普通高中课程标准实验教科书同步教学资源

教材解读

数学

选修 1-2

人民教育出版社教学资源编辑室
北京百川菁华科技有限公司 策划组编

人民教育出版社

A版

普通高中课程标准实验教科书同步教学资源

教材解读

数学

选修 1-2

人民教育出版社教学资源编辑室
北京百川菁华科技有限公司 策划组编

人民教育出版社

A版

图书在版编目(CIP)数据

教材解读：A版．数学．1-2：选修／人民教育出版社教学资源编辑室，北京百川菁华科技有限公司组编．—北京：人民教育出版社，2012.12

ISBN 978-7-107-25804-6

Ⅰ．①教… Ⅱ．①人… ②北… 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（2012）第303675号

人民教育出版社出版发行

网址：http://www.pep.com.cn

北京汇祥印务有限公司印装 全国新华书店经销

2012年12月第1版 2012年12月第1次印刷

开本：890毫米×1 240毫米 1/16 印张：10 字数：400千字

定价：22.80元

如发现印、装质量问题，影响阅读，请与本社出版科联系调换。

（联系地址：北京市海淀区中关村南大街17号院1号楼 邮编：100081）

《教材解读》编委会

前言

为了帮助广大师生更好地理解和把握教材，落实各学科课程标准要求，实现三维目标，人民教育出版社发挥教材研究编写的优势，组织教材编写专家、一线教研员和优秀教师，精心策划和编写了这套配合人教版教科书使用的同步系列丛书——《教材解读》。本丛书涵盖从小学到高中所有学科和学段。

本系列丛书综合了对教材的整体解读、单元解读和课节解读，并形成以下主要特色：

1. **高屋建瓴**。以“新、透、细、精”为编写原则，高屋建瓴地对教材知识点进行深入解读，系统总结教与学的规律方法，全方位拓展知识空间，融知识性、科学性、趣味性、针对性和实用性于一体，形成了基础与能力并重，综合与创新结合的科学体系。

2. **点面结合**。本书科学阐释了课节内容在整个单元或整套教材中的地位及《课程标准》对相关内容的具体要求，精细梳理各知识点，深入挖掘、研究教材中的重点、难点、易错易混点及其突破方法，关注教材所述内容的背景材料，形成对学生思维过程的策略引导，全面提升综合素养。

3. **活学活用**。本书在拓展应用中，注重典型例题和综合练习的对应性，突出题目的鲜活和示范特点，用最精练的题目、最科学的题型组合培养学生最具实效的解决问题的能力。

“不积跬步，无以至千里；不积小流，无以成江海。”本丛书化面为点、点面结合，通过精准地解读、巧妙地点拨，致力于打造知识梳理、方法指导、针对练习“三位一体”的多功效参考书。希望她能成为老师备课、讲课、教研、教改的好助手，成为学生自主学习、有效复习的好老师，成为家长辅导孩子的好帮手。

由于编写时间紧迫和水平有限，本丛书一定还存在不足，特诚挚地希望广大读者提出批评和建议，以便再版修订时参考。在本套丛书的编写过程中，引用了部分相关资料，有的已与原作者取得联系，但有些未能与原作者取得联系，希望原作者看到此书后与我们联系，以便支付相应的稿酬。在此，特向各位作者表示诚挚的感谢。

编 者

目录

CONTENTS

第一章 统计案例

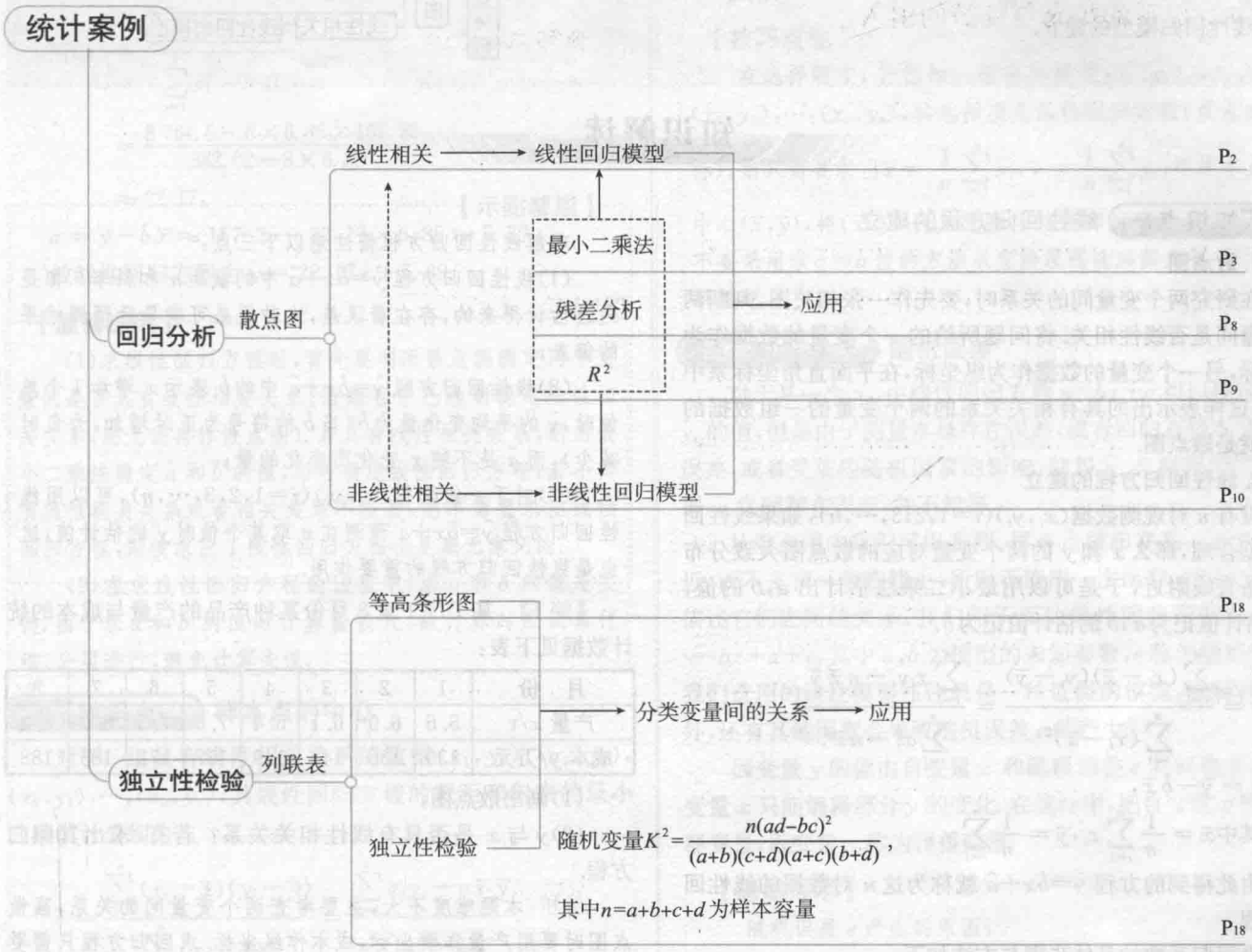
第二章 推理与证明

第三章　数系的扩充与复数的引入

第四章　框图

第一章 统计案例

思维导图



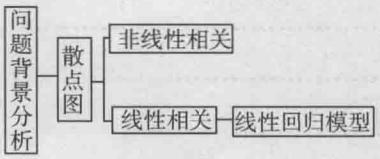
1.1 回归分析的基本思想及其初步应用

第1课时 线性回归模型

学习目标

1. 了解回归模型与函数模型的区别.
2. 了解任何模型只能近似描述实际问题.
3. 线性回归模型的建立.

知识结构



知识解读

知识点一 线性回归方程的建立

1. 散点图

在研究两个变量间的关系时,要先作一张相关图,判断两个变量间是否线性相关.将问题所给的一个变量的数据作为横坐标,另一个变量的数据作为纵坐标,在平面直角坐标系中描点,这样表示出的具有相关关系的两个变量的一组数据的图形就是散点图.

2. 线性回归方程的建立

设有 n 对观测数据$(x_i,y_i)(i=1,2,3,\cdots,n)$,如果线性回归模型合理,那么 x 和 y 的两个变量对应的散点图大致分布在一条直线附近,于是可以用最小二乘法估计出 a,b 的值,a 的估计值记为$\hat{a}$,b 的估计值记为$\hat{b}$.

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

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

其中$\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\overline{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$,

由此得到的方程 $\hat{y}=\hat{b}x+\hat{a}$ 就称为这 n 对数据的线性回归方程.

求回归系数的具体步骤与方法如下:

(1)先将所给的数据 x,y 列成相应的表格:

i	1	2	…	n	$\sum$
x	x_1	x_2	…	x_n	$\sum_{i=1}^{n}x_i$
y	y_1	y_2	…	y_n	$\sum_{i=1}^{n}y_i$
x^2	x_1^2	x_2^2	…	x_n^2	$\sum_{i=1}^{n}x_i^2$
xy	x_1y_1	x_2y_2	…	x_ny_n	$\sum_{i=1}^{n}x_iy_i$

(2) 计算$\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\overline{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$;

(3)代入公式计算 $\hat{b},\hat{a}$ 的值,即可得到线性回归方程.

【温馨提示】

理解线性回归方程需注意以下三点:

(1)线性回归方程 $\hat{y}=\hat{b}x+\hat{a}$ 中的截距 $\hat{a}$ 和斜率 $\hat{b}$ 都是通过估计得来的,存在着误差,这种误差可能导致预测结果的偏差;

(2)线性回归方程 $\hat{y}=\hat{b}x+\hat{a}$ 中的 $\hat{b}$ 表示 x 增加1个单位时,$\hat{y}$ 的平均变化量为 $\hat{b}$(当 $\hat{b}$ 的符号为正时增加,为负时减少),而 $\hat{a}$ 是不随 x 变化而变化的量;

(3)对于一组数据$(x_i,y_i)(i=1,2,3,\cdots,n)$,可以用线性回归方程 $\hat{y}=\hat{b}x+\hat{a}$ 预测在 x 取某个值时 y 的估计值,这也是线性回归方程的重要作用.

【例1】 某工厂1至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 x/t	5.6	6.0	6.1	6.4	7.0	7.5	8.0	8.2
成本 y/万元	130	136	143	149	157	172	183	188

(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.

分析:本题难度不大,主要考查两个变量间的关系,画散点图时要用产量作横坐标,成本作纵坐标.求回归方程只需要将表中的数据代入最小二乘法公式求出 $\hat{a}$ 和 $\hat{b}$ 的值即可.

解:(1)由表中所给的数据,画出散点图,如图1.1-1所示.

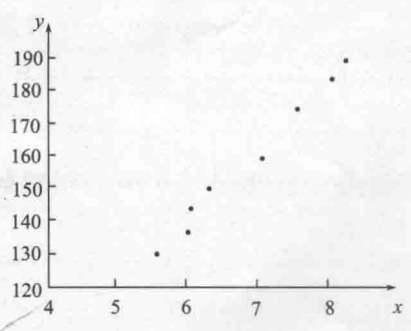


图1.1-1

(2)从散点图可以看出,这些点基本分布在一条直线附

近，可以认为变量 x 和 y 具有线性相关关系. 设线性回归方程为 $\hat{y}=\hat{b}x+\hat{a}$，要求线性回归方程，可先列出下表：

i	1	2	3	4	5	6	7	8	$\sum$
x_i	5.6	6.0	6.1	6.4	7.0	7.5	8.0	8.2	54.8
y_i	130	136	143	149	157	172	183	188	1 258
x_i^2	31.36	36.00	37.21	40.96	49.00	56.25	64.00	67.24	382.02
x_iy_i	728.0	816.0	872.3	953.6	1 099.0	1 290.0	1 464.0	1 541.6	8 764.5

$\bar{x}=6.85$，$\bar{y}=157.25$.

所以 $\hat{b}=\dfrac{\sum\limits_{i=1}^{8}x_iy_i-8\bar{x}\,\bar{y}}{\sum\limits_{i=1}^{8}x_i^2-8\bar{x}^2}$

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

≈22.17，

$\hat{a}=\bar{y}-\hat{b}\bar{x}\approx157.25-22.17\times6.85\approx5.39$，

故线性回归方程为 $\hat{y}=22.17x+5.39$.

【温馨提示】

(1)求线性回归方程时，首先要利用散点图确定两个变量间是否具有线性相关关系(若题目中已经明确是线性相关关系，则无需再作散点图). 若具有线性相关关系，则用最小二乘法确定 $\hat{a}$ 和 $\hat{b}$ 的值，即可确定线性回归方程；若不具有线性相关关系或者相关关系不显著，则不需要求出线性回归方程，即使求出了线性回归方程也是毫无意义的.

(2)在求线性回归方程的过程中，求 $\hat{a}$ 和 $\hat{b}$ 的值是关键. 由于求 $\hat{a}$ 和 $\hat{b}$ 的值时计算量较大，故计算时应认真仔细、分层进行，避免计算失误.

知识点二 样本点的中心

对于一组具有线性相关关系的数据 (x_1,y_1)，(x_2,y_2)，(x_3,y_3)，…，(x_n,y_n)，其线性回归方程的截距和斜率的最小二乘估计分别为

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

$\hat{a}=\bar{y}-\hat{b}\bar{x}$，其中 $\bar{x}=\dfrac{1}{n}\sum\limits_{i=1}^{n}x_i$，$\bar{y}=\dfrac{1}{n}\sum\limits_{i=1}^{n}y_i$，$(\bar{x},\bar{y})$ 称为样本点的中心.

【温馨提示】

(1)样本点的中心的坐标分别是两个变量的观测数据的算术平均数.

(2)由于 $\hat{a}=\bar{y}-\hat{b}\bar{x}$，即 $\bar{y}=\hat{b}\bar{x}+\hat{a}$，所以点 $(\bar{x},\bar{y})$ 在回归直线 $\hat{y}=\hat{b}x+\hat{a}$ 上，即回归直线过样本点的中心.

【例2】 在一次实验中，测得 (x,y) 的四组值分别是 $A(1,2)$，$B(2,3)$，$C(3,4)$，$D(4,5)$，则 y 与 x 之间的线性回归方程为 (　　)

A. $\hat{y}=x+1$　　B. $\hat{y}=x+2$

C. $\hat{y}=2x+1$　　D. $\hat{y}=x-1$

解析：由已知可得

$\bar{x}=\dfrac{1}{4}(1+2+3+4)=2.5$，

$\bar{y}=\dfrac{1}{4}(2+3+4+5)=3.5$，

所以样本点的中心为(2.5,3.5).

由于回归直线 $\hat{y}=\hat{b}x+\hat{a}$ 必过样本点的中心，

所以必过点(2.5,3.5).

经验证可知，3.5=2.5+1，

所以线性回归方程为 $\hat{y}=x+1$.

答案：A

【技巧点拨】

在选择题中，若已知一组观测数据 (x_1,y_1)，(x_2,y_2)，(x_3,y_3)，…，(x_n,y_n)，其选择项是线性回归方程(或点的坐标)，则只需要求出 $\bar{x}=\dfrac{1}{n}\sum\limits_{i=1}^{n}x_i$，$\bar{y}=\dfrac{1}{n}\sum\limits_{i=1}^{n}y_i$，即样本点的中心 $(\bar{x},\bar{y})$，将 $(\bar{x},\bar{y})$ 代入选择项验证即可得到答案. 尽量不要采用求 $\hat{a}$ 和 $\hat{b}$ 值的方法来直接求线性回归方程，除非是解答题.

知识点三 随机误差

对于某一个 x_i，由线性回归方程 $\hat{y}=\hat{b}x+\hat{a}$ 可以确定一个 $\hat{y}_i$ 的值，但是由于测量本身存在误差，或者回归直线本身存在误差，或者受某些随机因素的影响，使得 $\hat{y}_i$ 与测得的实际数据 y_i 之间存在误差，并不相等.

从散点图中我们可以看到，样本点散布在某一条直线附近，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，我们用下面的线性回归模型来表示：$y=bx+a+e$，其中 a，b 为模型的未知参数，e 称为随机误差. 我们选用的线性模型往往只是一种近似的模型，除解释变量外，还有其他因素会导致随机误差 e 的产生.

因变量 y 的值由自变量 x 和随机误差 e 共同确定，即自变量 x 只能解释部分 y 的变化. 在统计中，把自变量 x 称为解释变量，因变量 y 称为预报变量.

【温馨提示】

随机误差 e 产生的原因：

(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包含在 e 中；

(2)忽略了某些因素的影响. 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(例如，在描述身高和体重关系的模型中，体重不仅受身高的影响，还会受遗传基因、饮食习惯、生长环境等其他因素的影响)，它们的影响都体现在随机误差 e 中；

(3)观测误差. 由于测量工具等原因，导致 y 的观测值产生误差(比如一个人的体重是确定的数，不同的秤可能会得到不同的观测值，与真实值之间存在误差)，这样的误差也包含在随机误差 e 中.

典例精析

类型一 基本概念辨析

【例1】 下列说法中正确的个数为 ( )

①线性回归分析就是由样本点去寻找一条直线,使它贴近这些样本点的数学方法;

②利用样本点的散点图可以直观判断两个变量的关系是否可以用线性回归方程表示;

③通过线性回归方程 $\hat{y}=\hat{b}x+\hat{a}$,可以估计变量的取值和观测变量的变化趋势;

④能求出线性回归方程的一组数据一定是线性相关的.

A. 1 B. 2 C. 3 D. 4

解析:①反映的正是最小二乘法的思想,正确.

②反映的是散点图的作用,也正确.

③反映的是回归模型 $y=bx+a+e$,其中 e 为随机误差,也正确.

④一组观测数据能求出线性回归方程 $\hat{y}=\hat{b}x+\hat{a}$,但不能保证这些点在散点图中呈线性相关分布,可能是函数关系,也可能没有关系,错误.

答案:C

温馨提示:确定关系与相关关系之间并没有一条不可逾越的鸿沟,由于实验误差、测量误差的存在,变量之间的确定关系往往通过相关关系表现出来.反过来,在有些问题上,我们可以通过研究相关关系来深入了解变量的内在规律,从而找到它们的确定关系.

变式·拓展1 下列说法中错误的是 ( )

A. 如果两个变量 x 与 y 之间存在着线性相关关系,那么我们根据实验数据得到的点 $(x_i,y_i)(i=1,2,3,\cdots,n)$ 将散布在某条直线附近

B. 如果两个变量 x 与 y 之间不存在线性相关关系,那么根据它们的一组数据 $(x_i,y_i)(i=1,2,3,\cdots,n)$ 不能写出一个线性方程

C. 设 x,y 是具有相关关系的两个变量,且 y 关于 x 的线性回归方程为 $\hat{y}=\hat{b}x+\hat{a}$,则 $\hat{b}$ 叫做回归系数

D. 可以使用散点图大体判断两个变量 x 与 y 之间是否具有线性相关关系,若具有,则可求出线性回归方程

类型二 求线性回归方程

【例2】 炼钢是一个氧化降碳的过程,钢水含碳量的多少直接影响冶炼时间的长短,必须掌握钢水含碳量和冶炼时间的关系.如果已测得炉料熔化完毕时,钢水的含碳量 x 与冶炼时间 y(从炉料熔化完毕到出钢的时间)的一组数据如下表所示:

x/0.01%	104	180	190	177	147	134	150	191	204	121
y/min	100	200	210	185	155	135	170	205	235	125

(1)作出散点图,你能从散点图中发现含碳量与冶炼时间的一般规律吗?

(2)求线性回归方程;

(3)预测当钢水的含碳量为160时,应冶炼多少分钟.

分析:(1)题目所给的数据中已经明确了横坐标与纵坐标,则只需建立相应坐标系,描点;

(2)若通过散点图判断出变量 x 与 y 具有线性相关关系,则将数据代入公式 $\hat{b}=\dfrac{\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}=\dfrac{\sum\limits_{i=1}^{n}x_iy_i-n\overline{x}\,\overline{y}}{\sum\limits_{i=1}^{n}x_i^2-n\overline{x}^2}$,$\hat{a}=\overline{y}-\hat{b}\overline{x}$,求出线性回归方程;

(3)当钢水的含碳量为160时,代入所求出的回归方程中,即可预测冶炼时间.

解:(1)x 轴表示含碳量,y 轴表示冶炼时间,作出散点图.如图1.1-2所示.

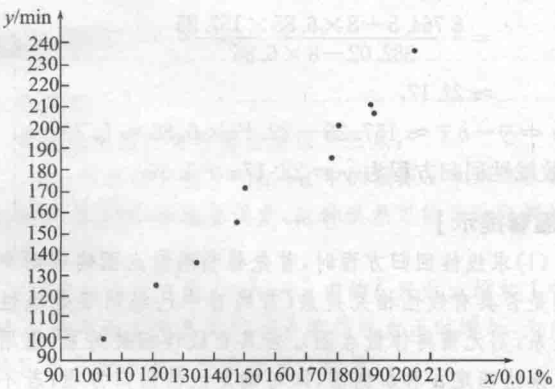


图1.1-2

从图中可以看出,各点分布在一条直线附近,所以它们线性相关,并且随着 x 值的增大,y 值也在增大.由此可以判断,含碳量越高,需要的冶炼时间越长.

(2)列出下表:

i	1	2	3	4	5
x_i	104	180	190	177	147
y_i	100	200	210	185	155
x_iy_i	10 400	36 000	39 900	32 745	22 785
i	6	7	8	9	10
x_i	134	150	191	204	121
y_i	135	170	205	235	125
x_iy_i	18 090	25 500	39 155	47 940	15 125

$\overline{x}=159.8,\overline{y}=172,\sum\limits_{i=1}^{10}x_i^2=265\,448,\sum\limits_{i=1}^{10}x_iy_i=287\,640$

设所求的线性回归方程为 $\hat{y}=\hat{b}x+\hat{a}$,其中 $\hat{a},\hat{b}$ 的值使 $Q=\sum\limits_{i=1}^{10}(y_i-bx_i-a)^2$ 的值最小,

$\hat{b}=\dfrac{\sum\limits_{i=1}^{10}x_iy_i-10\overline{x}\,\overline{y}}{\sum\limits_{i=1}^{10}x_i^2-10\overline{x}^2}\approx1.267,\hat{a}=\overline{y}-\hat{b}\overline{x}\approx-30.467$,

即所求的线性回归方程为 $\hat{y}=1.267x-30.467$.

(3)当 $x=160$ 时,$\hat{y}=1.267\times160-30.467\approx172$(min),即大约应冶炼172 min.

温馨提示：在求线性回归方程的题目中，最复杂的就是求 $\hat{a}$ 和 $\hat{b}$ 的值，只要确定了这两个值，则回归直线的方程即可确定. 但由于公式 $\hat{b}=\dfrac{\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}=\dfrac{\sum\limits_{i=1}^{n}x_iy_i-n\overline{x}\,\overline{y}}{\sum\limits_{i=1}^{n}x_i^2-n\overline{x}^2}$ 较复杂，所以在求解过程中一定要认真计算，切不可大意，一个数据计算错，最终的结果将“失之毫厘，谬以千里”，前功尽弃！也正因如此，考试中的题目一般只给 5 组数据左右，以减轻学生的运算负担. 另外，在选择公式时，一般采用 $\hat{b}=\dfrac{\sum\limits_{i=1}^{n}x_iy_i-n\overline{x}\,\overline{y}}{\sum\limits_{i=1}^{n}x_i^2-n\overline{x}^2}$ 来计算，其运算量相对要小一些.

变式·拓展 2　某班 5 名学生的数学和物理成绩如下表：

学科成绩＼学生	A	B	C	D	E
数学成绩 x	88	76	73	66	63
物理成绩 y	78	65	71	64	61

(1)画出散点图；

(2)求物理成绩 y 对数学成绩 x 的线性回归方程；

(3)一名学生的数学成绩是 96，试预测他的物理成绩.

易错点　混淆相关关系与确定性关系

【例 3】　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

错解

由三角函数知识可得答案 A 是函数关系；答案 B 中的身高与视力是没有关系的；根据生活常识可知，工资越高，交的税越多，所以答案 C 是相关关系；答案 D 中日照时间与水稻的亩产量关系不大，水和肥料才是关键.

答案：C

正解

答案 C 中的收入水平和纳税水平并不是相关关系，而是函数关系，下面是 2012 年实行的 7 级超额累进个人所得税税率表(工资薪金所得适用)：

级数	个税免征额 3 500 元 全月应纳税所得额 (含税级距)	税率/%	速算扣除数
1	不超过 1 500 元	3	0
2	超过 1 500 元至 4 500 元的部分	10	105
3	超过 4 500 元至 9 000 元的部分	20	555
4	超过 9 000 元至 35 000 元的部分	25	1 005
5	超过 35 000 元至 55 000 元的部分	30	2 755
6	超过 55 000 元至 80 000 元的部分	35	5 505
7	超过 80 000 元的部分	45	13 505

例如，某人某月工资减去社保个人缴纳金额和住房公积金个人缴纳金额后为 5 500 元，个税计算：(5 500－3 500)×10%－105＝95(元)，这事实上是一个函数关系.

根据生物学知识可得，农作物除了需要足够的水与肥料外，光合作用是生物生长的前提条件，而充足的光照正是光合作用的关键，所以，答案 D 中的日照时间和水稻的亩产量是相关关系.

答案：D

【误区警示】

误区：部分学生在学习相关关系的初期，容易混淆相关关系与确定性关系(函数关系就是一种确定性关系). 例 3 中，由于对纳税知识的缺乏，很容易认为收入与个人所得税是相关关系，也就是收入水平高则纳税水平高，忽略了其函数关系. 日照时间和水稻的亩产量之间是相关关系，日照时间长，亩产量一定高，但却不是确定的关系，因为水稻的生长条件还受水、肥料、温度、土壤等因素的影响.

悟区：学习相关关系的概念时，不仅要把教材本身学好，更要把与之相关的其他知识学好，利用所学的其他学科的知识来判断所遇到的问题是确定性关系还是相关关系. 函数关系是确定性关系的一种，也就是两个量的关系是确定的，是用对应法则联系在一起的.

变式·拓展 3　在下列各量之间，存在相关关系的是（　　）

①正方体的体积与棱长之间的关系；

②一块农田的小麦产量与施肥量之间的关系；

③人的身高与年龄之间的关系；

④家庭的支出与收入之间的关系；

⑤某户家庭用电量与电价之间的关系.

A. ②③　　B. ③④　　C. ④⑤　　D. ②③④

直击高考

考 点	命题趋势
散点图	散点图部分难度不大，部分省份的高考题有所涉及，估计单独考查的可能性并不是很大，即便是考查到，也是考查变量间的相关性，或要求画出散点图. 题型以选择题、填空题为主，有时也会考解答题
线性回归方程	求线性回归方程或者是与线性回归方程有关的题目经常出现在高考中，具体涉及到的内容有：求线性回归方程、求预报变量的值、样本点的中心、预报变量和解释变量与 $\hat{a}$ 及 $\hat{b}$ 的关系，预计今年考到其中之一的可能性较大. 题型既有选择题、填空题，又有解答题，但以选择题、填空题为主

【真题 1】 (2012·湖南·文)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回归方程为 $\hat{y}=0.85x-85.71$，则下列结论中不正确的是 (　　)

A. y 与 x 具有正的线性相关关系

B. 回归直线过样本点的中心 $(\overline{x},\overline{y})$

C. 若该大学某女生身高增加 1 cm，则其体重约增加 0.85 kg

D. 若该大学某女生身高为 170 cm，则可断定其体重必为 58.79 kg

解析：由线性回归方程为 $\hat{y}=0.85x-85.71$，知 y 随 x 的增大而增大，所以 y 与 x 具有正的线性相关关系，所以 A 正确；回归直线 $\hat{y}=\hat{b}x+\hat{a}$ 过样本点的中心 $(\overline{x},\overline{y})$，所以 B 也正确；在线性回归方程 $\hat{y}=\hat{b}x+\hat{a}$ 中，x 增加 1 个单位时，y 的平均变化量为 $\hat{b}$(当 $\hat{b}$ 的符号为正时增加，为负时减少)，所以 C 正确；由于线性回归方程仅仅是可以预测估计总体，所以 D 不正确.

答案：D

【真题 2】 (2012·福建·文)某工厂为了对新研发的一种产品进行合理定价，将该产品按事先拟定的价格进行试销，得到如下数据：

单价 x/元	8	8.2	8.4	8.6	8.8	9
销量 y/件	90	84	83	80	75	68

(1)求线性回归方程 $\hat{y}=\hat{b}x+\hat{a}$，其中 $\hat{b}=-20$，$\hat{a}=\overline{y}-\hat{b}\overline{x}$；

(2)预计在今后的销售中，销量与单价仍然服从(1)中的关系，且该产品的成本是 4 元/件，为使工厂获得最大利润，该产品的单价应定为多少元？(利润=销售收入-成本)

分析：利用最小二乘法求出 $\hat{a}$ 和 $\hat{b}$ 的值，即可确定线性回归方程，题设中已经给出了 $\hat{b}=-20$，计算就简单多了. 第(2)小题是二次函数问题，配方求最值.

解：(1)因为 $\overline{x}=\frac{1}{6}(x_1+x_2+x_3+x_4+x_5+x_6)=8.5$，

$\overline{y}=\frac{1}{6}(y_1+y_2+y_3+y_4+y_5+y_6)=80$，

所以 $\hat{a}=\overline{y}-\hat{b}\overline{x}=80+20\times 8.5=250$.

从而线性回归方程为 $\hat{y}=-20x+250$.

(2)设工厂获得的利润为 L 元，依题意得

$$L=x(-20x+250)-4(-20x+250)$$
$$=-20x^2+330x-1\,000$$
$$=-20\left(x-\frac{33}{4}\right)^2+361.25,$$

当且仅当 $x=8.25$ 时，L 取得最大值.

故当单价为 8.25 元时，工厂可获得最大利润.

【真题 3】 (2011·安徽·文)某地最近十年粮食需求量逐年上升，下表是部分统计数据：

年 份	2002	2004	2006	2008	2010
需求量/万吨	236	246	257	276	286

(1)利用所给数据求年需求量与年份之间的线性回归方程 $\hat{y}=\hat{b}x+\hat{a}$；

(2)利用(1)中所求出的直线方程预测该地 2012 年的粮食需求量.

分析：本题的关键就是求出 $\hat{a},\hat{b}$ 的值，但如果直接求解，数据较大，运算量也会相应增加，容易出错. 可以考虑将年份与需求量相应减去一定数量，以减小运算量.

解：(1)由所给数据看出，年需求量与年份之间是近似直线上升的，下面来求线性回归方程，为此对数据预处理如下：

年份-2 006	-4	-2	0	2	4
需求量-257	-21	-11	0	19	29

对预处理后的数据，容易算得：

$\overline{x}=0$，$\overline{y}=3.2$，

$$\hat{b}=\frac{(-4)\times(-21)+(-2)\times(-11)+2\times 19+4\times 29}{4^2+2^2+2^2+4^2}$$
$$=\frac{260}{40}=6.5,$$

$\hat{a}=\overline{y}-\hat{b}\overline{x}=3.2$.

由上述计算结果，知所求线性回归方程为

$\hat{y}-257=\hat{b}(x-2\,006)+\hat{a}=6.5(x-2\,006)+3.2$，

即 $\hat{y}=6.5(x-2\,006)+260.2$. ①

(2)利用线性回归方程①，可预测 2012 年的粮食需求量为 $6.5\times(2\,012-2\,006)+260.2=6.5\times 6+260.2=299.2$(万吨)$\approx$300(万吨)(近似值可不写).

全能训练

基础达标

1. 下列关于回归方程的叙述正确的是 ( )

A. 回归方程只适用于所研究的样本

B. 回归方程都有时间性

C. 样本的取值范围会影响回归方程的适用范围

D. 回归方程是反映总体的唯一的回归模型

2. 设有一个线性回归方程 $\hat{y}=2-3.5x$，则变量 x 增加 1 个单位时 ( )

A. y 平均增加 3.5 个单位　　B. y 平均增加 2 个单位

C. y 平均减少 3.5 个单位　　D. y 平均减少 2 个单位

3. 已知某车间加工零件的个数 x 与所花费时间 y(h)之间的线性回归方程为 $\hat{y}=0.01x+0.5$，则加工 600 个零件大约需要 ( )

A. 6.5 h　　B. 5.5 h　　C. 3.5 h　　D. 0.5 h

4. 某化工厂为了预测某产品的回归率 y，需要研究它和原料有效成分含量 x 之间的相关关系. 现取了 8 对观测数据，计算得 $\sum\limits_{i=1}^{8}x_i=52$，$\sum\limits_{i=1}^{8}y_i=228$，$\sum\limits_{i=1}^{8}x_i^2=478$，$\sum\limits_{i=1}^{8}x_iy_i=1\,849$，则 y 对 x 的线性回归方程为 ( )

A. $\hat{y}=11.47+2.62x$　　B. $\hat{y}=-11.47+2.62x$

C. $\hat{y}=2.62+11.47x$　　D. $\hat{y}=11.47-2.62x$

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

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

B. 直线 $\hat{y}=\hat{b}x+\hat{a}$ 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 $\dfrac{\sum\limits_{i=1}^{n}x_iy_i-n\overline{x}\,\overline{y}}{\sum\limits_{i=1}^{n}x_i^2-n\overline{x}^2}$

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

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 $\hat{y}=\hat{b}x+\hat{a}$ 必过定点_____.

能力提升

7. 若某地财政收入 x 与支出 y 满足线性回归方程 $\hat{y}=bx+a+e$(单位：亿元)，其中 $b=0.8$，$a=2$，$|e|\leqslant 0.5$. 如果今年该地区财政收入 10 亿元，则今年支出预计不会超过 ( )

A. 10 亿元　　B. 9 亿元

C. 10.5 亿元　　D. 9.5 亿元

8. 下表提供了某厂节能降耗技术改造后，生产甲产品过程中记录的产量 x(吨)与相应的能耗 y(吨标准煤)的几组对照数据.

x	3	4	5	6
y	2.5	3	4	4.5

(1)请画出上表数据的散点图；

(2)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 $\hat{y}=\hat{b}x+\hat{a}$；

(3)已知该厂技术改造前 100 吨甲产品的生产能耗为 90 吨标准煤，试根据(2)求出的线性回归方程，预测生产 100 吨甲产品的生产能耗比技术改造前降低多少吨标准煤.

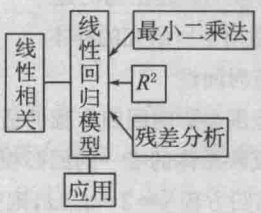
(参考值：$3\times2.5+4\times3+5\times4+6\times4.5=66.5$)

第2课时 相关指数 R^2、残差分析

学习目标

1. 了解模型拟合效果的分析工具——残差分析和 R^2.
2. 理解、分析残差变量.
3. 理解 R^2 的含义.

知识结构



知识解读

知识点一 残差

1. 定义

样本点(x_i,y_i)和它在回归直线上的位置差异 $\hat{e}_i=y_i-\hat{y}_i$ 称为残差. 它是随机误差 e 的效应.

在实际应用中,用回归方程 $\hat{y}=\hat{b}x+\hat{a}$ 中的 $\hat{y}$ 估计 $y=bx+a+e$ 中的 y,由于 $e=y-(bx+a)$,所以 $\hat{c}=y-\hat{y}$ 是 e 的估计量.

对于样本点$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$而言,它们的随机误差为 $e_i=y_i-bx_i-a,i=1,2,\cdots,n$,其估计值为 $\hat{e}_i=y_i-\hat{y}_i=y_i-\hat{b}x_i-\hat{a}(i=1,2,\cdots,n)$,$\hat{e}_i$ 称为相应于点(x_i,y_i)的残差.

【温馨提示】

残差 $\hat{e}$ 是随机误差 e 的估计量,$\hat{e}_i$ 是相应于点(x_i,y_i)的 e_i 的估计值. 可以通过残差发现原始数据中的可疑数据,判断所建立模型的拟合效果,残差的绝对值越小,拟合效果越好.

2. 残差图

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.

如果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,说明模型拟合的精度越高,回归方程的预报精度越高. 如果残差点分布不均匀,应首先确认采集的样本点是否有误. 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来拟合数据;如果数据的采集没有错误,则需要寻找其他的原因.

3. 利用残差图进行残差分析的步骤

(1)计算每组观察数据的残差 $\hat{e}_i=y_i-\hat{y}_i(i=1,2,\cdots,n)$,即残差等于观测值减去预测值. 当残差的绝对值比较小时,说明回归模型拟合数据较好.

(2)画残差图. 残差图的纵坐标为残差,横坐标通常可以是观测样本的编号,或自变量 x,或因变量的预测值等,分析结果是一致的.

(3)分析残差图. 几种常见的残差图如下所示:

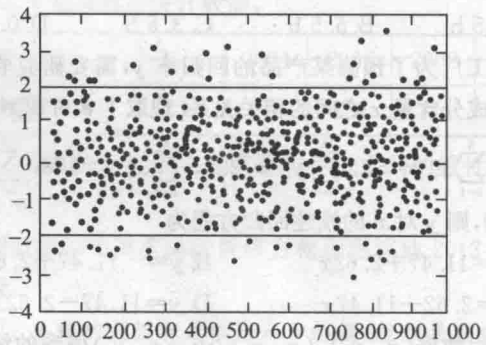


图 1.1-3

图 1.1-3:残差图中的点分布在以原点为中心的水平带状区域中,并且沿水平方向的分布规律相同,说明残差是随机的,所选择的回归模型是合理的.

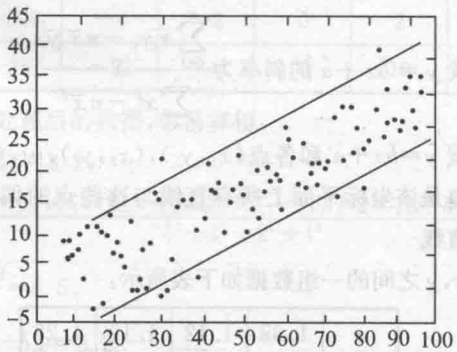


图 1.1-4

图 1.1-4:残差图中的点分布在一条倾斜的带状区域中,并且沿带形区域方向的分布规律相同,说明残差与其横坐标有线性关系,此时所选用的回归模型的效果不是最好的,有改进的余地.

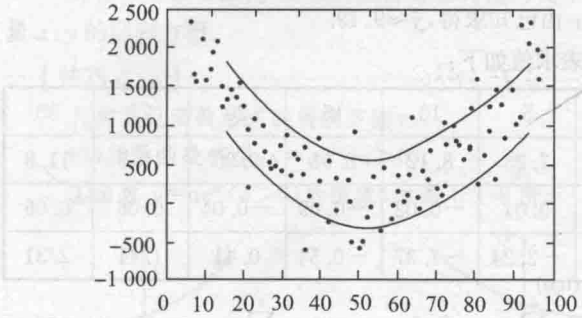


图 1.1-5

图 1.1-5：残差图中的点分布在一条二次曲线型的弯曲带状区域中，说明残差与其横坐标有二次关系，此时所选用的回归模型的效果不是最好的，有改进的余地.

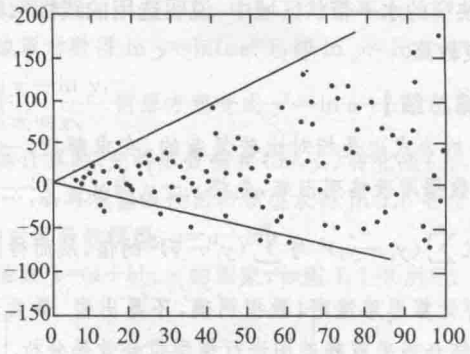


图 1.1-6

图 1.1-6：残差图中的点的分布范围随着横坐标的增加而增加，说明残差的方差与横坐标有关，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)查找异常样本数据(此时常以数据编号为横坐标，以方便查找异常数据). 根据计算的残差值和残差图，观察是否存在残差绝对值特别大的点，即远离横轴的点. 如果存在远离横轴的点，就要研究它出现的原因，如是否在数据收集和录入中发生了错误，如果有错误，改正后重新建立回归模型.

【例 1】　已知(1，1.1)，(2，2.1)，(3，3.3)，(4，3.9)，(5，5.05)，(6，5.9)是求出线性回归方程 $\hat{y}=x$ 的部分数据，就这些数据求残差并作残差图，作残差分析.

分析：本题先要求出残差，再用横轴表示编号、纵轴表示残差，作出残差图，最后利用残差图进行残差分析.

解：列表可求残差 $\hat{e}_i=y_i-\hat{y}_i$.

x_i	1	2	3	4	5	6
y_i	1.1	2.1	3.3	3.9	5.05	5.9
$\hat{y}_i$	1	2	3	4	5	6
$y_i-\hat{y}_i$	0.1	0.1	0.3	−0.1	0.05	−0.1

残差图如图 1.1-7 所示.

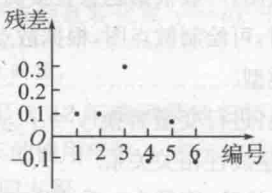


图 1.1-7

残差分析：由表中数据及残差图可以看出，残差比较均匀地落在宽度不超过 0.4 的水平带状区域内，说明线性回归模型拟合效果较好.

第 3 个样本点的残差绝对值较大，是可疑数据.

【规律总结】

作残差图从及进行残差分析类的题目难度不大，关键是要准确求出各点所对应的残差值，并选取合适的横轴、纵轴，描点即可. 在作残差分析时，要注意点的分布情况，特别要注意一些特殊点及其出现的原因.

知识点二　相关性检验

相关指数 R^2 可以用来刻画回归的效果，其计算公式为

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

对于已经获取的样本数据，R^2 表达式中的 $\sum_{i=1}^{n}(y_i-\bar{y})^2$ 为确定的数. 因此 R^2 越大，意味着残差平方和 $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ 越小，即模型的拟合效果越好；R^2 越小，残差平方和越大，即模型的拟合效果越差.

在线性回归模型中，R^2 表示解释变量对于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. R^2 是常用的选择模型的指标之一，在实际应用中应该尽量选择 R^2 大的回归模型.

【拓展延伸】

相关系数 r 与 R^2 的关系

在含有一个解释变量的线性回归模型中，R^2 恰好等于相关系数 r 的平方，推导如下：

由于 $\sum_{i=1}^{n}(y_i-\bar{y})^2$

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

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

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

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

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

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

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

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

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

即 $\sum_{i=1}^{n}(y_i-\bar{y})(y_i-\hat{y}_i)=0$.

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

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

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

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

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

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

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

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

$=\left[\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]^2\cdot\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$

$=\dfrac{\left[\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right]^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2\cdot\sum_{i=1}^{n}(y_i-\bar{y})^2}=r^2$.

由上式以及线性相关系数的性质知:在线性回归模型中有 $0\leqslant R^2\leqslant 1$. 因此在线性回归模型中,$R^2$ 和两个变量的相关系数都能刻画用线性回归模型拟合数据的效果. 相关系数的绝对值越大,R^2 就越大,线性回归模型拟合数据的效果就越好.

当 $r=\pm 0.8$ 时,$R^2=0.64$;当 $r=\pm 0.9$ 时,$R^2=0.81$. 通常当 $R^2>0.80$ 时,认为线性回归模型对该组数据是很有效的,这时两个变量的相关系数的绝对值几乎超过 0.9.

【例 2】 在研究弹簧伸长长度 y(cm)与拉力 x(N)的关系时,对不同拉力的 6 根弹簧进行测量,测得下表中的数据:

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

若依据散点图及最小二乘法求出线性回归方程为 $\hat{y}=0.18x+6.34$,求 R^2,并结合残差说明拟合效果.

分析:首先直接代入 R^2 公式求解,然后结合残差说明回归模型的拟合效果.

解:由已知求得,$\bar{y}\approx 9.49$.

列表求值如下:

x_i	5	10	15	20	25	30
y_i	7.25	8.12	8.95	9.90	10.9	11.8
$y_i-\hat{y}_i$	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

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

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

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拟合精度较高.

【规律总结】

R^2 的公式也是相对比较复杂的,在求解时,一定要将所给的数据用表格列出来,先将 $y_i-\hat{y}_i$ 与 $y_i-\bar{y}$ 一一对应求出,再求 $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ 与 $\sum_{i=1}^{n}(y_i-\bar{y})^2$ 的值,从而得到 R^2 的值,这样计算思路清晰,数据明确,不易出错. 最后通过 R^2 的值并结合残差或残差图进行模型拟合效果分析.

知识点三 建立回归模型的基本步骤

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

(3)由经验确定回归方程的类型(如我们观察到数据呈线性关系,则选用线性回归方程).

(4)按一定规则(如最小二乘法)估计回归方程中的参数.

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

知识点四 非线性回归问题的处理

在解决实际问题时,研究的两个变量不一定都呈线性相关关系. 对于这类问题,常采用适当的变量代换,把问题转化为线性回归问题,求出线性回归模型后,再通过相应的变换,得到非线性回归方程.

方法步骤:

(1)绘制散点图:一般根据题意直接确定变量满足的曲线类型,不能确定时,可绘制散点图,根据散点的分布,选择接近的、合适的曲线类型.

(2)变量替换:进行变量替换 $y'=f(y)$,$x'=g(x)$,使变换后的两个变量呈线性相关关系.

(3)求方程并检验:按最小二乘法原理求线性回归方程并进行相关性检验.