

第 02 讲 8.2 一元线性回归模型及其应用

(8.2.1 一元线性回归模型+8.2.2 一元线性回归模型参数的最小二乘法估计)

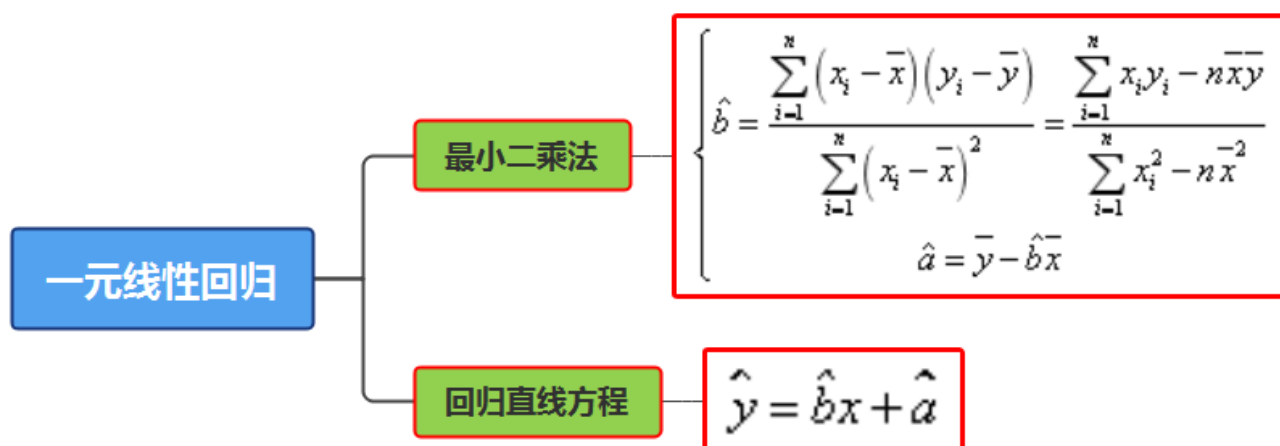
01

学习目标

课程标准	学习目标
①了解一元线性回归模型的含义，理解两个变量之间随机关系的一元线性回归模型的作用与意义。 ②了解残差在线性回归与非线性回归问题的作用及意义。 ③了解一元线性回归模型参数与最小二乘估计的推导过程，理解最小二乘估计的原理。 ④会结合题意求一元线性回归方程。 ⑤会用相关指数进行分析模型拟合的效果情况。	通过本节课的学习，要求会求一元线性回归方程，会进行残差分析，能判断回归模型的拟合效果，能利用样本数据建立统计模型并能进行预测

02

思维导图



知识点 1: 一元线性回归模型

(1) 一元线性回归模型

我们称
$$\begin{cases} Y = bx + a + e \\ E(e) = 0, D(e) = \sigma^2 \end{cases}$$

为 Y 关于 x 的一元线性回归模型, 其中 Y 称为因变量或响应变量, x 称为自变量或解释变量; a 和 b 为模型的未知参数, a 称为截距参数, b 称为斜率参数; e 是 Y 与 $bx + a$ 之间的随机误差.

(2) 随机误差

在线性回归模型 $Y = bx + a + e$ 中, a 和 b 为模型的未知参数, e 是 Y 与 $bx + a$ 之间的误差, 通常 e 为随机变量, 称为随机误差. 它的均值 $E(e) = 0$, 方差 $D(e) = \sigma^2 > 0$.

线性回归模型的完整表达式为
$$\begin{cases} Y = bx + a + e \\ E(e) = 0, D(e) = \sigma^2 \end{cases}$$
, 在此模型中, 随机误差 e 的方差 σ^2 越小, 用

$bx + a$ 预报真实值 Y 的精度越高.

知识点 2: 一元线性回归模型参数的最小二乘法

(1) 经验回归方程的求解法: 最小二乘法

回归直线方程过样本点的中心 $(\bar{x}, \bar{y})$, 是回归直线方程最常用的一个特征;

我们将 $\hat{y} = \hat{b}x + \hat{a}$ 称为 Y 关于 x 的线性回归方程, 也称经验回归函数或经验回归公式, 其图形称为经验回归直线. 这种求经验回归方程的方法叫做最小二乘法, 求得的 $\hat{b}, \hat{a}$, 叫做 b, a 的最小二乘估计, 其中 $\hat{b}$ 称为回归系数, 它实际上也就是经验回归直线的斜率, $\hat{a}$ 为截距.

其中
$$\begin{cases} \hat{b} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \\ \hat{a} = \bar{y} - \hat{b}\bar{x} \end{cases}$$

【即学即练 1】 (2024 上·全国·高三专题练习) 某校数学建模学生社团进行了一项实验研究, 采集了 x, y 的一组数据如下表所示:

x	2	3	4	5	6	7
y	52.5	45	40	30	25	17.5

该社团对上述数据进行了分析, 发现 y 与 x 之间具有线性相关关系.

(1) 画出表中数据的散点图, 并指出 y 与 x 之间的相关系数 r 是正还是负;

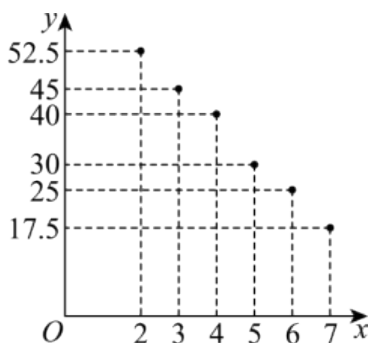
(2) 求出 y 关于 x 的线性回归方程, 并写出当 $x=9$ 时, 预测数据 y 的值.

附: 在线性回归方程 $\hat{y} = \hat{a} + \hat{b}x$ 中, $\hat{b} = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n(\bar{x})^2}$, $\hat{a} = \bar{y} - \hat{b}\bar{x}$, 其中 $\bar{x}, \bar{y}$ 为样本平均值.

【答案】(1) 散点图见解析, 负

(2) $\hat{y} = 66.5 - 7x$, $y = 3.5$

【详解】(1) 由题意得散点图如图所示:



由图可知 y 与 x 之间成负相关关系, 所以 r 是负.

(2) 因为 $\bar{x} = \frac{2+3+4+5+6+7}{6} = 4.5$, $\bar{y} = \frac{52.5+45+40+30+25+17.5}{6} = 35$,

$$\sum_{i=1}^6 x_i y_i = 882.5, \quad \sum_{i=1}^6 x_i^2 = 139,$$

$$\text{所以 } \hat{b} = \frac{\sum_{i=1}^6 x_i y_i - 6\bar{x}\bar{y}}{\sum_{i=1}^6 x_i^2 - 6\bar{x}^2} = \frac{882.5 - 6 \times 4.5 \times 35}{139 - 6 \times 4.5^2} = -7, \quad \hat{a} = \bar{y} - \hat{b}\bar{x} = 35 - (-7) \times 4.5 = 66.5,$$

$\therefore$ 关于 x 线性回归方程为 $\hat{y} = 66.5 - 7x$,

所以当 $x=9$ 时, $y = 3.5$.

(2) 求经验回归方程的步骤

① 作出散点图, 判断两变量是否具有线性相关关系, 若具有线性相关关系, 则可求其经验回归方程;

② 列表求出 $\bar{x}, \bar{y}$ 的值;

③ 利用公式先计算 $\hat{b}$, 再根据经验回归直线过样本点的中心 $(\bar{x}, \bar{y})$ 计算 $\hat{a}$;

④ 写出经验回归方程 $\hat{y} = \hat{b}x + \hat{a}$.

求经验回归方程, 关键在于正确求出系数 $\hat{b}, \hat{a}$, 由于计算量较大, 所以计算时要仔细谨慎、分层进行, 避免因计算产生错误要特别注意, 只有两个变量呈线性相关关系时, 求出的经验回归方程才有意义.

(3) 经验回归方程的性质

① 经验回归直线一定过点 $(\bar{x}, \bar{y})$, 点 $(\bar{x}, \bar{y})$ 通常称为样本点的中心;

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad R^2 \text{ 越大, 即拟合效果越好, } R^2 \text{ 越小, 模型拟合效果越差.}$$

【即学即练 3】 (2023 下·青海西宁·高二校考阶段练习) 甲、乙、丙、丁四位同学在建立变量 x, y 的回归模型时, 分别选择了 4 种不同模型, 计算可得它们的相关指数 R^2 分别如下表:

	甲	乙	丙	丁
R^2	0.98	0.78	0.50	0.85

建立的回归模型拟合效果最好的同学是_____.

【答案】 选甲 相关指数 R^2 越大, 表示回归模型拟合效果越好.

【详解】 相关指数 R^2 越大, 相关性越强, 回归模型拟合效果越好, 所以效果最好的是甲.

(3) 决定系数 R^2 与相关系数 r 的联系与区别

- ① 相关系数 r 反映两个变量的相关关系的强弱及正相关或负相关, 决定系数 R^2 反映回归模型的拟合效果.
- ② 在含有一个解释变量的线性模型中, 决定系数 R^2 的数值是相关系数 r 的平方, 其变化范围为 $[0, 1]$, 而相关系数的变化范围为 $[-1, 1]$.
- ③ 当相关系数 $|r|$ 接近于 1 时, 说明两变量的相关性较强, 当 $|r|$ 接近于 0 时, 说明两变量的相关性较弱; 而当 R^2 接近于 1 时, 说明经验回归方程的拟合效果较好.

04

题型精讲

题型 01 由散点图判断是否线性相关

【典例 1】 (2023 下·河南南阳·高二唐河县第一高级中学校考阶段练习) 2003 年春季, 我国部分地区 SARS 流行, 党和政府采取果断措施, 防治结合, 很快使病情得到控制, 下表是某同学记载的 5 月 1 日至 5 月 12 日每天北京市 SARS 治愈者数据, 以及根据这些数据绘制出的散点图

日期	5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8	5.9	5.10	5.11	5.12
人数	100	109	115	118	121	134	141	152	168	175	186	203

下列说法: ① 根据此散点图, 可以判断日期与人数具有线性相关关系; ② 根据此散点图, 可以判断日期与人数具有一次函数关系. 其中正确的个数为 ()

- A. 0 个 B. 1 个 C. 2 个 D. 以上都不对

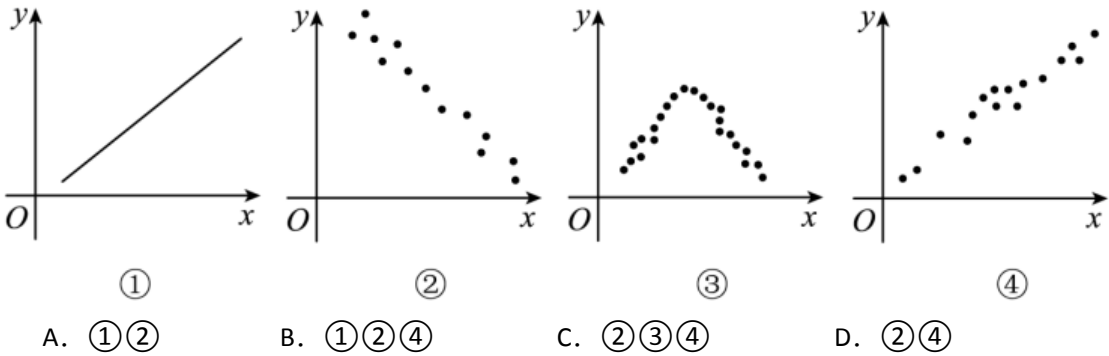
【典例 2】 (2023·全国·高二专题练习) 某个男孩的年龄与身高的统计数据如下表所示:

年龄 x (岁)	1	2	3	4	5	6
------------	---	---	---	---	---	---

身高 y (cm)	78	87	98	108	115	120
-------------	----	----	----	-----	-----	-----

- (1)画出散点图；
- (2)判断 y 与 x 是否具有线性相关关系，如果相关，是正相关还是负相关.

【变式 1】（2023 下·高二课时练习）下列四个图中，两个变量 x, y 具有线性相关关系的是（ ）



题型 02 求回归直线方程

【典例 1】（2024 上·江西赣州·高二统考期末）大气污染物 $\text{PM}_{2.5}$ （直径不大于 $2.5\mu\text{m}$ 的颗粒物）的浓度超过一定限度会影响人的身体健康. 为研究 $\text{PM}_{2.5}$ 浓度 y （单位： $\mu\text{g}/\text{m}^3$ ）与汽车流量 x （单位：千辆）的线性关系，研究人员选定了 10 个城市，在每个城市建立交通监测点，统计了 24h 内过往的汽车流量以及同时段空气中的 $\text{PM}_{2.5}$ 浓度，得到如下数据：

城市编号	1	2	3	4	5	6	7	8	9	10	总和
x	1.300	1.444	0.786	1.652	1.756	1.754	1.200	1.500	1.200	0.908	13.5
y	66	76	21	170	156	120	72	120	100	129	1030

并计算得 $\sum_{i=1}^{10} x_i^2 = 19.225$ ， $\sum_{i=1}^{10} y_i^2 = 124434$ ， $\sum_{i=1}^{10} x_i y_i = 1479.5$.

- (1)求变量 y 关于 x 的线性回归方程；
- (2)根据 24h 内 $\text{PM}_{2.5}$ 浓度确定空气质量等级， $\text{PM}_{2.5}$ 浓度在 $0\sim 35\mu\text{g}/\text{m}^3$ 为优， $35\sim 75\mu\text{g}/\text{m}^3$ 为良， $75\sim 115\mu\text{g}/\text{m}^3$ 为轻度污染， $115\sim 150\mu\text{g}/\text{m}^3$ 为中度污染， $150\sim 250\mu\text{g}/\text{m}^3$ 为重度污染，已知某城市 24h

内过往的汽车流量为 1360 辆，判断该城市的空气质量等级.

参考公式：线性回归方程为 $\hat{y} = \hat{b}x + \hat{a}$ ，其中以 $\hat{b} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$.

【典例 2】（2024 上·重庆沙坪坝·高三重庆八中校考阶段练习）在入室盗窃类案件中，出现频率最高的痕迹物证之一就是足迹．负重行走对足迹步伐特征影响的规律强，而且较为稳定．正在行走的人在负重的同时，步长变短，步宽变大，步角变大．因此，以身高分别为 170cm，175cm，180cm 的人员各 20 名作为实验对象，让他们采取双手胸前持重物的负重方式行走，得到实验对象在负重 0kg，5kg，10kg，15kg，20kg 状态下相对稳定的步长数据平均值．并在不同身高情况下，建立足迹步长 s (单位: cm)关于负重 x (单位: kg)的三个经验回归方程．根据身高 170cm 组数据建立线性回归方程①: $\hat{s}_1 = -0.402x + 55.728$ ；根据身高 175cm 组数据建立线性回归方程②: $\hat{s}_2 = -0.497x + 64.404$ ；根据身高 180cm 组数据建立线性回归方程③: $\hat{s}_3 = \hat{b}x + \hat{a}$ ．

(1)根据身高 180cm 组的统计数据，求 $\hat{a}$ ， $\hat{b}$ 的值，并解释参数 $\hat{b}$ 的含义；

身高 180cm 不同负重情况下的步长数据平均值					
负重 x/kg	0	5	10	15	20
足迹步长 s/cm	74.35	73.50	71.80	68.60	65.75

(2)在一起盗窃案中，被盗窃物品重为 9kg，在现场勘查过程中，测量得犯罪嫌疑人往返时足迹步长的差值为 4.464cm，推测该名嫌疑人的身高，并说明理由．

附: $\hat{s} = \hat{b}x + \hat{a}$ 为回归方程, $\hat{b} = \frac{\sum_{i=1}^n x_i s_i - n\bar{x} \cdot \bar{s}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$, $\hat{a} = \bar{s} - \hat{b}\bar{x}$, $\sum_{i=1}^5 s_i = 354$, $\sum_{i=1}^5 x_i s_i = 3429.5$.

【典例 3】（2024 上·全国·高三专题练习）某种产品的广告费支出 x （单位：万元）与销售额 y （万元）之间有如下的一组数据：

广告费支出 x	2	4	5	6	8
销售额 y	30	40	60	50	70

(1) 求出样本点中心

(2) 求回归直线方程（其中 $\hat{b} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x} \cdot \bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$ ， $\hat{a} = \bar{y} - \hat{b}\bar{x}$ ）

【变式 1】（2024 上·黑龙江牡丹江·高三牡丹江市第二高级中学校联考期末）近期，一些地方中小学生“课间 10 分钟”问题受到社会广泛关注，国家号召中小学要增加学生的室外活动时间。但是进入 12 月后，天气渐冷，很多学生因气温低而减少了外出活动次数。为了解本班情况，一位同学统计了一周（5 天）的气温变化和某一固定课间该班级的学生出楼人数，得到如下数据：

温度 x_i （零下 $^{\circ}\text{C}$ ）	7	10	11	15	17
出楼人数 y_i	20	16	17	10	7

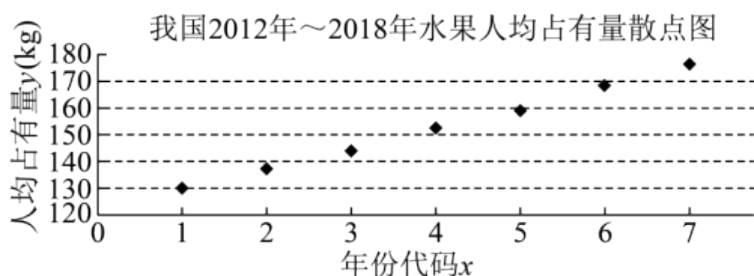
(1) 利用最小二乘法，求变量 x, y 之间的线性回归方程；

附：用最小二乘法求线性回归方程 $\hat{y} = \hat{b}x + \hat{a}$ 的系数： $\hat{b} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$ $\hat{a} = \bar{y} - \hat{b}\bar{x}$

(2) 预测当温度为 -8°C 时，该班级在本节课间的出楼人数（人数：四舍五入取整数）。

(3) 为了号召学生能够增加室外活动时间，学校举行拔河比赛，采取 3 局 2 胜制（无平局）。在甲、乙两班的较量中，甲班每局获胜的概率均为 $\frac{3}{5}$ ，设随机变量 X 表示甲班获胜的局数，求 X 的分布列和期望。

【变式 2】（2024 上·全国·高三专题练习）下面给出了根据我国 2012 年：2018 年水果人均占有量 y （单位：kg）和年份代码 x 绘制的散点图（2012 年：2018 年的年份代码 x 分别为 1~7）。



(1) 根据散点图分析 y 与 x 之间的相关关系；

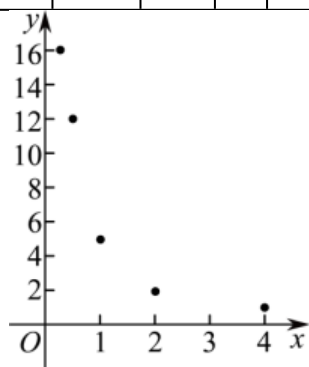
(2) 根据散点图相应数据计算得 $\sum_{i=1}^7 y_i = 1074$, $\sum_{i=1}^7 x_i y_i = 4517$, 求 y 关于 x 的线性回归方程. (精确到 0.01)

附：回归方程 $\hat{y} = \hat{b}x + \hat{a}$ 中斜率和截距的最小二乘估计公式分别为：

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

【变式 3】（2024·全国·高三专题练习）在一次抽样调查中测得 5 个样本点，得到下表及散点图.

x	0.25	0.5	1	2	4
y	16	12	5	2	1



(1) 根据散点图判断 $y = a + bx$ 与 $y = c + k \cdot x^{-1}$ 哪一个适宜作为 y 关于 x 的回归方程；

(给出判断即可，不必说明理由)

(2) 根据 (1) 的判断结果试建立 y 与 x 的回归方程；(计算结果保留整数)

参考公式： $b = \frac{\sum_{i=1}^n x_i y_i - n \cdot \bar{x} \cdot \bar{y}}{\sum_{i=1}^n x_i^2 - n \cdot \bar{x}^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \hat{a} = \bar{y} - \hat{b}\bar{x}$

题型 03 求样本中心(根据样本中心求参数)

【典例 1】(2024 上·全国·高三专题练习) 具有线性相关关系的变量 x, y 的一组数据如下:

x	0	1	2	3
y	-5	-4.5	-4.2	-3.5

其线性回归直线方程为 $\hat{y} = \hat{b}x + \hat{a}$, 则回归直线经过 ()

- A. 第一、二、三象限 B. 第二、三、四象限
C. 第一、二、四象限 D. 第一、三、四象限

【典例 2】(2024 下·全国·高二随堂练习) 已知 x, y 取表中的数值, 若 x, y 具有线性相关关系, 线性回归方程为 $\hat{y} = 0.95x + 2.6$, 则 $a =$ ()

x	0	1	3	4
y	a	4.3	4.8	6.7

- A. 2.2 B. 2.4 C. 2.5 D. 2.6

【典例 3】(2024 下·全国·高二随堂练习) 某公司一种型号的产品近期销售情况如表:

月份 x	2	3	4	5	6
销售额 y (万元)	15.1	16.3	17.0	17.2	18.4

根据上表可得到回归直线方程 $\hat{y} = 0.75x + \hat{a}$, 据此估计, 该公司 7 月份这种型号产品的销售额为 ()

- A. 18.85 万元 B. 19.3 万元 C. 19.25 万元 D. 19.05 万元

【典例 4】(多选) (2024 上·浙江宁波·高三统考期末) 某电商平台为了对某一产品进行合理定价, 采用不同的单价在平台试销, 得到的数据如下表所示:

单价 x /元	8	8.5	9	9.5	10
销量 y /万件	89	85	80	78	68

根据以上数据得到 y 与 x 具有较强的线性关系, 若用最小二乘估计得到经验回归方程为 $\hat{y} = -19.8x + \hat{a}$, 则 ()

- A. 相关系数 $r > 0$ B. 点 (9, 80) 一定在经验回归直线上
C. $\hat{a} = 258.2$ D. $x = 9.5$ 时, 对应销量的残差为 -7.9

【变式 1】(2024 上·四川绵阳·高二绵阳南山中学实验学校校考期末) 已知 x 与 y 之间的一组数据:

x	0	1	2	3
y	1	3	5	7

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/976041221135010134>