线性回归分析

〔美〕G. A. F. 塞伯 著

方开泰　张永光　冯士雍 译

冯士雍 校

科学出版社

1987

内 容 简 介

本书系统地介绍了线性回归分析的理论和方法．内容包括线性回归模型参数的最小二乘估计与假设检验、基本假定的检验、直线与多项式回归、方差分析与协方差分析模型、回归变量的筛选以及相应的计算方法．本书取材新颖，内容丰富，深入浅出，每章给出许多例题和习题，书末附有习题答案，可作为大学教科书或教学参考书．

G. A. F. Seber
LINEAR REGRESSION ANALYSIS
John Wiley & Sons, 1977

线 性 回 归 分 析
〔美〕G. A. F. 塞伯 著
方开泰 张永光 冯士雍 译
冯士雍 校
责任编辑 杨贤英 刘嘉善
科学出版社出版
北京朝阳门内大街137号
中国科学院印刷厂印刷
新华书店北京发行所发行 各地新华书店经售
*
1987年12月第一版 开本：787×1092 1/32
1987年12月第一次印刷 印张：17
印数：0001—5,300 字数：384,000
ISBN 7-03-000051-X/O·13
统一书号：13031·3936
定价：4.00元

译 者 的 话

数理统计在国民经济中有着重要的应用，其中以回归分析(包括方差分析和协方差分析)的应用为最广泛．早在十八世纪，利用最小二乘原理就产生了回归分析．从那时起，回归分析一直成为数据分析的最重要方法．随后发展起来的数据分析的许多其他统计方法，大多与回归分析有着密切的联系，比如：时间序列分析、判别分析、典型相关分析、主成分分析、因子分析等．有趣的是，这样一个古典的方法，一百多年来经久不衰，其理论和方法不断得到发展．例如，关于变量筛选、稳健回归、回归诊断等都是七十年代以后发展迅速的分支．现在，回归分析仍然是数理统计中最活跃的分支之一．

塞伯教授所著的《线性回归分析》一书，兼顾了理论和应用两个方面．它深入浅出地系统地阐述了回归分析的理论和方法，并用大量的例子来说明，既总结了古典的线性模型理论，又概述了近代的发展．因此，这本书在国际上曾得到一致的好评．世界上许多第一流的大学都选用它作为教科书．我们也曾选用本书中的部分内容在国内讲授，收到了很好的效果．

本书第一章至第六章由中国科学院系统科学研究所张永光翻译，第七章至第八章由系统科学研究所冯士雍翻译，第九章至第十二章及附录由应用数学研究所方开泰翻译．全书由冯士雍负责校订．由于我们的水平有限，特别是许多新的专门术语国内尚未统一，缺点和错误在所难免，恳请专家和广大读者批评指正．

方开泰　张永光　冯士雍

一九八六年六月于北京

序

回归分析是统计学家手中的一件常用的工具. 由于它理论完善,计算富有魅力,因而, 无论是从事纯理论研究还是从事应用的统计学家,对此都不陌生. 例如,理论工作者对最小二乘的各种推广及其各种特殊情形,至今仍感兴趣;而与此同时, 实际工作者在继续发展用于检验模型和考察基本假定的各种图解方法. 随着数值分析与统计学的逐渐相互渗透, 统计学家已意识到某些曾一度认为相当可靠的计算方法在实施中有很大的困难. 在今天, 发展有效而又精确的计算机回归程序已被认为是统计研究中的一个重要课题.

然而对回归分析的这种长盛不衰而又广泛 (从纯理论到应用) 的研究兴趣却使教科书的编写者为难. 从已有的回归分析的一些教科书情况来看, 它们在讨论问题时所涉及的数学程度相差甚大. 例如 Seber [1966], Searle [1971] 与 Rao [1973],这三本书是对一般情形加以讨论的;而 Williams [1959] 与 Sprent [1969]两书所讨论的问题就比较复杂.由于读者的程度不同, 所以要求的差异很大, 因此不同层次的书都是需要的. 在过去十年的教学过程中,我愈来愈感到,只介绍结果而不予证明与给出最一般证明这两种写法之间的大学用的教科书是十分必要的. 由于回归分析与许多方差分析都是考虑满秩模型的, 因而为顾及理论的一般性而对不满秩情形的讨论容易被过分强调. 尤其是对广义逆的过多讨论反而会掩盖最小二乘的几何意义. 当然,广义逆是有用的,但对它所起的作用的评价也应恰如其分.

回归分析是描述处理数据方法的一门应用学科．当然，任何一种理论都应有实际工作作支柱．于是产生了这样一个问题：一本教科书是否应在理论和计算两方面都作详尽的讨论？显然，要做到这一点是不容易的，再说现在已是广泛使用统计程序包的时代，因此，将上述两个方面分开处理也许更为适宜．譬如说，一种理论上十分漂亮的方法不一定就有令人满意的计算方法；而一些繁复的算法却可能是有效而精确的．由于程序包随计算机设施的差异有不同的形式，因此较好的解决办法是既有包括计算概要的理论方面的教科书，也有给出数值例子和详细计算程序的实用手册．例如 Daniel 与 Woole［1971］就是属于后一类型的一本早期的著作．

带着上述想法，我力图使本书成为一本主要适合关心数学推理读者的理论性著作，但又尽可能照顾到读者的广泛性．我也尽力对目前流行的一些计算方法给出最新解释，但又不陷入具体的计算细节．鉴于回归分析的研究文献仍在迅速增加，在本书中我引用了较为著名的统计杂志，希望本书也能作为一本通用的参考书使用．为阅读本书，读者必须有较好的矩阵代数基础，并对直线回归及简单的方差分析模型有初步的了解．

本书前四章以标准的正规形式讨论了多元线性回归模型的最小二乘拟合及假设检验．其中第一章引进了随机向量的期望和协方差算子．第二章讨论多元正态分布与关于二次型的某些结果，第三章讨论最小二乘估计问题，包括广义最小二乘估计以及不满秩情形和带约束的估计．第四章则详细地讨论了对线性假设的 F 检验．第五章讨论了置信区间及应用于回归模型的联合推断问题，也讨论了预报及反预报（即判别）区间问题．第六章研究了最小二乘理论的基本假定，提出了检验这些假定的各种方法．鉴于直线拟合和多项式拟合的重要

性，在第七、八两章对此分别作了研究．第九章揭示了回归模型与方差分析模型之间的密切关系，给出了进行方差分析的步骤，主要侧重于平衡(正交)设计．第十章用回归分析观点来处理协方差分析，也详细讨论了与协方差分析有密切联系的丢失观测值的问题．最后两章是关于回归分析的计算方面的，其中第十一章为最小二乘拟合的计算机算法，第十二章则考虑了从一组可能的回归因子(自变量)中选择"最好的"回归子集的问题．

附录A和附录B包含了一些证明较难的矩阵代数的结果；附录C介绍了概率纸点图法；附录D，附录E及附录F给出了几个用于联合推断的统计表．最后是练习答案及提示．

汇编一套互相联系的理论方面的练习题，并不是一件易事，但也不至于太困难．我希望本书中的200个左右的习题，不仅对学生有所帮助，而且也给教师提供一些思路．

本书是我过去十年中在新西兰奥克兰大学教这门课程的基础上形成的．我谨对曾激励我对这门学科教学产生浓厚兴趣的学生表示谢意，我还要特别感谢 Heather Lucas 阅读了我的手稿．最后，对于 Peggy Haworth 出色的打字表示感谢．

致谢——承蒙允许，引用他们已出版的数表，谨对以下杂志中有关作者和编者表示谢意：Biometrika (附录E)，Journal of the American Statistical Association (表 5.1，5.2 和附录 D)，Journal of the Royal Statistical Society，Series B (附录 F) 和 Technometrics (表 5.3)．

G. A. F. 塞伯

1976 年 7 月于新西兰奥克兰

目　　录

第一章 随机向量

1.1 记号

本书用黑正体字母$\mathbf{A}$和$\mathbf{a}$分别表示矩阵和向量，用斜体字母表示标量．用大写字母表示随机变量，而用小写字母表示它们的值(例如相应地用Y和y表示)．用大写字母表示随机变量的方法看来已被广泛接受，在回归分析中为区分确定的和随机的回归因子(自变量)时这特别有用． 但是,在处理随机向量时会产生一些问题，例如$\mathbf{Y}$可能会被误认为是一个矩阵．在第十一章中,因为字母不够用,也偶然地使用黑体小写字母表示一个随机向量.

如果X和Y都是随机变量，则记号 $E[Y]$, $\mathrm{var}[Y]$, $\mathrm{cov}[X, Y]$ 及 $E[X|Y=y]$ (或更简单地 $E[X|Y]$)分别表示期望,方差,协方差及条件期望.

用 $\mathrm{diag}(d_1, d_2, \cdots, d_n)$ 表示对角线元素为 $d_1, d_2, \cdots, d_n$ 而其他元素均为 0 的一个 $n \times n$ 矩阵,当所有的 d_i 全是 1 时,就成为单位矩阵 $\mathbf{I}_n$.

如果$\mathbf{a}$是$n \times 1$的列向量，其分量（也称元素）为 $a_1, a_2, \cdots, a_n$，则记 $\mathbf{a}=[(a_i)]$，且用 $\|\mathbf{a}\|$ 表示$\mathbf{a}$的“长度”或“模数”．于是

$$\|\mathbf{a}\| = \sqrt{\mathbf{a}'\mathbf{a}} = (a_1^2 + a_2^2 + \cdots + a_n^2)^{1/2}$$

其分量都是 1 的向量表示为$\mathbf{1}_n$.

如果$m \times n$矩阵$\mathbf{A}$的元素为a_{ij}，则记$\mathbf{A}=[(a_{ij})]$，其对角线元素之和称为$\mathbf{A}$的迹，记为 $\mathrm{tr}\mathbf{A}(=a_{11}+a_{22}+\cdots+a_{kk}$，这里$k$表示$m$与$n$中较小的一个)．$\mathbf{A}$的转置记为$\mathbf{A}'=$

$[(a'_{ij})]$，这里 $a'_{ij} = a_{ji}$. 如果$\mathbf{A}$是方阵，则用 $|\mathbf{A}|$ 表示它的行列式；又若$\mathbf{A}$为非奇异的，用 $\mathbf{A}^{-1}$ 表示它的逆矩阵. 由$\mathbf{A}$的列向量张成的空间称之为 $\mathbf{A}$ 的值空间，用 $\mathscr{R}[\mathbf{A}]$ 表示. 又用 $\mathscr{N}[\mathbf{A}]$ 表示 $\mathbf{A}$ 的零空间或 $\mathbf{A}$ 的核 $(=\{\mathbf{x}: \mathbf{Ax} = \mathbf{0}\})$.

如果 Y 服从正态分布，其均值为 θ，方差为 σ^2，则记为 $Y \sim N(\theta, \sigma^2)$，当 $\theta = 0$，$\sigma^2 = 1$ 时，则称 Y 为标准正态分布. 自由度为 k 的 t 分布和 χ^2 分布分别表示为 t_k 和 χ^2_k，自由度为 m 及 n 的 F 分布表示为 $F_{m,n}$.

最后，我们用记号"·"和"-"分别表示"和"与"平均数"，例如，

$$a_{i\cdot} = \sum_{j=1}^{J} a_{ij} \qquad 及 \qquad \bar{a}_{i\cdot} = a_{i\cdot}/J$$

对于单重脚标的情形，通常略去点号，而直接记为 $\bar{a}$.

假定读者已经掌握了线性代数的有关知识，若需要简单回顾，有几本书可供使用(例如 Scheffe [1959:附录], Graybill [1961,1969], Rao [1973:第一章]). 另外，本书末的附录A和附录 B 中也包括了有关矩阵的若干结果. 当需要让读者参阅附录中有关段落时，即注出如 A2.3 等.

1.2 线性回归模型

统计学中一个常见问题是估计两个随机变量 X 与 Y 之间所存在的关系(如果有的话)；例如，人的高度与重量，收入与智商 (IQ)，一对夫妇在结婚时的年龄，树叶的长度与宽度，一定体积气体的温度与压力，金属棒的长度与它本身的温度等. 如果有 n 对观测值 (x_i, y_i) $(i = 1, 2, \cdots, n)$，可以把它们点在纸上，画出所谓散布图，且在这些点之间尽量拟合一条光滑的曲线，使这些点尽可能地"趋近"这条曲线. 显然，我们并不期望这些点能和曲线完全拟合，因为上面所例举的

每对变量由于许多不可控因素都会发生随机波动．即使像一定体积气体的温度和压力这二者间存在着确定关系，由于测量误差，这种偶然性的波动也仍然会在散布图上表现出来．

通常所要拟合曲线的类型，可象下述例子那样按经验的依据或推理的论证而加以选定．

例 1.1 欧姆定律可以表示为 $Y = rX$，其中 X 是通过 r 欧姆电阻的电流安培数，Y 是加在该电阻上的电压伏特数*．上述 Y 与 X 的关系是通过原点的一条直线，因而呈直线型的散布图将能印证这个定律，从这条拟合的直线的斜率还可估计出电阻 r 的值．

例 1.2 由力学定律可知，为了不让一个 w 克重的物体从一个倾角为 θ 的光滑斜面上滑下去，必须对它施加 $Y = w \cdot \sin\theta$ 克重的力．若令 $X = \sin\theta$，我们将又一次得到一条通过原点的直线．在此情形由于 Y 和 θ 的测量误差，以及在该物体和斜面之间还有摩擦力的存在，使得观测值 (x_i, y_i) 会稍微偏离线性关系．

例 1.3 理论化学指出，对于保持恒温的一定量的气体，其体积 V 和压力 P 近似满足关系 $PV = c$．若记 $Y = P$ 及 $X = 1/V$，就能得到 $Y = cX$．

例 1.4 更精密的实验表明，上例的压力和体积的关系应是 $PV^\gamma = c$ 的形式，其中 $\gamma \neq 1$．若取对数我们仍然能得到一个线性关系：即

$$\log P = \log c - \gamma \log V$$

或

$$Y = a + bX$$

因此，$\log c$ 和 $-\gamma$ 可由拟合这些实验数据的直线上估计出

---

* 原文将 X 与 Y 所表示的量颠倒了，这里已改正——译者注．

来.

例 1.5 逆平方定律表明相距为 D 的两个物体之间的引力 F 由下式给定，

$$F = \frac{c}{D^{\beta}}$$

其中 $\beta = 2$. 两边取对数就得到

$$\log F = \log c - \beta \log D$$

从试验数据我们可以估计出 β，并验证是否 $\beta = 2$.

例 1.6 实验表明金属棒受热时会伸长，伸长量与温升成正比. 另外，把两个等长的金属棒对接起来，伸长总量是单根金属棒的两倍，所以伸长量又正比于金属棒原来的长度. 这告诉我们应考虑直线模型 $Y_T = Y_0(1 + \alpha T)$，其中 Y_T 是在温度 T 时的长度（T 是从一个适当的起点进行度量的），α 是所谓的线性膨胀系数. 由更精确的研究表明，可以采用一个二次模型，即

$$Y_T = Y_0(1 + \alpha T + \beta T^2)$$

当没有有关的理论和试验结果帮助我们时，要决定应该拟合哪一类型的曲线有时是困难的，例如图 1.1 所配的直线似乎不会差于任何其它的曲线，因为它所包含的参数很少，但为了比较不同曲线拟合的好坏，必需要有某种尺度. 有时，散布图上的点看不出存在有规律的可供拟合的趋势. 例如，图 1.2 的散布图实际上等于没有提供关于 X 和 Y 之间的关系.

在许多情形下，变量之一（比如说 X）不是随机的而是确定的或可控制的. 例如，X 可以看作是生产年份，Y 是某家公司在一年中生产的产品数量. 作为 X 是可控变量的一个例子，我们考察这样一个试验，其中 X 是某种化肥的施加量，而 Y 是在每单位面积上施加固定肥量的产量. 在这两个例中，对每个 $X = x$ 的值我们认为 Y 是随机变量且均值为 $\varphi(x)$，

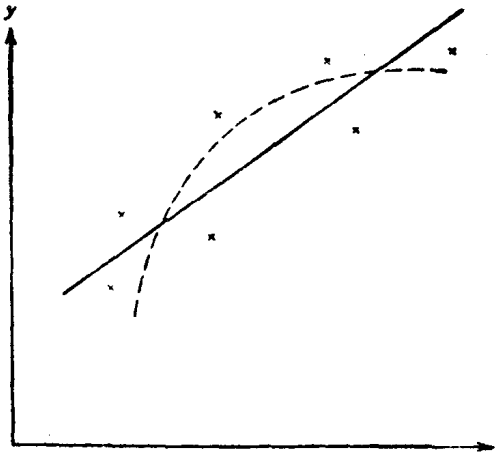

图 1.1　拟合同一组数据的两条不同曲线

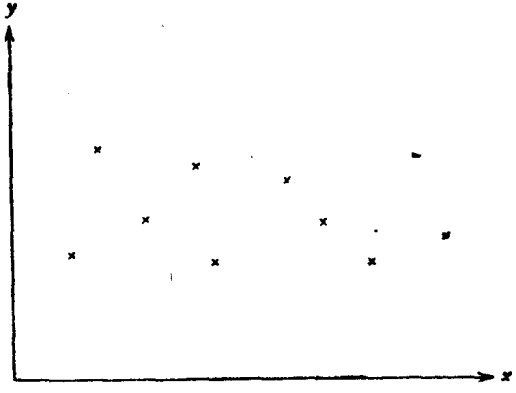

图 1.2　X和Y不相关的一种散布图

即 $Y=\varphi(x)+\varepsilon$，其中 $E[\varepsilon]=0$. 这里 $\varphi(x)$ 称为Y关于X的回归曲线或回归函数.

为了说明怎样从数据对 (x_i, y_i) 估计出 φ，我们考虑 φ为简单直线情形,即 $\varphi(x)=\beta_0+\beta_1 x$. 我们的模型是

$$Y_i=\beta_0+\beta_1 x_i+\varepsilon_i \qquad (i=1,2,\cdots,n)$$

有一种估计 β_0 和 β_1 的精巧方法就是所谓的最小二乘法. 据

此所得的估计量其所以具有某种最优性，是由于最小二乘法出色的构想：它选择 β_0 和 β_1 时，使数据点与被拟合的曲线（见图 1.3）的纵向偏差的平方达到极小；即对于 β_0 和 β_1 极小化 $\sum_i \varepsilon_i^2 = \sum_i (Y_i - \beta_0 - \beta_1 x_i)^2$. 显然，如若 φ 关于未知参数不是线性的，极小化可能是困难的，但上述最小二乘原理可以应用到任意的回归曲线 $\varphi(x)$. 例如，$\varphi(x) = \beta_0 e^{x\beta_1}$ 关于 β_1 是非线性的，$\varphi(x) = \beta_0 + \beta_1 x + \beta_2 x^2$ 关于 β_j 是线性的.

由上述例子，我们已看到 X 和 Y 二者都可以是随机的，从而 X 和 Y 有联合分布，此时可以定义两个回归函数 $E[Y|X=x]$ 及 $E[X|Y=y]$. 例如假定存在线性关系，我们得到

$$E[Y|X=x] = \beta_0 + \beta_1 x$$

我们可以象对确定性的 X 那样进行处理. 当然，任何推断在此时都只是在 X 所取观测值的条件下进行的.

例 1.7 设对由 N 个动物组成的总体进行 n 次连续捕捉. 假定每个动物在每次特定的捕捉中被捕到的概率为常数 p.

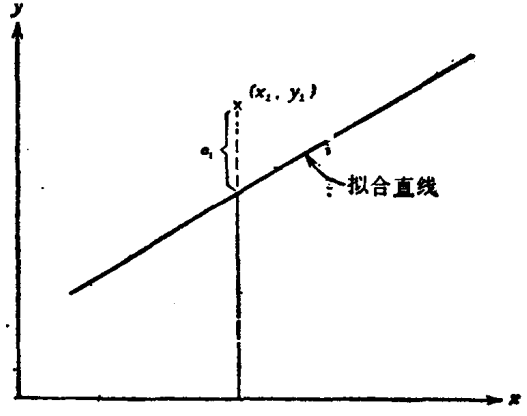

图 1.3 由极小化 $\sum \varepsilon_i^2$ 表示的最小二乘法