

第四章 概率与统计

4.3 统计模型

4.3.1 一元线性回归模型

知识梳理

一. 相关关系

1. 变量之间的常见关系

(1) 函数关系：两个变量之间的关系可以用函数表示.

(2) 相关关系：如果两个变量中一个变量的取值一定时，另一个变量的取值带有一定的随机性，那么这两个变量之间的关系，叫做相关关系.

(3) 不相关：两个变量间没有任何关系.

2. 散点图

(1) 在讨论两个变量 x 和 y 之间的关系时，常把它们写成点 (x, y) 的形式，以便利用平面直角坐标系来考虑它们之间的关系，此时 x 和 y 可以看成是描述同一个体的两个不同的特征量.

(2) 将各数据在平面直角坐标系中的对应点画出来，得到表示两个变量的一组数据的图形，这样的图形叫散点图.

3. 线性相关关系

(1) 线性相关：如果由变量的成对数据、散点图或直观经验可知，变量 x 与变量 y 之间的关系可以近似地用一次函数来刻画，则称 x 与 y 线性相关.

(2) 正相关：在线性相关中，如果一个变量增大，另一个变量大体上也增大，则称这两个变量正相关.

(3) 负相关：在线性相关中，如果一个变量增大，另一个变量大体上减少，则称这两个变量负相关.

二. 回归直线方程

1. 回归直线方程

一般地，已知变量 x 与 y 的 n 对成对数据 (x_i, y_i) , $i=1, 2, 3, \dots, n$. 任意给定一个一次函数

$y = bx + a$, 对每一个已知的 x_i , 由直线方程可以得到一个估计值 $\hat{y}_i = bx_i + a$, 如果一次函数 $y = \hat{b}x + \hat{a}$ 能

使 $(\hat{y}_1 - y_1)^2 + (\hat{y}_2 - y_2)^2 + \dots + (\hat{y}_n - y_n)^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ 取得最小值, 则 $\hat{y} = \hat{b}x + \hat{a}$ 称为 y 关于 x 的回归直线方程 (对应的直线称为回归直线).

2.最小二乘法确定回归直线方程: 给定两个变量 y 与 x 的一组数据之后, 回归直线方程 $\hat{y} = \hat{b}x + \hat{a}$ 总

是存在, 而且 $\hat{b} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$, $\hat{a} = \bar{y} - \hat{b}\bar{x}$.

其中, $\hat{b}$ 称为回归系数. 它实际上也就是回归直线方程的斜率.

3.回归直线的性质

(1)回归直线过点 $(\bar{x}, \bar{y})$.

(2)回归直线方程 $\hat{y} = \hat{b}x + \hat{a}$, $\hat{b} > 0$ 时, y 与 x 正相关; $\hat{b} < 0$ 时, y 与 x 负相关.

三. 相关系数

1.相关系数 r 的计算公式

假设两个随机变量的数据分别为 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, 则变量间相关系数 r 的计算公式如下:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^n x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^n y_i^2 - n\bar{y}^2\right)}}.$$

2.相关系数 r 的性质

(1) $|r| \leq 1$, 且 y 与 x 正相关的充要条件是 $r > 0$, y 与 x 负相关的充要条件是 $r < 0$.

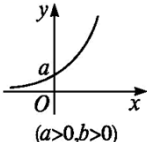
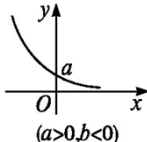
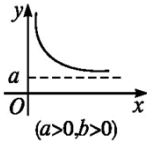
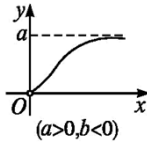
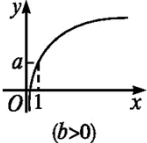
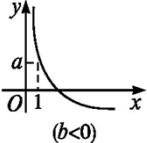
(2) $|r|$ 越小, 说明两个变量之间的线性相关性越弱, 也就是得出的回归直线方程越没有价值, 即方程越不能反映真实的情况; $|r|$ 越大, 说明两个变量之间的线性相关性越强, 也就是得出的回归直线方程越有价值.

(3) $|r| = 1$ 的充要条件是成对数据构成的点都在回归直线上.

四、非线性回归

常见的非线性回归模型转化为线性回归模型

曲线方程	曲线图形	变换公式	变换后的 线性函数
$y = ax^b$		$c = \ln a$ $v = \ln x$ $u = \ln y$	$u = c + bv$

$y=ae^{bx}$	 	$c=\ln a$ $u=\ln y$	$u=c+bx$
	 	$c=\ln a$ $v=\frac{1}{x}$ $u=\ln y$	$u=c+bv$
$y=a+b\ln x$	 	$v=\ln x$ $u=y$	$u=a+bv$

常见考点

考点一 线性回归

典例 1. 关于回归分析，下列说法错误的是（ ）

- A. 回归分析是研究两个具有相关关系的变量的方法
- B. 散点图中，解释变量在 x 轴，预报变量在 y 轴
- C. 回归模型中一定存在随机误差
- D. 散点图能明确反映变量间的关系

【答案】D

【分析】

由回归分析的概念逐项辨析即可得解.

【详解】

由回归分析的概念可得 A B C 正确；

对于 D 选项，由于用散点图反映两个变量间的关系时，存在误差，因此不能准确的反映变量间的关系

故选：D

变式 1-1. 在对两个变量 x , y 进行回归分析时有下列步骤：

- ①对所求出的回归方程作出解释；②收集数据 (x_i, y_i) , $i=1, 2, \dots, n$;

③求回归方程；④根据所收集的数据绘制散点图.

则下列操作顺序正确的是 ()

- A. ①②④③ B. ③②④① C. ②③①④ D. ②④③①

【答案】D

【分析】

先收集数据，再画散点图，然后求回归方程即可

【详解】

根据回归分析的思想，可知对两个变量 x, y 进行回归分析时，应先收集数据 (x_i, y_i) ，然后绘制散点图，再求回归方程，最后对所求的回归方程作出解释.

故选：D

变式 1-2. 已知两个变量 x 和 y 之间有线性相关关系，经调查得到如下样本数据，

	3	4	5	6	7
y	3.5	2.4	1.1	-0.2	-1.3

根据表格中的数据求得回归方程 $\hat{y} = \hat{b}x + \hat{a}$ ，则下列说法正确的是 ()

- A. $a > 0, b > 0$ B. $a > 0, b < 0$
C. $a < 0, b > 0$ D. $a < 0, b < 0$

【答案】B

【分析】

根据已知数据，可知变量 x 和变量 y 之间存在负相关的关系，即可判断出 $\hat{b} < 0$ ，再令 $x = 0$ ，可求出 $\hat{a} = \hat{y} > 3.5 > 0$ ，即可得出答案.

【详解】

解：由已知数据，可知 y 随着 x 的增大而减小，

则变量 x 和变量 y 之间存在负相关的关系， $\therefore \hat{b} < 0$ ，

当 $x = 0$ 时，则 $\hat{a} = \hat{y} > 3.5 > 0$ ，

即： $a > 0, b < 0$.

故选：B.

【点睛】

本题考查变量间的相关关系和线性回归方程的应用问题.

变式 1-3. 对具有线性相关关系的变量 x, y , 测得一组数据如下表, 根据表中数据, 利用最小二乘法得到回归直线方程, 据此模型预测当 $x = 20$ 时, y 的估计值为 ()

x	2	4	5	6	8
y	20	40	60	70	80

- A. 210 B. 210.5 C. 211.5 D. 212.5

【答案】C

【分析】

求出样本中心, 然后确定回归直线方程, 即可求解预测当 $x = 20$ 时, y 的估计值.

【详解】

由题意可知: $\bar{x} = \frac{2+4+5+6+8}{5} = 5$,

$\bar{y} = \frac{20+40+60+70+80}{5} = 54$.

$\therefore$ 回归直线方程经过样本中心, $\therefore 54 = 10.5 \times 5 + \hat{a}$, $\hat{a} = 1.5$,

回归直线方程为: $\hat{y} = 10.5x + 1.5$,

当 $x = 20$ 时, y 的估计值为: $10.5 \times 20 + 1.5 = 211.5$.

故选: C.

考点二 相关系数的意义及辨析

典例 2. 有一组数据统计了 2013 年至 2020 年中国高铁每年的运营里程, 它反映了中国高铁近几年的飞速发展, 甲同学用线性回归模型 $y = bx + a$ 来拟合, 并算得相关系数 $r_1 = 0.70$, 乙同学用指数函数模型 $y = ce^{dx}$ 来拟合, 并算得转化为线性回归方程所对应的相关系数 $r_2 = 0.99$, 武问哪个模型拟合效果更好? ()

- A. 线性回归模型拟合效果更好
B. 指数函数模型拟合效果更好
C. 两种模型拟合效果都不好
D. 不能确定

【答案】B

【分析】

根据相关系数越接近拟合效果越好即可得出答案.

【详解】

因为相关系数的绝对值越接近 1 拟合效果越好，

由题意得 r_2 更接近 1，

所以指数函数模型拟合效果更好

故选：B

变式 2-1. 已知变量与 Y 相对应的一组数据为 $(10,1)$, $(11.3,2)$, $(11.8,3)$, $(12.5,4)$, $(13,5)$, 变量 U 与 V 相对应的一组数据为 $(10,5)$, $(11.3,4)$, $(11.8,3)$, $(12.5,2)$, $(13,1)$. 表示变量与 Y 之间的线性相关系数, r_2 表示变量 U 与 V 之间的线性相关系数, 则下列结论中正确的是 ()

- A. $r_2 < r_1 < 0$ B. $0 < r_2 < r_1$ C. $r_2 < 0 < r_1$ D. $r_2 = r_1$

【答案】C

【分析】

根据变量对应数据可确定与 Y 之间正相关, U 与 V 之间负相关, 由此可得相关系数的大小关系.

【详解】

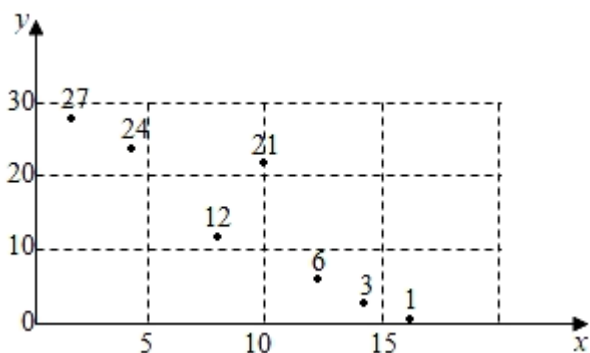
由变量与 Y 相对应的一组数据为 $(10,1)$, $(11.3,2)$, $(11.8,3)$, $(12.5,4)$, $(13,5)$, 可得变量与 Y 之间正相关, $\therefore r_1 > 0$;

由变量 U 与 V 相对应的一组数据为 $(10,5)$, $(11.3,4)$, $(11.8,3)$, $(12.5,2)$, $(13,1)$, 可知变量 U 与 V 之间负相关, $\therefore r_2 < 0$;

综上所述: 与 r_2 的大小关系是 $r_2 < 0 < r_1$.

故选: C.

变式 2-2. 如图是相关变量, y 的散点图, 现对这两个变量进行线性相关分析, 方案一: 根据图中所有数据, 得到线性回归方程 $\hat{y} = \hat{\beta}_1 x + \hat{\beta}_1$, 相关系数为; 方案二: 剔除点 $(10,21)$, 根据剩下数据得到线性回归直线方程 $\hat{y} = \hat{\beta}_2 x + \hat{\beta}_2$, 相关系数为 r_2 . 则 ()



- A. $0 < r_1 < r_2 < 1$ B. $0 < r_2 < r_1 < 1$
 C. $-1 < r_1 < r_2 < 0$ D. $-1 < r_2 < r_1 < 0$

【答案】D

【分析】

根据散点图知变量、 y 具有负线性相关关系，且点(10,21) 是离群值；
 剔除离群值后，线性相关性强些，是负相关，由此得出正确的结论。

【详解】

根据相关变量、 y 的散点图知，变量、 y 具有负线性相关关系，且点(10,21) 是离群值；
 方案一中，没剔除离群值，线性相关性弱些，成负相关；
 方案二中，剔除离群值，线性相关性强些，也是负相关；
 所以相关系数 $-1 < r_2 < r_1 < 0$.

故选：D

变式 2-3. 对四组数据进行统计，获得以下散点图，关于其相关系数的比较，正确的是（ ）

- A. $r_2 < r_4 < 0 < r_3 < r_1$ B. $r_4 < r_2 < 0 < r_1 < r_3$
 C. $r_4 < r_2 < 0 < r_3 < r_1$ D. $r_2 < r_4 < 0 < r_1 < r_3$

【答案】A

【分析】

由给出的四组数据的散点图，结合相关系数的概念，逐图判定，即可求解.

【详解】

由给出的四组数据的散点图可以看出，题图 1 和题图 3 是正相关，相关系数大于 0，
 题图 2 和题图 4 是负相关，相关系数小于 0，

题图 1 和题图 2 的点相对更加集中，所以相关性更强，所以接近于 1， r_2 接近于，

由此可得 $r_2 < r_4 < 0 < r_3 < r_1$.

故选：A.

考点三 残差与相关指数的计算

典例 3. 根据一组样本数据 (x_1, y_1) , (x_2, y_2) , $\dots$, (x_n, y_n) 的散点图分析 x 与 y 之间是否存在线性相关关系，求得其线性回归方程为 $\hat{y} = 0.85x - 85.7$ ，则在样本点 $(165, 57)$ 处的残差为 ()

A. 54.55 B. 2.45 C. 3.45 D. 111.55

【答案】B

【分析】

根据样本点的横坐标和回归直线方程得出 y 的估计值，根据残差定义计算.

【详解】

把 $x = 165$ 代入 $\hat{y} = 0.85x - 85.7$ ，得 $\hat{y} = 0.85 \times 165 - 85.7 = 54.55$ ，

所以在样本点 $(165, 57)$ 处的残差 $\hat{e} = y - \hat{y} = 57 - 54.55 = 2.45$.

故选：B.

变式 3-1. 变量 U 与 V 相对应的一组样本数据为 $(1, 1.4)$ 、 $(2, 2.2)$ 、 $(3, 3)$ 、 $(4, 3.8)$ ，由上述样本数据得到 U 与 V 的线性回归分析， R^2 表示解释变量对于预报变量变化的贡献率，则 $R^2 =$ ()

附：决定系数公式 $R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$.

A. $\frac{3}{5}$ B. $\frac{4}{5}$
C. 1 D. 3

【答案】C

【分析】

利用决定系数公式可求得结果.

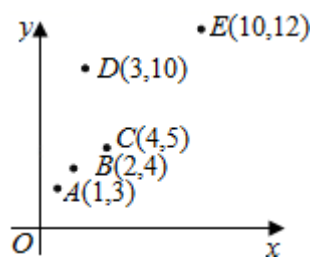
【详解】

易知，点 $(1,1.4)$ 、 $(2,2.2)$ 、 $(3,3)$ 、 $(4,3.8)$ 都在直线 $y=0.8x+0.6$ 上，

所以， $y_i = \hat{y}_i (i=1,2,3,4)$ ，所以， $R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1$.

故选：C.

变式 3-2. 如图，5 个 (x,y) 数据，去掉 $D(3,10)$ 后，下列说法错误的是（ ）



- A. 与 y 的相关性变强
- B. 残差平方和变大
- C. 相关指数 R^2 变大
- D. 解释变量与预报变量 y 的相关性变强

【答案】B

【分析】

在去掉点 $D(3,10)$ 后，结合散点图可对各选项的正误进行判断.

【详解】

由散点图知，去掉 $D(3,10)$ 后， y 与 x 的线性相关性加强，A 正确；

残差平方和变小，B 错误；

相关系数 r 变大，相关指数 R^2 变大，C 正确；

解释变量与预报变量 y 的相关性变强，D 正确.

故选：B.

变式 3-3. 某种产品的广告支出费用（单位：万元）与销售量 y （单位：万件）之间的对应数据如下表所示：根据表中的数据可得回归直线方程 $\hat{y}=2.27x-1.08$ ， $R^2 \approx 0.96$ ，以下说法正确的是（ ）

广告支出费用	2.2	2.6	4.0	5.3	5.9
销售量 y	3.8	5.4	7.0	11.6	12.2

- A. 第三个样本点对应的残差 $\hat{e}_3 = -1$ ，回归模型的拟合效果一般
- B. 第三个样本点对应的残差 $\hat{e}_3 = 1$ ，回归模型的拟合效果较好

C. 销售量 y 的多少有 96%是由广告支出费用引起的

D. 销售量 y 的多少有 4%是由广告支出费用引起的

【答案】C

【分析】

利用回归模型可计算残差，结合相关指数可判断 AB 的正误，根据相关指数的意义可判断 CD 的正误.

【详解】

$x=4$ 时，残差为 $7-2.27\times 4+1.08=-1$ ，而 $R^2\approx 0.96$ ，故拟合效果较好，故 AB 错误.

因为 $R^2\approx 0.96$ ，故销售量 y 的多少有 96%是由广告支出费用引起的，故 C 正确，D 错误.

故选：C.

考点四 线性回归方程的应用

典例 4. 某科技公司研发了一项新产品 A，经过市场调研，对公司 1 月份至 6 月份销售量及销售单价进行统计，销售单价（千元）和销售量 y （千件）之间的一组数据如下表所示：

月份 <i>i</i>	1	2	3	4	5	6
销售单价	9	9.5	10	10.5	11	8
销售量	11	10	8	6	5	15

(1) 试根据 1 至 5 月份的数据，建立 y 关于的回归直线方程；

(2) 若由回归直线方程得到的估计数据与剩下的检验数据的误差不超过 0.65 千元，则认为所得到的回归直线方程是理想的，试问 (1) 中所得到的回归直线方程是否理想？

参考公式：回归直线方程 $\hat{y} = \hat{b}x + \hat{a}$ ，其中 $\hat{b} = \frac{\sum_{i=1}^n x_i y_i - n \cdot \bar{x} \cdot \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2}$.

参考数据： $\sum_{i=1}^5 x_i y_i = 392$ ， $\sum_{i=1}^5 x_i^2 = 502.5$.

【答案】(1) $\hat{y} = -3.2x + 40$ ；(2) 是.

【分析】

(1) 先由表中的数据求出 $\bar{x}, \bar{y}$ ，再利用已知的数据和公式求出 $\hat{b}, \hat{a}$ ，从而可求出 y 关于

的回归直线方程：

(2) 当 $x=8$ 时，求出 $\hat{y}$ 的值，再与 15 比较即可得结论

【详解】

(1) 因为 $\bar{x} = \frac{1}{5}(9+9.5+10+10.5+11) = 10$ ， $\bar{y} = \frac{1}{5}(11+10+8+6+5) = 8$ ，

所以 $\hat{b} = \frac{392 - 5 \times 10 \times 8}{502.5 - 5 \times 10^2} = -3.2$ ，

得 $\hat{a} = 8 - (-3.2) \times 10 = 40$ ，

于是 y 关于的回归直线方程为 $\hat{y} = -3.2x + 40$ ；

(2) 当 $x=8$ 时， $\hat{y} = -3.2 \times 8 + 40 = 14.4$ ，

则，

故可以认为所得到的回归直线方程是理想的.

变式 4-1. 某电脑公司有 6 名产品推销员，其工作年限与年推销金额数据如下表：

推销员编号	1	2	3	4	5
工作年限 x /年	3	5	6	7	9
推销金额 y /万元	2	3	3	4	5

(1) 求年推销金额 y 关于工作年限 x 的线性回归方程；

(2) 若第 6 名推销员的工作年限为 11 年，试估计他的年推销金额.

附：回归直线的斜率和截距的最小二乘法估计公式分别为 $\hat{b} = \frac{\sum_{i=1}^n (t_i - \bar{t})(y_i - \bar{y})}{\sum_{i=1}^n (t_i - \bar{t})^2}$ ， $\hat{a} = \bar{y} - b\bar{t}$.

【答案】(1) $\hat{y} = 0.5x + 0.4$ ；(2) 5.9 万元.

【分析】

(1) 根据表中的数据求出 $\bar{y}$ ，再利用公式可求出 $\hat{b}$ ， $\hat{a}$ ，从而可求出推销金额 y 关于工作年限 x 的线性回归方程；

(2) 将 $x=11$ 化入回归方程中求解即可

【详解】

解（1）设所求的线性回归方程为 $\hat{y} = \hat{b}x + \hat{a}$ ，

$$\bar{x} = \frac{1}{5}(3+5+6+7+9) = 6, \quad \bar{y} = \frac{1}{5}(2+3+3+4+5) = 3.4,$$

$$\text{所以 } \hat{b} = \frac{\sum_{i=1}^5 (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^5 (x_i - \bar{x})^2} = \frac{10}{20} = 0.5,$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 0.4.$$

所以年推销金额 y 关于工作年限 x 的线性回归方程为 $\hat{y} = 0.5x + 0.4$ 。

（2）当 $x=11$ 时， $\hat{y} = 0.5x + 0.4 = 0.5 \times 11 + 0.4 = 5.9$ （万元）。

所以可以估计第 6 名推销员的年推销金额为 5.9 万元

变式 4-2. 如图是某地 2014 年至 2020 年生活垃圾无害化处理量（单位：万吨）的折线图。

注：年份代码 1~7 分别对应年份 2014~2020.

（1）由折线图看出，可用线性回归模型拟合 y 与 t 的关系，请用相关系数加以证明；

（2）建立 y 关于 t 的回归方程（系数精确到 0.01），预测 2022 年某地生活垃圾无害化处理量。

附注：

参考数据： $\sum_{i=1}^7 y_i = 9.32$ ， $\sum_{i=1}^7 t_i y_i = 40.17$ ， $\sqrt{\sum_{i=1}^7 (y_i - \bar{y})^2} = 0.55$ ， $\sqrt{7} \approx 2.646$ 。

参考公式：相关系数 $r = \frac{\sum_{i=1}^n (t_i - \bar{t})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (t_i - \bar{t})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$ ，回归方程 $\hat{y} = \hat{a} + \hat{b}t$ 中斜率和截距的最小二乘法估计

公式分别为 $\hat{b} = \frac{\sum_{i=1}^n (t_i - \bar{t})(y_i - \bar{y})}{\sum_{i=1}^n (t_i - \bar{t})^2}$ ， $\hat{a} = \bar{y} - \hat{b}\bar{t}$ 。

【答案】

（1）存在较强的正相关关系，理由见解析

（2） $\hat{y} = 0.10t + 0.92$ ，1.82 万吨

【分析】

(1)、结合参考数据及参考公式 $r = \frac{\sum_{i=1}^n (t_i - \bar{t})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (t_i - \bar{t})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$ 求出相关系数，进而可以得出结论；

(2)、根据参考公式求出回归直线方程，进而可以根据回归直线方程进行数据统计。

(1)

由折线图看出， y 与 t 之间存在较强的正相关关系，理由如下：

$$Q \sum_{i=1}^7 y_i = 9.32, \quad \sum_{i=1}^7 t_i y_i = 40.17, \quad \sqrt{\sum_{i=1}^7 (y_i - \bar{y})^2} = 0.55, \quad \bar{t} = \frac{1+2+3+4+5+6+7}{7} = 4,$$

$$\therefore r = \frac{\sum_{i=1}^7 (t_i - \bar{t})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^7 (t_i - \bar{t})^2 \sum_{i=1}^7 (y_i - \bar{y})^2}} = \frac{\sum_{i=1}^7 t_i y_i - 7\bar{t}\bar{y}}{\sqrt{\sum_{i=1}^7 (t_i - \bar{t})^2 \sum_{i=1}^7 (y_i - \bar{y})^2}} \approx \frac{40.17 - 4 \times 9.32}{2\sqrt{7} \times 0.55} \approx 0.993.$$

$Q 0.993 > 0.75$ ，故 y 与 t 之间存在较强的正相关关系。

(2)

$$\text{由 (1) 结合题中数据可得 } \hat{b} = \frac{\sum_{i=1}^7 (t_i - \bar{t})(y_i - \bar{y})}{\sum_{i=1}^7 (t_i - \bar{t})^2} = \frac{\sum_{i=1}^7 t_i y_i - 7\bar{t}\bar{y}}{\sum_{i=1}^7 t_i^2 - 7\bar{t}^2} \approx \frac{2.89}{28} \approx 0.103,$$

$$\hat{a} = \bar{y} - \hat{b}\bar{t} \approx 1.331 - 0.103 \times 4 \approx 0.92,$$

$\therefore y$ 关于 t 的回归方程 $\hat{y} = 0.10t + 0.92$ ，2022 年对应的 t 值为 9，故 $\hat{y} = 0.10 \times 9 + 0.92 = 1.82$ ，

预测 2022 年该地生活垃圾无害化处理量为 1.82 万吨。

变式 4-3. 某公司为了预测下月产品销售情况，找出了近 7 个月的产品销售量 y （单位：万件）的统计表：

月份代码 t	1	2	3	4	5	6	7
销售量 y / 万件		y_2			y_5	y_6	

但其中数据污损不清，经查证 $\sum_{i=1}^7 y_i = 9.32, \sum_{i=1}^7 t_i y_i = 40.17, \sqrt{\sum_{i=1}^7 (y_i - \bar{y})^2} = 0.55$ 。

(1) 请用相关系数 r 说明销售量 y 与月份代码 t 之间有很强线性相关关系（当 $|r| > 0.75$ 时认为两个变量有很强线性相关关系）；

(2) 求 y 关于 t 的回归直线方程 (结果中 $\hat{b}$ 保留两位小数);

(3) 公司经营期间的广告宣传费 (单位: 万元) $x_i = \sqrt{t_i}$ ($i=1,2,\dots,7$), 每件产品的销售价为 10 元, 预测第 8 个月的毛利润能否突破 15 万元, 请说明理由 (毛利润=销售金额-广告宣传费).

参考数据: $\sqrt{7} \approx 2.65$, $\sqrt{2} \approx 1.41$.

【答案】(1) 答案见解析; (2) $\hat{y} = 0.10t + 0.93$; (3) 不能, 理由见解析

【分析】

(1) 由已知数据利用相关系数公式求得 r , 由 $r > 0.75$, 可知销售量 y 与月份代码 t 有很强的线性相关系数.

(2) 求出 $\bar{y}$ 与 $\hat{b}$ 的值, 就可以得到 y 关于 t 的回归方程;

(3) 在 (2) 的线性回归方程中, 取 $t=8$, 求得 y , 进一步得到第 8 个月份的毛利润, 与 15 万元比较大小可得结论.

【详解】

由题意, 得, $\sum_{i=1}^7 (t_i - \bar{t}) = 28$,

$$r = \frac{2.89}{2\sqrt{7} \times 0.55} \approx \frac{2.89}{2 \times 2.65 \times 0.55} \approx 0.99.$$

$$|r| = 0.99 > 0.75,$$

销售量 y 与月份代码 t 之间有很强的线性相关关系。

$$(2) \quad \bar{y} = \frac{9.32}{7} \approx 1.33, \quad \hat{b} = \frac{\sum_{i=1}^7 (t_i - \bar{t})(y_i - \bar{y})}{\sum_{i=1}^7 (t_i - \bar{t})^2} = \frac{2.89}{28} \approx 0.10$$

$$\hat{a} = \bar{y} - \hat{b}\bar{t} = 1.33 - 0.10 \times 4 = 0.93,$$

y 关于 t 的回归直线方程为 $\hat{y} = 0.10t + 0.93$.

(3) 当 $t=8$ 时, $\hat{y} = 0.10 \times 8 + 0.93 = 1.73$, 而,

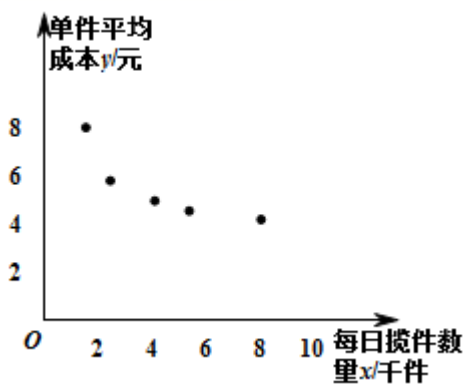
第 8 个月的毛利润约为 14.48 万元

又 $14.48 < 15$,

第 8 个月的毛利润不能突破 15 万元

考点五 非线性回归方程的应用

典例 5. 近几年，快递业的迅速发展导致行业内竞争日趋激烈.某快递网点需了解一天中收发一件快递的平均成本 y (单位：元)与当天揽收的快递件数 x (单位：千件)之间的关系，对该网点近 5 天的每日揽件量(单位：千件)与当日收发一件快递的平均成本(单位：元)($i=1, 2, 3, 4, 5$)数据进行了初步处理，得到下面的散点图及一些统计量的值.



	$\bar{y}$		$\sum_{i=1}^5 (x_i - \bar{x})(y_i - \bar{y})$	$\sum_{i=1}^5 (w_i - \bar{w})(y_i - \bar{y})$	$\sum_{i=1}^5 (x_i - \bar{x})^2$	$\sum_{i=1}^5 (w_i - \bar{w})^2$
4	5.16	0.415	-13.2	2.028	30	0.507

表中， $\bar{w} = \frac{1}{5} \sum_{i=1}^5 w_i$.

(1) 根据散点图判断，与 $y = c + \frac{d}{x}$ 哪一个适宜作为 y 关于 x 的回归方程类型？并根据判断结果及表中数据求出 y 关于 x 的回归方程；

(2) 各快递业为提高快递揽收量并实现总利润的增长，除了提升服务质量、提高时效保障外，价格优惠也是重要策略之一.已知该网点每天揽收快递的件数 x (单位：千件)与单件快递的平均价格 t (单位：元)之间的关系是 $x = 25 - 2t$ ($5 \leq t \leq 12$)，收发一件快递的利润等于单件的平均价格减去平均成本，根据 (1) 中建立的回归方程解决以下问题：

- ①预测该网点某天揽收 2000 件快递可获得的总利润；
- ②单件快递的平均价格 t 为何值时，该网点一天内收发快递所获利润的预报值最大？

附：对于一组数据 $(u_1, v_1), (u_2, v_2), \dots, (u_n, v_n)$ ，其回归直线 $\hat{v} = \alpha + \beta u$

的斜率和截距的最小二乘估计分别为 $\hat{b} = \frac{\sum_{i=1}^n (u_i - \bar{u})(v_i - \bar{v})}{\sum_{i=1}^n (u_i - \bar{u})^2}$, $\hat{a} = \bar{v} - \hat{b}\bar{u}$.

【答案】(1) $y = c + \frac{d}{x}$ 适宜作为 y 关于 x 的回归方程类型, 回归方程为 $\hat{y} = \frac{4}{x} + 3.5$; (2) ①总利润约为 12000 元; ②平均价格 t 为 8 元.

【分析】

(1) 点不在一条直线的近旁, 但与双曲线类似, 可得回归曲线类型. 令 $w = \frac{1}{x}$, 根据已知数据求得回归方程 $y = c + dw$, 即可得结论.

(2) ①利用 (1) 的结论求出利润函数, 令 $x = 2$ 可得估计利润值; ②由二次函数性质可得.

【详解】

解: (1) $y = c + \frac{d}{x}$ 适宜作为 y 关于 x 的回归方程类型.

$$\text{令 } \frac{1}{x} = w, \text{ 则 } y = dw + c, \quad \hat{d} = \frac{2.028}{0.507} = 4,$$

$$\hat{c} = \bar{y} - \hat{d} \cdot \bar{w} = 5.16 - 4 \times 0.415 = 3.5,$$

$$\therefore \hat{y} = 4w + 3.5, \text{ 即所求回归方程为 } \hat{y} = \frac{4}{x} + 3.5;$$

$$(2) \text{ 设收发 } x \text{ 千件快递获利 } z \text{ 千元, 则 } z = (t - y)x = \left(\frac{25 - x}{2} - \frac{4}{x} - 3.5 \right)x = 9x - \frac{1}{2}x^2 - 4, \quad x \in [1, 15],$$

①当 $x = 2$ 时, $z = 12$, 故该网点某天揽收 2000 件快递可获得的总利润约为 12000 元;

② $z = -\frac{1}{2}(x - 9)^2 + \frac{73}{2}$, $\therefore$ 当 $x = 9$ 即 $t = 8$ 时, z 取最大值, 故单件快递的平均价格 t 为 8 元时, 该网点一天内收发快递所获利润的预报值最大.

变式 5-1. 如图是某市 2011 年至 2020 年当年在售二手房均价 (单位: 千元/平方米) 的散点图 (图中年份代码 1~10 分别对应 2011 年~2020 年). 现根据散点图选择用和 $y = e^{c+dx}$ 两个模型对年份代码和房价 y 的关系进行拟合, 经过数据处理得到两个模型对应回归方程的相关指数 R^2 和一些统计量的值, 如下表:

模型		$y = e^{c+dx}$
相关指数 R^2	0.8821	0.9046

$\bar{y}$	$\bar{w}$	$\sum_{i=1}^{10} (x_i - \bar{x})^2$	$\sum_{i=1}^{10} (x_i - \bar{x})(y_i - \bar{y})$	$\sum_{i=1}^{10} (x_i - \bar{x})(w_i - \bar{w})$
6.81	1.89	82.5	44.55	6.6

表中 $w_i = \ln y_i$, $\bar{w} = \frac{1}{10} \sum_{i=1}^{10} w_i$.

(1) 请利用相关指数 R^2 判断: 哪个模型的拟合效果更好; 并求出该模型对应的回归方程 (参数估计值精确到 0.01);

(2) 根据 (1) 得到的方程预计: 到哪一年, 该市的当年在售二手房均价能超过 10.5 千元/平方米.

参考公式: 对于一组数据 $(u_1, v_1), (u_2, v_2), \dots, (u_n, v_n)$, 其回归线 $v = \alpha + \beta u$ 的斜率和截距的最小二

乘估计分别为: $\hat{\beta} = \frac{\sum_{i=1}^n (u_i - \bar{u})(v_i - \bar{v})}{\sum_{i=1}^n (u_i - \bar{u})^2}$, 参考数据: $e^{2.35} \approx 10.49$, $e^{2.36} \approx 10.59$.

【答案】

(1) 模型 $y = e^{c+dx}$ 的拟合效果更好, $\hat{y} = e^{1.45+0.08x}$

(2) 到 2022 年, 该市的当年在售二手房均价能超过 10.5 千元/平方米

【分析】

(1) 根据相关指数的数值可知模型 $y = e^{c+dx}$ 的拟合效果更好, 从而可得 $\ln y = c + dx$, 利用最小二乘法即可求解.

(2) 由 (1) 将 $x=11, x=12$ 代入即可求解.

(1)

由相关指数 R^2 : $0.9046 > 0.8821$, 知模型 $y = e^{c+dx}$ 的拟合效果更好.

$\because y = e^{c+dx}$, $\therefore \ln y = c + dx$, 令 $w = \ln y$,

可知 w 与满足线性模型回归方程 $\hat{w} = \hat{c} + \hat{d}x$,

$\bar{x} = \frac{1}{10}(1+2+\dots+10) = 5.5$,

则 $\hat{d} = \frac{\sum_{i=1}^{10} (x_i - \bar{x})(w_i - \bar{w})}{\sum_{i=1}^{10} (x_i - \bar{x})^2} = \frac{6.60}{82.5} = 0.08$,

,

所以回归方程为 $\hat{w} = 1.45 + 0.08x$, 即 $\hat{y} = e^{1.45+0.08x}$.

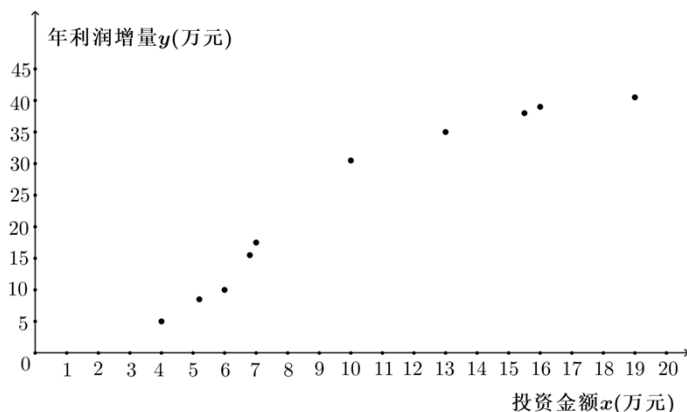
(2)

将 $x=11$ 代入, 可得 $\hat{y} = e^{2.33} < e^{2.35} < 10.5$,

将 $x=12$ 代入, 可得 $\hat{y} = e^{2.41} > e^{2.36} > 10.5$,

所以, 根据方程预计: 到 2022 年, 该市的当年在售二手房均价能超过 10.5 千元/平方米.

变式 5-2. 某投资公司 2012 年至 2021 年每年的投资金额 (单位: 万元) 与年利润增量 y (单位: 万元) 的散点图如图. 该投资公司为了预测 2022 年投资金额为 20 万元时的年利润增量, 建立了 y 关于的两个回归模型; 模型①: 由最小二乘公式可求得 y 与 x 的线性回归方程: $\hat{y} = 2.50x - 2.50$; 模型②: 由图中样本点的分布, 可以认为样本点集中在由线 $y = b \ln x + a$ 的附近, 对投资金额做换元, 令 $t = \ln x$, 则 $y = b \cdot t + a$, 且有,



(1) 根据所给的统计量, 求模型②中 y 关于的回归方程;

(2) 分别利用这两个回归模型, 预测投资金额为 20 万元时的年利润增量 (结果保留两位小数);

附: 样本 $(t_i, y_i) (i=1, 2, \dots, n)$ 的最小乘估计公式为 $\hat{b} = \frac{\sum_{i=1}^n (t_i - \bar{t})(y_i - \bar{y})}{\sum_{i=1}^n (t_i - \bar{t})^2}$, $\hat{a} = \bar{y} - \hat{b}\bar{t}$; 参考数据:

$\ln 2 \approx 0.6931, \ln 5 \approx 1.6094$.

【答案】

(1) $\hat{y} = 25 \ln x - 32$

(2) 模型①的年利润增量的预测值为 47.50 (万元), 模型②的年利润增量的预测值为 42.89 (万元)

【分析】

(1) 结合已知数据和公式求出 $\hat{a}, \hat{b}$ 这两个系数即可得回归方程;

(2) 把 $x = 20$ 代入模型①、②的回归方程, 算出 $\hat{y}$ 即可.

(1)

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/948101042007007023>