

应用统计方法

古华民 主编
张斌东 胡 男 编
刘郁强 王毅刚

暨南大学出版社

应用统计方法

古华民

主编

张斌东

胡男

刘郁强

王毅刚

编

暨南大学出版社

粤新登字 13 号

应用统计方法

古华民 主编

张斌东

胡 男 审校

暨南大学出版社出版

广东省新华书店经销

华南师范大学印刷厂印刷

开本: 850×1168 1/32

印张: 10.75

字数: 27 万

1994 年 10 月第 1 版

1994 年 10 月第 1 次印刷

印数: 1—3000 册

书号: ISBN7-81029-296-X/0·15

定价: 9.50 元

前 言

数理统计是一门应用相当广泛的学科。从产品的设计到生产的管理乃至国民经济的预测、控制及决策，都离不开应用数理统计的理论和方法。

本书内容以应用统计方法为主，包括了应用于系统分析与决策、质量控制与管理、预测预报等方面的应用统计方法。在写法上采用通过大量实例引入并介绍有关方法，侧重阐明有关的统计思想，力求做到方法程序化，适合于从事生产、管理、设计和科学研究等方面的实际工作者的需要。本书可作为各个领域实际工作者的参考读物，也可以作为大专院校有关专业“应用统计方法”课程的参考教材。全书共分五章，各章内容基本上自成单元，便于读者根据实际需要选读。

本书由古华民教授主编。第一章由胡男、古华民执笔，二至五章依次分别由古华民、张斌东、刘郁强和王毅刚执笔。本书作为华南师范大学数学系高年级数理统计选修课的教材实际上已使用了4年。每一次使用时授课老师都作了修改，增删了部分内容并配备了习题，使本书不断得以完善。特别是张斌东在综合整理修改，润色全书及促成本书出版方面做了大量工作。

本书得以出版，除了概率统计教研室全体教师的共同努力外，还得到许多朋友的关心与支持，在此谨表衷心的感谢。由于编者水平有限，书中缺点错误肯定不少，恳请读者批评指正。

编 者

1994年1月

目 录

第一章	回归方法	(1)
§ 1.1	多元线性回归	(2)
§ 1.2	多项式回归与正交多项式回归.....	(19)
§ 1.3	逐步回归.....	(30)
§ 1.4	加权回归.....	(48)
§ 1.5	主成分回归.....	(53)
§ 1.6	路径分析.....	(61)
第二章	试验设计法	(83)
§ 2.1	一般回归的正交设计.....	(83)
§ 2.2	二次回归旋转设计	(111)
§ 2.3	三次设计简介	(120)
第三章	时间序列分析	(165)
§ 3.1	长期趋势和季节变动	(166)
§ 3.2	时间序列模型及预测方法	(177)
§ 3.3	平稳时间序列预测	(206)

第四章	马尔柯夫链方法	(243)
§ 4.1	离散马尔柯夫链	(243)
§ 4.2	状态概率分布	(247)
§ 4.3	马尔柯夫链预测方法	(251)
§ 4.4	马尔柯夫链决策方法	(266)
§ 4.5	马尔柯夫链分析	(274)
第五章	抽样调查方法	(282)
§ 5.1	基本概念	(282)
§ 5.2	简单随机抽样	(284)
§ 5.3	分层抽样	(296)
§ 5.4	整群抽样	(305)
§ 5.5	等距抽样	(310)
§ 5.6	多阶段抽样	(315)
附表 1	正态分布表	(324)
附表 2	t 分布的双侧分位数 (t_{α}) 表	(326)
附表 3	F 分布表	(327)
附表 4	正交多项式表	(330)
附表 5	随机数表	(337)

第一章 回归方法

常用的数理统计方法中应用最为广泛的是回归方法。在自然科学、工程技术以至社会科学中许多问题的研究，往往归结为弄清楚一些有关变量的联系，这种联系具体反映为寻求其中一个目标变量 y （为方便计以下称之为因变量）通过其它的变量 $x_1, x_2, \dots, x_m$ （以下称之为自变量）表达出来。在这方面可分为两种基本类型：第一类的特征是：只要知道了自变量 $x_1, \dots, x_m$ 所取的值，因变量 y 所取之值就唯一地确定了。这种关系叫做确定性关系，即数学上的所谓函数关系，如物理学中的欧姆定律， $V=IR$ ，它确切地描述了电压、电流、电阻三者之间的确定关系。另一类关系叫做非确定性关系，有时也称为相关关系，它的特征是因变量 y 所取的值与自变量 $x_1, \dots, x_m$ 所取的值有关系，但这种关系没有密切到可以唯一决定的程度。例如人的年龄 x 与血压 y ，一般来说，随着年龄的增大，血压也升高，表明两者之间存在着某种关系，但显然不是函数关系。又如在一项农业研究中，可能要考虑每亩播种量 x_1 、每亩施肥量 x_2 对每亩产量 y 的影响，当然， y 的值与 x_1, x_2 ，有很大关系，但是， x_1, x_2 并不能唯一地决定 y 。回归方法是研究变量间相关关系的一种手段，它是在对客观事物进行大量试验和观察的基础上，用来寻找隐藏在那些看上去是不确定的现象中的统计规律性的数理统计方法，它在预测预报中有着广泛的应用。

本章介绍几个应用最为广泛或较新的回归方法，要求读者具备一元线性回归的知识。一元线性回归的内容在数理统计教科书中均有叙述。本章第三节使用了逆矩阵，第五节使用了矩阵的特征根、特征向量，这些内容在线性代数或高等代数教科书中均有叙述。

§ 1.1 多元线性回归

当我们所讨论的回归问题有一个因变量 y 和一个自变量 x 时，是一元回归问题；在很多实际问题中，影响因变量 y 的自变量不是一个而是多个，我们称这类回归问题为多元回归。多元回归中最简单的是多元线性回归。许多非线性回归和多项式回归都可化为多元线性回归问题来研究，因而多元线性回归有着广泛的应用，多元线性回归与一元线性回归在原理上是相同的，但在计算上前者比后者复杂得多，当自变量个数较多或观测数据数量较大时，往往要使用电子计算机来进行计算。

1.1.1 多元线性回归方程的一般求法

我们先从两个具体的例子谈起：

例 1.1.1 一个国家服务业发达的程度可用其从业人员的多少来衡量。我们在研究中发现服务业从业人员的数量 (y) 与人均国民收入 (x_1)，服务业投资额 (x_2) 之间有直接的关系。从《中国经济年鉴》中整理出我国有关的资料数据如表 1.1.1，试建立 y 对 x_1, x_2 的线性回归方程。

表 1.1.1

年份	人均国民收入 x_1 (元)	投资额 x_2 (亿元)	从业人员数 y (万人)
1957	142	38.79	77.0
1965	130	28.95	74.0
1978	315	104.75	56.0
1979	340	158.34	88.4
1980	376	199.61	112.4
1981	396	190.48	148.3
1982	423	252.63	182.4
1983	458	274.69	227.4
1984	548	299.75	297.5
1985	656	463.03	353.9

例 1.1.2 影响我国粮食总产量 (y) 的因素有粮食单位面积产量 (x_1)、农业机械总动力 (x_2)、化肥施用量 (x_3) 等。现收集到 1978 至 1985 共 8 年的统计数据如表 1.1.2 所示。试建立 y 对 x_1 、 x_2 、 x_3 的线性回归方程。

表 1.1.2

年份	粮食单位面积 产量(公斤/亩) x_1	农机总动力 (万马力) x_2	化肥施用量 (万吨) x_3	年粮食总产量 (万吨) y
1978	169	15975	884	30477
1979	189	18191	1086.3	33212
1980	183	20049	1269.4	32056
1981	189	21319	1334.9	32502
1982	209	22589	1513.4	35450
1983	227	24503	1659.8	38728
1984	241	26509	1739.8	40731
1985	232	28433	1775.8	37911

一般地，设有 m 个自变量 $x_1, \dots, x_m$ 。经试验或观测得到 n 组数据，数据资料如表 1.1.3 所示。

表 1.1.3

序 号	x_1	x_2		x_m	y
1	x_{11}	x_{12}		x_{1m}	y_1
2	x_{21}	x_{22}		x_{2m}	y_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
n	x_{n1}	x_{n2}		x_{nm}	y_n

表中 x_{ij} ($i=1, 2, \dots, n; j=1, 2, \dots, m$) 表示第 j 个自变量 x_j ，在第 i 次试验中的观测值。 y_i ($i=1, 2, \dots, n$) 表示因变量 y 在第 i 次试验中的观测值。我们把试验结果 y 看成是由两部分叠加而成的。一部分是由 $x_1, x_2, \dots, x_m$ 的线性函数关系引起的，记为 $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m$ 。另一部分是由不能控制的随机因素引起的，记为 ε 。假设 $\varepsilon \sim N(0, \sigma^2)$ 。

多元线性回归模型为

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2) \quad (1.1.1)$$

其中 $x_1, x_2, \dots, x_m$ 是一般变量，它们可以精确测量或可以加以控制， y 是可观测量值的随机变量， $\beta_0, \beta_1, \dots, \beta_m$ 是 $m+1$ 个待估参数， ε 是不可观测的随机变量，假定它服从 $N(0, \sigma^2)$ 分布。

为了估计未知参数 $\beta_0, \beta_1, \beta_2, \dots, \beta_m$ 和 ε 的方差 σ^2 ，我们进行 n 次观测，得 n 组观测数据 $(x_{i1}, x_{i2}, \dots, x_{im}, y_i), i=1, 2, \dots, n$ ($n > m$)，它们应有回归关系 (1.1.1)，可写成如下形式

[illegible]

其中 $\varepsilon_i \sim N(0, \sigma^2)$ $i=1, 2, \dots, n$

估计 $\beta_0, \beta_1, \beta_2, \dots, \beta_m$ 最常用的方法是最小二乘法, 设 $b_0, b_1, b_2, \dots, b_m$ 分别是 $\beta_0, \beta_1, \beta_2, \dots, \beta_m$ 的最小二乘估计, 从而得到方程

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_m x_m \quad (1.1.3)$$

称此方程为 m 元线性回归方程, $b_i (i=0, 1, \dots, m)$ 称为回归系数, 确定 $b_0, b_1, b_2, \dots, b_m$ 的原则是使得全部观测值 y_i 与相应的回归值 $\hat{y}_i$ 的偏差平方和 Q 达到最小, 即使

$$Q = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_{i1} - \dots - b_m x_{im})^2 \quad (1.1.4)$$

达最小值，用微积分中令偏导数等于零求函数极值的方法，便可求出 $b_0, b_1, \dots, b_n$ 之值。

具体步骤如下:

1. 求 l_{ij} 、 l_{iv} 之值

其中 $l_{ij} = \sum_{k=1}^n x_{ik}x_{jk} - n\bar{x}_i\bar{x}_j \quad i, j=1, \dots, m \quad (1.1.5)$

$$l_{ij} = \sum_{n=1}^n x_n y_n - n \bar{x}_i \bar{y}_j \quad i=1, \dots, m \quad (1.1.6)$$

$$\bar{x}_i = \frac{1}{n} \sum_{j=1}^n x_{ij} \quad i=1, \dots, m \quad (1.1.7)$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad (1.1.8)$$

$$l_{11} = \sum_{i=1}^{10} x_{i1}^2 - 10 \cdot \bar{x}_1^2 = 1669414 - 10 \times 378.4^2$$

$$= 237548.4$$

$$l_{22} = \sum_{i=1}^{10} x_{i2}^2 - 10 \cdot \bar{x}_2^2 = 558037.0236 - 10 \times 201.102^2$$

$$= 153616.8796$$

$$l_{12} = l_{21} = \sum_{i=1}^{10} x_{i1}x_{i2} - 10 \cdot \bar{x}_1\bar{x}_2$$

$$= 947268.16 - 10 \times 378.4 \times 201.102$$

$$= 186298.192$$

$$l_{1y} = \sum_{i=1}^{10} x_{i1}y_i - 10\bar{x}_1\bar{y}$$

$$= 745732 - 10 \times 378.4 \times 161.73$$

$$= 133745.68$$

$$l_{2y} = \sum_{i=1}^{10} x_{i2}y_i - 10\bar{x}_2\bar{y}$$

$$= 437262.894 - 10 \times 201.103 \times 161.73$$

$$= 112020.6294$$

正规方程是

$$\begin{cases} 237548.4b_1 + 186298.192b_2 = 133745.68 \\ 186298.192b_1 + 153616.8796b_2 = 112020.6294 \end{cases}$$

解正规方程，得

$$b_1 = -0.181372 \quad b_2 = 0.94918$$

代入 (1.1.10) 得

$$b_0 = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2$$

$$= 161.73 + 0.181372 \times 378.4 - 0.94918 \times 201.102$$

$$= 39.4792$$

$$\hat{y} = 39.4792 - 0.181372x_1 + 0.94918x_2$$

是所求的二元线性回归方程，利用此方程可由 x_1, x_2 的取值来预测 y 的值。

对于例 1.1.2，我们先计算表 1.1.5，其中 $m=3, n=8$ 。

表 1.1.5

年份	x_1	x_2	x_3	y	x_1^2	x_2^2	x_3^2	y^2
78	169	15975	884	30477	28561	255200625	781456.00	928847529
79	189	18191	1086.3	33212	35721	330912481	1180047.69	1103036944
80	183	20049	1267.4	32056	33489	401962401	1611376.36	1027587136
81	189	21319	1334.9	32502	35721	454499761	1781958.01	1056380004
82	209	22589	1513.4	35450	43681	510262921	2290379.56	1256702500
83	227	24503	1659.8	38728	51529	600397009	2754936.04	1499857984
84	241	26509	1739.8	40731	58081	702727081	3026904.04	1659014361
85	232	28433	1775.8	37911	53824	808435489	3153465.64	1437243921
Σ	1639	177568	11263.4	281067	340607	4064397768	16580523.34	9968670379

年份	x_1x_2	x_1x_3	x_2x_3	x_1y	x_2y	x_3y
78	2699775	149396.0	14121900.0	5150613	486870075	26941668.0
79	3438099	205310.7	19760883.3	6277068	604159492	36078195.6
80	3668967	232300.2	25450200.6	5866248	642690744	40691886.4
81	4029291	252296.1	28458733.1	6142878	692910138	43386919.8
82	4721101	316300.6	34186192.6	7409050	800780050	53650030.0
83	5562181	376774.6	40670079.4	8791256	948952184	64280734.4
84	6388669	419291.8	46120358.2	9816171	1079738079	70863793.8
85	6596456	411985.6	50491321.4	8795352	1077923463	67322353.8
Σ	37104539	2363655.6	259259668.6	58248636	6334024225	403215581.8

由表 1.1.5 算得

$$l_{11} = \sum_{i=1}^8 x_{i1}^2 - \frac{1}{8} \left(\sum_{i=1}^8 x_{i1} \right)^2 = 340607 - (1639)^2/8$$

$$= 4816.875$$

$$l_{22} = \sum_{i=1}^8 x_{i2}^2 - \frac{1}{8} \left(\sum_{i=1}^8 x_{i2} \right)^2 = 4064397768 - (177568)^2/8$$

$$= 123098441$$

$$l_{33} = \sum_{i=1}^8 x_{i3}^2 - \frac{1}{8} \left(\sum_{i=1}^8 x_{i3} \right)^2 = 16580523.34 - (11263.4)^2/8$$

$$= 722500.91$$

$$l_{12} = l_{21} = \sum_{i=1}^8 x_{i1}x_{i2} - \frac{1}{8} \left(\sum_{i=1}^8 x_{i1} \right) \left(\sum_{i=1}^8 x_{i2} \right)$$

$$= 37104539 - \frac{1639 \times 177568}{8} = 725295$$

$$\begin{aligned} l_{13} = l_{31} &= \sum_{t=1}^8 x_{t1} x_{t3} - \frac{1}{8} \left(\sum_{t=1}^8 x_{t1} \right) \left(\sum_{t=1}^8 x_{t3} \right) \\ &= 2363655.6 - \frac{1639 \times 11263.4}{8} = 56066.525 \end{aligned}$$

$$\begin{aligned} l_{23} = l_{32} &= \sum_{t=1}^8 x_{t2} x_{t3} - \frac{1}{8} \left(\sum_{t=1}^8 x_{t2} \right) \left(\sum_{t=1}^8 x_{t3} \right) \\ &= 259259668.6 - \frac{177568 \times 11263.4}{8} = 9257242.3 \end{aligned}$$

$$\begin{aligned} l_{1y} &= \sum_{t=1}^8 x_{t1} y_t - \frac{1}{8} \left(\sum_{t=1}^8 x_{t1} \right) \left(\sum_{t=1}^8 y_t \right) \\ &= 58248636 - \frac{1639 \times 281067}{8} = 665034.38 \end{aligned}$$

$$\begin{aligned} l_{2y} &= \sum_{t=1}^8 x_{t2} y_t - \frac{1}{8} \left(\sum_{t=1}^8 x_{t2} \right) \left(\sum_{t=1}^8 y_t \right) \\ &= 6334024225 - \frac{177568 \times 281067}{8} = 95461094 \end{aligned}$$

$$\begin{aligned} l_{3y} &= \sum_{t=1}^8 x_{t3} y_t - \frac{1}{8} \left(\sum_{t=1}^8 x_{t3} \right) \left(\sum_{t=1}^8 y_t \right) \\ &= 403215581.8 - \frac{11263.4 \times 281067}{8} = 7494326 \end{aligned}$$

正规方程是

$$\begin{cases} 4816.875b_1 + 725295b_2 + 56066.525b_3 = 665034.38 \\ 725295b_1 + 123098441b_2 + 9257242.3b_3 = 95461094 \\ 56066.525b_1 + 9257242.3b_2 + 7225003.91b_3 = 7494326 \end{cases}$$

我们用矩阵法解正规方程，顺便把系数矩阵 L 的逆矩阵 C 求出，供下面 1.1.3 之用

$$\text{其中 } L = \begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1m} \\ l_{21} & l_{22} & \cdots & l_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ l_{m1} & l_{m2} & \cdots & l_{mm} \end{bmatrix} \quad L^{-1} = C$$

求解方法是作一个正规方程的增广矩阵 $L \begin{bmatrix} L_{1y} \\ \vdots \\ L_{my} \end{bmatrix}$ 对此增广矩阵的行施行初等变换, 把 L 变为 I_m 。这相当于用 L 阵的逆阵 C 左乘增广矩阵, 则变换后增广矩阵的第 $m+1$ 列就是正规方程的解, I_m 变换为 C 。计算按表 1.1.6 进行。

由表 1.1.6 的第 4 列得

$$b_1 = 185.200908; \quad b_2 = -0.411132671;$$

$$b_3 = 1.268800341$$

并得 L 阵的逆阵 C 为

$$C = \begin{bmatrix} 0.0021959515 & -0.000003389848 & -0.00012697398 \\ -0.000003389848 & 0.000000228075515 & -0.0000026592284 \\ -0.00012697398 & -0.0000026592284 & 0.000045309439 \end{bmatrix}$$

将 b_1 、 b_2 、 b_3 代入 (1.1.10) 得

$$\begin{aligned} b_0 &= \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2 - b_3 \bar{x}_3 \\ &= 35133.375 - 185.200908 \times 204.875 + 0.411132671 \times \\ &\quad 22196 - 1.268800341 \times 1407.925 \\ &= 4529.464025 \end{aligned}$$

所求的三元线性回归方程是

$$\hat{y} = 4529.46 + 185.201x_1 + 0.411133x_2 + 1.2688x_3$$

利用所求的线性回归方程, 就可由 x_1 、 x_2 、 x_3 的取值来预测 y 的取值。

表 1.1.6 (用矩阵法解正规方程, 并求出 L 阵的逆阵 C)

L			l_y			l_z		
4816.875	725295	58066.525	6650347.38	1	0	0	0	(1) 4816.875
725295	123088441	9257242.3	95461094	0	1	0	0	
58066.525	9257242.3	722599.51	7494326	0	0	0	1	
1	150.573764	11.6396056	128.66345	0.000207603477	0	0	0	-725295X(1)+(2)
725295	123088441	9257242.3	95461094	0	1	0	0	-58066.525X(1)+(3)
58066.525	9257242.3	722599.51	7494326	0	0	0	1	(2) 13888042.9
1	150.573764	11.6396056	138.06345	0.000207603477	0	0	0	-150.573764X(2)+(1)
0	13888042.9	815094.593	-4675635.9	-150.573764	1	1	0	-815094.593X(2)+(3)
0	815094.593	69908.6717	-246411.871	-11.6396056	0	0	1	(3) 22070.45623
1	0	2.80237373	188.7565606	0.0018401237	-0.00001084197	0	0	-2.80237373X(3)+(1)
0	1	0.058690385	-0.336666291	-0.000010841971	0.00000072004385	0	0	-0.058690385X(3)+(2)
0	0	22070.45623	28003.0024	-2.80237378	-0.058690385	1	1	
1	0	0	185.200908	0.0021959515	-0.000003389848	-0.00012697398	0	
0	1	0	-0.411132671	-0.00000228075515	0.000000228075515	0.0000026592284	0	
0	0	1	1.268800341	-0.00012697398	-0.0000026592284	0.000045309439	0	
			b_i			C		
			l_z					