究竟谁是回归直线的方程

花大庆
(安徽省泾县二中,242500)

问题 某5名学生的数学和化学成绩如下表:

数学成绩 x	88	76	73	66	63
化学成绩 y	78	65	71	64	61

若 x,y 成线性相关,求:(1) y 对 x 的线性回归方程;(2) x 对 y 的线性回归方程.

解 (1) 设 y 对 x 的线性回归方程为

$$\hat{y}=\hat{b}_1x+\hat{a}_1.$$

$\because \bar{x}=\frac{1}{5}\times(88+76+73+66+63)=73.2,$

$\bar{y}=\frac{1}{5}\times(78+65+71+64+61)=67.8,$

$\sum_{i=1}^{5}x_iy_i=88\times78+76\times65+73\times71+66\times64+63\times61=25054,$

$\sum_{i=1}^{5}x_i^2=88^2+76^2+73^2+66^2+63^2=27174,$

$$\therefore \hat{b}_1=\frac{\sum_{i=1}^{5}x_iy_i-5\overline{xy}}{\sum_{i=1}^{5}x_i^2-5\bar{x}^2}=\frac{25054-5\times73.2\times67.8}{27174-5\times73.2^2}\approx0.62,$$

$\therefore \hat{a}_1=\bar{y}-\hat{b}_1\bar{x}=67.8-0.62\times73.2=22.42.$

所以 y 对 x 的线性回归方程为

$$\hat{y}=0.62x+22.42.$$

(2) 解法一:

设 x 对 y 的线性回归方程为 $\hat{x}=\hat{b}_2y+\hat{a}_2.$

$\because \sum_{i=1}^{5}y_i=78^2+65^2+71^2+64^2+61^2=23167,$

$$\therefore \hat{b}_2=\frac{\sum_{i=1}^{5}x_iy_i-5\overline{xy}}{\sum_{i=1}^{5}y_i^2-5\bar{y}^2}=\frac{25054-5\times73.2\times67.8}{23167-5\times67.8^2}\approx1.31.$$

$\hat{a}_2=\bar{x}-\hat{b}_2\bar{y}=73.2-1.31\times67.8=-15.62.$

所以 x 对 y 的线性回归方程为

$$\hat{x}=1.31y=15.62 \qquad (Ⅰ)$$

课堂教学时,有学生认为问题(2)有更简单的解法.

解法二:

由问题(1)的结论,得

$$x=\frac{\hat{y}-22.42}{0.62}=1.61\hat{y}-36.16 \qquad (Ⅱ)$$

所以 x 对 y 的线性回归方程为

$$\hat{x}=1.61y=36.16 \qquad (Ⅱ)$$

方程(Ⅰ)(Ⅱ)明显不同,到底哪个才是 x 对 y 的线性回归方程呢?解法一和解法二哪个正确?

很简单!只需比较方程(Ⅰ)(Ⅱ)的残差平方和的大小,即可判断方程(Ⅰ)(Ⅱ)哪个方程拟合效果更好.

记由方程(Ⅰ)得到各 x 的估计值为 $\hat{x}_{Ⅰ}$,由方程(Ⅱ)得到各 x 的估计值为 $\hat{x}_{Ⅱ}$,则由方程(Ⅰ)(Ⅱ)得到的残差平方和分别为 $(\sum\hat{x}_{Ⅰ}-x)^2$、$\sum(\hat{x}_{Ⅱ}-x)^2$.

列表如下:

y	78	65	71	64	61
x	88	76	73	66	63
$\hat{x}_{Ⅰ}$	86.56	69.53	77.39	68.22	64.29
$\hat{x}_{Ⅱ}$	89.42	68.49	78.15	66.88	62.05
$(x-\hat{x}_{Ⅰ})^2$	2.07	41.86	19.27	4.93	1.66
$(x-\hat{x}_{Ⅱ})^2$	2.02	56.40	26.52	0.77	0.90

$\therefore \sum(\hat{x}_{Ⅰ}-x)^2=69.80,\quad \sum(\hat{x}_{Ⅱ}-x)^2=86.61.$

显然,方程(Ⅰ)的残差平方和较小.

结论:通过方程(Ⅰ),由 y 的值估计 x 的值回归效果更好;以上问题(2)的解法一正确,而解法二是错误的.

解法二错在何处?通过图1,可以作出几何上的解释.虽然回归直线 $l:y=bx+a$ 使得各散点 $P_i(i=1,2,3,\cdots,n)$ 到它的"竖直距离"和最小,但未必能使各散点到它的"水平距离"和最小.例如,上图中,点 $P_i(x_i,y_i)$ 到直线 l 的"水平距离" $|P_iR|$ 大于它到直线 l 的"竖直距离" $|P_iQ|$.这就使得通过方程 $y=bx+a$ $(b\neq0)$ 解得的方程 $x=\frac{1}{b}y-\frac{a}{b}$ 并不能作为 x 对 y 的线性回归方程,因为各残差的绝对值 $|e_i|=|x_i-\hat{x}_i|=|P_iR|$ 之和并不一定最小,从而残差的平方和也不一定最小.

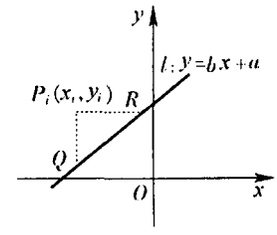

图 1

(收稿日期:2010-04-26)

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net