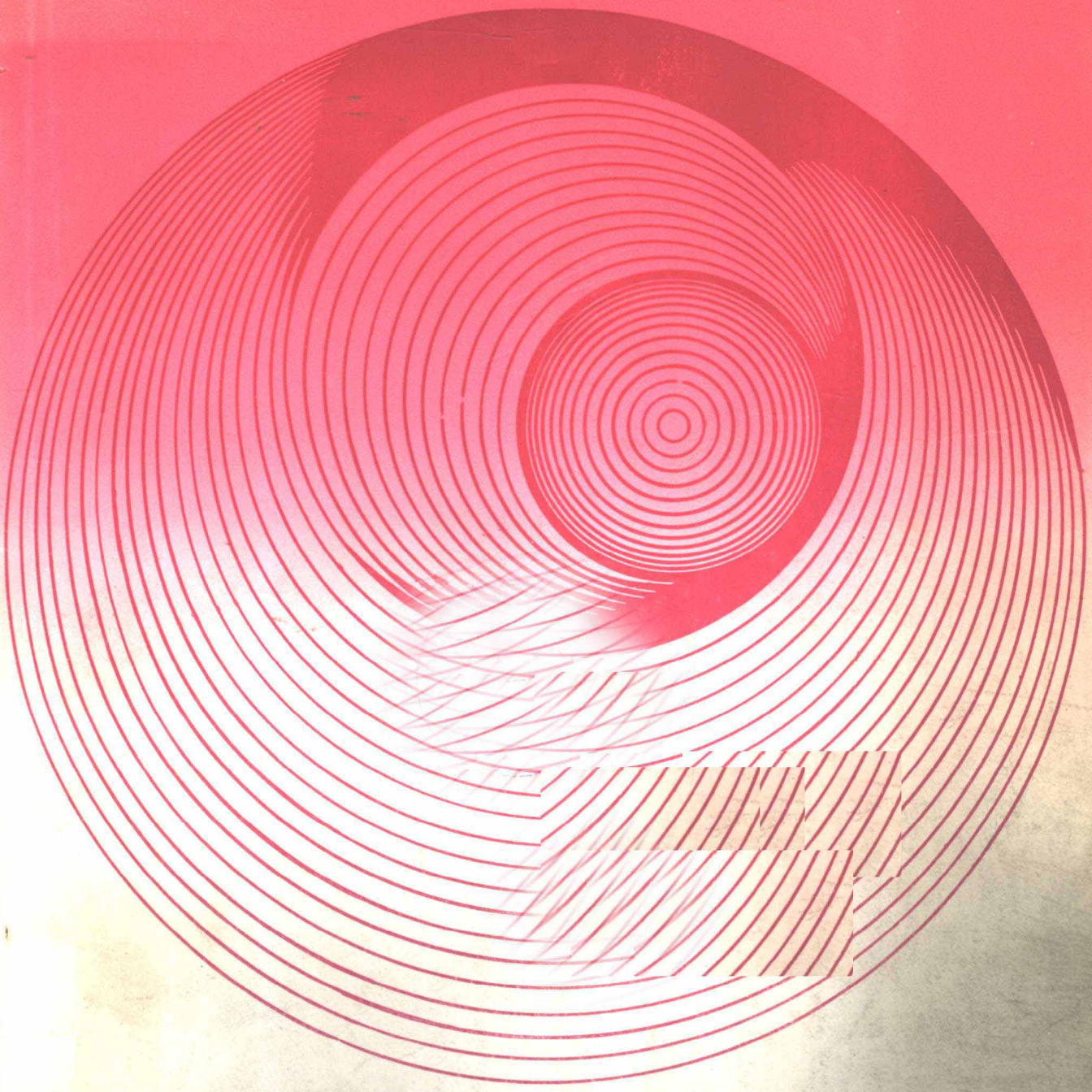
应用回归分析

王学仁 温忠赍 编译

重庆大学出版社



YINGYONG HUIGUI FENXI

内 容 简 介

本书是回归分析的一本基础读物，具有基本的数理统计知识及初步的矩阵和微积分知识的读者都可以阅读。

书中收集了目前解决回归问题的主要方法及其应用。除了包含其它同类教科书中常见的线性回归参数估计和假设检验外，还详细讨论了残差检验、预处理数据、最优回归方程的选择、大型回归设计方案、方差分析的回归处理，最后介绍了非线性模型的一些结果。

读者对象为理、工、农、医、经济、管理等有关专业的本科生和研究生以及应用统计工作者。

应用回归分析

[redacted]学仁　温忠麟　编译

[redacted]任编辑　周　任

重庆大学出版社出版发行

新 华 书 店 经 销

重庆大学出版社印刷厂印刷

开本：787×1092　1/16　印张：16.625　字数：415千

1989年5月第1版　1991年5月第1次印刷

印数：1——3500

标准书号：ISBN 7-5624-0133-0/O·21　定价：4.35 元

序

由美国芝加哥大学、威斯康星大学麦迪逊分校、加州伯克力大学和哈佛大学等联合来华招收统计学研究生考试所指定的应用统计的主要参考书《Applied Regression Analysis》(《应用回归分析》)是由美国N.R.Draper和H.Smith所著，由美国John Wiley & Sons出版公司于1966年出第一版。作为美国许多院校和工厂应用统计的教材，普遍受到欢迎。为了反映七十年代回归分析的新思想、新理论和新方法，作者在多年多校教学实践的基础上为第一版增加了许多回归分析的新的和实用的内容，1981年又出了第二版。该书着重统计思想、统计背景和统计方法的阐述，内容丰富，实用性强。

为了满足国内有关专业学生和应用统计工作者的需要，现将原著第二版编译出版。除了删去了原著第七章（两个特殊问题）外，本书基本上保留了原著的全部内容。为了节省篇幅，删去了原著书末所附的参考文献和出现于正文中诸多详尽的引文出处。在内容编排上作了不少调整，文字上也作了大量处理和加工，使本书在保持原著基本思想和内容的前提下比原著更简炼、更连贯，条理更清楚，节目归类更合理。我们选留了有代表性的习题，其中不少是生产或科研中的实际问题，如何处理和分析这些具体问题，不一定都能从正文中找到答案，因而保留了习题答案，供读者参考。

本书编译过程中，参考了云南大学数学系84级研究生班数理统计专业学员的原著翻译稿，本书第三章部分小节的编译还参考了陈希孺、王松桂著的《近代回归分析》，我们向这些译者和作者表示衷心感谢。由于水平所限，本书难免有不妥甚至谬误之处，敬请专家和读者批评指正。

编译者

1990年4月

目 录

第一章　最小二乘法拟合的直线

1.0　概　论

在当今社会生产生活的各个领域，可以说不存在数据资料缺乏的问题。在工业、农业、生物、医学、气象、水文、地质等部门的生产和科研中，人们利用各种各样的仪器和方法收集的数据日益增多；社会科学、人文科学中的许多学科也有将定性指标数量化的趋势。以工业生产为例，人们经常收集诸如输入的温度、反应物浓度、催化剂的百分比、消耗率、压力等数据。对成品的抽样检查则可以获取成品的寿命、光洁度、承受力、颜色等信息。在许多工厂内都积累了大量这些类型的数据，但他们只是简单地利用它们计算平均值、百分比等，而没有对数据进行整理、分析，因而没能从数据中获取更多的有用信息。

虽然回归分析对于收集数据有时也可以提供一定的方法（见1.8节），但是本书的目的不是想说明对于某种给定的程序该不该收集某种类型的数据，而是要说明从以上提到的那些数据堆中抽取信息的技术，并揭示出隐含在那些数据中的相互关系的主要特征。

在含有变量的系统中，考察一些变量对另一些变量的作用是很有趣的。它们之间可能存在一种简单的函数关系，也可能存在一种非常复杂的函数关系。对于后一种情形，或许希望用一些简单的函数，例如多项式去近似这种关系。通过考察这种近似函数，往往能够对隐含的内在关系有更多的了解，并且在某些重要变量变化时评价它们单独的或共同的作用。

即使在变量之间存在着感觉不到的物理关系，也期望用某些类型的数学方程式去描述这种关系。也许这种方程式并没什么物理意义，但在某些限制下，却可以用方程式通过一些熟知的变量去预测另一些变量。

在本书中，将采用一种获得数学关系的特殊方法。该方法基于一个原始假设：即某种关于未知参数的线性关系成立。在某些有助于利用数据资料的其他假设下，可以进行未知参数的估计，并找到一个拟合方程；可以衡量这个方程的价值，还可以检验上述那些假设正确与否。这个过程的一个最简单的例子是：当获得了观察值(X_1, Y_1)，(X_2, Y_2)，…，(X_n, Y_n)后建立一条拟合直线。本章将用一种简单的代数方法解决这个问题。为了处理包含多个变量的问题，矩阵的方法是最基本的，将在第二章中介绍。此外，第二章的内容还包含了许多关于最一般的回归问题的基本结果，其中有些结果在第四章讨论由一个平面方程给出的两个自变量X_1和X_2与一个因变量Y的关系时得到了应用。第四章涉及了更复杂的模型。第五章讨论了选择一个最优拟合方程的方法。第六章解决了建立模型的步骤以及由此产生的各种问题。第七章讨论了方差分析的回归处理。许多章节都会用到的残差检验在第三章给出。第八章简介了非线性模型。

假设本书的读者已学过简单的矩阵知识及一些初步的统计课程，掌握了一些基本的统计概念，如随机变量的数学期望（均值）和方差，两个随机变量之间的协方差，分布（正态分布、t-分布、χ^2-分布、F-分布），参数估计，简单的假设检验（包括单侧和双侧的t-检验和F-检验）。然而，即使这些知识已经生疏了或掌握得不好的读者，读过本书后也会有所

收益。

本书着重统计背景和方法的叙述，而不拘泥于严密的数学推导。同时，对于涉及的数学知识不深的问题和结论，则尽量给予分析和证明。我们并不打算将本书编写成一本面面俱到的回归分析手册，而是希望为读者提供一本解决实际回归问题所必备的基本教材。

在一个问题中，首先要区分两种主要类型的变量。一种变量相当于通常函数关系中的自变量，对这样的变量能够赋予其一个需要的值（如加入催化剂的比率）或者能够取到一个可观察但不能人为控制的值（如室外温度）。这种变量称为**预报变量**，或称为自变量。预报变量的变化能波及另一些变量（如化工产品的纯度或颜色），这样的变量称为**响应变量**，或称为因变量。人们通常感兴趣的问题是，预报变量的变化对响应变量的取值有什么样的影响。预报变量与响应变量之间的区别并无明显的界线，往往与考虑的课题有关。例如，某因素在产品生产过程的中期可能认为是响应变量，但对于生产出来的产品的颜色却可以认为是预报变量。然而，在实际问题中，两种变量是容易区分的。在一组特殊的数据中，可能由于试验方法所致，其中的两个或多个预报变量一起发生变化，这是不期望的事，因为这样一来得不到单个变量作用的信息，但这是难于避免的。

现在看一下附录A的数据，这些数据是从一个庞大的工业中心的一个蒸汽工厂按一定时间间隔取到的观测值，其中记录了如下十个变量：

1. 每月蒸汽用量；
2. 每月纯脂肪酸贮存量；
3. 生甘油产量；
4. 平均风速；
5. 每月天数；
6. 每月生产天数；
7. 低于32°F的天数；
8. 平均气温（°F）；
9. （平均风速）^{2}；
10. 起动次数。

人们的主要兴趣在于每月的蒸汽用量以及该变量如何随另一些变量而变化。因此，将变量1视为响应变量Y，其余的X_2，X_3，…X_{10}作为预报变量。

下面将用一种称为最小二乘法的分析方法处理数据，并获得一些有意义的结论。这种分析方法通常称为**回归分析**。“回归”（regression）一词是由英国著名人类学家和气象学家Francis Galton爵士引入的。他在“身高遗传中的平庸回归”的论文中使用了“回归”的术语。在该文中，Galton阐述了他的重大发现：子代的身高并不像其父代，而是趋向于比他们的父代更加平庸（更加平均）。就是说，如果父代身材高大，则子代的身材要比父代小一些；如果父代身材矮小，则子代要比父代高大一些。换言之，子代的身材有向平均值靠拢的趋向，因此他用回归一词来描述子代身高与父代身高的关系。虽然今天的“回归”一词已没多少原始含义，但人们仍然继续使用这个词。

下面先把最小二乘法用于一个最简单的情形，即对于已知的数据，拟合一条直线来说明两个变量X和Y之间的关系，然后再推广到多个变量的情形。

1.1 两个变量之间的直线关系

在许多试验工作中，都希望弄清楚一个变量的变化对另一个变量有些什么影响。有时候，两个变量恰好由一条直线联系起来。例如，当一条简单电路的电阻R保持常量不变时，电流I的变化与电压V的变化就是直线关系，这由欧姆定律$I=V/R$便知。假如不懂欧姆定律，由V的变化和I的观测值，可以凭经验获得这种关系。当R固定时，观测I相对于V的变化，可得到一条过原点的近似直线。虽然它们之间的关系是确切的直线关系，但在测量过程中可能会出现误差，因此在描点作图时观测点也就不会恰好落在一条直线上。然而，在利用V来预报I的值时，应该用过原点的直线。

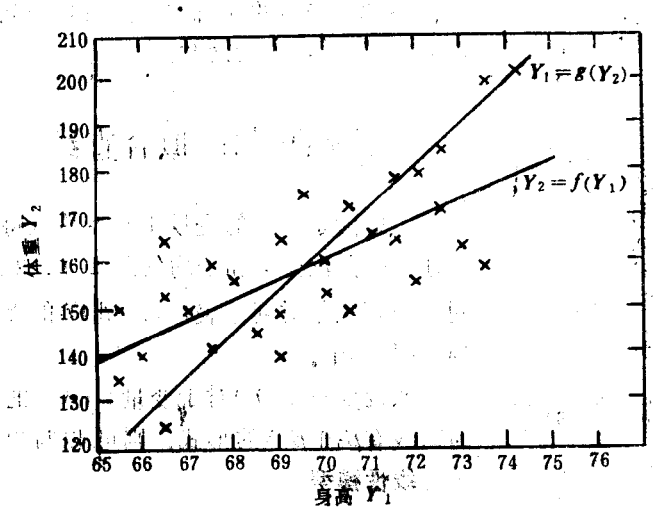


图1.1 三十个男子的身高和体重

有时候，即使不考虑误差，某种关系也不是精确的直线关系，不过这时考虑直线关系仍有很大意义。例如，假设要考察某一给定总体成年男子的身高和体重，描出点对（Y_1，Y_2）=（身高，体重），便得到图1.1的图象。

注意到对于已知的身高，其对应的体重观测点有一个范围，反之亦然。这种量的变化，部分地是由测量误差引起的，但主要还是由各个变量单独变化引起的。因此不能指望实际身高和体重之间存在唯一的关系，但也可以注意到，随着身高观测值的增加，对应的体重观测值的平均值也会增加，这种对于给定的身高观测值所得的体重观测平均值的轨迹称为体重对于身高的**回归曲线**，记为$Y_2=f(Y_1)$。身高对体重的回归曲线也同样存在，记为$Y_1=g(Y_2)$。假设这两条曲线均为直线（一般情况下可能不是），这两条直线一般来说是不相同的，如图中所示的两条直线。

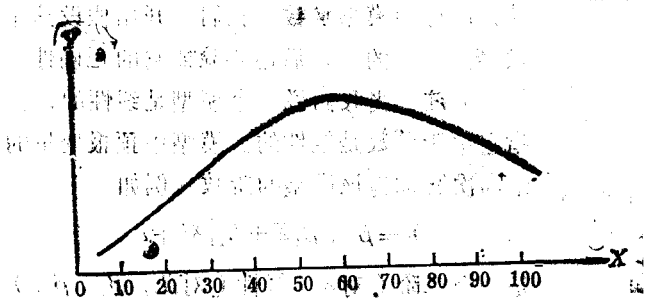


图1.2 响应关系

现在，假设已经有了每个人身高的记录，但不知道他们各自的体重，该如何将它们估计出来？利用体重对身高的回归线，可以得到各个给定身高所对应的体重平均观测值，人们就用这个值去估计体重。

通常，两个变量之间不是直线关系，但在一个小范围内却近似于直线关系。如图1.2中所示的响应关系，在$0\leqslant X\leqslant 100$区间内，显然不是直线关系，但如果只对$0\leqslant X\leqslant 45$范围内

的关系感兴趣，就可以看作是直线关系。当然这种直线关系不适合 $0 \leqslant X \leqslant 45$ 以外的场合，这一点在做预报时是值得注意的，就是说，利用直线关系对 $0 \leqslant X \leqslant 45$ 以外的点作出的预报将是不可靠的。预报变量不止一个时，也要注意类似情况。

当要考察一个随机变量Y与一个非随机变量X之间的相关性时，描述Y与X的关系的方程通常称为**回归方程**。几乎在本书的各章，都假设响应变量是随机变量而预报变量不发生随机变化。然而，就实际而言，这种假设几乎是不真的，但如果不这样做，其拟合程序要复杂得多。为了避免这个问题，在那些可以认为任何一个预报变量的随机变化与预报变量域相比很小的场合，才可以使用最小二乘法，此时忽略预报变量的随机变化。对这个问题的进一步说明见2.8节。

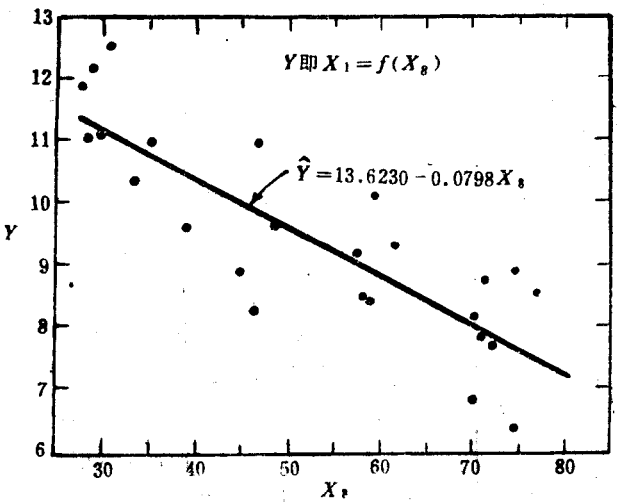


图1.3 数据和拟合的直线

表1.1 变量1和8的25组观测值

观测序号	变量	
	1(Y)	8(X)
1	10.98	35.3
2	11.13	29.7
3	12.51	30.8
4	8.40	58.8
5	9.27	61.4
6	8.73	71.3
7	6.36	74.4
8	8.50	76.7
9	7.82	70.7
10	9.14	57.5
11	8.24	46.4
12	12.19	28.9
13	11.88	28.1
14	9.57	39.1
15	10.94	46.8
16	9.58	48.5
17	10.09	59.3
18	8.11	70.0
19	6.83	70.0
20	8.88	74.5
21	7.68	72.1
22	8.47	58.1
23	8.86	44.6
24	10.36	33.4
25	11.08	28.6

1.2 线性回归：拟合直线

考虑附录A中的变量1和变量8(每月使用的蒸汽磅数和平均气温°F)，观测值的相应数对列于表1.1，相应的点描在图1.3中。

假设变量1(Y)对于变量8(X)的回归线有$\beta_0+\beta_1X$的形式，则可以写出一阶**线性模型**

$$Y=\beta_0+\beta_1X+\varepsilon \qquad (1.2.1)$$

β_0，β_1称为模型**参数**。我们一开始假设这个模型是成立的，以后还要验证它的正确性。

（**注** 当我们说一个模型是线性的，是指它关于参数是线性的。模型中预报变量的最高次幂称为该模型的阶数。例如

$$Y=\beta_0+\beta_1X+\beta_{11}X^2+\varepsilon$$

是一个二阶（对X）线性（对β_0，β_1，β_{11}）回归模型。除了特别申明一个模型是非线性的外，给出的模型都理解为是线性的。）

模型中β_0，β_1和ε是未知的。ε随着每个观测值Y变化，是随机变量。β_0和β_1是固定的，但不能求出它们的精确值，除非对Y和X发生的一切情况都进行观测。不过可以利用表1.1提供的数据，作出β_0和β_1的估计b_0和b_1。于是可以写出

$$\hat{Y}=b_0+b_1X \qquad (1.2.2)$$

其中$\hat{Y}$表示b_0和b_1确定之后对于给定的X相应的Y的**预报值**（也称为Y的**拟合值**或**回归值**），（1.2.2）可作为**预报方程**（也称为**拟合方程**或**回归方程**），对应的直线称为**拟合直线**（或**回归直线**）。代入一个X的值，就得到对应于这个X的Y的均值的预报。

为了估计模型的参数，现在使用最小二乘法。假设有n组观测值（X_1，Y_1），（X_2，Y_2），…，（X_n，Y_n）（在本例中$n=25$），则由（1.2.1）有

$$Y_i=\beta_0+\beta_1 X_i+\varepsilon_i \qquad (1.2.3)$$

$i=1, 2\cdots, n$，从而得到偏离真实直线的偏差平方和

$$S(\beta_0, \beta_1)=\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}(Y_i-\beta_0-\beta_1 X_i)^2 \qquad (1.2.4)$$

现在选择估计b_0和b_1取代（1.2.4）中的β_0和β_1使得S达到最小值，见图1.4（注意图中X_i，Y_i是常数）。为此将（1.2.4）分别对β_0，β_1求偏导数得

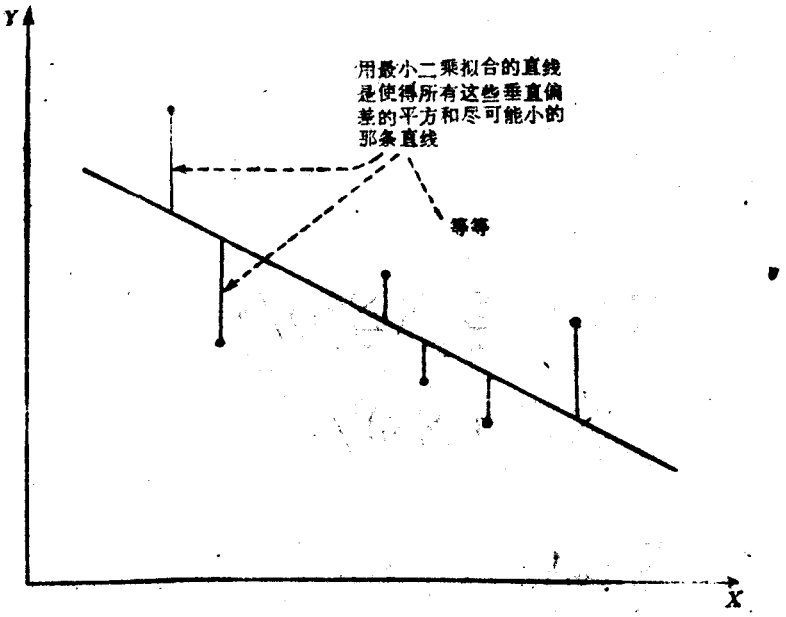


图1.4 最小二乘法拟合的直线

$$\frac{\partial S}{\partial \beta_0}=-2\sum_{i=1}^{n}(Y_i-\beta_0-\beta_1 X_i)$$

$$\frac{\partial S}{\partial \beta_1}=-2\sum_{i=1}^{n}X_i(Y_i-\beta_0-\beta_1 X_i) \qquad (1.2.5)$$

令（1.2.5）等于0，并用b_0，b_1取代β_0，β_1：

$$\sum_{i=1}^{n}(Y_i-b_0-b_1 X_i)=0$$

$$\sum_{i=1}^{n}X_i(Y_i-b_0-b_1 X_i)=0 \qquad (1.2.6)$$

于是有

$$\sum_{i=1}^{n} Y_i - nb_0 - b_1 \sum_{i=1}^{n} X_i = 0$$

$$\sum_{i=1}^{n} X_i Y_i - b_0 \sum_{i=1}^{n} X_i - b_1 \sum_{i=1}^{n} X_i^2 = 0 \qquad (1.2.7)$$

即

$$nb_0 + b_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i$$

$$b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i \qquad (1.2.8)$$

(1.2.8) 称为**正规方程**。

由正规方程解得

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)/n}{\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2/n}$$

$$= \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \qquad (1.2.9)$$

$$b_0 = \overline{Y} - b_1 \overline{X} \qquad (1.2.10)$$

其中 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$，$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ 分别称为 X 的均值和 Y 的均值。

用这种方法求出的参数 β_0，β_1 的估计 b_0，b_1 称为**最小二乘估计**，简称为**LS估计**。

将 (1.2.10) 代入 (1.2.2) 便得到回归方程

$$\hat{Y} = \overline{Y} + b_1 (X - \overline{X}) \qquad (1.2.11)$$

其中 b_1 由 (1.2.9) 决定。

显然，b_1是拟合直线的斜率，b_0是拟合直线在$X=0$处的截距。不难看出，点$(\overline{X}, \overline{Y})$落在拟合直线上。

人们经常使用下列方便的记号及等式的各种变形：

$$S_{XX}=\sum_{i=1}^{n}(X_i-\overline{X})^2=\sum_{i=1}^{n}(X_i-\overline{X})X_i$$

$$=\sum_{i=1}^{n}X_i^2-\Big(\sum_{i=1}^{n}X_i\Big)^2\Big/n=\sum_{i=1}^{n}X_i^2-n\overline{X}^2$$

$$S_{XY}=\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})=\sum_{i=1}^{n}(X_i-\overline{X})Y_i$$

$$=\sum_{i=1}^{n}X_i(Y_i-\overline{Y})=\sum_{i=1}^{n}X_iY_i-\Big(\sum_{i=1}^{n}X_i\Big)\Big(\sum_{i=1}^{n}Y_i\Big)\Big/n$$

$$=\sum_{i=1}^{n}X_iY_i-n\overline{X}\,\overline{Y}$$

$\sum_{i=1}^{n}X_i^2$称为X的**总平方和**，S_{XX}称为X的（总）**校正平方和**。

计算b_1较易记住的公式是

$$b_1=S_{XY}/S_{XX} \qquad (1.2.9a)$$

用袖珍计算器计算b_1，通常用(1.2.9)的第一等式。然而，为了避免舍入误差，计算时应尽可能多地采用有效数字。用(1.2.9)的第二等式在电子计算机上计算b_1，可以得到更加精确的结果。

现在将上述结果用于表1.1列出的数据，这里：

$$n=25$$

$$\sum Y_i=10.98+11.13+\cdots 11.08=235.60$$

$$\overline{Y}=235.60/25=9.424$$

$$\sum X_i=35.3+29.7+\cdots+28.6=1315$$

$$\overline{X}=1315/25=52.60$$

$$\sum X_iY_i=(10.98)(35.3)+(11.13)(29.7)+\cdots$$

$$+(11.08)(28.6)=11821.4320$$

$$\sum X_i^2=(35.3)^2+(29.7)^2+\cdots+(28.6)^2=76323.42$$

$$b_1=\frac{11821.4320-(1315)(235.60)/25}{76323.42-(1315)^2/25}$$

$$=-0.079829$$

拟合方程是

$$\hat{Y}=9.4240-0.079829(X-52.60)$$

即

$$\hat{Y}=13.623005-0.079829X$$

拟合的回归线见图1.3。对每组X_i，Y_i，可以求出拟合值$\hat{Y}_i$以及**残差**$Y_i-\hat{Y}_i$，见表1.2。

表1.2 观测值，拟合值和残差

序 号	Y_i	$\hat{Y}_i$	$Y_i-\hat{Y}_i$
1	10.98	10.81	0.17
2	11.13	11.25	−0.12
3	12.51	11.17	1.34
4	8.40	8.93	−0.53
5	9.27	8.72	0.55
6	8.73	7.93	0.80
7	6.36	7.68	−1.32
8	8.50	7.50	1.00
9	7.82	7.98	−0.16
10	9.14	9.03	0.11
11	8.24	9.92	−1.68
12	12.19	11.32	0.87
13	11.88	11.38	0.50

续表

序 号	Y_i	$\hat{Y}_i$	$Y_i-\hat{Y}_i$
14	9.57	10.50	−0.93
15	10.94	9.89	1.05
16	9.58	9.75	−0.17
17	10.09	8.89	1.20
18	8.11	8 03	0.08
19	6.83	8.03	−1.20
20	8.88	7.68	1.20
21	7.68	7.87	−0.19
22	8.47	8.98	−0.51
23	8.86	10.06	−1.20
24	10 36	10.96	−0.60
25	11.08	11.34	−0.26

因为

$$\hat{Y}_i=\overline{Y}+b_1(X_i-\overline{X})$$

所以

$$Y_i-\hat{Y}_i=(Y_i-\overline{Y})-b_1(X_i-\overline{X})$$

$$\sum_{i=1}^{n}(Y_i-\hat{Y}_i)=\sum_{i=1}^{n}(Y_i-\overline{Y})-b_1\sum_{i=1}^{n}(X_i-\overline{X})=0$$

这说明残差之和为零，但在实际计算中，残差之和可能不为零，这是由舍入误差引起的。在任何一个回归问题中，如果β_0是由正则方程的第一个式子解出，则残差之和总为零。

1.3 回归估计的精度

现在讨论回归直线的估计精度问题。

考察等式

$$Y_i-\hat{Y}_i=Y_i-\overline{Y}-(\hat{Y}_i-\overline{Y}) \qquad (1.3.1)$$

它的几何意义如图1.5所示，残差$e_i=Y_i-\hat{Y}_i$是如下两个量之差：i)观测值Y_i与均值$\overline{Y}$的偏差；ii)拟合值$\hat{Y}_i$与均值$\overline{Y}$的偏差。注意到

$$\sum_{i=1}^{n}\hat{Y}_i/n=\sum_{i=1}^{n}(b_0+b_1X_i)/n$$

$$=(nb_0+b_1n\overline{X})/n$$

$$=b_0+b_1\overline{X}=\overline{Y}$$

即$\hat{Y}_i$的平均值与Y_i的平均值相同。这个事实又一次证明了$\sum_{i=1}^{n}e_i=\sum_{i=1}^{n}(Y_i-\hat{Y})=n\overline{Y}-n\overline{Y}=0$

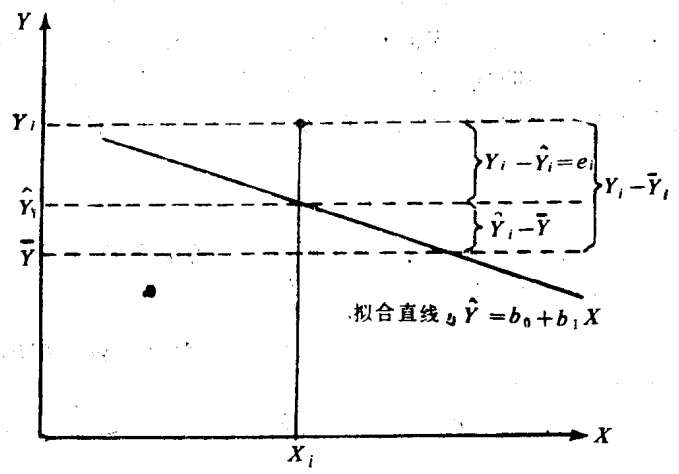


图1.5 (1.3.1)式的几何意义

将(1.3.1)写成$Y_i-\overline{Y}=(\hat{Y}_i-\overline{Y})+(Y_i-\hat{Y}_i)$，等式两边平方，再从$i=1$到$n$求和，得到

$$\sum(Y_i-\overline{Y})^2-\sum(\hat{Y}_i-\overline{Y})^2+\sum(Y_i-\hat{Y}_i)^2 \qquad (1.3.2)$$

其中交叉项

$$2\sum(\hat{Y}_i-\overline{Y})(Y_i-\hat{Y}_i)$$

$$=2\sum b_1(X_i-\overline{X})\{(Y_i-\overline{Y})-b_1(X_i-\overline{X})\}$$

$$=2b_1(S_{XY}-b_1S_{XX})=0$$

这里用了等式$\hat{Y}_i-\overline{Y}=b_1(X_i-\overline{X})$，由此等式还可得到

$$\sum(\hat{Y}_i-\overline{Y})^2=\sum b_1^2(X_i-\overline{X})^2-b_1^2S_{XX}=b_1S_{XY} \qquad (1.3.3)$$

(1.3.2)左边是Y的校正平方和，简写为校正SS。$\hat{Y}_i-\overline{Y}$是第i次观测的预报值与均值的偏差，其平方和称为**回归平方和**，简写为回归SS。$Y_i-\hat{Y}_i$是第i次观测值与它的预报值的偏差(残差)，其平方和称为**残差平方和**，简写为残差SS。这样，可以将(1.3.2)表示为

校正平方和＝回归平方和＋残差平方和

这就将Y关于其均值的方差(即校正平方和)分解为两部分，前一部分是由回归线引起的，后一部分则是由于实际观测值没有落在回归线上引起的(否则残差平方和为零)。由此找到了一种判别回归线拟合程度好坏的方法：看校正SS中，包含了多少回归SS和残差SS。如果回归SS远比残差SS大，或者$R^2=$(回归SS)/(校正SS)接近于1，则将感到满意。

每一个平方和都与一个称为**自由度**(记为df)的数联系在一起。自由度表示在平方和中独立的项数。例如，在校正SS中，因为$Y_1-\overline{Y}$，$Y_2-\overline{Y}$，…，$Y_n-\overline{Y}$之和为零，所以只有

$n-1$个是独立的，故自由度为$n-1$。可以用Y_1，Y_2，…，Y_n的一个函数即b_1来计算回归SS（见（1.3.3）），因而这个平方和的自由度是1。用变量代换，可以求出残差SS的自由度是$n-2$，这表明残差是从需要估计两个参数的直线模型的拟合中出现的。一般地，残差平方和有“观测次数－需要估计的参数个数”个自由度。与（1.3.2）相应，可得自由度的分解为

$$n-1=1+(n-2) \qquad (1.3.4)$$

从（1.3.2）和（1.3.4）可以列出**方差分析**表，形式见表1.3，其中的“**均方**”栏表示，每个平方和/各自的自由度。

表1.3　方差分析（ANOVA）表

方差来源	自由度 (df)	平方和 (SS)	均方 (MS)
回归	1	$\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2$	
残差	$n-2$	$\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$	$S^2=\frac{\text{残差SS}}{n-2}$
总和（校正）	$n-1$	$\sum_{i=1}^{n}(Y_i-\bar{Y})^2$	

表1.4　编入SS（b_0）后的方差分析表

来源	df	SS	MS
b_0	1	$SS(b_0)=n\bar{Y}^2$	
$b_1 \mid b_0$	1	$SS(b_1 \mid b_0)=\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y}_i)^2$	
残差	$n-2$	$\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$	s^2
总和(校正)	$n-1$	$\sum_{i=1}^{n}(Y_i-\bar{Y})^2$	
总和	n	$\sum_{i=1}^{n}Y_i^2$	

方差分析的更一般的表格形式见表1.4，它是在表1.3的基础上加入因子$n\bar{Y}^2$后得到的，记$SS(b_0)=n\bar{Y}^2$（其含义见2.3节）。又将回归平方和记为$SS(R|b_0)$（参见(2.1.14)），对本章讨论的情形，$SS(R|b_0)=SS(b_1|b_0)$。($SS(b_1|b_0)$表示模型含有β_0项时加入项β_1X引起的额外平方和，见2.3节。)

对表1.3和表1.4的实际计算可在袖珍计算器上进行。残差SS难于直接计算，通常是由校正SS减去回归SS得到。因为在拟合直线时已算出了b_1和S_{XY}，故利用

$$\sum (\hat{Y}_i - \overline{Y})^2 = b_1 S_{XY} \qquad (1.3.5)$$

很容易算出回归SS，但为了减少舍入误差，下面带除法的式子是经常使用的：

$$\sum (\hat{Y}_i - \overline{Y})^2 = S_{XY}^2 / S_{XX} \qquad (1.3.6)$$

残差均方s^2提供了一个关于回归的方差（记为$\sigma^2_{Y\cdot X}$）的具有$n-2$个自由度的估计。如果回归方程是由大量的观测数据拟合出来的，则回归的方差就表示了“误差”的一个量度。这里，误差是指对每一个给定的X值，由回归方程预报Y时产生的误差（见1.4节注1）。

现在将这一节的公式用于前面的例子。由（1.3.6），回归SS$=(-571.1280)^2/7154.42$ $=45.5924$，校正SS$=2284.1102-(235.60)^2/25=63.8158$，$\sigma^2_{Y\cdot X}$的估计是$s^2=0.7923$。

表1.5 例题的方差分析表

来　源	df	SS	MS	F值
回　归	1	45.5924	45.5924	57.54
残　差	23	18.2234	$s^2=0.7923$	
总　和（校正）	24	63.8158		

1.4 回归方程的考核

现在对模型$Y_i=\beta_0+\beta_1 X_i+\varepsilon_i$，$i=1, 2, \cdots, n$提出一些基本假定：

1. ε_i是随机变量，$E(\varepsilon_i)=0$，$V(\varepsilon_i)=\sigma^2$（未知），从而$E(Y_i)=\beta_0+\beta_1 X_i$，$V(Y_i)=\sigma^2$；

2. 当$i\neq j$时，$\mathrm{cov}(\varepsilon_i, \varepsilon_j)=0$，即$\varepsilon_i$与$\varepsilon_j$不相关，这时$Y_i$与$Y_j$也不相关；

3. ε_i服从正态分布，即$\varepsilon_i \sim N(0, \sigma^2)$，在这个假定下，$\varepsilon_i$与$\varepsilon_j$独立（$i\neq j$），见图1.6。

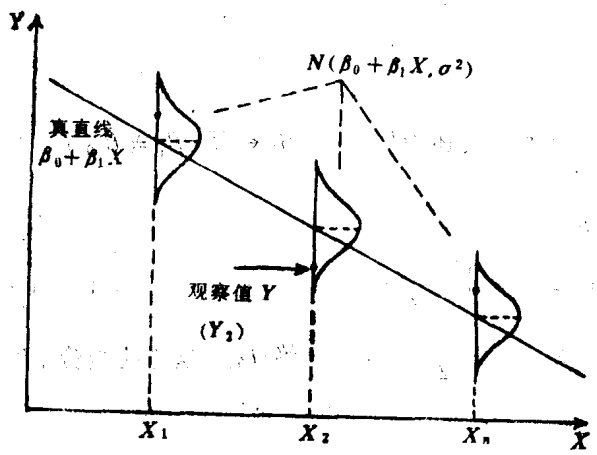


图1.6 假设每个响应观测值都取自正态分布，方差都是σ^2

注1 σ^2与$\sigma^2_{Y\cdot X}$可能相等也可能不相等，$\sigma^2_{Y\cdot X}$就是上节叙述过的关于回归的方差。如果假定的模型为真，则$\sigma^2=\sigma^2_{Y\cdot X}$，这时残差均方$s^2$可作为$\sigma^2$的估计；如果假定的模型不真，则$\sigma^2<\sigma^2_{Y\cdot X}$，这时不能用$s^2$来估计$\sigma^2$。当$\sigma^2<\sigma^2_{Y\cdot X}$时，就认为原来假定的模型是不对的，或者说存在拟合不足，判别方法将在后面讨论。

注2 由中心极限定理，在许多实际问题中，误差的分布是渐近正态分布。如果误差ε是多种来源的误差之和，则不论