模块四 成对数据的统计分析

第 1 节 一元线性回归模型及其应用（★★★）

内容提要

本节主要归纳变量的相关关系、一元线性回归模型及其应用等内容，先梳理一些基本概念.

1．样本相关系数 r：用于判断线性相关关系的强弱，以及是正相关还是负相关.

①计算公式：$r=\dfrac{\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}\cdot\sqrt{\sum\limits_{i=1}^{n}(y_i-\overline{y})^2}}=\dfrac{\sum\limits_{i=1}^{n}x_iy_i-n\overline{x}\,\overline{y}}{\sqrt{\sum\limits_{i=1}^{n}x_i^2-n\overline{x}^2}\cdot\sqrt{\sum\limits_{i=1}^{n}y_i^2-n\overline{y}^2}}$，且求得的 r 必在 $[-1,1]$ 上；

②当 $r>0$ 时，变量 x 和 y 正相关，当 $r<0$ 时，变量 x 和 y 负相关；

③$|r|$ 越接近 1，变量 x 和 y 的线性相关程度越大；$|r|$ 越接近 0，变量 x 和 y 的线性相关程度越小.

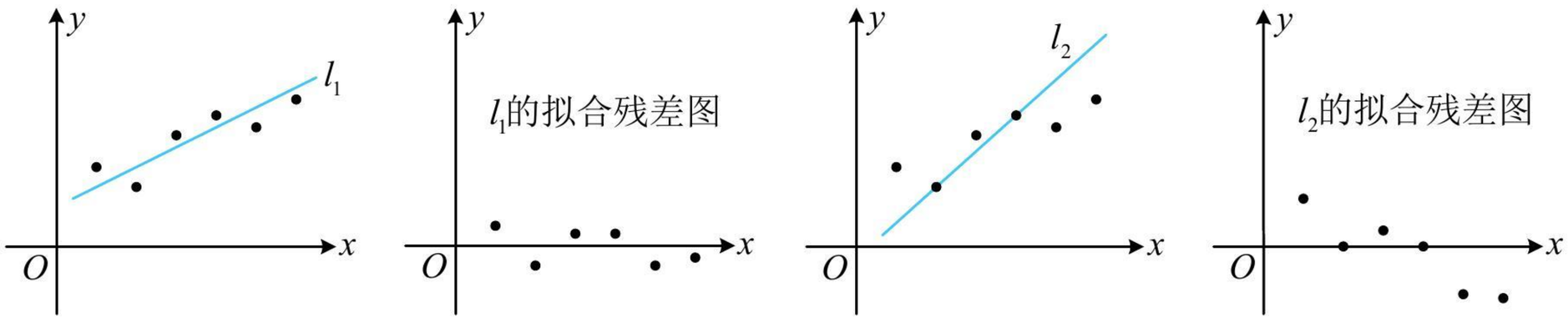
2．一元线性回归模型参数的最小二乘估计：在 y 关于 x 的经验回归方程 $\hat{y}=\hat{b}x+\hat{a}$ 中，参数 $\hat{b}$ 和 $\hat{a}$ 的最小二乘估计公式分别为 $\hat{b}=\dfrac{\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}=\dfrac{\sum\limits_{i=1}^{n}x_iy_i-n\overline{x}\,\overline{y}}{\sum\limits_{i=1}^{n}x_i^2-n\overline{x}^2}$，$\hat{a}=\overline{y}-\hat{b}\overline{x}$.

注：上式中 $\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})=\sum\limits_{i=1}^{n}x_iy_i-n\overline{x}\,\overline{y}$，$\sum\limits_{i=1}^{n}(x_i-\overline{x})^2=\sum\limits_{i=1}^{n}x_i^2-n\overline{x}^2$，这种公式的转换需熟悉，可能出现给的是其中一种形式，但计算时却必须用另一种的情况，下面给出 $\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})=\sum\limits_{i=1}^{n}x_iy_i-n\overline{x}\,\overline{y}$ 的证明过程.

证明：$\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})=\sum\limits_{i=1}^{n}(x_iy_i-\overline{y}x_i-\overline{x}y_i+\overline{x}\,\overline{y})=\sum\limits_{i=1}^{n}x_iy_i-\sum\limits_{i=1}^{n}\overline{y}x_i-\sum\limits_{i=1}^{n}\overline{x}y_i+\sum\limits_{i=1}^{n}\overline{x}\,\overline{y}$

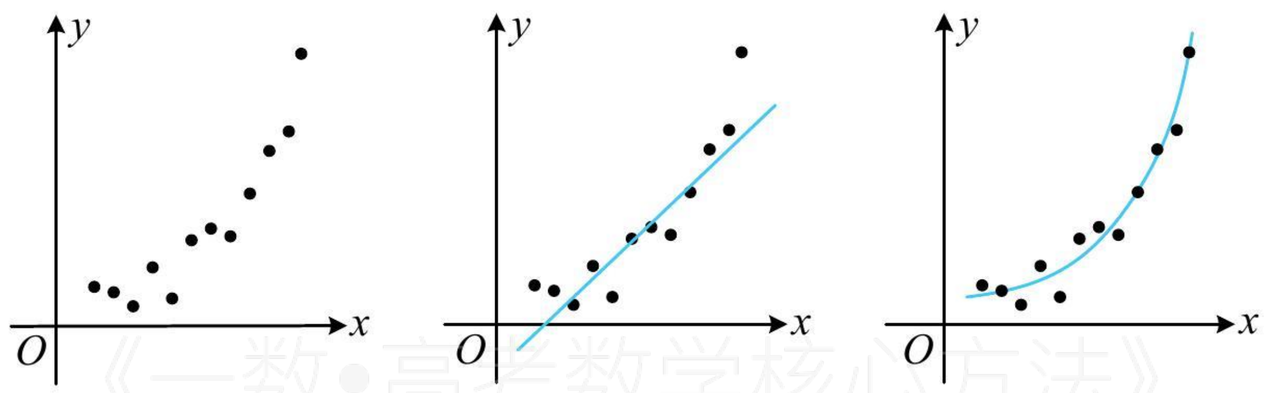
$=\sum\limits_{i=1}^{n}x_iy_i-\overline{y}\sum\limits_{i=1}^{n}x_i-\overline{x}\sum\limits_{i=1}^{n}y_i+n\overline{x}\,\overline{y}=\sum\limits_{i=1}^{n}x_iy_i-\overline{y}\cdot n\overline{x}-\overline{x}\cdot n\overline{y}+n\overline{x}\,\overline{y}=\sum\limits_{i=1}^{n}x_iy_i-n\overline{x}\,\overline{y}$.

3．残差：用回归方程拟合两个变量 x 和 y 的关系时，对于样本点 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$，称观测值 y_i 与预测值 $\hat{y}_i$ 的差 $y_i-\hat{y}_i$ 为相对于样本点 (x_i,y_i) 的残差，其中 $i=1,2,\cdots,n$．将所有样本点的残差绘制成图形即可得到残差图，残差点比较均匀地落在水平带状区域中，且这样的区域越窄，模型的拟合效果越好. 例如，下面是用两个不同的线性回归模型 l_1 和 l_2 对同一组观测数据进行拟合以及对应的残差图，对比可得线性回归模型 l_1 的残差点分布在 x 轴附近狭窄的带状区域内，拟合效果比 l_2 好.



4．决定系数：$R^2=1-\dfrac{\sum\limits_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum\limits_{i=1}^{n}(y_i-\bar{y})^2}$，对于已经获取的样本数据，分母部分$\sum\limits_{i=1}^{n}(y_i-\bar{y})^2$是确定的数，分子部分$\sum\limits_{i=1}^{n}(y_i-\hat{y}_i)^2$是残差平方和，决定系数$R^2$越大，意味着残差平方和越小，拟合效果越好．当要比较两种回归模型的拟合效果时，可计算决定系数R^2加以对比.

5．非线性回归模型：通过变换（取对数、取指数、平方等）转化为线性回归模型计算，有关考题一般会给出参考数据．例如下图的这组观测数据$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$，若用线性回归模型$\hat{y}=\hat{b}x+\hat{a}$拟合，效果就比用指数模型$\hat{y}=\hat{a}\mathrm{e}^{\hat{b}x}$拟合差．而欲求模型$\hat{y}=\hat{a}\mathrm{e}^{\hat{b}x}$中的$\hat{a}$和$\hat{b}$，可两端取自然对数，得到$\ln\hat{y}=\hat{b}x+\ln\hat{a}$，若设$\begin{cases}\hat{z}=\ln\hat{y}\\ \hat{c}=\ln\hat{a}\end{cases}$，则$\hat{z}=\hat{b}x+\hat{c}$，这样就将$y$关于$x$的非线性拟合转化成了$z$关于$x$的线性拟合．这里用到的变换，就是取对数，我们可以将观测数据$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$变换成$(x_1,z_1),(x_2,z_2),\cdots,(x_n,z_n)$，再用最小二乘法求得$z$关于$x$的线性回归方程，最后将$z$换回成$\ln y$即可.



典型例题

类型Ⅰ：概念、计算类小题

【例 1】关于线性回归的描述，下列说法不正确的是（　　）

（A）经验回归方程$\hat{y}=-1.8x+2.1$中变量x，y成正相关关系

（B）相关系数r的绝对值越接近 1，线性相关程度越强

（C）经验回归方程$\hat{y}=1.1x-1.6$中变量x，y成正相关关系

（D）残差平方和越小，拟合效果越好

解析：A 项，经验回归方程$\hat{y}=-1.8x+2.1$的斜率$\hat{b}=-1.8<0\Rightarrow x$，$y$成负相关关系，故 A 项错误；

B 项，$|r|$越接近 1，成对样本数据的线性相关程度越强，越接近 0，线性相关程度越弱，故 B 项正确；

C 项，经验回归方程$\hat{y}=1.1x-1.6$的斜率$\hat{b}=1.1>0$，所以x，y成正相关关系，故 C 项正确；

D 项，残差平方和越小，则回归模型的拟合效果越好，故 D 项正确.

答案：A

【例 2】某同学在研究性学习中，收集到某制药厂今年前 5 个月甲胶囊生产产量（单位：万盒）的数据如下表所示：

x	1	2	3	4	5
y	5	6	5	6	8

若 x，y 线性相关，经验回归方程为 $\hat{y}=0.7x+\hat{a}$，则以下判断正确的是（　　）

（A）x 增加 1 个单位，则 y 必增加 0.7 个单位

（B）x 减少 1 个单位，则 y 必减少 0.7 个单位

（C）当 $x=6$ 时，y 的预测值为 8.1 万盒

（D）经验回归直线 $\hat{y}=0.7x+\hat{a}$ 经过点 $(2,6)$

解析：A 项，由经验回归方程 $\hat{y}=0.7x+\hat{a}$ 可知当 x 增加 1 个单位时，y 大约增加 0.7 个单位，不是必增加 0.7 个单位，故 A 项错误，同理也可得 B 项错误；

C 项，要求 $x=6$ 时 y 的预测值，需先求出 $\hat{a}$，可将样本中心点 $(\overline{x},\overline{y})$ 代入经验回归方程，

由表中数据，$\overline{x}=\dfrac{1+2+3+4+5}{5}=3$，$\overline{y}=\dfrac{5+6+5+6+8}{5}=6$，将 $(3,6)$ 代入 $\hat{y}=0.7x+\hat{a}$ 可得 $6=0.7\times3+\hat{a}$，

解得：$\hat{a}=3.9$，所以 $\hat{y}=0.7x+3.9$，当 $x=6$ 时，$\hat{y}=0.7\times6+3.9=8.1$，故 C 项正确；

D 项，经验回归直线经过点 $(3,6)$，不过点 $(2,6)$，故 D 项错误.

答案：C

【反思】$(\overline{x},\overline{y})$ 叫做样本数据 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$ 的样本中心点，样本中心点一定在经验回归直线上.

【例 3】某部门统计了某地区今年前 7 个月在线外卖的规模如下表：

月份代号 x	1	2	3	4	5	6	7
在线外卖规模 y（百万元）	11	13	18	★	28	★	35

其中 4、6 两个月的在线外卖规模数据模糊，但这 7 个月的平均值为 23，若利用经验回归方程 $\hat{y}=\hat{b}x+\hat{a}$ 来拟合预测，且 7 月相应于点 $(7,35)$ 的残差为 -0.6，则 $\hat{a}-\hat{b}=$（）

（A）1　　（B）2　　（C）3　　（D）4

解析：要求 $\hat{a}-\hat{b}$，应建立关于 $\hat{a}$ 和 $\hat{b}$ 的方程组，由题意，$\overline{x}=\dfrac{1+2+3+4+5+6+7}{7}=4$，$\overline{y}=23$，

所以样本中心点 $(4,23)$ 满足经验回归方程 $\hat{y}=\hat{b}x+\hat{a}$，故 $4\hat{b}+\hat{a}=23$ ①，

题干还给了一个残差数据，可由此建立第二个方程，

因为 7 月相应于点 $(7,35)$ 的残差为 -0.6，所以 $35-(7\hat{b}+\hat{a})=-0.6$ ②，

联立①②解得：$\hat{a}=6.2$，$\hat{b}=4.2$，所以 $\hat{a}-\hat{b}=2$.

答案：B

【例 4】据一组样本数据 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$ 求得经验回归方程为 $\hat{y}=1.2x+0.4$，且 $\overline{x}=3$，现发现这组样本数据中有两个样本点 $(1.2,0.5)$ 和 $(4.8,7.5)$ 误差较大，去除后重新求得的经验回归直线 l 的斜率为 1.1，则（　　）

（A）去除两个误差较大的样本点后，y 的估计值增加速度变快

（B）去除两个误差较大的样本点后，重新求得的回归方程对应直线一定过点 $(3,5)$

（C）去除两个误差较大的样本点后，重新求得的回归方程为 $\hat{y}=1.1x+0.7$

（D）去除两个误差较大的样本点后，相应于样本点 $(2,2.7)$ 的残差为 0.1

解析：A 项，去除两个误差较大的样本点后，经验回归直线的斜率由 1.2 变成了 1.1，所以 y 的估计值增加速度变慢了，故 A 项错误；

B 项，经验回归直线一定过样本中心点，所以只需看新样本数据的 $\overline{x}$ 和 $\overline{y}$ 是否分别为 3 和 5，

由题意，原样本数据中，$\overline{x}=\dfrac{x_1+x_2+\cdots+x_n}{n}=3$，所以 $x_1+x_2+\cdots+x_n=3n$，

又 $\overline{y}=\dfrac{y_1+y_2+\cdots+y_n}{n}=1.2\overline{x}+0.4=4$，所以 $y_1+y_2+\cdots+y_n=4n$，

故去除两个误差较大的样本点后，$\overline{x}=\dfrac{x_1+x_2+\cdots+x_n-1.2-4.8}{n-2}=\dfrac{3n-6}{n-2}=3$，

$\overline{y}=\dfrac{y_1+y_2+\cdots+y_n-0.5-7.5}{n-2}=\dfrac{4n-8}{n-2}=4$，所以重新求得的回归方程对应直线一定过点 $(3,4)$，故 B 项错误；

C 项，给出了新的经验回归直线的斜率为 1.1，故只需再求出截距，可将新的样本中心点 $(3,4)$ 代入，

将 $(3,4)$ 代入 $\hat{y}=1.1x+\hat{a}$ 可得 $4=1.1\times3+\hat{a}$，解得：$\hat{a}=0.7$，所以新回归方程为 $\hat{y}=1.1x+0.7$，故 C 项正确；

D 项，新回归方程中，当 $x=2$ 时，$\hat{y}=1.1\times2+0.7=2.9$，所以残差为 $2.7-2.9=-0.2$，故 D 项错误.

答案：C

类型Ⅱ：线性回归综合大题

【例 5】某种机械设备随着使用年限的增加，它的使用功能逐渐减退，使用价值逐年减少，通常把它使用价值逐年减少的“量”换算成费用，称之为“失效费”. 该种机械设备的使用年限 x（单位：年）与失效费 y（单位：万元）的统计数据如下表所示：

使用年限 x（年）	1	2	3	4	5	6	7
失效费 y（万元）	2.9	3.3	3.6	4.4	4.8	5.2	5.9

（1）由上表数据可知，可用线性回归模型拟合 y 与 x 的关系，请用样本相关系数加以说明；

（2）求出 y 关于 x 的经验回归方程，并估算该种机械设备使用 10 年的失效费.

参考公式：相关系数 $r=\dfrac{\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}\cdot\sqrt{\sum\limits_{i=1}^{n}(y_i-\overline{y})^2}}$，经验回归方程 $\hat{y}=\hat{b}x+\hat{a}$ 中的 $\hat{b}$ 和 $\hat{a}$ 的最小二乘估计公式为 $\hat{b}=\dfrac{\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}$，$\hat{a}=\overline{y}-\hat{b}\overline{x}$.

参考数据：$\sum\limits_{i=1}^{7}(x_i-\overline{x})(y_i-\overline{y})=14$，$\sum\limits_{i=1}^{7}(y_i-\overline{y})^2=7.08$，$\sqrt{198.24}\approx14.08$.

解：（1）（相关系数的公式中，只有 $\sum\limits_{i=1}^{7}(x_i-\overline{x})^2$ 这部分没给参考数据，需用表中数据来计算它）

由表中数据，$\overline{x}=\dfrac{1+2+3+4+5+6+7}{7}=4$，故 $\sum\limits_{i=1}^{7}(x_i-\overline{x})^2=(1-4)^2+(2-4)^2+\cdots+(7-4)^2=28$，

结合所给参考数据可得相关系数 $r=\dfrac{\sum\limits_{i=1}^{7}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum\limits_{i=1}^{7}(x_i-\overline{x})^2}\cdot\sqrt{\sum\limits_{i=1}^{7}(y_i-\overline{y})^2}}=\dfrac{14}{\sqrt{28}\times\sqrt{7.08}}=\dfrac{14}{\sqrt{198.24}}\approx\dfrac{14}{14.08}\approx0.99$，

因为 r 很接近 1，所以 y 与 x 有很强的线性相关关系，可用线性回归模型拟合 y 与 x 的关系.

（2）（从公式看，算 $\hat{b}$ 需要用到的量齐了，故先求 $\hat{b}$ ）由（1）得 $\hat{b}=\dfrac{\sum\limits_{i=1}^{7}(x_i-\overline{x})(y_i-\overline{y})}{\sum\limits_{i=1}^{7}(x_i-\overline{x})^2}=\dfrac{14}{28}=0.5$，

（算 $\hat{a}$ 要用 $\overline{x}$ 和 $\overline{y}$，故再算 $\overline{y}$ ）由表中数据，$\overline{y}=\dfrac{2.9+3.3+3.6+4.4+4.8+5.2+5.9}{7}=4.3$，

所以 $\hat{a}=\overline{y}-\hat{b}\overline{x}=4.3-0.5\times4=2.3$，故 y 关于 x 的经验回归方程为 $\hat{y}=0.5x+2.3$，

当 $x=10$ 时，$\hat{y}=0.5\times10+2.3=7.3$，所以该种机械设备使用 10 年的失效费约为 7.3 万元.

【例 6】下面给出了根据我国 2016 年~2022 年水果人均占有量 y（单位：kg）和年份代码 x 绘制的散点图（图 1）和线性回归方程的残差图（图 2），其中 2016 年~2022 年的年份代码 x 分别为 1~7.

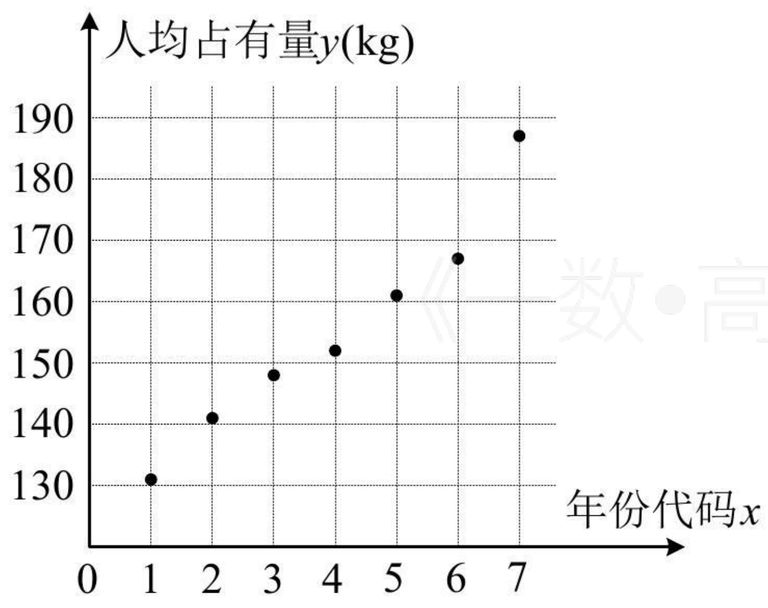


图1

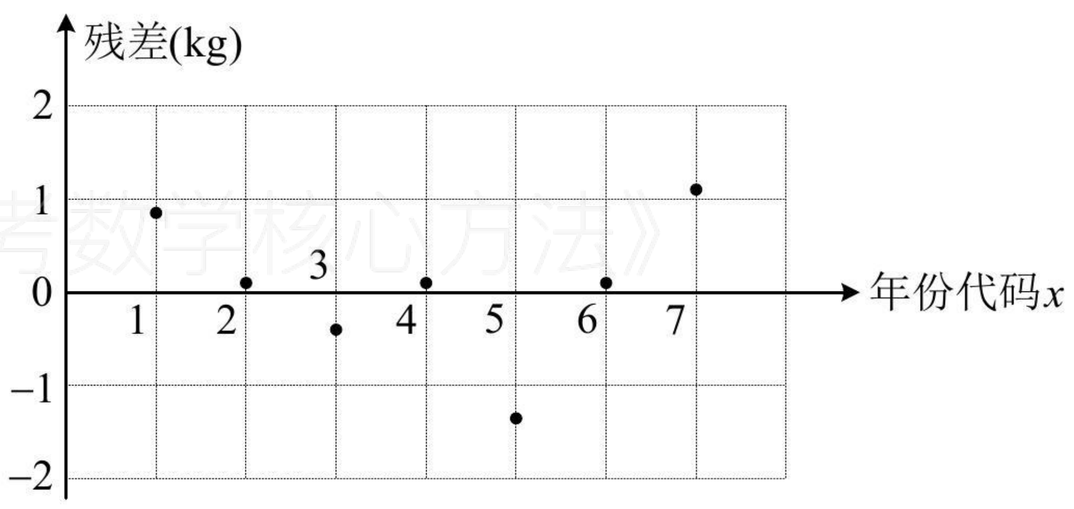


图2

（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 $\sum\limits_{i=1}^{7}y_i=1074$，$\sum\limits_{i=1}^{7}x_iy_i=4517$，求 y 关于 x 的经验回归方程；（精确到 0.01）

（3）根据经验回归方程的残差图，分析线性回归方程的拟合效果.

附：经验回归方程 $\hat{y}=\hat{b}x+\hat{a}$ 中的 $\hat{b}$ 和 $\hat{a}$ 的最小二乘估计公式为 $\hat{b}=\dfrac{\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}$，$\hat{a}=\overline{y}-\hat{b}\overline{x}$.

解：（1）由散点图可知，样本点均匀地分布在一条直线附近，且随着 x 的增大，y 也增大，所以 y 与 x 有较强的线性相关关系，且为正相关.

（2）（先由 $\hat{b}=\dfrac{\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}$ 求 $\hat{b}$，观察参考数据发现分子分母都没给，故都得算，先算较简单的分母）

由题意，$\overline{x}=\dfrac{1+2+3+4+5+6+7}{7}=4$，所以$\sum\limits_{i=1}^{7}(x_i-\overline{x})^2=(1-4)^2+(2-4)^2+(3-4)^2+\cdots+(7-4)^2=28$，

（再算分子，从图中我们无法读出准确的样本点纵坐标$y_i(i=1,2,\cdots,7)$，故应结合参考数据来算，参考数据中给了$\sum\limits_{i=1}^{7}x_iy_i=4517$，于是将$\sum\limits_{i=1}^{7}(x_i-\overline{x})(y_i-\overline{y})$转换成$\sum\limits_{i=1}^{7}x_iy_i-n\overline{x}\,\overline{y}$来算，见反思）

由所给参考数据，$\overline{y}=\dfrac{1}{7}\sum\limits_{i=1}^{7}y_i=\dfrac{1074}{7}$，$\sum\limits_{i=1}^{7}(x_i-\overline{x})(y_i-\overline{y})=\sum\limits_{i=1}^{7}x_iy_i-7\overline{x}\overline{y}=4517-7\times4\times\dfrac{1074}{7}=221$，

所以$\hat{b}=\dfrac{\sum\limits_{i=1}^{7}(x_i-\overline{x})(y_i-\overline{y})}{\sum\limits_{i=1}^{7}(x_i-\overline{x})^2}=\dfrac{221}{28}\approx7.89$，$\hat{a}=\overline{y}-\hat{b}\overline{x}=\dfrac{1074}{7}-\dfrac{221}{28}\times4=\dfrac{853}{7}\approx121.86$，

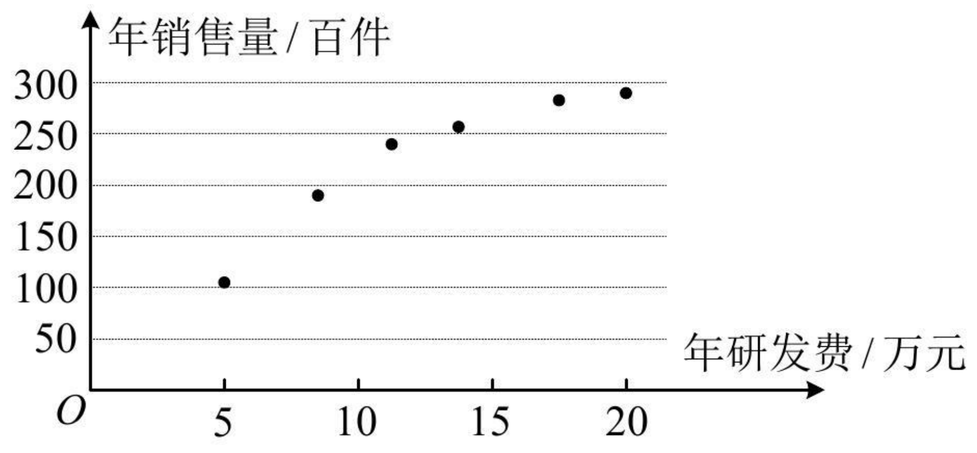
故y关于x的经验回归方程为$\hat{y}=7.89x+121.86$.

（3）由残差图可以看出，观测数据的残差点分布在水平带状区域内，且宽度较窄，说明拟合效果较好.

【反思】$\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})$和$\sum\limits_{i=1}^{n}x_iy_i-n\overline{x}\overline{y}$的转化务必熟悉，本题公式给的是$\hat{b}=\dfrac{\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}$，但实际计算却必须把分子转化成$\sum\limits_{i=1}^{n}x_iy_i-n\overline{x}\overline{y}$.

类型Ⅲ：非线性回归综合大题

【例 7】某公司为确定下一年度投入某种产品的研发费，需了解年研发费 x（单位：万元）对年销售量 y（单位：百件）和年利润 z（单位：万元）的影响，现对近 6 年的年研发费 x_i 和年销售量 $y_i(i=1,2,\cdots,6)$ 数据做了初步处理，得到下面的散点图以及一些统计量的值.



$\overline{x}$	$\overline{y}$	$\overline{u}$	$\sum\limits_{i=1}^{6}(x_i-\overline{x})^2$	$\sum\limits_{i=1}^{6}(u_i-\overline{u})^2$	$\sum\limits_{i=1}^{6}(x_i-\overline{x})(y_i-\overline{y})$	$\sum\limits_{i=1}^{6}(u_i-\overline{u})(y_i-\overline{y})$
12.5	222	3.5	157.5	4.5	1854	270

表中$u_i=\ln x_i$，$\overline{u}=\dfrac{1}{6}\sum\limits_{i=1}^{6}u_i$.

（1）根据散点图判断$\hat{y}=\hat{a}+\hat{b}x$与$\hat{y}=\hat{c}+\hat{d}\ln x$哪一个更适合作为年销售量$y$关于年研发费$x$的回归方程类型；（给出判断即可，不必说明理由）

（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；

（3）已知这种产品的年利润$z=0.5y-x$，根据（2）的结果，求出当年研发费为多少时，年利润z的预测

值最大？

附：经验回归方程 $\hat{y}=\hat{b}x+\hat{a}$ 中的 $\hat{b}$ 和 $\hat{a}$ 的最小二乘估计公式为 $\hat{b}=\dfrac{\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}$，$\hat{a}=\overline{y}-\hat{b}\overline{x}$.

解：（1）（由散点图选回归模型，就看哪种回归模型的函数图象和散点图的趋势比较接近即可）

由散点图可知，$\hat{y}=\hat{c}+\hat{d}\ln x$ 更适合作为年销售量 y 关于年研发费 x 的回归方程类型.

（2）（把 $\ln x$ 看作整体 u，则 $\hat{y}=\hat{c}+\hat{d}\ln x$ 即为 $\hat{y}=\hat{c}+\hat{d}u$，故可按线性回归方程的最小二乘估计求 $\hat{d}$ 和 $\hat{c}$）

设 $u=\ln x$，则 $\hat{y}=\hat{c}+\hat{d}u$，由所给表中数据，$\hat{d}=\dfrac{\sum\limits_{i=1}^{6}(u_i-\overline{u})(y_i-\overline{y})}{\sum\limits_{i=1}^{6}(u_i-\overline{u})^2}=\dfrac{270}{4.5}=60$，$\hat{c}=\overline{y}-\hat{d}\overline{u}=222-60\times3.5=12$，

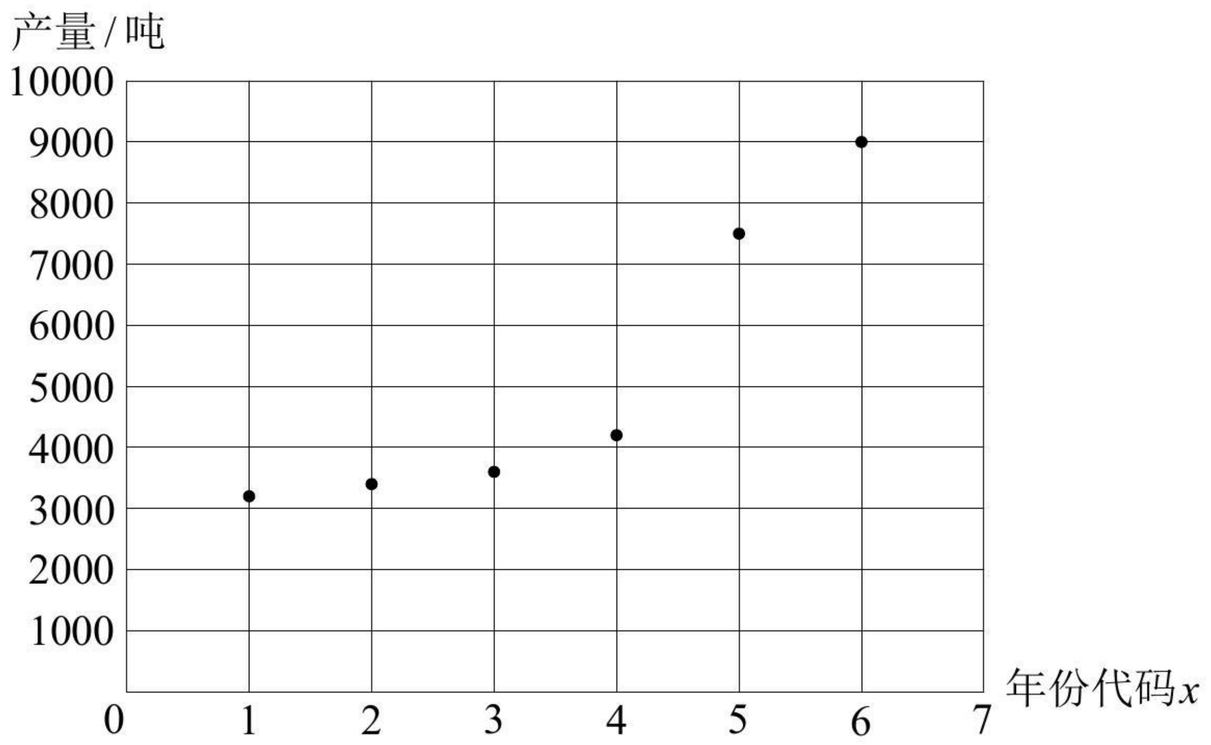
所以 $\hat{y}=12+60u$，故 y 关于 x 的回归方程为 $\hat{y}=12+60\ln x$.

（3）由（2）可知 $z=0.5y-x=0.5(12+60\ln x)-x=6+30\ln x-x$，所以 $z'=\dfrac{30}{x}-1=\dfrac{30-x}{x}$，

从而 $z'>0\Leftrightarrow0<x<30$，$z'<0\Leftrightarrow x>30$，故 $z=6+30\ln x-x$ 在 $(0,30)$ 上单调递增，在 $(30,+\infty)$ 上单调递减，所以当年研发费为 30 万元时，年利润 z 的预测值最大.

【例 8】近年来，云南省保山市龙陵县紧紧围绕打造“中国石斛之乡”的发展定位，大力发展石斛产业，该产业带动龙陵县近四分之一人口脱贫致富. 2022 年 8 月，龙陵紫皮石斛获国家地理标志运用促进工程重点项目，并被评为优秀等次. 在政府的大力扶持下，龙陵紫皮石斛产量逐年增长，2017 年底到 2022 年底龙陵县石斛产量统计表和散点图如下.

年份	2017	2018	2019	2020	2021	2022
年份代码 x	1	2	3	4	5	6
紫皮石斛产量 y（吨）	3200	3400	3600	4200	7500	9000



（1）根据散点图判断，$\hat{y}=\hat{b}x+\hat{a}$ 与 $\hat{y}=\hat{c}\mathrm{e}^{\hat{d}x}$（$\hat{a}$，$\hat{b}$，$\hat{c}$，$\hat{d}$ 均为常数）哪一个更适合作为龙陵县紫皮石斛产量 y 关于年份代码 x 的回归模型？（给出判断即可，不必说明理由）

（2）经计算得下表中数据，根据（1）中结果，求出 y 关于 x 的回归方程；

$\overline{x}$	$\overline{y}$	$\overline{u}$	$\sum_{i=1}^{6}(x_i-\overline{x})^2$	$\sum_{i=1}^{6}(x_i-\overline{x})(y_i-\overline{y})$	$\sum_{i=1}^{6}(x_i-\overline{x})(u_i-\overline{u})$
3.5	5150	8.46	17.5	20950	3.85

其中 $u_i=\ln y_i(i=1,2,3,4,5,6)$.

（3）龙陵县计划到 2025 年底实现紫皮石斛年产量达 1.5 万吨，根据（2）所求得的回归方程，预测该目标是否能完成？（参考数据：$e^{9.45}\approx12708$，$e^{9.67}\approx15835$）

附：经验回归方程 $\hat{y}=\hat{b}x+\hat{a}$ 中的 $\hat{b}$ 和 $\hat{a}$ 的最小二乘估计公式为 $\hat{b}=\dfrac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$，$\hat{a}=\overline{y}-\hat{b}\overline{x}$.

解：（1）（由散点图选回归模型，就看哪种回归模型的函数图象和散点图更接近）

由散点图可知，$\hat{y}=\hat{c}e^{\hat{d}x}$ 更适合作为龙陵县紫皮石斛产量 y 关于年份代码 x 的回归模型.

（2）（要作变换 $\hat{u}=\ln\hat{y}$，故先在回归方程两端取对数）由 $\hat{y}=\hat{c}e^{\hat{d}x}$ 可得 $\ln\hat{y}=\ln(\hat{c}e^{\hat{d}x})=\ln\hat{c}+\hat{d}x$，

令 $\hat{u}=\ln\hat{y}$，$\hat{h}=\ln\hat{c}$，则 $\hat{u}=\hat{h}+\hat{d}x$，（u 关于 x 即为线性回归模型，可代最小二乘估计公式算 $\hat{h}$ 和 $\hat{d}$）

由表中数据，$\hat{d}=\dfrac{\sum_{i=1}^{6}(x_i-\overline{x})(u_i-\overline{u})}{\sum_{i=1}^{6}(x_i-\overline{x})^2}=\dfrac{3.85}{17.5}=0.22$，$\hat{h}=\overline{u}-\hat{d}\overline{x}=8.46-0.22\times3.5=7.69$，

所以 u 关于 x 的回归方程为 $\hat{u}=7.69+0.22x$，故 y 关于 x 的回归方程为 $\ln\hat{y}=7.69+0.22x$，即 $\hat{y}=e^{7.69+0.22x}$.

（3）2025 年对应年份代码 9，将 $x=9$ 代入 $\hat{y}=e^{7.69+0.22x}$ 可得 $\hat{y}=e^{7.69+0.22\times9}=e^{9.67}\approx15835>15000$，

所以预测到 2025 年底实现紫皮石斛年产量达 1.5 万吨的目标能够完成.

【总结】从上面两道题可以看出，求非线性回归方程，常通过代换转化为线性回归方程，用最小二乘估计公式来算，且这类题往往会给出大量参考数据，所以实际计算量不大.

强化训练

1.（2023·河南模拟·★）为了研究汽车减重对降低油耗的作用，对一组样本数据 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$ 进行分析，其中 x_i 表示减重质量（单位：kg），y_i 表示每行驶一百公里降低的油耗（单位：升），$i=1,2,\cdots,n$，由此得到的经验回归方程为 $\hat{y}=\hat{b}x+\hat{a}(\hat{b}>0)$，有下列四个说法：

①$\hat{a}$ 的值一定为 0；②$\hat{b}$ 越大，减重对降低油耗的作用越大；③决定系数 R^2 越大，拟合效果越好；④至少有一个数据点在经验回归直线上. 其中所有正确说法的编号是（　　）

（A）①④　　（B）②③　　（C）②③④　　（D）①②④

2.（2023·湖南模拟·★）（多选）设某大学的女生体重 y（单位：kg）与身高 x（单位：cm）具有线性相关关系，根据一组样本数据 $(x_i,y_i)(i=1,2,\cdots,n)$，用最小二乘法建立的经验回归方程为 $\hat{y}=0.85x-85.71$，则

下列结论中正确的是（　　）

（A）y 与 x 有正的线性相关关系

（B）若该大学女生的平均身高为 168cm，则平均体重约为 57.09kg

（C）若该大学某女生身高增加 1cm，则其体重约增加 0.85kg

（D）若该大学某女生身高为 170cm，则可断定其体重必为 58.79kg

3．（2023·吉林模拟·★★）某地以“绿水青山就是金山银山”理念为引导，推进绿色发展，现要订购一批苗木，苗木长度与售价如下表：

苗木长度 x（cm）	38	48	58	68	78	88
售价 y（元）	16.8	18.8	20.8	22.8	24	25.8

若苗木长度 x（cm）与售价 y（元）之间存在线性相关关系，其经验回归方程为 $\hat{y}=\hat{b}x+8.9$，则当售价大约为 38.9 元时，苗木长度大约为（　　）

（A）148cm　　（B）150cm　　（C）152cm　　（D）154cm

4．（2022·贵州模拟·★★★）某企业新研发了一种产品，产品的成本由原料成本及非原料成本组成，每件产品的非原料成本 y（元）与生产的产品数量 x（千件）有关，经统计得到如下数据：

x	2	5	8	9	11
y	12	10	8	8	7

（1）根据表中的数据，运用相关系数进行分析说明，可以用一元线性回归模型拟合 y 与 x 的关系，并指出是正相关还是负相关；

（2）求 y 关于 x 的经验回归方程，并预测生产该产品 13 千件时，每件产品的非原料成本为多少元？

参考公式：相关系数 $r=\dfrac{\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}\cdot\sqrt{\sum\limits_{i=1}^{n}(y_i-\overline{y})^2}}$，经验回归方程 $\hat{y}=\hat{b}x+\hat{a}$ 中的 $\hat{b}$ 和 $\hat{a}$ 的最小二乘估计公式为 $\hat{b}=\dfrac{\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}$，$\hat{a}=\overline{y}-\hat{b}\overline{x}$；参考数据：$\sqrt{2}\approx1.414$.

5.（2023·苏州模拟·★★★）新能源汽车作为战略新兴产业，代表汽车产业的发展方向，发展新能源汽车，对改善能源消费结构、减少空气污染、推动汽车产业和交通运输业转型升级具有积极意义，经过十多年的精心培育，我国新能源汽车产业取得了显著成绩，产销量连续四年全球第一，保有量居全球首位.

（1）已知某公司生产的新能源汽车电池的使用寿命 ξ（单位：万公里）服从正态分布 $N(60,16)$，问：该公司每月生产的 2 万块电池中，大约有多少块电池的使用寿命可以超过 68 万公里？

参考数据：若随机变量 $\xi\sim N(\mu,\sigma^2)$，则 $P(\mu-\sigma<\xi<\mu+\sigma)\approx0.683$，$P(\mu-2\sigma<\xi<\mu+2\sigma)\approx0.955$，$P(\mu-3\sigma<\xi<\mu+3\sigma)\approx0.997$.

（2）下表给出了我国 2017~2021 年新能源汽车保有量 y（单位：万辆）的数据.

年份	2017	2018	2019	2020	2021
年份代码 x	1	2	3	4	5
新能源汽车保有量 y	153	260	381	492	784

经计算，变量 x，y 的样本相关系数 $r_1\approx0.946$，变量 x^2 与 y 的样本相关系数 $r_2\approx0.985$.

①试判断 $\hat{y}=\hat{b}x+\hat{a}$ 和 $\hat{y}=\hat{b}x^2+\hat{a}$ 哪一个更适合作为 y 与 x 之间的回归模型？

②根据①的判断结果，求出 y 关于 x 的回归方程（精确到 0.1），并预测 2023 年我国新能源汽车的保有量.

参考数据：令 $t_i=x_i^2(i=1,2,3,4,5)$，计算得 $\overline{y}=414$，$\sum\limits_{i=1}^{5}x_iy_i=7704$，$\sum\limits_{i=1}^{5}t_iy_i=32094$，$\sum\limits_{i=1}^{5}t_i^2=979$.

参考公式：在回归方程 $\hat{y}=\hat{b}x+\hat{a}$ 中，$\hat{b}=\dfrac{\sum\limits_{i=1}^{n}x_iy_i-n\overline{x}\,\overline{y}}{\sum\limits_{i=1}^{n}x_i^2-n\overline{x}^2}$，$\hat{a}=\overline{y}-\hat{b}\overline{x}$.

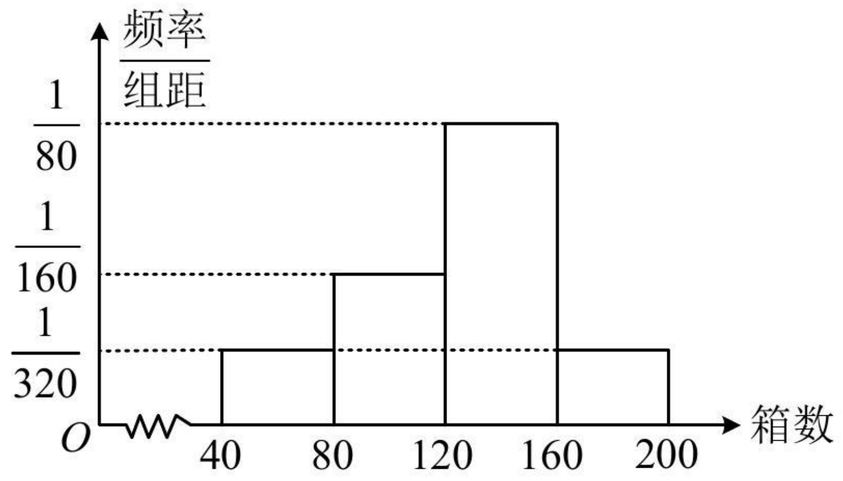
6.（2023·长沙雅礼中学模拟·★★★）为贯彻中共中央、国务院 2023 年一号文件，某单位在当地定点帮扶某村种植一种草莓，并把这种露天种植的草莓搬到了大棚里，收到了很好的经济效益. 根据资料显示，产出的草莓的箱数 x（单位：箱）与成本 y（单位：千元）的关系如下：

x	1	3	4	6	7
y	5	6.5	7	7.5	8

可用回归方程 $\hat{y}=\hat{b}\lg x+\hat{a}$（其中 $\hat{a}$，$\hat{b}$ 为常数）来拟合 y 与 x 的关系.

（1）若农户卖出该草莓的价格为 150 元/箱，试预测该草莓 100 箱的利润是多少元；（利润 = 售价 – 成本）

（2）据统计，1 月份的连续 16 天中农户每天为甲地可配送的该草莓的箱数的频率分布直方图如图，用这 16 天的情况来估计相应的概率. 一个运输户拟购置 n 辆小货车专门运输农户为甲地配送的该草莓，一辆货车每天只能运一趟，每辆车每趟最多只能装载 40 箱该草莓，满载发车，否则不发车. 若发车，则每辆车每趟可获利 500 元；若未发车，则每辆车每天平均亏损 200 元. 试比较 $n=3$ 和 $n=4$ 时，此项业务每天的利润平均值的大小.



参考数据与公式：线性回归直线 $\hat{y}=\hat{b}t+\hat{a}$ 中，$\hat{b}=\dfrac{\sum\limits_{i=1}^{n}(t_i-\overline{t})(y_i-\overline{y})}{\sum\limits_{i=1}^{n}(t_i-\overline{t})^2}$，$\hat{a}=\overline{y}-\hat{b}\overline{t}$；设 $t=\lg x$，则

$\overline{t}$	$\overline{y}$	$\sum\limits_{i=1}^{5}(t_i-\overline{t})(y_i-\overline{y})$	$\sum\limits_{i=1}^{5}(t_i-\overline{t})^2$
0.54	6.8	1.53	0.45

X	[40,80)	[80,120)	[120,160)	[160,200]
P	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{8}$

7.（2023·潍坊一模·★★★★）某学校研究性学习小组在学习生物遗传学的过程中，为验证高尔顿提出的关于儿子成年后身高 y（单位：cm）与父亲身高 x（单位：cm）之间的关系及存在的遗传规律，随机抽取了 5 对父子的身高数据，如下表：

父亲身高 x	160	170	175	185	190
儿子身高 y	170	174	175	180	186

（1）根据表中数据，求出 y 关于 x 的线性回归方程，并利用回归直线方程分别确定儿子比父亲高和儿子比父亲矮的条件，由此可得到怎样的遗传规律？

（2）记 $\hat{e}_i=y_i-\hat{y}_i=y_i-\hat{b}x_i-\hat{a}(i=1,2,\cdots,n)$，其中 y_i 为观测值，$\hat{y}_i$ 为预测值，$\hat{e}_i$ 为对应 (x_i,y_i) 的残差. 求（1）中儿子身高的残差的和，并探究此结果是否对任意具有线性相关关系的两个变量都成立？若是，则加以证明；若不是，说明理由.

参考公式：$\hat{b}=\dfrac{\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}$，$\hat{a}=\overline{y}-\hat{b}\overline{x}$.