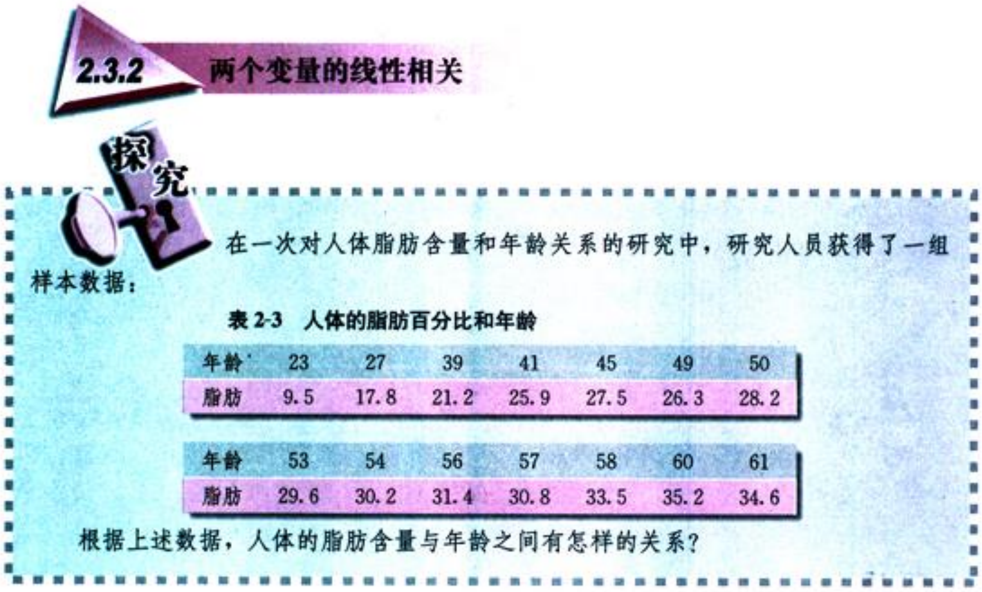
第 144 期
高中教材配套课件创作

课　　题	变量间的相关关系（线性回归方程）
册别 单元	高中数学 人教 A 版 必修 3　 第二章 2.3 变量间的相关关系 人教 A 版选修 2—3 第三章 3.1 回归分析的基本思想及其初步应用
教材所在页码	必修 3 P85~ P93，选修 2—3 P80~ P85

教材对应截图

由原始数据➜散点图➜最小二乘法➜线性回归方程



2.3.2 两个变量的线性相关

探究

在一次对人体脂肪含量和年龄关系的研究中，研究人员获得了一组样本数据：

表 2-3　人体的脂肪百分比和年龄

年龄	23	27	39	41	45	49	50
脂肪	9.5	17.8	21.2	25.9	27.5	26.3	28.2

年龄	53	54	56	57	58	60	61
脂肪	29.6	30.2	31.4	30.8	33.5	35.2	34.6

根据上述数据，人体的脂肪含量与年龄之间有怎样的关系？

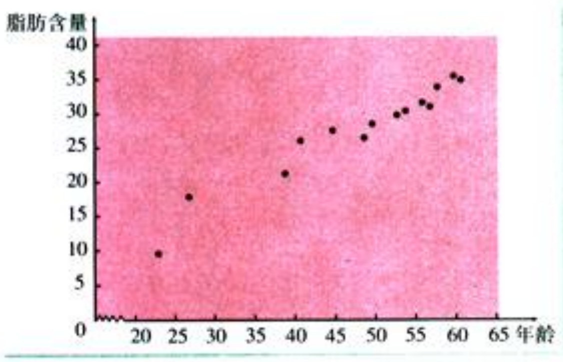


图 2.3-1

线性回归方程➞残差图➞相关指数

上面这些方法虽然有一定的道理，但总让人感到可靠性不强.

实际上，求回归方程的关键是如何用数学的方法来刻画“从整体上看，各点与此直线的距离最小”.

> 你能解释一下“从整体上看，各点与此直线的距离最小”的含义吗？

假设我们已经得到两个具有线性相关关系的变量的一组数据

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n),$$

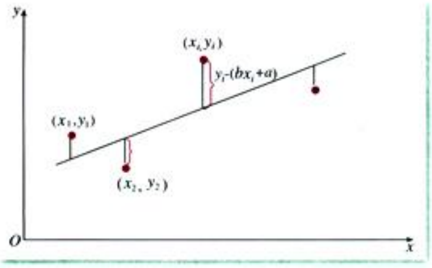
下面探讨如何表达这些点与一条直线

$$y=bx+a$$

之间的距离. 我们可以用点 (x_i, y_i) 与这条直线上横坐标为 x_i 的点之间的距离来刻画点 (x_i, y_i) 到直线的远近，即用

$$|y_i-(bx_i+a)| \quad (i=1, 2, 3, \cdots, n)$$

表示点 (x_i, y_i) 到直线的远近（图 2.3-6）. 这样，用这 n 个距离之和来刻画各点与此直线的“整体距离”是比较合适的，即可以用 $\sum\limits_{i=1}^{n}|y_i-(bx_i+a)|$ 表示各点到直线 $y=bx+a$ 的“整体距离”.



这样，问题就归结为：当 a，b 取什么值时 Q 最小，即点到直线 $y=bx+a$ 的“整体距离”最小. 经过数学上的推导（参见《选修 2-3》），a，b 的值由下列公式给出

$$\begin{cases}\hat{b}=\dfrac{\sum\limits_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}=\dfrac{\sum\limits_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}}{\sum\limits_{i=1}^{n}x_i^2-n\bar{x}^2}, \\ \hat{a}=\bar{y}-\hat{b}\bar{x}.\end{cases} \qquad ②$$

这样，回归方程的斜率为 $\hat{b}$，截距为 $\hat{a}$，即回归方程为

$$\hat{y}=\hat{b}x+\hat{a}.$$

这种通过求①式的最小值而得到回归直线的方法，即使得样本数据的点到回归直线的距离的平方和最小的方法叫做**最小二乘法**（method of least square）.

选修 2—3 P80~ P82

例 1　从某大学中随机选取 8 名女大学生，其身高和体重数据如表 3-1 所示.

表 3-1

编号	1	2	3	4	5	6	7	8
身高/cm	165	165	157	170	175	165	155	170
体重/kg	48	57	50	54	64	61	43	59

求根据女大学生的身高预报体重的回归方程，并预报一名身高为 172 cm 的女大学生的体重.

解：由于问题中要求根据身高预报体重，因此选取身高为自变量 x，体重为因变量 y.

作散点图（图 3.1-1）.

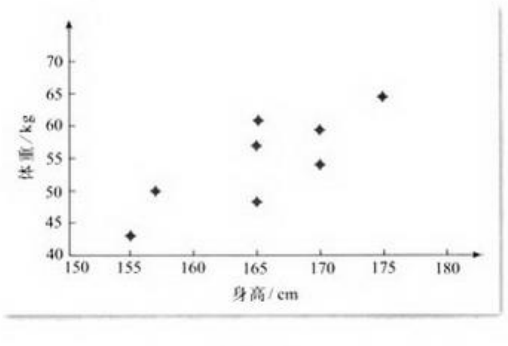


图 3.1-1

从图 3.1-1 中可以看出，样本点呈条状分布，身高和体重有比较好的线性相关关系，因此可以用回归直线 $y=bx+a$ 来近似刻画它们之间的关系.

根据探究中的公式（1）和（2），可以得到

$$\hat{b}=0.849,\ \hat{a}=-85.712.$$

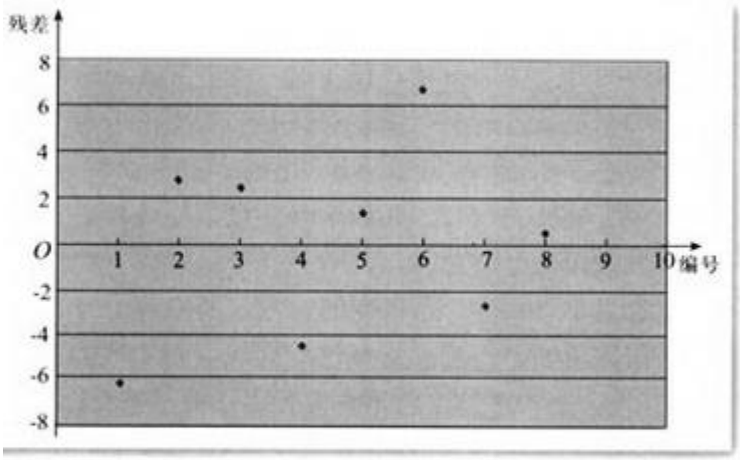
于是得到回归方程

$$\hat{y}=0.849x-85.712.$$

$\hat{b}=0.849$ 是回归直线的斜率的估计值，说明身高 x 每增加 1 个单位时，体重 y 就增加 0.849 个单位，这表明体重与身高具有正的线性相关关系.

因此，对于身高 172 cm 的女大学生，由回归方程可以预报其体重为

$$\hat{y}=0.849\times172-85.712=60.316(\text{kg}).$$



外，残差点比较均匀地落在水平的带状区域中，说明选用的模型比较合适. 这样的带状区域的宽度越窄，说明模型拟合精度越高，回归方程的预报精度越高.

另外，我们还可以用

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

> 在含有一个解释变量的线性模型中，R^2 恰好等于相关系数 r 的平方.

来刻画回归的效果. 对于已经获取的样本数据，R^2 表达式中的 $\sum_{i=1}^{n}(y_i-\bar{y})^2$ 为确定的数. 因此 R^2 越大，意味着残差平方和 $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ 越小，即模型的拟合效果越好；R^2 越小，残差平方和越大，即模型的拟合效果越差. 在线性回归模型中，R^2 表示解释变量对于预报变量变化的贡献率，R^2 越接近于 1，表示回归的效果越好. 在例 1 中，$R^2\approx0.64$，表明“女大学生的身高解释了 64%的体重变化”，或者说“女大学生的体重差异有 64%是由身高引起的”. R^2 是常用的选择模型的指标之一，在实际应用中应该尽量选择 R^2 大的回归模型.

对应的学习目标	1、利用散点图直观认识两个变量之间的相关系，具体地会作散点图，并由此对变量的正相关或负相关关系作出直观的判断； 2、经历描述两个变量线性相关关系的过程.
教学/学习难点	1、利用散点图直观认识两个变量之间的相关系； 2、了解最小二乘法的思想（动态统计各点与直线间的“整体距离”）； 3、根据给出的线性回归方程的系数公式建立回归方程； 4、建立回归思想，理解回归直线与观测数据的关系.
课件设计说明	1、绘制散点图（正相关、负相关）； 2、根据散点图寻找回归直线（各点到直线的整体距离最小）并介绍最小二乘法： ①用 n 个距离之和刻画各点与此直线间的“整体距离”（含绝对值运算不太方便）； ②修正各点与此直线间的“整体距离”（绝对值改为平方）. 3、解释散点图、线性回归方程、样本点中心之间的关系： 两次假设：①第一次假设：（体重和所有因素无关），每个人身高相同（平均身高）➞②第二次假设：（体重只和身高有关）由于身高的原因，各点推离水平直线到一条新的直线（回归直线）➞③由于其他原因（体重和身高以外的因素有关）各点偏离回归直线（散点图）. 4、绘制残差图； 5、动态计算回归直线的相关指数.
使用说明	利用课件按钮提示和变量尺进行操作，可动态改变数据个数、动态绘制表格、动态作散点图（残差图）、动态统计各点与直线间的“整体距离”、动态统计回归直线模型的相关指数，可很方便的重复动态多次操作.
备　　注	