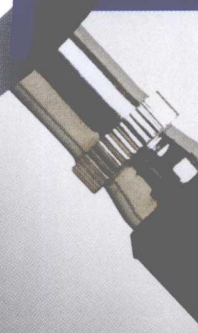
经纶学典

总主编○李朝东

修订版

教材解析

JIAOCAIJIEXI



YZLI0890147001

高中数学

选修 1-2

读者出版集团
DUZHE PUBLISHING GROUP
甘肃少年儿童出版社

经纶学典

总主编◎李朝东

教材解析

JIAOCAIJIEXI

本册主编：田志玲　牛海峰

人教A版

高中数学

选修1-2

YZLI0890147001

读者出版集团
DPGC.L
甘肃少年儿童出版社

图书在版编目(CIP)数据

教材解析:人教版.高中数学.1-2:选修/李朝东总主编.
—兰州:甘肃少年儿童出版社,2011.9
ISBN 978-7-5422-3010-2

Ⅰ.① 教… Ⅱ. ①李… Ⅲ. ①中学数学课-高中-教学参考资料 Ⅳ. ①G634

中国版本图书馆 CIP 数据核字(2011)第 173808 号

责任编辑:伏文东
封面设计:杭永鸿

教材解析·高中数学
选修 1-2 人教 A 版
李朝东 总主编
甘肃少年儿童出版社出版发行
(730030 兰州市读者大道 568 号)
0931-8773255
合肥市美格印务有限公司
开本 880 毫米×1230 毫米 1/16 印张 10.25 字数 205 千
2011 年 9 月第 1 版 2011 年 9 月第 1 次印刷
印数:1~5 000
ISBN 978-7-5422-3010-2 定价:21.00 元

前言

当一道道疑似难题摆在你面前时，是胸有成竹，还是找不着头绪？如果是前者，那恭喜你，你已经跨越了教材与考试之间的差距；如果是后者，那你也别急，《经纶学典·教材解析》在教材与考试间为你搭建一个沟通平台。

不少同学有这样的感觉：教材都熟悉了，课堂上也听懂了，但考试却取不到好成绩。原因在于教材内容与考试要求有差距，课堂教学与选拔性考试有差别。这就需要在教材之上、课堂之外能够得到补充、提升，直至达到高考的选拔要求。本书就是从以下两个方面填补这种差距。

首先是对教材的深度挖掘。教材内容通俗易懂，但里面包含着丰富的信息，我们把教材所包含的信息挖掘出来，并进行系统整理，让知识内涵和外延、知识间的联系充分展现。

第二是对课堂教学的补充和拓展。本书不是对课堂教学的重复，而是在课堂教学基础上，对课堂教学进行补充、提高，挖掘那些学生难以理解、难以掌握的内容，进行归纳和总结，为学生串起一条规律性的“线”。数学侧重解题方法、解题技巧、解题思路的整理，注重方法的拓展，找出最优的解题方法，对本节内容与其他小专题内容进行归纳总结。这些由于课堂教学时间限制或教师水平发挥的问题，在课堂上并没有全部传授给学生，而这些恰恰就是考试中要考查的，学生拉开差距的所在。

正是本着上述编写理念，本丛书以学生为中心，用最易理解的表现形式呈现学习中难以理解的部分。希望本书为你的成长助力，有更好的想法和意见请登录：www.jing-lun.cn。

编 者

QIANYAN

读者反馈表

尊敬的读者：

您好！感谢您使用《经纶学典·教材解析》！

为了不断提高图书质量，恳请您写下使用本书的体会与感受，我们将真诚地吸纳。**在修订时将刊登您的意见，并予以一定的奖励，以表达我们诚挚的谢意。**

<table>
<tr><td rowspan="3">读者简介</td><td>姓　名</td><td></td><td>性　别</td><td></td><td>出生年月</td><td colspan="2"></td></tr>
<tr><td>所在学校</td><td colspan="2"></td><td>通讯地址</td><td colspan="3"></td></tr>
<tr><td>联系方式</td><td colspan="5">(H)：　　(O)：
手机：　　E－mail：</td></tr>
<tr><td colspan="2">本书情况</td><td>学科</td><td></td><td>版本</td><td></td><td>年级</td><td></td></tr>
<tr><td colspan="3">您对本书栏目的评价：
1. 教材梳理：
全面□　一般□　不全面□
2. 教材拓展：
难□　合理□　易□
3. 典型题解：
全面□　不全面□
4. 针对性练习：
难□　合理□　易□
5. 拓展阅读：
需要□　不需要□
6. 五年高考回放：
需要□　不需要□</td><td colspan="3">您对本书体例形式的评价：
1. 栏目设置：
过多□　适中□　过　少□
2. 题空：
过大□　正好□　过　小□
3. 版式：
美观□　一般□　不美观□
4. 封面：
美观□　一般□　不美观□</td><td colspan="2">您的购买行为：
1. 您购买本书的途径：
广　告□　教师推荐□
家长购买□　学校统一购买□
自己购买□　同学推荐□
2. 您购买本书的主要原因（可多选）：
广告宣传□　包装形式□
内　容□　图书价格□
封面设计□　书　名□</td></tr>
<tr><td colspan="8">您对本书的其他意见：</td></tr>
</table>

欢迎登录：www. jing-lun. cn

通信地址：南京红狐教育传播研究所（南京市租用 16－02#信箱）　　**邮编：**210016

目录

MULU

第一章 统计案例

1.1 回归分析的基本思想及其初步应用

A 教材梳理

知识点一 回归分析

1. 回归分析是处理变量之间相关关系的一种统计方法. 若两个变量之间具有线性相关关系,则称相应的回归分析为线性回归分析.

2. 散点图

当确定两个事物之间有相关关系后,要先作一张相关图,判断事物间是否线性相关,然后才能计算相关系数,在相关图中,横坐标代表一个变量,纵坐标代表另一个变量,将各对应量依次用坐标点绘于图上,这个图便称为散点图.

散点图可以说明变量间有无线性相关关系、相关的方向,但不能精确地说明两个变量之间关系的密切程度,因此需要计算相关系数来描述两个变量之间关系的密切程度.

3. 随机误差

从散点图中我们可以看到,样本点分布在某一条直线的附近,而不是在一条直线上,所以不能用一次函数 $y=bx+a$ 来描述它们之间的关系,例如我们把身高与体重的关系用线性回归模型 $y=bx+a+e$ 来表示,其中 a,b 为模型的未知参数,e 称为随机误差. 在实际问题中,一个人的体重除了受身高的影响外,还受许多其他因素的影响,例如饮食习惯、是否喜欢运动、度量误差等,而且我们选用的线性回归模型往往只是一种近似的模型,所有这些因素都会导致随机误差 e 的产生.

4. 回归直线方程 $y=\hat{b}x+\hat{a}$,其中

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

$\hat{a}=\bar{y}-\hat{b}\bar{x}\left(\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i\right)$. $(\bar{x},\bar{y})$ 称为样本点的中心. 回归直线过样本点的中心.

5. 线性回归分析

一般情况下,在尚未断定两个变量之间是否具有线性相关关系的情况下,应先进行相关性检验,在确认具有线性相关关系后,再求回归直线方程. 回归分析的一般步骤为:

(1)从一组数据出发,求出两个变量的相关系数 r,确定两者之间是否具有线性相关关系.

(2)如果具有线性相关关系,求出回归方程 $\hat{y}=\hat{b}x+\hat{a}$. 其中 $\hat{a}$ 是常数项,$\hat{b}$ 是回归系数.

(3)根据回归方程,由一个变量的值,预测或控制另一个变量的值.

6. 在利用回归直线方程进行预报时,要注意下列问题:

(1)回归方程只适用我们所研究的样本的总体.

(2)所建立的回归方程一般都有时间性.

(3)样本取值的范围会影响回归方程的适用范围.

(4)不能期望回归方程得到的预报值就是预报变量的精确值,事实上,它是预报变量的可能取值的平均值.

知识点二 相关系数与相关性检验

1. 相关系数 $r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}=\frac{\sum_{i=1}^{n}x_iy_i-n\bar{x}\,\bar{y}}{\sqrt{(\sum_{i=1}^{n}x_i^2-n\bar{x}^2)(\sum_{i=1}^{n}y_i^2-n\bar{y}^2)}}$,用它来衡量两个变量之间的线性相关关系.

2. 当 $r>0$ 时,表明两个变量正相关;当 $r<0$ 时,表明两个变量负相关. r 的绝对值越接近 1,表明两个变量的线性相关性越强;r 的绝对值接近于 0 时,表明两个变量之间几乎不存在线性相关关系,通常当 $|r|$ 大于 0.75 时,认为两个变量有很强的线性相关关系.

知识点三 回归模型的误差分析

1. 在数学上,把每个效应(观测值减去总的平均值)的平

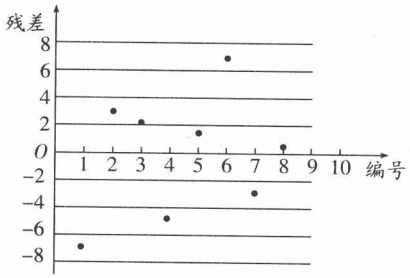
方加起来，即用 $\sum_{i=1}^{n}(y_i-\bar{y})^2$ 表示总的效应，称为总偏差平方和.

2. 数据点和它在回归直线上相应位置的差异 $(y_i-\hat{y}_i)$ 是随机误差的效应，称 $\hat{e}_i=y_i-\hat{y}_i$ 为残差，将所得的值平方后加起来，用数学符号表示为 $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$，称为残差平方和，它代表了随机误差的效应.

3. 用相关指数 R^2 来刻画回归的效果，其计算公式是 $R^2=1-\dfrac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$. R^2 的值越大，说明残差平方和越小，也就是说模型的拟合效果越好. 在线性回归模型中，R^2 表示解释变量对预报变量变化的贡献率. R^2 越接近于1，表示回归的效果越好(因为 R^2 越接近于1，表示解释变量和预报变量的线性相关性越强).

4. 在研究两个变量间的关系时，首先要根据散点图来粗略判断它们是否线性相关，是否可以用线性回归模型来拟合数据，然后，通过残差 $\hat{e}_1,\hat{e}_2,\cdots,\hat{e}_n$ 来判断模型拟合的效果，判断原始数据中是否存在可疑数据，这方面的分析工作称为残差分析.

5. 利用图形来分析残差特性，作图时纵坐标为残差，横坐标可以选为样本编号，或身高数据，或体重估计值等，这样作出的图形称为残差图，如图是以样本编号为横坐标的残差图.



从图中可以看出，第1个样本点和第6个样本点的残差比较大，需要确认在采集这两个样本点的过程中是否有人为的错误，如果数据采集有错误，就予以纠正，然后再重新利用线性回归模型拟合数据；如果数据采集没有错误，则需要寻找其他的原因. 另外，残差点比较均匀地落在水平的带状区域中，说明选用的模型比较合适，这样的带状区域的宽度越窄，说明模型拟合精度越高，回归方程的预报精度越高.

知识点四　建立回归模型的基本步骤

一般地，建立回归模型的基本步骤为：

(1)确定研究对象，明确哪个变量是解释变量，哪个变量是预报变量.

(2)画出确定好的解释变量和预报变量的散点图，观察它们之间的关系(如是否存在线性关系等).

(3)由经验确定回归方程的类型(如观察到数据呈线性关系，则选用线性回归方程 $y=bx+a$).

(4)按一定规则估计回归方程中的参数(如最小二乘法).

(5)得出结果后分析残差图是否有异常(个别数据对应残差过大，或残差呈现不随机的规律性等等)，若存在异常，则检查数据是否有误，或模型是否合适等.

B 教材拓展

拓展点一　相关关系与函数关系

1. 两者之间的不同点

(1)相关关系是一种非确定性关系，即相关关系是非随机变量与随机变量之间的关系. 如人的身高与年龄；商品的销售额与广告费等都是相关关系，而函数关系中的两个变量是一种确定性关系. 如正方形的面积 S 与边长 x 之间的关系 $S=x^2$ 就是函数关系，即对于边长 x 的每一个确定的值，都有面积 S 的唯一确定的值与之对应.

(2)函数关系是一种因果关系，而相关关系不一定是因果关系. 如有人发现，对于在校儿童，身高与阅读能力有很强的相关关系，然而学会新词并不能使儿童马上长高，而是涉及第三个因素——年龄，当儿童长大一些，他们的阅读能力会提高，而由于长大身高也会高一些.

2. 两者之间的联系

相关关系与函数关系有着密切的联系，在一定条件下可以相互转化. 例如正方形的面积 S 与其边长 x 之间虽然是一种确定性关系，但在每次测量时，由于测量误差等原因，其数值大小又表现出一种随机性，而对于具有线性关系的两个变量来说，当求得其回归直线方程后，我们又可以用一种确定的关系对这两个变量间的关系进行估计.

例题 1　在下列各组量中：①正方体的体积与棱长；②一块农田的水稻产量与施肥量；③人的身高与年龄；④家庭的支出与收入；⑤某户家庭的用电量与电价. 其中量与量之间的关系是相关关系的是　(　　)

A. ①②　　B. ③④

C. ③④　　D. ②③④

[解析]　①是函数关系 $V=a^3$；⑤电价是统一规定的，与用电量有一定的关系，但这种关系是确定的关系；②③④中的两个量之间的关系都是相关关系，因为水稻产量与施肥量在一定范围内是正比，反比或其他关系，并不确定；人的身高一开始随着年龄的增加而增大，之后则不变化或降低，在身高增大

时,也不是均匀增大的;家庭的支出与收入有一定的关系,在一开始,支出会随着收入的增加而增加,而当收入增大到一定的值后,家庭支出趋向于一个常数值,也不是确定关系.

[答案] D

拓展点二 线性回归模型

1. 线性回归模型研究的是具有线性相关关系的样本点,而不是一次函数的关系. 用最小二乘法求出线性回归直线方程,相关系数 r 反映两个变量之间相关关系的强弱;用相关指数 R^2 来刻画回归的效果,即模型的拟合效果;再通过残差分析纠正错误数据,重新利用线性回归模型拟合数据,并预测或控制预报变量.

2. 随机误差 e 是通过残差及残差平方和反映出来的. 也就是样本中的每个数据点和它在回归直线上的相应位置的差异,就是一个随机误差——残差,样本中所有样本点的残差的平方和代表了随机误差的效应.

3. 产生随机误差项 e 的原因

随机误差也叫残差变量,产生的原因是:(1)用线性回归模型近似真实模型(真实模型是客观存在的,通常我们并不知道真实模型到底是什么)所引起的误差. 可能存在非线性的函数能够更好地描述 y 与 x 之间的关系,但是现在却用线性函数来表述这种关系,结果会产生误差. 这种由模型近似所引起的误差是 e 的一部分;(2)忽略了某些因素的影响,通常影响变量 y 的因素不只是变量 x,可能还包括其他许多因素,而为了研究问题的方便,常常忽略一些次要因素,由此而产生误差;(3)观测引起的误差,即由于测量工具等原因,导致 y 的观察值产生误差.

4. 如何刻画预报变量的变化? 这个变化在多大程度上与解释变量有关? 在多大程度上与随机误差有关?

预报变量的变化程度可以分解为解释变量引起的变化程度与残差变量的变化程度之和. 其中这个变化与解释变量和随机误差(即残差平方和)有关的程度是由相关指数 R^2 的值决定的. 在线性回归模型中,R^2 表示解释变量对预报变量变化的贡献率. R^2 越接近于 1,表示解释变量和预报变量的线性相关性越强;反之,R^2 越小,随机误差对预报变量的效应越大.

拓展点三 回归系数

1. 回归系数 $\hat{a}$、$\hat{b}$ 的简单推导

对于一组具有线性相关关系的数据 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$,其回归方程的截距和斜率的最小二乘估计公式分别为:

$$\hat{a}=\bar{y}-\hat{b}\bar{x}, \qquad ①$$

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

其推导过程为:

我们知道,截距 $\hat{a}$ 和斜率 $\hat{b}$ 分别是使 $Q(a,b)=\sum_{i=1}^{n}(y_i-bx_i-a)^2$ 取最小值时 a,b 的值.

$$\begin{aligned}\because Q(a,b)&=\sum_{i=1}^{n}[y_i-bx_i-(\bar{y}-b\bar{x})+(\bar{y}-b\bar{x})-a]^2\\&=\sum_{i=1}^{n}\{[y_i-bx_i-(\bar{y}-b\bar{x})]^2+2[y_i-bx_i-(\bar{y}-b\bar{x})]\cdot[(\bar{y}-b\bar{x})-a]+[(\bar{y}-b\bar{x})-a]^2\}\\&=\sum_{i=1}^{n}[y_i-bx_i-(\bar{y}-b\bar{x})]^2+2\sum_{i=1}^{n}[y_i-bx_i-(\bar{y}-b\bar{x})]\cdot(\bar{y}-b\bar{x}-a)+n(\bar{y}-b\bar{x}-a)^2,\end{aligned}$$

注意到 $\sum_{i=1}^{n}[y_i-bx_i-(\bar{y}-b\bar{x})](\bar{y}-b\bar{x}-a)$

$$\begin{aligned}&=(\bar{y}-b\bar{x}-a)\sum_{i=1}^{n}[y_i-bx_i-(\bar{y}-b\bar{x})]\\&=(\bar{y}-b\bar{x}-a)\left[\sum_{i=1}^{n}y_i-b\sum_{i=1}^{n}x_i-n(\bar{y}-b\bar{x})\right]\\&=(\bar{y}-b\bar{x}-a)[n\bar{y}-nb\bar{x}-n(\bar{y}-b\bar{x})]\\&=0,\end{aligned}$$

$$\begin{aligned}\therefore Q(a,b)&=\sum_{i=1}^{n}[y_i-bx_i-(\bar{y}-b\bar{x})]^2+n(\bar{y}-b\bar{x}-a)^2\\&=b^2\sum_{i=1}^{n}(x_i-\bar{x})^2-2b\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})+\sum_{i=1}^{n}(y_i-\bar{y})^2+n(\bar{y}-b\bar{x}-a)^2\\&=n(\bar{y}-b\bar{x}-a)^2+\sum_{i=1}^{n}(x_i-\bar{x})^2\left[b-\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]^2-\\&\quad\frac{\left[\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right]^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}+\sum_{i=1}^{n}(y_i-\bar{y})^2.\end{aligned}$$

在上式中,后两项和 a,b 无关,而前两项为非负数,因此要使 Q 取得最小值,当且仅当前两项的值均为 0,即有

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

$$\hat{a}=\bar{y}-\hat{b}\bar{x},$$

这正是我们要推导的公式.

2. 求回归系数 $\hat{a}$、$\hat{b}$ 的具体步骤和方法

(1)列表,将所给的数据 x、y 列成相应的表格. 如下表所示:

序号	x_i	y_i	x_i^2	y_i^2	x_iy_i
1	x_1	y_1	x_1^2	y_1^2	x_1y_1

2	x_2	y_2	x_2^2	y_2^2	x_2y_2
…	…	…	…	…	…
n	x_n	y_n	x_n^2	y_n^2	x_ny_n
Σ	Σx_i	Σy_i	Σx_i^2	Σy_i^2	Σx_iy_i

(2)计算$\bar{x}$,$\bar{y}$,$\sum_{i=1}^{n}x_i^2$,$\sum_{i=1}^{n}y_i^2$,$\sum_{i=1}^{n}x_iy_i$.

(3)代入公式计算$\hat{b}$,$\hat{a}$的值.

拓展点四　残差分析与回归模型的选取

1. 残差分析主要用于检查样本数据中的异常点,以保证回归模型具有较高的拟合精度. 注意所建立的回归方程的适用条件和范围,以保证预测值对预报变量的较好的精确度.

2. 残差图的效果与选用的横坐标无关. 作残差分析时,一般从以下几个方面予以说明:(1)散点图;(2)相关系数r;(3)相关指数R^2;(4)残差图中的异常样本点和样本点的带状分布区域的宽窄.

例题 2　一个车间为了规定工时定额,需要确定加工零件所花费的时间,为此进行了10次试验,测得数据如下:

零件数x(个)	10	20	30	40	50	60	70	80	90	100
加工时间y(分)	62	68	75	81	89	95	102	108	115	122

(1)计算总偏差平方和,残差及残差平方和;

(2)求出相关指数R^2;

(3)作出残差图;

(4)进行残差分析.

[**答案**]　解:(1)列出残差表(由知识点一,通过计算可知:$\hat{y}=0.668x+54.96$,$\bar{y}=91.7$)

y_i	62	68	75	81	89	95	102	108	115	122
$\hat{y}_i$	61.6	68.3	75.0	81.7	88.4	95.0	101.7	108.4	115.1	121.8
$y_i-\bar{y}$	-29.7	-23.7	-16.7	-10.7	-2.7	3.3	10.3	16.3	23.3	30.3
$y_i-\hat{y}_i$	0.40	-0.30	0	-0.70	0.60	0	0.30	-0.40	-0.10	0.20

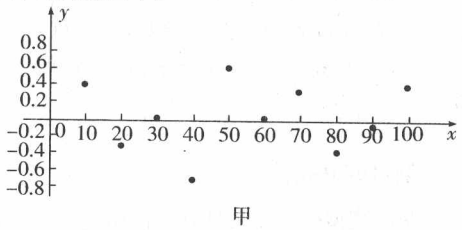
$$\therefore \sum_{i=1}^{10}(y_i-\bar{y})^2=(-29.7)^2+(-23.7)^2+\cdots+30.3^2$$
$$=3\,688.1.$$

$$\sum_{i=1}^{10}(y_i-\hat{y}_i)^2=0.40^2+(-0.30)^2+\cdots+0.24^2$$
$$=1.4.$$

即总偏差平方和为3 688.1,残差平方和为1.4,残差值如表中第四行的值.

(2)$R^2=1-\frac{1.4}{3\,688.1}\approx1-0.000\,38=0.999\,62$,相关指数$R^2$非常接近于1,回归直线模型拟合效果较好.

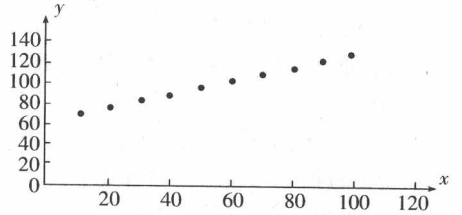
(3)作出残差图甲:



甲

图甲:横坐标为零件个数,纵坐标为残差.

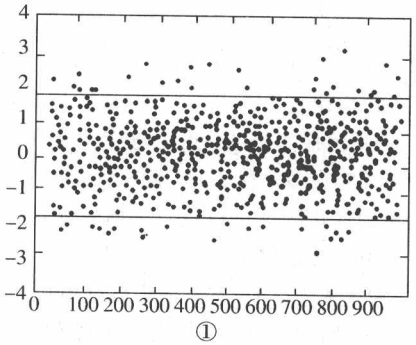
(4)残差分析:



乙

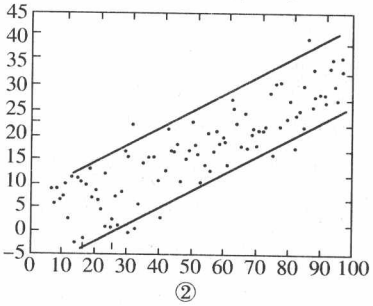
由散点图乙和r的值(由知识点二,通过计算有$r=0.999\,8$)可以说明x与y有很强的相关性,由R^2的值可以看出回归直线模型的拟合效果很好. 由残差图可以观察到,第4个样本点和第5个样本点的残差比较大,需要确认在采集这两个样本点的过程中是否有人为的失误,如果有则需要纠正数据,重新利用线性回归模型拟合数据,由残差图中的残差点比较均匀地落在水平的带状区域中(在两条直线$y=-0.65$和$y=0.67$之间),也说明选用的线性回归模型较为合适,带状区域的宽度仅为1.32,比较狭窄,说明模型拟合精度较高.

3. 几种常见的残差图如下所示:



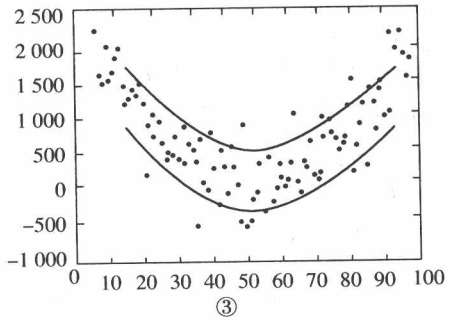
①

图①:残差散点图中的点分布在以原点为中心的水平带形区域上,并且沿水平方向散点的分布规律相同,说明残差是随机的,所选择的回归模型建模是合理的.



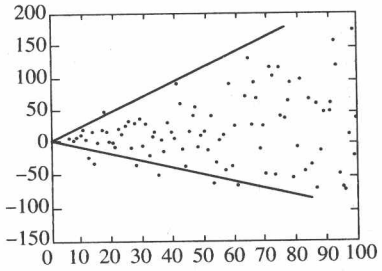
②

图②:残差散点图中的点分布在一条倾斜的带形区域上,并且沿带形区域方向散点的分布规律相同,说明残差与横坐标有线性关系,此时所选择的回归模型的效果不是最好的,有改进的余地.



③

图③:残差散点图中的点分布在一条二次曲线的弯曲带形区域上,说明残差与坐标横轴变量有二次关系,此时所选用的回归模型的效果不是最好的,有改进的余地.



④

图④:残差散点图中的点的分布范围随着横坐标的增加而增加,说明残差的方差与坐标横轴变量有关,不是一个常数,此时所选用的回归模型的效果不是最好的,有改进的余地.

4. 残差平方和代表了随机误差的效应

解释变量的效应叫回归平方和(regression sum of squares),即

回归平方和 = 总偏差平方和 - 残差平方和

$$=\sum_{i=1}^{n}(y_i-\bar{y})^2-\sum_{i=1}^{n}(y_i-\hat{y}_i)^2.$$

由此可以得到两个比值:残差平方和与回归平方和占总偏差平方和的百分比,即

$$\frac{\text{残差平方和}}{\text{总偏差平方和}}=\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

$$\frac{\text{回归平方和}}{\text{总偏差平方和}}=\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2-\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

$$=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}.$$

5. 预报变量受解释变量和随机误差的影响,影响的大小是用残差平方和与回归平方和各占总偏差平方和的百分比来表示的,这两个百分比的和为1.

残差对预报变量影响越小,则回归模型的拟合效果越好.

6. 对于给定的样本(样本点为$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$),如果建立了含有未知参数的两个回归模型:$\hat{y}_1=f(x)$,$\hat{y}_2=g(x)$,可以按如下步骤来比较它们的拟合效果:

(1)分别建立对应于两个回归模型的回归方程$\hat{y}_1=f(x)$,$\hat{y}_2=g(x)$,即求出回归方程中未知参数的估计值.

(2)计算出两个回归方程中各自的残差平方和:$\hat{Q}_1=\sum_{i=1}^{n}(y_i-\hat{y}_{1i})^2$与$\hat{Q}_2=\sum_{i=1}^{n}(y_i-\hat{y}_{2i})^2$.(也可以求出相关指数$R^2$)

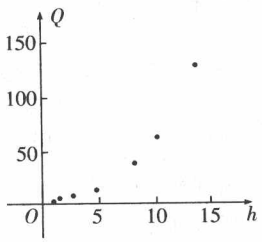
(3)若$\hat{Q}_1<\hat{Q}_2$,则$\hat{y}_1=f(x)$的拟合效果比$\hat{y}_2=g(x)$的好;反之则是$\hat{y}_2=g(x)$的拟合效果比$\hat{y}_1=f(x)$的好.

例题 3 有一个测量水流量的实验装置,测得试验数据如下表:

i	1	2	3	4	5	6	7
水高h(cm)	0.7	1.1	2.5	4.9	8.1	10.2	13.5
流量Q(L/min)	0.082	0.25	1.8	11.2	37.5	66.5	134

根据表中数据,建立Q与h之间的回归方程.

[答案] 解:由表中测得的数据可以作出散点图,如图所示.



观察散点图中样本点的分布规律,可以判断出样本点分布在某一条曲线上,表示该曲线的函数模型是$Q=m\cdot h^n$(m,n是正常数).

两边取常用对数,则$\lg Q=\lg m+n\cdot\lg h$,令$y=\lg Q$,$x=\lg h$,那么:$y=nx+\lg m$即为线性函数模型$y=bx+a$(其中$b=n$,$a=\lg m$).

由下面的数据表,用最小二乘法可求得

$\hat{b}\approx 2.509\,7$,$\hat{a}\approx -0.707\,7$,

$\therefore n\approx 2.509\,7$,$m\approx 0.196$.

于是所求得的回归方程为:$Q\approx 0.196\cdot h^{2.51}$.

i	h_i	Q_i	$x_i=\lg h_i$	$y_i=\lg Q_i$	x_i^2	x_iy_i
1	0.7	0.082	−0.154 9	−1.086 2	0.024 0	0.168 3
2	1.1	0.25	0.041 4	−0.602 1	0.001 7	−0.024 9
3	2.5	1.8	0.397 9	0.255 3	0.158 3	0.101 6
4	4.9	11.2	0.690 2	1.049 2	0.476 4	0.724 2
5	8.1	37.5	0.908 5	1.574 0	0.825 4	1.430 0
6	10.2	66.5	1.008 6	1.822 8	1.017 3	1.838 5
7	13.5	134	1.130 3	2.127 1	1.277 6	2.404 3
Σ			$\sum_{i=1}^{7}x_i$ =4.022	$\sum_{i=1}^{7}y_i$ =5.140 1	$\sum_{i=1}^{7}x_i^2$ =3.780 7	$\sum_{i=1}^{7}x_iy_i$ =6.642

[点评] 在建立回归方程时，选择合适的函数类型很重要，需要熟练掌握学过的函数，如反比例函数，二次函数，指数型函数，对数型函数，幂函数等模型，再进行线性转换，最后求得回归模型的方程.

拓展点五　易错点、易忽略点归纳

求回归直线方程，应注意首先对样本点是否线性相关进行检验，因为对于任何一组样本点，都可以根据最小二乘法求得一个回归直线方程，但这条回归直线方程是否较好地反映了样本点的分布呢，显然不一定，特别是对于不呈线性相关的回归模型. 可以通过散点图或求相关系数 r 首先作出是否线性相关的检验，然后再选择恰当的回归模型进行模拟.

例题 4　在一次抽样调查中测得样本的 5 个样本点，数值如下表：

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

试建立 y 与 x 之间的回归方程.

[错解] 解：由已知条件列表如下：

序号	x_i	y_i	x_iy_i	x_i^2	y_i^2
1	0.25	16	4	0.062 5	256
2	0.5	12	6	0.25	144
3	1	5	5	1	25
4	2	2	4	4	4
5	4	1	4	16	1
Σ	7.75	36	23	21.312 5	430

$\therefore \bar{x}=1.55, \bar{y}=7.2.$

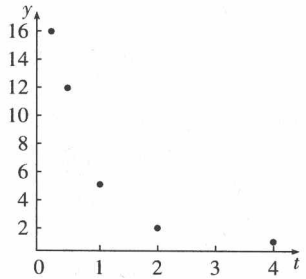
$$\hat{b}=\frac{\sum_{i=1}^{5}x_iy_i-5\bar{x}\,\bar{y}}{\sum_{i=1}^{5}x_i^2-5(\bar{x})^2}\approx -3.53.$$

$\hat{a}=\bar{y}-\hat{b}\bar{x}\approx 12.67.$

所求的 y 与 x 之间的回归方程是

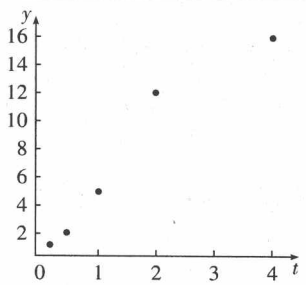
$\hat{y}=-3.53x+12.67.$

[错解分析] 本题直接取已知数据求回归直线方程，没有画出散点图或求相关系数 r，进行线性相关性的检验，而本题的样本点恰好不是线性相关的. 根据散点图可以发现 y 与 x 近似地呈反比例函数关系，即 $y=\frac{k}{x}$ 的关系（如图），令 $t=\frac{1}{x}$，则 $y=kt$，即 y 与 $\frac{1}{x}$ 呈线性相关的关系.



[正解] 解：根据散点图可知 y 与 x 近似地呈反比例函数关系，设 $y=\frac{k}{x}$，令 $t=\frac{1}{x}$，则 $y=kt$，原数据变为：

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



由散点图也可以看出 y 与 t 呈近似的线性相关关系. 列表如下：

序号	t_i	y_i	t_iy_i	t_i^2	y_i^2
1	4	16	64	16	256
2	2	12	24	4	144
3	1	5	5	1	25
4	0.5	2	1	0.25	4
5	0.25	1	0.25	0.062 5	1
Σ	7.75	36	94.25	21.312 5	430

$\therefore \bar{t}=1.55, \bar{y}=7.2.$

$$\hat{b}=\frac{\sum_{i=1}^{5}t_iy_i-5\bar{t}\,\bar{y}}{\sum_{i=1}^{5}t_i^2-5(\bar{t})^2}\approx 4.134\,4.$$

$\hat{a}=\bar{y}-\hat{b}\bar{t}\approx 0.8$.

$\therefore \hat{y}=4.1344t+0.8$.

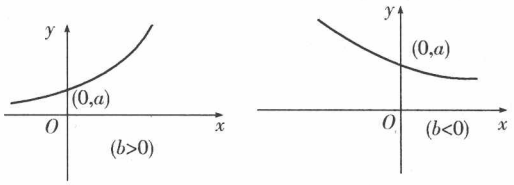
$\therefore y$ 与 x 的回归方程是 $\hat{y}=\dfrac{4.1344}{x}+0.8$.

拓展点六　几种常见非线性回归模型

在大量的实际问题中，研究的两个变量不一定都呈线性相关关系，它们之间可能呈指数关系或对数关系等非线性关系. 在某些情况下可以借助线性回归模型研究呈非线性关系的两个变量的关系.

(1)指数函数型 $y=ae^{bx}(a>0)$

①函数 $y=ae^{bx}(a>0)$ 的图象，如图所示.



②处理方法

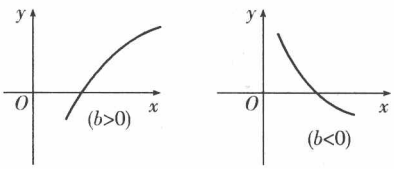
两边取对数得 $\ln y=\ln(ae^{bx})$，即 $\ln y=\ln a+bx$.

设 $\begin{cases}y'=\ln y,\\x'=x,\end{cases}$ 则原方程变成 $y'=\ln a+bx'$.

具体计算时，先将原数据点 (x_i,y_i) 转化成 $(x_i,\ln y_i)$，$i=1,2,\cdots,n$，再根据一次线性回归模型的方法得出 $\ln a$ 和 b.

(2)对数函数型 $y=a+b\ln x$

①函数 $y=a+b\ln x$ 的图象，如图所示.



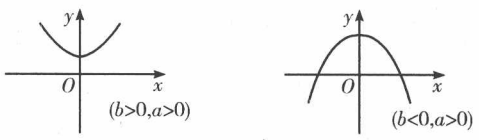
②处理方法

设 $\begin{cases}x'=\ln x,\\y'=y,\end{cases}$ 则原方程就转化成 $y'=a+bx'$，

然后按一次线性回归模型求出 a,b 的值.

(3)$y=bx^2+a$ 型

①函数 $y=bx^2+a$ 的图象，如图所示.



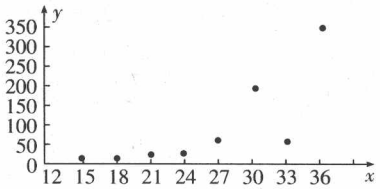
②处理方法

设 $\begin{cases}x'=x^2,\\y'=y,\end{cases}$ 则原方程就转化成 $y'=bx'+a$，然后按一次线性回归模型求出 a,b 的值.

例题 5　在某一化学反应过程中，其化学物质的反应速度 y(g/min)与一种催化剂的量 x(g)有关，现收集了 8 组测验数据列于下表中，试建立 y 与 x 之间的回归方程.

催化剂量 x(g)	15	18	21	24	27	30	33	36
物质反应速度 y(g/min)	6	8	30	27	70	205	65	350

[答案]　解：由表中的数据可以作出散点图，如图所示.

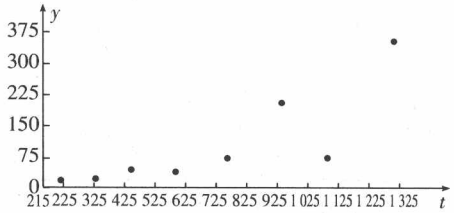


根据样本点的分布情况，可以选用两种曲线模型来拟合.

(1)可以认为样本点集中在某二次曲线 $y=c_1x^2+c_2$ 的附近. 令 $t=x^2$，则 $y=c_1t+c_2$，即线性回归模型：$y=bt+a(b=c_1,a=c_2)$，由此得 t 与 y 的样本数据表：

t	225	324	441	576	729	900	1 089	1 296
y	6	8	30	27	70	205	65	350

画出 y 与 t 的散点图，如图所示.



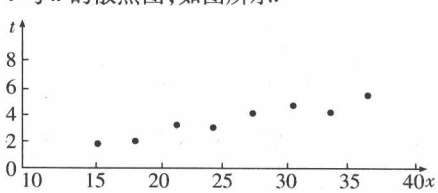
由 y 与 t 的散点图可以观察到样本数据点并不分布在一条直线的附近，样本点的分布仍然呈现曲线的变化趋势，因此不宜用 $y=bt+a$ 来拟合，即不宜用二次曲线 $y=c_1x^2+c_2$ 来拟合 y 与 x 之间的关系.

(2)根据 y 与 x 的散点图，也可以认为样本点集中在某一条指数函数曲线 $y=c_1e^{c_2x}$ 的附近，令 $t=\ln y$，则 $t=\ln y=c_2x+\ln c_1$，即变换变量后样本点应分布在直线 $t=bx+a(a=\ln c_1,b=c_2)$ 附近.

t 与 x 的关系的数据如下表所示：

x	15	18	21	24	27	30	33	36
t	1.792	2.079	3.401	3.296	4.248	5.323	4.174	5.858

画出 t 与 x 的散点图，如图所示.



由散点图可观察到样本点大致在一条直线上，所以可以用线性回归方程来拟合它.

用最小二乘法可得 t 与 x 的回归方程是：

$\hat{t}=0.1812x-0.8485$，

所求的非线性回归方程是 $\hat{y}=e^{0.1812x-0.8485}$.

即 y 与 x 之间的回归方程是 $\hat{y}=e^{0.1812x-0.8485}$.

[点评] 对于用二次曲线拟合时，也可以求出 y 与 x 之间的回归方程，但其拟合效果不好. 还可以计算出这两种模型的残差平方和，从而可看出指数型函数模型的拟合效果好. 计算它们的相关指数 R^2，也可以比较出这两种回归模型的优劣.

C 典型题解

▶问题一 线性回归方程

例题 1 某工厂 1～8 月份某种产品的产量与成本的统计数据见下表：

月份	1	2	3	4	5	6	7	8
产量（吨）	5.6	6.0	6.1	6.4	7.0	7.5	8.0	8.2
成本（万元）	130	136	143	149	157	172	183	188

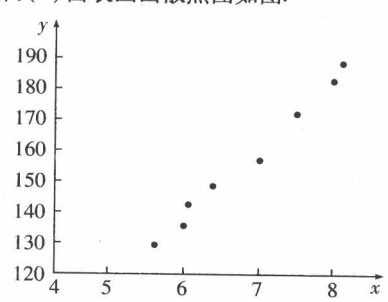
以产量为 x，成本为 y.

(1)画出散点图；

(2)y 与 x 是否具有线性相关关系？若有，求出其回归方程.

[解析] 本题主要考查对两个变量的回归分析，画出散点图，代入回归系数公式即可得.

[答案] 解：(1)由表画出散点图如图.



(2)从图上可看出，这些点基本上散布在一条直线附近，可以认为 x 和 y 线性相关关系显著，下面求其回归方程，首先列出下表.

序号	x_i	y_i	x_i^2	y_i^2	x_iy_i
1	5.6	130	31.36	16 900	728.0
2	6.0	136	36.00	18 496	816.0
3	6.1	143	37.21	20 449	872.3
4	6.4	149	40.96	22 201	953.6
5	7.0	157	49.00	24 649	1 099.0
6	7.5	172	56.25	29 584	1 290.0
7	8.0	183	64.00	33 489	1 464.0
8	8.2	188	67.24	35 344	1 541.6
Σ	54.8	1 258	382.02	201 112	8 764.5

$\bar{x}=6.85$，$\bar{y}=157.25$.

$$\therefore \hat{b}=\frac{\sum_{i=1}^{8}x_iy_i-8\bar{x}\,\bar{y}}{\sum_{i=1}^{8}x_i^2-8\bar{x}^2}$$

$$=\frac{8\,764.5-8\times6.85\times157.25}{382.02-8\times6.85^2}$$

≈22.17，

$\hat{a}=\bar{y}-\hat{b}\bar{x}=157.25-22.17\times6.85$

≈5.39，

故线性回归方程为 $\hat{y}=22.17x+5.39$.

[点评] (1)回归分析是定义在具有相关关系的两个变量基础上的，对于关系不明确的两组数据，可先作散点图，在图上看它们有无关系及关系的密切程度，然后再进行相关回归分析.

(2)求回归直线方程时，首先应注意到，只有在散点图大致呈线性分布时，求出的回归直线方程才有实际意义，否则，求出的回归直线方程毫无意义.

例题 2 假设关于某设备的使用年限 x（年）和所支出的维修费用 y（万元）有如下的统计资料：

x	2	3	4	5	6
y	2.2	3.8	5.5	6.5	7.0

若由资料可知 y 对 x 呈线性相关关系. 试求：

(1)y 与 x 之间的线性回归方程；

(2)估计使用年限为 10 年时，维修费用是多少万元.

[解析] 本题主要考查对两个变量的回归分析和预测说明，代入公式即可得.

[答案] 解：(1)列表如下：

i	1	2	3	4	5
x_i	2	3	4	5	6
y_i	2.2	3.8	5.5	6.5	7.0
x_iy_i	4.4	11.4	22.0	32.5	42.0
x_i^2	4	9	16	25	36
$\bar{x}=4,\bar{y}=5$ $\sum_{i=1}^{5}x_i^2=90,\sum_{i=1}^{5}x_iy_i=112.3$					

$$\therefore \hat{b}=\frac{\sum_{i=1}^{5}x_iy_i-5\bar{x}\,\bar{y}}{\sum_{i=1}^{5}x_i^2-5\bar{x}^2}=\frac{112.3-5\times4\times5}{90-5\times4^2}=1.23,$$

$$\hat{a}=\bar{y}-\hat{b}\bar{x}=5-1.23\times4=0.08.$$

$\therefore$ 线性回归方程为:$\hat{y}=\hat{b}x+\hat{a}=1.23x+0.08$.

(2)当 $x=10$ 时,$\hat{y}=1.23\times10+0.08=12.38$(万元).

即估计使用 10 年时维修费用是 12.38 万元.

[点评] 本题若没有告诉我们 y 与 x 之间是呈线性相关的,应首先进行相关性检验.如果两个变量本身不具备线性相关关系,或者说它们之间相关关系不显著,即使求出回归方程也是没有意义的,而且其估计与预测也是不可信的.

▶问题二 回归分析

例题 3 已知某地每单位面积菜地年平均使用氮肥量 x(kg)与每单位面积蔬菜年平均产量 y(t)之间的关系有如下数据:

年份	1985	1986	1987	1988	1989	1990	1991	1992
x(kg)	70	74	80	78	85	92	90	95
y(t)	5.1	6.0	6.8	7.8	9.0	10.2	10.0	12.0
年份	1993	1994	1995	1996	1997	1998	1999	
x(kg)	92	108	115	123	130	138	145	
y(t)	11.5	11.0	11.8	12.2	12.5	12.8	13.0	

(1)求 x 与 y 之间的相关系数,并检验是否线性相关;

(2)若线性相关,求蔬菜产量 y 与使用氮肥量 x 之间的回归直线方程,并估计每单位面积施氮肥 150 kg 时,每单位面积蔬菜的年平均产量.

[解析] 本题主要考查对两个变量的相关性检验和回归分析.(1)使用样本相关系数计算公式来完成;(2)先作统计假设,由小概率 0.05 与 $n-2$ 在附表中查得相关系数临界值 $r_{0.05}$,若 $r>r_{0.05}$ 则线性相关,否则不线性相关.

[答案] 解:(1)列出下表,并用科学计算器进行相关计算:

i	1	2	3	4	5	6	7	8
x_i	70	74	80	78	85	92	90	95
y_i	5.1	6.0	6.8	7.8	9.0	10.2	10.0	12.0
x_iy_i	357	444	544	608.4	765	938.4	900	1 140
i	9	10	11	12	13	14	15	
x_i	92	108	115	123	130	138	145	
y_i	11.5	11.0	11.8	12.2	12.5	12.8	13.0	
x_iy_i	1 058	1 188	1 357	1 500.6	1 625	1 766.4	1 885	

$$\bar{x}=\frac{1\,515}{15}=101,\bar{y}=\frac{151.7}{15}\approx10.11,$$

$$\sum_{i=1}^{15}x_i^2=161\,125,\sum_{i=1}^{15}y_i^2=1\,628.55,\sum_{i=1}^{15}x_iy_i=16\,076.8.$$

故蔬菜产量与施用氮肥量的相关系数

$$r=\frac{16\,076.8-15\times101\times10.11}{\sqrt{(161\,125-15\times101^2)(1\,628.55-15\times10.11^2)}}\approx0.8643.$$

$\because |r|>0.75$,

$\therefore$ 认为这两个变量有很强的线性相关关系.

(2)设所求的回归直线方程为 $\hat{y}=\hat{b}x+\hat{a}$,则

$$\hat{b}=\frac{\sum_{i=1}^{15}x_iy_i-15\bar{x}\,\bar{y}}{\sum_{i=1}^{15}x_i^2-15\bar{x}^2}=\frac{16\,076.8-15\times101\times10.11}{161\,125-15\times101^2}\approx0.0937,$$

$$\hat{a}=\bar{y}-b\bar{x}\approx10.11-0.0937\times101=0.6463,$$

$\therefore$ 直线回归方程为 $\hat{y}=0.0937x+0.6463$.

$\therefore$ 当每单位面积施氮肥 150 kg 时,每单位面积蔬菜年平均产量为 $0.0937\times150+0.6463\approx14.701$(t).

[点评] 求解两个变量的相关系数及它们的回归直线方程的计算量较大,需要细心、谨慎地计算.如果会使用含统计的科学计算器,能简单得到 $\sum_{i=1}^{n}x_i,\sum_{i=1}^{n}y_i,\sum_{i=1}^{n}x_i^2,\sum_{i=1}^{n}y_i^2,\sum_{i=1}^{n}x_iy_i$ 这些量,也就无需制表这一步,直接算出结果就行了.另外,利用计算机中有关应用程序也可以对这些数据进行处理.

例题 4 为了研究三月下旬的平均气温 x(℃)与四月二十日前棉花害虫化蛹高峰日 y(日)的关系,某地观察了 2000 年至 2005 年间的情况,得到下面数据表:

年份	2000	2001	2002	2003	2004	2005
x	24.4	29.5	32.9	28.7	30.3	28.9
y	19	6	1	10	1	8

(1)求 y 对 x 的线性回归方程,并说明线性回归模型拟合

的效果；

(2)根据规律推断，若该地区2007年三月下旬平均气温为27 ℃，试估计2007年四月化蛹高峰日为哪一天？

[答案] 解：(1)$\bar{x}=\frac{1}{6}(24.4+29.5+\cdots+28.9)\approx 29.12$.

$\bar{y}=\frac{1}{6}(19+6+\cdots+8)=7.5$.

$\sum_{i=1}^{6}x_i^2=24.4^2+\cdots+28.9^2=5\,125.01$.

$\sum_{i=1}^{6}y_i^2=19^2+\cdots+8^2=563$.

$\sum_{i=1}^{6}x_iy_i=24.4\times 19+\cdots+28.9\times 8=1\,222$.

$\therefore \hat{b}=\frac{1\,222-6\times 7.5\times 29.12}{5\,125.01-6\times 29.12^2}\approx -2.4$,

$\hat{a}=\bar{y}-\hat{b}\bar{x}\approx 7.5+2.4\times 29.12=77.388$.

所求直线回归方程是$\hat{y}=-2.4x+77.388$.

列出残差表：

$y_i-\hat{y}_i$	0.172	−0.588	2.572	1.492	−3.668	−0.028
$y_i-\bar{y}$	11.5	−1.5	−6.5	2.5	−6.5	0.5

$\therefore \sum_{i=1}^{6}(y_i-\hat{y}_i)^2\approx 22.67$，$\sum_{i=1}^{6}(y_i-\bar{y})^2=225.5$.

$R^2=1-\frac{22.67}{225.5}\approx 0.899$.

∴线性回归模型的拟合效果较好，x与y之间的线性相关性较强.

(2)当$x=27$时，$\hat{y}=-2.4\times 27+77.388=12.588$.

据此估计该地2007年4月12日或13日为化蛹高峰日.

[点评] 本题如果求相关系数r，$|r|>0.75$，同样可得两个变量有较强的线性相关关系. 通常在有着较强的线性相关关系的前提下，所求得的直线回归方程才有实际意义，据此所得的预测值也才有一定的价值.

例题5 为研究拉力x(N)对弹簧长度y(cm)的影响，对不同拉力的6根弹簧进行测量，测得如下表中的数据：

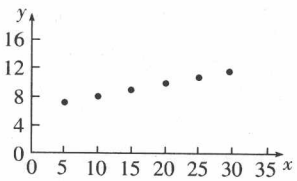
x	5	10	15	20	25	30
y	7.25	8.12	8.95	9.90	10.9	11.8

(1)画出散点图；

(2)如果散点图中的各点大致分布在一条直线的附近，求y与x之间的回归直线方程；

(3)求出残差，进行残差分析.

[答案] 解：(1)散点图如图所示.



(2)将已知表中的数据列成下表：

x_i	5	10	15	20	25	30
y_i	7.25	8.12	8.95	9.90	10.9	11.8
x_iy_i	36.25	81.2	134.25	198	272.5	354
x_i^2	25	100	225	400	625	900

$\bar{x}=17.5$，$\bar{y}\approx 9.49$，$\sum_{i=1}^{6}x_iy_i=1\,076.2$，$\sum_{i=1}^{6}x_i^2=2\,275$.

$\therefore \hat{b}=\frac{\sum_{i=1}^{6}x_iy_i-6\bar{x}\,\bar{y}}{\sum_{i=1}^{6}x_i^2-6(\bar{x})^2}\approx\frac{1\,076.2-6\times 17.5\times 9.49}{2\,275-6\times 17.5^2}\approx 0.18$,

$\hat{a}=\bar{y}-\hat{b}\bar{x}\approx 9.49-0.18\times 17.5=6.34$.

∴所求回归直线方程为$\hat{y}=0.18x+6.34$.

(3)列出残差表：

$y_i-\hat{y}_i$	0.01	−0.02	−0.09	−0.04	0.06	0.06
$y_i-\bar{y}$	−2.24	−1.37	−0.54	0.41	1.41	2.31

$\therefore \sum_{i=1}^{6}(y_i-\hat{y}_i)^2=0.017\,4$，$\sum_{i=1}^{6}(y_i-\bar{y})^2=14.678\,4$.

$R^2=1-\frac{0.017\,4}{14.678\,4}\approx 0.998\,81$，回归模型的拟合效果较好.

残差分析：由散点图可以看出两个变量之间有很强的相关性，由R^2的值可知回归直线模型的拟合效果很好；由残差表中的数值可以看出第3个样本点的残差比较大，需要确认在采集这个数据的时候的误差所在，是人为的，还是系统(弹簧)本身产生的，如果有的话需要纠正数据，重新建立回归模型；由表中数据可以看出残差比较均匀地落在宽度不超过0.15的狭窄的水平带状区域中，说明选用的线性回归模型的精度较高，由以上分析可知，弹簧长度与拉力成线性关系.

[点评] 计算数据要认真细心，残差分析要全面.

例题6 在英语教学中，为了了解学生的词汇量，设计了一份包含100个单词的试卷，现抽取15名学生进行测试，得到学生掌握试卷中单词个数x与该生实际掌握单词量y的对应数据如下：

x	61	65	70	69	83	75	58	73
y	2 030	2 140	2 270	2 250	2 240	2 220	1 970	2 330
x	63	72	71	68	65	67	74	
y	2 100	2 300	2 300	2 200	2 200	2 200	2 370	

(1)对变量y与x进行相关性检验；

(2)如果 y 与 x 之间具有线性相关关系,

①求 y 对 x 的直线回归方程;

②求 x 对 y 的直线回归方程.

[解析] 本题主要考查对两个变量进行回归分析,先进行相关性检验,由相关系数公式可得.

[答案] 解:(1)列出下表,并用科学计算器进行有关计算.

i	1	2	3	4	5	6	7	8
x_i	61	65	70	69	83	75	58	73
y_i	2 030	2 140	2 270	2 250	2 240	2 220	1 970	2 330
x_iy_i	123 830	139 100	158 900	155 250	185 920	166 500	114 260	170 090
i	9	10	11	12	13	14	15	
x_i	63	72	71	68	65	67	74	
y_i	2 100	2 300	2 300	2 200	2 200	2 200	2 370	
x_iy_i	132 300	165 600	163 300	149 600	143 000	147 400	175 380	

$\bar{x}\approx 68.93, \bar{y}=2\,208,$

$\sum_{i=1}^{15}x_i^2=71\,822, \sum_{i=1}^{15}y_i^2=73\,298\,600, \sum_{i=1}^{15}x_iy_i=2\,290\,430.$

$\therefore \sum_{i=1}^{15}x_i^2-15\bar{x}^2=71\,822-15\times 68.93^2\approx 551.83,$

$\sum_{i=1}^{15}y_i^2-15\bar{y}^2=73\,298\,600-15\times 2\,208^2=169\,640,$

$\sum_{i=1}^{15}x_iy_i-15\bar{x}\,\bar{y}=2\,290\,430-15\times 68.93\times 2\,208=7\,468.4,$

$$r=\frac{\sum_{i=1}^{15}x_iy_i-15\bar{x}\,\bar{y}}{\sqrt{\left(\sum_{i=1}^{15}x_i^2-15\bar{x}^2\right)\left(\sum_{i=1}^{15}y_i^2-15\bar{y}^2\right)}}$$

$$=\frac{7\,468.4}{\sqrt{551.83\times 169\,640}}$$

$\approx 0.772.$

查相关系数检验的临界值表,得 $r_{0.05}=0.514.$

$\because |r|>r_{0.05}$,故 y 与 x 有线性相关关系.

(2)①设 y 对 x 的直线回归方程 $\hat{y}=\hat{b}x+\hat{a}$,则

$$\hat{b}=\frac{\sum_{i=1}^{15}x_iy_i-15\bar{x}\,\bar{y}}{\sum_{i=1}^{15}x_i^2-15\bar{x}^2}=\frac{7\,468.4}{551.83}\approx 13.5,$$

$\hat{a}=\bar{y}-\hat{b}\bar{x}=2\,208-13.5\times 68.93=1\,277.445,$

即所求 y 对 x 的直线回归方程为 $\hat{y}=13.5x+1\,277.445.$

②设 x 对 y 的直线回归方程为 $\hat{x}=\hat{d}y+\hat{c}$,则

$$\hat{d}=\frac{\sum_{i=1}^{15}x_iy_i-15\bar{x}\,\bar{y}}{\sum_{i=1}^{15}y_i^2-15\bar{y}^2}=\frac{7\,468.4}{169\,640}\approx 0.044,$$

$\hat{c}=\bar{x}-\hat{d}\bar{y}=68.93-0.044\times 2\,208\approx -28.22,$

即所求的 x 对 y 的直线回归方程为 $\hat{x}=0.044y-28.22.$

D 针对性练习

1. 下列两个变量之间的关系是相关关系的是 ( )

A. 单位圆中角的度数和所对弧长

B. 人的身高与视力

C. 收入水平和纳税水平

D. 日照时间和水稻的亩产量

2. 由一组数据 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$ 得到的直线回归方程为 $\hat{y}=\hat{b}x+\hat{a}$,则下列说法不正确的是 ( )

A. 直线 $\hat{y}=\hat{b}x+\hat{a}$ 必过点 $(\bar{x},\bar{y})$

B. 直线 $\hat{y}=\hat{b}+\hat{a}$ 至少经过点 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$ 中的一个点

C. 直线 $\hat{y}=\hat{b}x+\hat{a}$ 的斜率为 $\dfrac{\sum_{i=1}^{n}x_iy_i-n\bar{x}\,\bar{y}}{\sum_{i=1}^{n}x_i^2-n\bar{x}^2}$

D. 直线 $\hat{y}=\hat{b}x+\hat{a}$ 和各点 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$ 的偏差是该坐标平面上所有直线与这些点的偏差中最小的直线

3. 部门所属的 10 个工业企业生产性固定资产价值与工业增加值资料如下表(单位:百万元):

固定资产价值	3	3	5	6	6	7	8	9	9	10
工业增加值	15	17	25	28	30	36	37	42	40	45

根据上表资料计算的相关系数为 ( )

A. 0　　B. $-0.897\,3$

C. 1.022 8　　D. 0.991 8

4. 在两个变量 y 与 x 的回归模型中,分别选择了 4 个不同模型,它们的相关指数 R^2 如下,其中拟合效果最好的模型是 ( )

A. 模型一的相关指数 R^2 为 0.98

B. 模型二的相关指数 R^2 为 0.85

C. 模型三的相关指数 R^2 为 0.61

D. 模型四的相关指数 R^2 为 0.31

5. 设有一个回归方程为 $\hat{y}=3-5x$,变量 x 增加一个单位时,则有 ( )

A. y 平均增加 3 个单位　　B. y 平均减少 5 个单位

C. y 平均增加 5 个单位　　D. y 平均减少 3 个单位

6. 已知 x,y 之间的一组数据如下表所示:

x	1.08	1.12	1.19	1.28
y	2.25	2.37	2.40	2.55

则 y 与 x 之间的线性回归方程 $\hat{y}=\hat{a}+\hat{b}x$ 必过定点_____.