- 文化

以上分解可见，不考虑回归时，随机误差是 Y 的总变异 $SS_{总}$

$SS_{残}$

假如两变量间总体回归关系确实存在，回归的贡献就要不小于随机误差，大到何种程度时能够以为具有统计意义，可计算统计量F。

$$F=\frac{SS_{回}/\nu_{回}}{SS_{残}/\nu_{残}}=\frac{MS_{回}}{MS_{残}},\qquad \nu_{回}=1,\quad \nu_{残}=n-2 \tag{9-8}$$

式中

$MS_{回}$ 为回归均方

$MS_{残}$ 为残差均方。

F 服从自由度为 $\nu_{回}$、$\nu_{残}$ 的 F 分布。

$$SS_{回}=bl_{XY}=l_{XY}^2/l_{XX}=b^2l_{XX} \tag{9-9}$$

2. t 检验

对 $\beta = 0$ 这一假设是否成立还可进行如下 t

$$t_b = \frac{b-0}{S_b}, \quad \nu = n-2 \tag{9-10}$$

$$S_b = \frac{S_{Y \cdot X}}{\sqrt{l_{XX}}} \tag{9-11}$$

$$S_{Y \bullet X} = \sqrt{\frac{SS_{残}}{n-2}} \tag{9-12}$$

例9-2 检验例9-1数据得到的直线回归方程是否成立？

（1）方差分析

H_0：$\beta = 0$，即尿肌酐含量与年龄之间无直线关系

H_1：$\beta \neq 0$，即尿肌酐含量与年龄之间有直线关系

$\alpha = 0.05$

$SS_{回} = l_{XY}^2 / l_{XX} = 5.845^2 / 42 = 0.8134$

$SS_{残} = SS_{总} - SS_{回} = 1.0462 - 0.8134 = 0.2328$

列出方差分析表如表9-2。

表9-2　方差分析表

变异来源	自由度	*SS*	*MS*	*F*	*P*
总 变 异	7	1.0462			
回　　归	1	0.8134	0.8134	20.97	<0.01
残　　差	6	0.2328	0.0388		

$\nu_1=1$、$\nu_2=6$，查F 界值表，得$P<0.01$。按$\alpha=0.05$ 水准拒绝H_0，接受H_1，可以认为尿肌酐含量与年龄之间有直线关系。

（2）*t* 检验

H_0、H_1 及α同上

本例$n=8$，$SS_{残}=0.2328$，$l_{XX}=42$，$b=0.1392$

按公式(9-10)、(9-11)和(9-12)

$$S_{Y\cdot X}=\sqrt{\frac{0.2328}{8-2}}=0.1970\text{，}\quad S_b=\frac{0.1970}{\sqrt{42}}=0.0304$$

$$t=\frac{0.1392}{0.0304}=4.579$$

$\nu=6$，查t界值表，得$0.002<P<0.005$。按$\alpha=0.05$水准，拒绝H_0，接受H_1，结论同上。

注意：

本例$\sqrt{F}=\sqrt{20.97}=4.579=t$，即直线回归中对回归系数的$t$检验与$F$检验等价，类似于两样本均数比较可以作$t$检验亦可作方差分析。