

第九章 统计与统计模型

9.2.1 统计模型（题型战法）

知识梳理

一 线性回归方程及其应用

1. 相关系数:
$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum_{i=1}^n x_i^2 - n\bar{x}^2)(\sum_{i=1}^n y_i^2 - n\bar{y}^2)}}$$

用它来衡量两个变量之间的线性相关程度。

(1) 当 $|r| \leq 1$ ，当 $r > 0$ 时，变量 y 与 x 正相关，当 $r < 0$ 时，变量 y 与 x 负相关。

(2) $|r| \leq 1$ ，且 $|r|$ 越接近 1，相关程度越强；且 $|r|$ 越接近 0，相关程度越弱，几乎不存在。

2. 求线性回归方程的步骤：

(1) 利用散点图或进行相关性检验判断两个变量具有线性相关关系；

(2) 列表求出 $\bar{x}, \bar{y}$ 等。

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

二 独立性检验在实际中的应用

解题步骤：(1) 独立性检验原理只能解决两个对象，且每个对象有两类属性的问题，所以对于一个实际问题，我们要明确能否用独立性检验的思想加以解决；

(2) 如果确实属于这类问题，要科学的抽取样本，样本容量要适当，不可太小；

(3) 根据数据列出 2×2 列联表；

(4) 提出假设 H_0 ：所研究的两类对象无关；

(5) 根据公式计算 $K^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$ 的值；

(6) 比较观测值 k

与临界值表中相应的检验水平，根据小概率原理肯定或否定假设，判断是否相关。

题型战法

题型战法一 相关系数与误差分析

典例 1. 对两个变量 y 与 x 进行回归分析，分别选择不同的模型，它们的相关系数 r 如下，其中拟合效果最好的模型是 ()

- ①模型Ⅰ的相关系数 r 为 0.25； ②模型Ⅱ的相关系数 r 为 0.80；
③模型Ⅲ的相关系数 r 为 -0.50； ④模型Ⅳ的相关系数 r 为 -0.90；

A. Ⅰ B. Ⅱ C. Ⅲ D. Ⅳ

【答案】D

【分析】根据相关系数的大小对相关关系强弱的判定，即可解出.

【详解】因为 $|r|$ 越趋近于 1，相关性越强，模型拟合效果越好，
所以拟合效果最好的模型是Ⅳ.

故选：D.

变式 1-1. 下列有关样本线性相关系数 r 的说法，错误的是 ()

- A. 相关系数 r 可用来衡量 x 与 y 之间的线性相关程度
B. $|r| \leq 1$ ，且 $|r|$ 越接近 0，相关程度越小
C. $|r| \leq 1$ ，且 $|r|$ 越接近 1，相关程度越大
D. $|r| \leq 1$ ，且 $|r|$ 越接近 1，相关程度越小

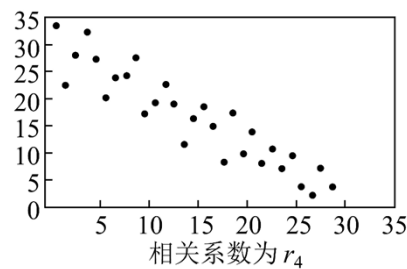
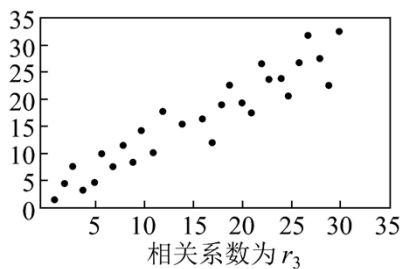
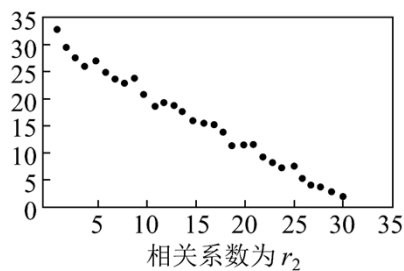
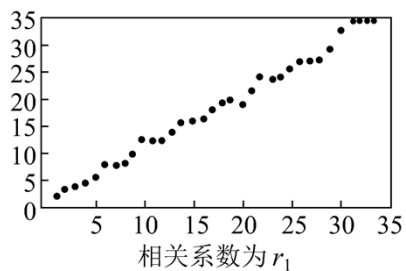
【答案】D

【分析】根据相关系数的定义，即可判断选项.

【详解】相关系数是用来衡量两个变量之间的线性相关程度的，线性相关系数是一个绝对值小于等于 1 的量，并且它的绝对值越大就说明相关程度越大，所以不正确的只有 D.

故选：D.

变式 1-2. 某统计部门对四组数据进行统计分析后，获得如图所示的散点图.



下面关于相关系数的比较，正确的是（ ）

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

【答案】C

【分析】根据散点图的分布可得相关性的强弱，即可比较大小.

【详解】由图可知： r_3, r_1 所对应的图中的散点呈现正相关，而且 r_1 对应的相关性比 r_3 对应的相关性要强，故 $0 < r_3 < r_1$ ， r_2, r_4 所对应的图中的散点呈现负相关，且根据散点的分布情况可知 $r_2 < r_4 < 0$ ，因此 $r_2 < r_4 < r_3 < r_1$ ，

故选：C

变式 1-3. 下列命题是真命题的有（ ）

- A. 经验回归方程 $\hat{y} = \hat{b}x + \hat{a}$ 至少经过其样本数据点 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ 中的一个
- B. 可以用相关系数 r 来刻画两个变量 x 和 y 线性相关程度的强弱， r 的值越小，说明两个变量线性相关程度越弱
- C. 在回归分析中，决定系数 $R^2 = 0.80$ 的模型比决定系数 $R^2 = 0.98$ 的模型拟合的效果要好
- D. 残差平方和越小的模型，拟合的效果越好

【答案】D

【分析】根据经验回归方程、相关系数、决定系数、残差等知识确定正确答案.

【详解】对于 A，经验回归方程 $\hat{y} = \hat{b}x + \hat{a}$

是由最小二乘法计算出来的，它不一定经过其样本数据点，一定经过 $(\bar{x}, \bar{y})$ ，所以 A 是假命题；

对于 B，由相关系数的意义，当 $|r|$ 越接近 1 时，表示变量 y 与 x 之间的线性相关程度越强，所以 B 是假命题；

对于 C，用决定系数 R^2 的值判断模型的拟合效果， R^2 越大，模型的拟合效果越好，所以 C 是假命题；

由残差的统计学意义知，D 为真命题.

故选： D

变式 1-4. 关于线性回归的描述，下列命题错误的是 ()

- A. 回归直线一定经过样本点的中心 B. 残差平方和越小，拟合效果越好
- C. 决定系数 R^2 越接近 1，拟合效果越好 D. 残差平方和越小，决定系数 R^2 越小

【答案】D

【分析】根据线性回归的性质判断即可

【详解】对 A，回归直线一定经过样本点的中心正确；

对 B，残差平方和越小，拟合效果越好正确；

对 C，决定系数 R^2 越接近 1，拟合效果越好正确；

对 D，残差平方和越小，拟合效果越好，决定系数 R^2 越接近 1，故 D 错误；

故选： D

题型战法二 回归分析

典例 2. 哈三中高二数学备课组对学生的记忆力 x 和判断力 y 进行统计分析，所得数据如下表所示

x	4	6	8	10
y	2	3	5	6

(1) 请根据上表提供的数据，用最小二乘法求出 y 关于 x 的线性回归方程 $\hat{y} = \hat{b}x + \hat{a}$ ；

(2) 根据 (1) 中求出的线性回归方程，预测记忆力为 9 的学生的判断力.

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

【答案】(1) $y = 0.7x - 0.9$; (2) 判断力为 5.4.

【分析】(1) 直接利用公式求解即可

(2) 把 $x = 9$ 代入回归方程中求解

【详解】解：(1) 由表中数据可得 $\bar{x} = \frac{1}{4}(4 + 6 + 8 + 10) = 7$, $\bar{y} = \frac{1}{4}(2 + 3 + 5 + 6) = 4$,

$$\sum_{i=1}^4 x_i y_i - 4\bar{x}\bar{y} = 4 \times 2 + 6 \times 3 + 8 \times 5 + 10 \times 6 - 4 \times 7 \times 4 = 14,$$

$$\sum_{i=1}^4 x_i^2 - 4\bar{x}^2 = 4^2 + 6^2 + 8^2 + 10^2 - 4 \times 7^2 = 20,$$

$$\text{所以 } \hat{b} = \frac{\sum_{i=1}^4 x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^4 x_i^2 - n\bar{x}^2} = \frac{14}{20} = 0.7,$$

$$\text{所以 } \hat{a} = \bar{y} - \hat{b}\bar{x} = 4 - 0.7 \times 7 = -0.9,$$

所以 y 关于 x 的线性回归方程为 $y = 0.7x - 0.9$,

(2) 当 $x = 9$ 时, $y = 0.7 \times 9 - 0.9 = 5.4$,

所以记忆力为 9 的学生的判断力约为 5.4

变式 2-1. 防疫抗疫, 人人有责, 随着奥密克戎的全球肆虐, 防疫形势越来越严峻, 防疫物资需求量急增. 下表是某口罩厂今年的月份 x 与订单 y (单位: 万元) 的几组对应数据:

月份 x	1	2	3	4	5
订单 y	20	24	m	43	52

(1) 求 y 关于 x 的线性回归方程, 并估计 6 月份该厂的订单数;

(2) 求相关系数 r (精确到 0.01), 说明 y 与 x 之间具有怎样的相关关系.

参考数据: $\sum_{i=1}^5 y_i = 175$, $\sum_{i=1}^5 x_i y_i = 608$, $\sum_{i=1}^5 (y_i - \bar{y})^2 = 700$. $\sqrt{7} \approx 2.646$, $\sqrt{10} \approx 3.162$. 参考公式: 相关系数

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}}; \text{ 回归直线的方程是 } y = bx + a, \text{ 其中 } \begin{cases} \hat{b} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \\ \hat{a} = \bar{y} - \hat{b}\bar{x} \end{cases}$$

【答案】(1) $y = 8.3x + 10.1$, 6 月份该厂的订单数为 59.9 万元;

(2) 0.99, y 与 x 之间具有很强的正相关关系.

【分析】(1) 求出 $\hat{b}$ 与 $\hat{a}$ 的值, 可得 y 关于 x 的线性回归方程, 取 $x=6$ 求得 $\hat{y}$ 值得答案;

(2) 由已知数据求得 r 值, 可得 y 与 x 的相关系数近似为 0.99, 故 y 与 x 之间的线性相关程度相当高.

(1)

解: 由题可得: $\bar{x} = \frac{1+2+3+4+5}{5} = 3, \bar{y} = \frac{\sum_{i=1}^5 y_i}{5} = \frac{175}{5} = 35, \sum_{i=1}^5 x_i^2 = 55$

$$\therefore \hat{b} = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} = \frac{608 - 5 \times 3 \times 35}{55 - 5 \times 3^2} = 8.3,$$

$$\hat{a} = \bar{y} - \hat{b} \bar{x} = 35 - 8.3 \times 3 = 10.1,$$

$\therefore y$ 关于 x 的线性回归方程为 $\hat{y} = 8.3x + 10.1$,

2022 年 6 月对应的变量为 6, 将 $x=6$ 代入 $\hat{y} = 8.3x + 10.1$,

$$\text{得 } \hat{y} = 8.3 \times 6 + 10.1 = 59.9,$$

$\therefore$ 估计 6 月份该厂的订单数为 59.9 万元.

(2)

$$\text{相关系数 } r = \frac{\sum_{i=1}^5 (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^5 (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^5 (y_i - \bar{y})^2}} = \frac{83}{\sqrt{10} \sqrt{700}} \approx 0.99.$$

$\therefore y$ 与 x 之间具有很强的正相关关系.

变式 2-2. 某服装企业采用服装个性化设计为客户提供服务, 即由客户提供身材的基本数据用于个人服装设计. 该企业为了设计所用的数据更精准, 随机地抽取了 10 位男子的身高和臂长的数据, 数据如下表所示:

身高 x / cm	164	165	168	172	173	176	178	181	182	191
臂长 y / cm	160	164	161	170	175	181	170	182	180	187

(1) 根据表中的数据, 求男子的身高预报臂长的线性回归方程 $\hat{y} = \hat{b}x + \hat{a}$, 并预报身高为 170cm 的男子的臂长 (男子臂长计算结果精确到 0.01);

(2) 统计学认为, 两个变量 x 、 y 的相关系数 r 的大小可表明两变量间的相关性强弱. 一般地, 如果 $|r| \in [0.75, 1]$, 那么相关性很强; 如果 $|r| \in [0.30, 0.75]$, 那么相关性一般; 如果 $|r| \in$

$[0, 0.30)$, 那么没有相关性. 求出 r 的值, 并判断变量 x, y 的相关性强弱 (结果精确到 0.01).

附: 线性回归方程 $\hat{y} = \hat{b}x + \hat{a}$ 其中 $\hat{a} = \bar{y} - \hat{b}\bar{x}$, $\hat{b} \approx 1.022$, $\sum_{i=1}^{10} x_i = 1750$, $\sum_{i=1}^{10} y_i = 1730$,

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}, \quad \sum_{i=1}^{10} (x_i - \bar{x})(y_i - \bar{y}) = 648, \quad \sqrt{634 \times 806} \approx 715, \quad \sqrt{635 \times 817} \approx 720$$

【答案】(1) $\hat{y} = 1.022x - 5.85$; 167.89cm; (2) $r \approx 0.91$; 变量 x, y 间的相关性很强.

【分析】(1) 根据表中的数据求出 $\bar{x}, \bar{y}$, 从而利用 $\hat{a} = \bar{y} - \hat{b}\bar{x}$ 可求出 $\hat{a}$, 进而可得回归方程, 然后当 $x = 170$ 时, 代入回归方程可求出身高为 170cm 的男子的臂长;

(2) 直接利用公式和已知的数据求解相关系数, 再根据所给数据判断强弱

【详解】(1) 解: $\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = 175$, $\bar{y} = \frac{\sum_{i=1}^{10} y_i}{10} = 173$

由 $\hat{b} \approx 1.022$, 得 $\hat{a} = \bar{y} - \hat{b}\bar{x} = 173 - 1.022 \times 175 = -5.85$

所以所求线性回归方程为 $\hat{y} = 1.022x - 5.85$

当 $x = 170$ 时, $\hat{y} = 1.022 \times 170 - 5.85 = 167.89$

所以身高为 170cm 的男性臂长约为 167.89cm

(2) $\sqrt{\sum_{i=1}^{10} (x_i - \bar{x})^2} = \sqrt{634}$, $\sqrt{\sum_{i=1}^{10} (y_i - \bar{y})^2} = \sqrt{806}$

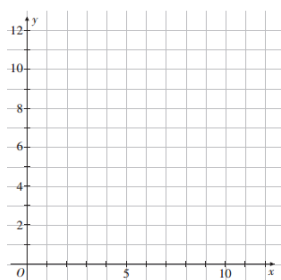
$$r = \frac{\sum_{i=1}^{10} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{10} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{10} (y_i - \bar{y})^2}} = \frac{648}{\sqrt{634} \times \sqrt{806}} \approx 0.91$$

因为 $r \in [0.75, 1]$, 所以变量 x, y 间的相关性很强.

变式 2-3. 某市为吸引大学生人才来本市就业, 大力实行人才引进计划, 提供现金补贴, 为了解政策的效果, 收集了 2011-2020 年人才引进就业人数数据 (单位: 万), 统计如下 (年份代码 1-10 分别代表 2011-2020 年) 其中 $z_i = \ln y_i$, $w_i = \ln x_i$, $e^{2.8} \approx 16.44, e^{2.92} \approx 18.54, e^{3.04} \approx 20.91$,

$\ln 11 \approx 2.40, \ln 12 \approx 2.48, \ln 13 \approx 2.56, \ln 2022 \approx 7.61$.

年份代码 x	1	2	3	4	5	6	7	8	9	10
引进人数 y	3.4	5.7	7.3	8.5	9.6	10.2	10.8	11.3	11.6	11.8



(1)根据数据画出散点图，并判断， $y=a+bx$ ， $y=e^{c+dx}$ ， $y=m+n\ln x$ 哪一个适合作为该市人才引进就业人数 y 关于年份 代码 x 的回归方程类型？（给出判断即可，不必说明理由）

$\bar{x}$	$\bar{y}$	$\bar{z}$	$\bar{w}$	$\sum_{i=1}^{10}(x_i - \bar{x})^2$
5.5	9.02	2.14	1.51	82.5
$\sum_{i=1}^{10}(w_i - \bar{w})^2$	$\sum_{i=1}^{10}(x_i - \bar{x})(y_i - \bar{y})$	$\sum_{i=1}^{10}(x_i - \bar{x})(z_i - \bar{z})$	$\sum_{i=1}^{10}(w_i - \bar{w})(y_i - \bar{y})$	
4.84	72.2	9.67	18.41	

(2)根据（1）的判断结果及表中数据，建立 y 关于 x 的回归方程；（所有过程保留两位小数）

(3)试预测该市 2022 年的人才引进就业人数.

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

【答案】(1)答案见解析

(2) $\hat{y} = 3.80 \ln x + 3.28$

(3)12.704

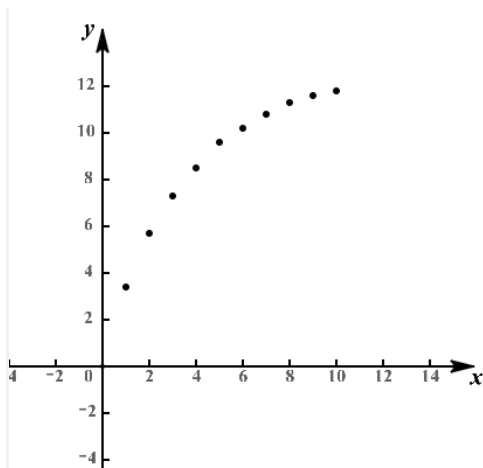
【分析】(1) 根据数据直接做图，并根据图判断函数类型；

(2) 根据回归方程计算公式和已知数据求解即可得出方程；

(3) 将 $x=12$ 代入回归方程即可求解.

(1)

图像



$y = m + n \ln x$ 适合作为该市人才引进就业人数 y 关于年份 代码 x 的回归方程类型

(2)

$$(2) \quad \hat{n} = \frac{\sum_{i=1}^{10} (w_i - \bar{w})(y_i - \bar{y})}{\sum_{i=1}^{10} (w_i - \bar{w})^2} = \frac{18.41}{4.84} \approx 3.80$$

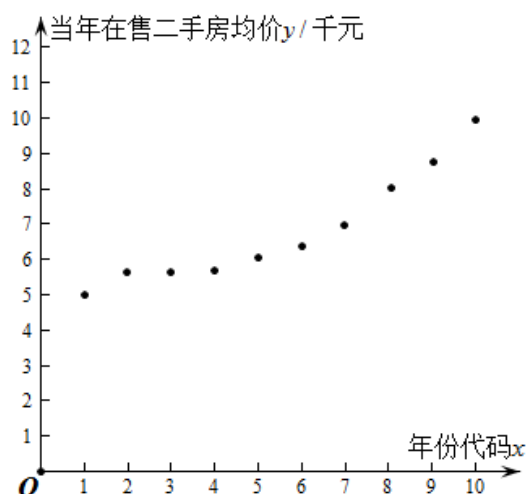
$$\hat{m} = \bar{y} - \hat{n}\bar{w} \approx 9.02 - 3.80 \times 1.51 = 3.28$$

$$\therefore \hat{y} = 3.80 \ln x + 3.28$$

(3)

(3) 将 $x=12$ 代入得 $\hat{y} = 3.80 \times 2.48 + 3.28 = 12.704$.

变式 2-4. 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 $y = a + bx$ 和 $y = e^{c+dx}$ 两个模型对年份代码 x 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	$y = a + bx$		$y = e^{c+dx}$	
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： $\hat{\beta} = \frac{\sum_{i=1}^n (u_i - \bar{u})(v_i - \bar{v})}{\sum_{i=1}^n (u_i - \bar{u})^2}$, $\hat{\alpha} = \bar{v} - \hat{\beta} \bar{u}$. 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 x 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 $\hat{d} = \frac{\sum_{i=1}^{10} (x_i - \bar{x})(w_i - \bar{w})}{\sum_{i=1}^{10} (x_i - \bar{x})^2} = \frac{6.60}{82.5} = 0.08$,

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/097113146105010041>