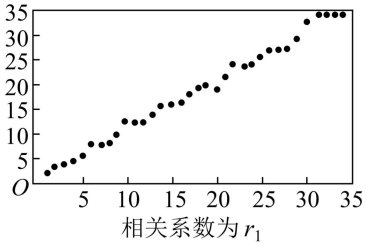
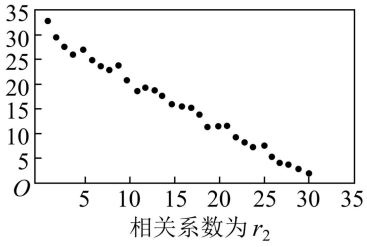
成对数据分析专项靶题

回归分析专项：

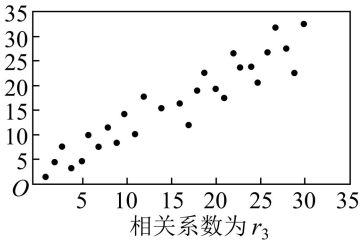
1.某统计部门对四组成对样本数据进行统计分析后，获得如图所示的散点图，关于样本相关系数的比较，其中正确的是（　　）



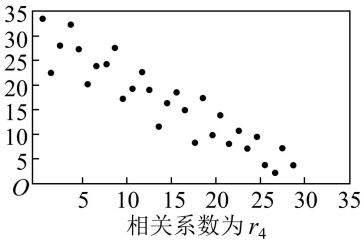
(1)



(2)



(3)



(4)

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

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

2.对于相关系数 r，下列说法中错误的是_____.

①$r=0$时，成对样本数据线性相关程度较弱；

②$r>0$时，表明成对样本数据正相关；

③若线性回归方程中的回归系数 $\hat{b}<0$，则相关系数 $r>0$；

④$|r|$越接近 1，线性相关程度越强，$|r|$越接近 0，线性相关程度越弱.

3.近年来，“考研热”持续升温，2022 年考研报考人数官方公布数据为 457 万，相比于 2021 年增长了 80 万之多，增长率达到 21%以上.考研人数急剧攀升原因较多，其中，本科毕业生人数增多、在职人士考研比例增大，是两大主要因素.据统计，某市各大高校近几年的考研报考总人数如下表：

年份	2018	2019	2020	2021	2022
年份序号 x	1	2	3	4	5
报考人数 y（万人）	1. 1	1.6	2	2.5	m

根据表中数据，可求得 y 关于 x 的线性回归方程为 $\hat{y}=0.43x+0.71$，则 m 的值为_____.

4.（多选）为了研究 y 关于 x 的线性相关关系，收集了 5 组样本数据(见下表)：

x	1	2	3	4	5
y	0.5	0.8	1	1.2	1.5

假设经验回归方程为 $\hat{y}=\hat{b}x+0.28$，则（　）

A．$\hat{b}=0.24$

B．当 $x=8$ 时，y 的预测值为 2.2

C．样本数据 y 的 40%分位数为 0.8

D．去掉样本点 $(3,1)$ 后，x 与 y 的样本相关系数 r 不变

5.已知变量 x 和 y 的统计数据如下表：

x	6	7	8	9	10
y	3.5	4	5	5.5	7

如果由表中数据可得经验回归直线方程为 $\hat{y}=0.85x+\hat{a}$，那么，当 $x=10$ 时，残差为_____.

（注：残差=观测值-预测值）

6.已知回归方程 $\hat{y}=2x+1$，而试验中的一组数据是 $(2,5.1)$，$(3,6.9)$，$(4,8.9)$，则其残差平

方和是_____.

7.（多选）对具有相关关系的两个变量 x 和 y 进行回归分析时，经过随机抽样获得成对的样本点数据 $(x_i,y_i)(i=1,2,\cdots,n)$，则下列结论正确的是（　　）

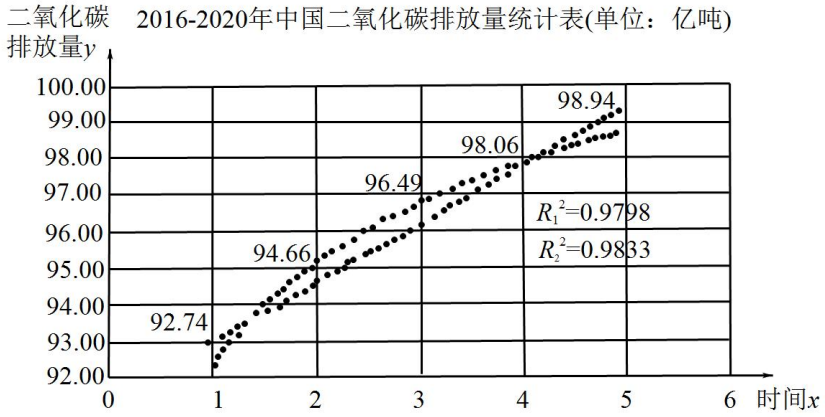
A．若两变量 x，y 具有线性相关关系，则回归直线至少经过一个样本点

B．若两变量 x，y 具有线性相关关系，则回归直线一定经过样本点中心 $(\overline{x},\overline{y})$

C．若以模型 $y=ae^{hx}(a>0)$ 拟合该组数据，为了求出回归方程，设 $z=\ln y$，将其变换后得到线性方程 $z=6x+\ln 3$，则 a，h 的估计值分别是 3 和 6

D．回归分析中常用残差平方和来刻画拟合效果好坏，残差平方和越小，拟合效果越好

8（多选）进入 21 世纪以来，全球二氧化碳排放量增长迅速，自 2000 年至今，全球二氧化碳排放量增加了约 40%，我国作为发展中国家，经济发展仍需要大量的煤炭能源消耗．下图是 2016—2020 年中国二氧化碳排放量的统计图表（以 2016 年为第 1 年）．利用图表中数据计算可得，采用某非线性回归模型拟合时，$R_1^2=0.9798$；采用一元线性回归模型拟合时，线性回归方程为 $\hat{y}=1.58x+91.44$，$R_2^2=0.9833$．则下列说法正确的是（　）



A．由图表可知，二氧化碳排放量 y 与时间 x 正相关

B．由决定系数可以看出，线性回归模型的拟合程度更好

C．利用线性回归方程计算 2019 年所对应的样本点的残差为-0.30

D．利用线性回归方程预计 2025 年中国二氧化碳排放量为 107.24 亿吨

9.从非洲蔓延到东南亚的蝗虫灾害严重威胁了国际农业生产，影响了人民生活.世界性与区域性温度的异常、旱涝频繁发生给蝗灾发生创造了机会.已知蝗虫的产卵量 y 与温度 x 的关系

可以用模型 $y=c_1e^{c_2x}$（其中 e 为自然对数的底数）拟合，设 $z=\ln y$，其变换后得到一组数据：

x	20	23	25	27	30
z	2	2.4	3	3	4.6

由上表可得经验回归方程 $z=0.2x+a$，则当 x=35 时，蝗虫的产卵量 y 的估计值为（　）

A．e^5　　B．e^6　　C．8　　D．e^{10}

10.某县依托种植特色农产品，推进产业园区建设，致富一方百姓．已知该县近 5 年人均可支配收入如下表所示，记 2017 年为 $x=1$，2018 年为 $x=2$，…以此类推．

年份	2017	2018	2019	2020	2021
年份代号 x	1	2	3	4	5
人均可支配收入 y（万元）	0.8	1.1	1.5	2.4	3.7

(1)使用两种模型：①$\hat{y}=\hat{b}x+\hat{a}$；②$\hat{y}=\hat{m}x^2+\hat{n}$ 的相关指数 R^2 分别约为 0.92，0.99，请选择一个拟合效果更好的模型，并说明理由；

(2)根据（1）中选择的模型，试建立 y 关于 x 的回归方程．（保留 2 位小数）

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

$\hat{a}=\bar{y}-\hat{b}\bar{x}$．

参考数据：$\sum_{i=1}^{5}(x_i-\bar{x})(y_i-\bar{y})=7.1$，令 $u_i=x_i^2$，$\sum_{i=1}^{5}(u_i-\bar{u})(y_i-\bar{y})=45.1$．

11.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

参考数据及公式：$\sum_{i=1}^{5}x_i=880$，$\sum_{i=1}^{5}x_i^2=155450$，$\sum_{i=1}^{5}y_i=885$，$\sum_{i=1}^{5}x_iy_i=156045$，

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

(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 $\hat{e}_i=y_i-\hat{y}_i=y_i-b\hat{x}_i-\hat{a}$，$(i=1,2,\cdots,n)$，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、并探究这个结果是否对任意具有线性相关关系的两个变量都成立？若成立加以证明；若不成立说明理由．

12.随机选取变量 x 和变量 Y 的 5 对观测数据，选取的第 $i(i=1,2,3,4,5)$ 对观测数据记为 (x_i,y_i)，其数值对应如下表所示：

编号 i	1	2	3	4	5
x_i	9	8	7	6	5
y_i	75	95	110	135	150

计算得：$\bar{x}=\frac{1}{5}\sum_{i=1}^{5}x_i=7$，$\bar{y}=\frac{1}{5}\sum_{i=1}^{5}y_i=113$，$\sum_{i=1}^{5}x_i^2-5\bar{x}^2=10$，$\sum_{i=1}^{5}y_i^2-5\bar{y}^2=3630$，

$\sum_{i=1}^{5}x_iy_i=3765$．

(1)求变量 x 和变量 Y 的样本相关系数（小数点后保留 4 位），判断这两个变量是正相关还是负相关，并推断它们的线性相关程度；

(2)假设变量 Y 关于 x 的一元线性回归模型为 $\begin{cases}Y=bx+a+e\\E(e)=0,D(e)=\sigma^2\end{cases}$．

（i）求 Y 关于 x 的经验回归方程，并预测当 $x=10$ 时 Y 的值；

（ii）设 $\widehat{e_i}$ 为 $x=x_i\left(i=1,2,3,4,5\right)$ 时该回归模型的残差，求 $\widehat{e_1}$、$\widehat{e_2}$、$\widehat{e_3}$、$\widehat{e_4}$、$\widehat{e_5}$ 的方差.

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

13.身高体重指数(*BMI*)的大小直接关系到人的健康状况，某高中高三(1)班班主任为了解该班学生的身体健康状况，从该班学生中随机选取 5 名学生，测量其身高、体重的数据如下表.

学生编号	1	2	3	4	5
身高 *x*/cm	165	170	175	170	170
体重 *y*/kg	58	67	67	65	63

(1)求体重关于身高的线性回归方程，并预测身高为 180cm 的同学的体重；

(2)试分析学生的体重差异约有多少是由身高引起的？(注:结果保留两位小数)参考公式:线性回归方程 $\hat{y}=\hat{b}x+\hat{a}$ 中，$\hat{b}=\dfrac{\sum\limits_{i=1}^{n}x_iy_i-n\overline{xy}}{\sum\limits_{i=1}^{n}x_i^2-n\overline{x}^{-2}}=\dfrac{\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}$，$\hat{a}=\overline{y}-\hat{b}\overline{x}$，其中 $\overline{x}$，$\overline{y}$ 为样本平均值，$R^2=1-\dfrac{\sum\limits_{i=1}^{n}(y_i-\hat{y})^2}{\sum\limits_{i=1}^{n}(y_i-\overline{y})^2}$.

14.为了加快实现我国高水平科技自立自强，某科技公司逐年加大高科技研发投入.下图 1 是该公司 2013 年至 2022 年的年份代码 *x* 和年研发投入 *y*（单位：亿元）的散点图，其中年份代码 1~10 分别对应年份 2013~2022.

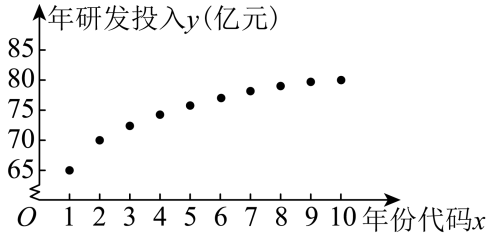


图1

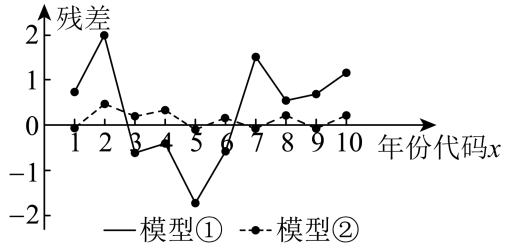


图2

根据散点图，分别用模型①$y=bx+a$，②$y=c+d\sqrt{x}$ 作为年研发投入 *y*（单位：亿元）关

于年份代码 x 的经验回归方程模型，并进行残差分析，得到图 2 所示的残差图.结合数据，计算得到如下表所示的一些统计量的值：

$\bar{y}$	$\bar{t}$	$\sum_{i=1}^{10}(x_i-\bar{x})^2$	$\sum_{i=1}^{10}(t_i-\bar{t})^2$	$\sum_{i=1}^{10}(y_i-\bar{y})(x_i-\bar{x})$	$\sum_{i=1}^{10}(y_i-\bar{y})(t_i-\bar{t})$
75	2.25	82.5	4.5	120	28.35

表中 $t_i=\sqrt{x_i}$，$\bar{t}=\frac{1}{10}\sum_{i=1}^{10}t_i$.

(1)根据残差图，判断模型①和模型②哪一个更适宜作为年研发投入 y（单位：亿元）关于年份代码 x 的经验回归方程模型?并说明理由；

(2)（i）根据（1）中所选模型，求出 y 关于 x 的经验回归方程；

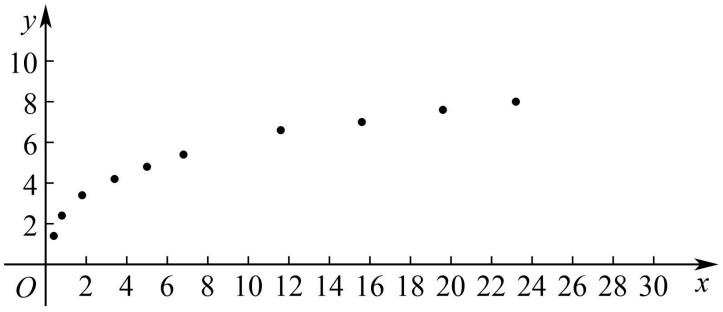
（ii）设该科技公司的年利润 L（单位：亿元）和年研发投入 y（单位：亿元）满足 $L=(111.225-y)\sqrt{x}$（$x\in\mathrm{N}^*$ 且 $x\in[1,20]$），问该科技公司哪一年的年利润最大?

附：对于一组数据 (x_1,y_1)，(x_2,y_2)，…，(x_n,y_n)，其经验回归直线 $\hat{y}=\hat{a}+\hat{b}x$ 的斜率和截距的最小二乘估计分别为 $\hat{b}=\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$，$\hat{a}=\bar{y}-\hat{b}\bar{x}$.

15.在正常生产条件下，根据经验，可以认为化肥的有效利用率近似服从正态分布 $N(0.54,0.02^2)$，而化肥施肥量因农作物的种类不同每亩也存在差异.

(1)假设生产条件正常，记 X 表示化肥的有效利用率，求 $P(X\geq 0.56)$;

(2)课题组为研究每亩化肥施用量与某农作物亩产量之间的关系，收集了 10 组数据，并对这些数据作了初步处理，得到了如图所示的散点图及一些统计量的值.其中每亩化肥施用量为 x（单位：公斤），粮食亩产量为 y（单位：百公斤）



参考数据：

$\sum_{i=1}^{10} x_i y_i$	$\sum_{i=1}^{10} x_i$	$\sum_{i=1}^{10} y_i$	$\sum_{i=1}^{10} x_i^2$	$\sum_{i=1}^{10} t_i z_i$	$\sum_{i=1}^{10} t_i$	$\sum_{i=1}^{10} z_i$	$\sum_{i=1}^{10} t_i^2$
650	91.5	52.5	1478.6	30.5	15	15	46.5

$t_i = \ln x_i$，　　$zi = \ln y_i (i = 1，2，$　　...，　　$10)$.

（i）根据散点图判断，　　$y = a + bx$与　　$y = cx^d$，哪一个适宜作为该农作物亩产量　　y关于每亩化肥施用量　　x的回归方程（给出判断即可，不必说明理由）；

（ii）根据（i）的判断结果及表中数据，建立　　y关于　　x的回归方程；并预测每亩化肥施用量为 27 公斤时，粮食亩产量　　y的值.　　$(e \approx 2.7)$

附：①对于一组数据　　$(u_i, v_i)(i = 1，2，3，$　　...，　　$n)$，其回归直线　　$\hat{v} = \hat{\beta}u + \hat{\alpha}$的斜率和截距的最小二乘估计分别为　　$\hat{\beta} = \frac{\sum_{i=1}^{n} u_i v_i - n\bar{u}\bar{v}}{\sum_{i=1}^{n} u_i^2 - n\bar{u}^2}$，　　$\hat{\alpha} = \hat{v} - \hat{\beta}\bar{u}$；

② 若随机变量　　$X \sim N(\mu, \sigma^2)$，则　　$P(\mu - \sigma < X < \mu + \sigma) \approx 0.6827$，$P(\mu - 2\sigma < X < \mu + 2\sigma) \approx 0.9545$.

独立性检验专项：

1．为学习贯彻中央农村工作会议精神“强国必先强农，农强方能国强”，某市在某村积极开展香菇种植，助力乡村振兴.香菇的生产可能受场地、基料、水分、菌种等因素的影响，现已知香菇有菌种甲和菌种乙两个品种供挑选，菌种甲在温度 20°C 时产量为 28 吨/亩，在温度 30℃时产量为 20 吨/亩；菌种乙在温度 20℃时产量为 22 吨/亩，在气温 30°C 时产量为 30 吨/亩.

(1)请补充完整 2×2 列联表，根据 2×2 列联表和小概率值 $\alpha = 0.1$ 的独立性检验，判断菌种甲、乙的产量与温度是否有关？

	20°C	30°C	合计

菌种甲			
菌种乙			
合计			

(2)某村选择菌种甲种植，已知菌种甲在气温为20°C时的发芽率为$\frac{5}{6}$，从菌种甲中任选 3 个，若设 X 为菌种甲发芽的个数，求 X 的分布列及数学期望.

附：参考公式：$\chi^2=\frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$，其中 $n=a+b+c+d$.

临界值表：

α	0.10	0.05	0.01
x_α	2.706	3.841	6.635

2.2023 年实行新课标新高考改革的省市共有 29 个，选科分类是高级中学在校学生生涯规划的重要课题，某高级中学为了解学生选科分类是否与性别有关，在该校随机抽取 100 名学生进行调查.统计整理数据得到如下的2×2列联表：

	选物理类	选历史类	合计
男生	35	15	
女生	25	25	
合计			100

(1)依据小概率值$\alpha=0.05$的独立性检验，能否据此推断选科分类与性别有关联？

(2)在以上随机抽取的女生中，按不同选择类别同比例分层抽样，共抽取 6 名女生进行问卷调查，然后在被抽取的 6 名女生中再随机抽取 4 名女生进行面对面访谈.设面对面访谈的女生中选择历史类的人数为随机变量X，求随机变量X的分布列和数学期望.

附：$\chi^2=\frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$，其中$n=a+b+c+d$.

α	0.10	0.05	0.025	0.010	0.005	0.001
x_α	2.706	3.841	5.024	6.635	7.879	10.828