

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 $r < 0$.

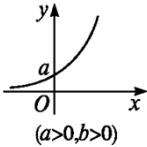
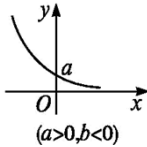
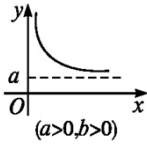
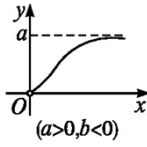
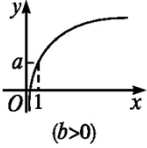
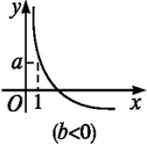
(2) $|r|$ 越小, 说明两个变量之间的线性相关性越弱, 也就是得出的回归直线方程越没有价值, 即方程越不能反映真实的情况; $|r|$ 越大, 说明两个变量之间的线性相关性越强, 也就是得出的回归直线方程越有价值.

(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

变式 1-1. 在对两个变量 x, y 进行回归分析时有下列步骤：

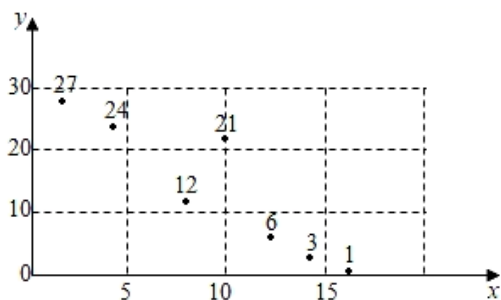
- ①对所求出的回归方程作出解释；②收集数据 $(x_i, y_i), i=1, 2, \dots, n$ ；
- ③求回归方程；④根据所收集的数据绘制散点图.

则下列操作顺序正确的是（ ）

- A. ①②④③
- B. ③②④①
- C. ②③①④
- D. ②④③①

变式 1-2. 已知两个变量和 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



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

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. $r_2 < r_4 < 0 < r_1 < r_3$

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 ()

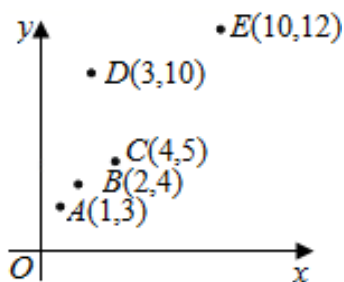
- A. 54.55 B. 2.45 C. 3.45 D. 111.55

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

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

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

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	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： $\sum_{i=1}^5 x_i y_i = 392$ ， $\sum_{i=1}^5 x_i^2 = 502.5$ 。

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

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

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	

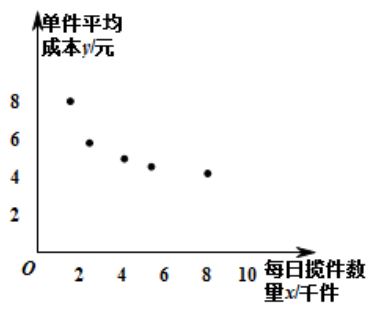
但其中数据污损不清，经查证 $\sum_{i=1}^7 y_i = 9.32, \sum_{i=1}^7 t_i y_i = 40.17, \sqrt{\sum_{i=1}^7 (y_i - \bar{y})^2} = 0.55$.

- （1）请用相关系数 r 说明销售量 y 与月份代码 t 之间有很强的线性相关关系（当 $|r| > 0.75$ 时认为两个变量有很强的线性相关关系）；
- （2）求 y 关于 t 的回归直线方程（结果中 $\$$ 保留两位小数）；
- （3）公司经营期间的广告宣传费（单位：万元） $x_i = \sqrt{t_i}$ （ $i = 1, 2, \text{L}$ ），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： $\sqrt{7} \approx 2.65$ ， $\sqrt{2} \approx 1.41$.

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 $\hat{\beta} = \frac{\sum_{i=1}^n (u_i - \bar{u})(v_i - \bar{v})}{\sum_{i=1}^n (u_i - \bar{u})^2}$, $\hat{\alpha} = \bar{v} - \hat{\beta} \bar{u}$.

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.

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/958101075007007023>