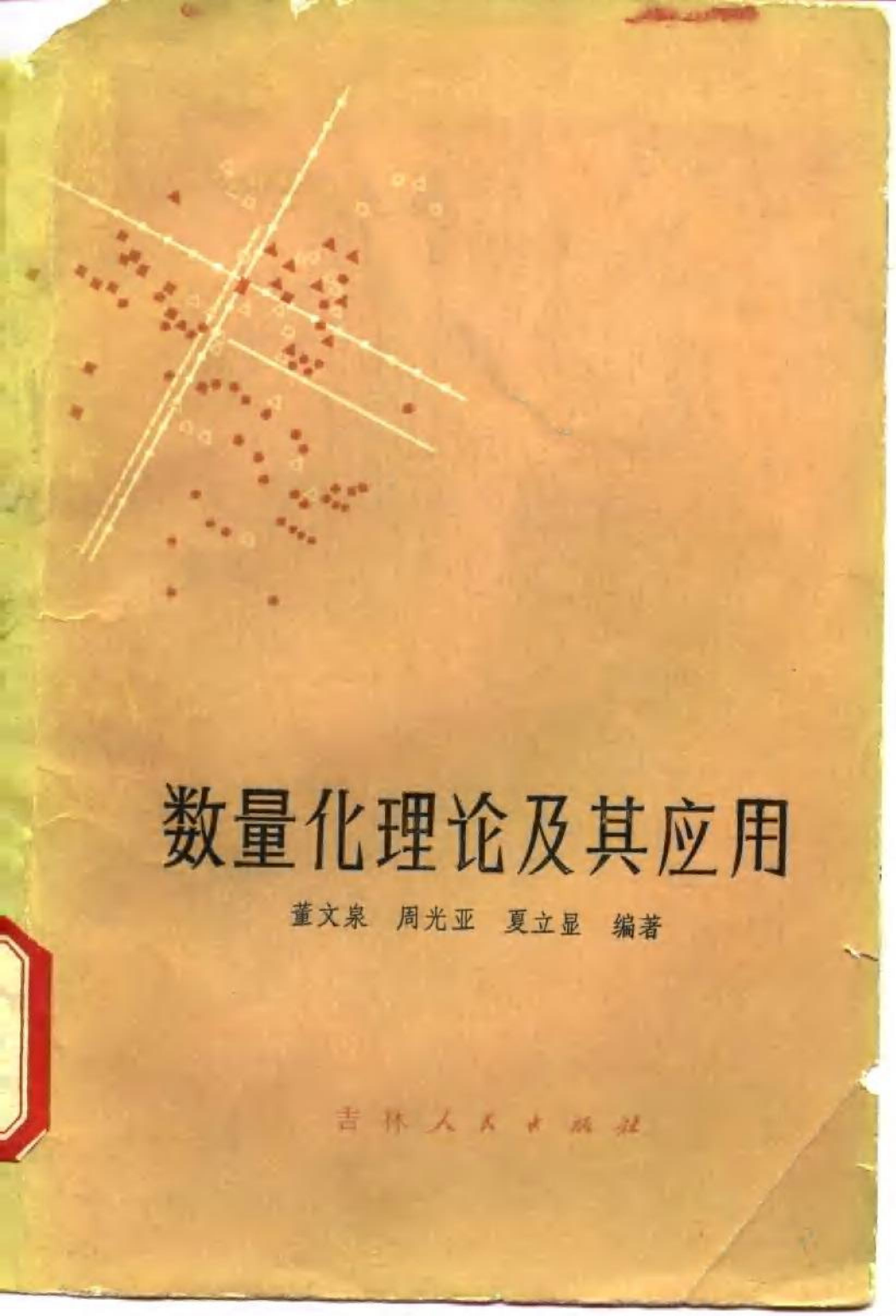


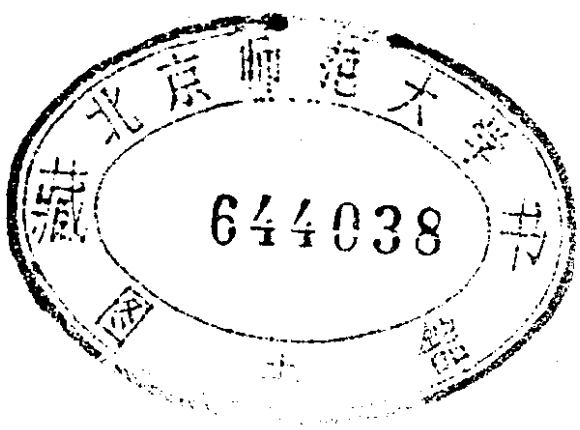
数量化理论及其应用

董文泉　周光亚　夏立显　编著

吉林人民出版社

数量化理论及其应用

董文泉 周光亚 夏立显 编著



吉林人民出版社

内 容 提 要

本书介绍多元统计分析的一个分支——处理定性数据的数量化理论，以及它在地质、气象、林业、环境保护、医学、生物学、企业管理、产品设计等方面的应用·可供有关方面的科技人员和大专院校师生参阅。

数量化理论及其应用

董文泉　周光亚　夏立显　编著

*

吉林人民出版社出版　吉林省新华书店发行

长春新华印刷厂印刷

*

850×1168毫米32开本　6½印张　146,000字

1979年8月第1版　1979年8月第1次印刷

印数：1—15,800册

书号：16091·311　定价：0.83元

前　　言

本书主要介绍林　知已夫的数量化理论，以及国内外在地质、气象、林业、环境保护、医学、生物学、企业管理等方面的具体应用，同时叙述作者在探讨数量化理论的数学模型、数量化理论与其它多元分析方法的关系，以及简化算法等方面的一些工作。

为了适应各方面读者的需要，书中列举了多方面的应用实例，并增添了“矩阵”、“TQ—16机 BCY 语言程序” 两个附录。

随着数量化理论的不断发展，它的应用也日趋广泛，除了上述提及的领域外，在工农业生产，社会科学调查等方面也有着重要的应用。近年来，国内不少单位应用数量化理论做了许多工作，因限于我们的见闻，未能广为收编。

在我们编写本书的过程中，曾得到吉林大学、长春地质学院、吉林省林科所、地质科学院地矿所、科学院地质研究所、辽宁省中心气象台等单位的一些同志的帮助和鼓励，对此我们表示衷心的感谢。

由于我们理论水平不高，实践经验不多，错误和不妥之处在所难免，欢迎读者批评指正。

概　　论

数量化理论（theory of quantification）作为多元分析的一个分支，始于五十年代。起初它的应用仅限于“计量社会学”方面，随着电子计算机的广泛应用，六十年代以后，它在自然科学领域中的应用日益增多。

多元分析是根据观测数据研究多个（随机）变量间关系的一个数理统计分支。由于已知条件和研究目的之不同，多元分析又可分为许多小的分支。在多个变量中，由于它在我们所研究的问题当中所处的地位之不同，常可分为两种情况：一些变量被视为变化的原因，称之为**说明变量**（explanatory variable）或自变量；另一种变量被视为变化的结果，称之为**基准变量**（criterion variable）或因变量。

另方面，在所考虑的变量中，由于其变化情况之不同，又可分为两种：一种就是我们通常所说的变量，例如长度、重量、人口、产量等，称之为**定量变量**或间隔尺度变量；另一种变量并非真有数量上的变化，而只有性质上的差异，例如天气（阴、晴）、性别（男、女）、职业（工人、职员、教员等）、品种等，称之为**定性变量**或名义尺度变量。定量变量和定性变量相应的数据也分别称为**定量数据**和**定性数据**。这种定性变量，在多元分析的许多实际问题中的作用是不可忽视的，某些定性变量甚至起决定性作用，如在生物的分类、地质年代的划分等类问题中。

定量变量与定性变量之间不是不可以转化的。如果我们将数轴划分为互不相交的若干个区间，当一定量变量取值于同一区间时认为是同一等级，这样便将此定量变量转化为定性变量，相应的数据也转化为定性数据。反之，对于定性变量及其数据，设法按某一合理的原则，实现向定量方面的转化，并以得到的定量数据为基础进行予测或分类等研究，这就是数量化理论的内容和目的。由此可见，数量化理论使得我们不仅可以利用定量变量，而且可以利用定性变量，从而可以更充分地利用可能搜集到的信息，更全面地研究并发现事物间的联系和规律性，因而它是多元分析中的一种有力工具，应用也非常广泛。

数量化理论按其所研究问题目的之不同，可分为四种，分别称为数量化理论Ⅰ、Ⅱ、Ⅲ、Ⅳ，后面将分章加以介绍。为了对这几种方法有个概括的了解，并且搞清它们与多元分析中其它方法间的关系，列出以下表格以供参考：

主要目的	基准变量	说明变量	主要方法
予测、发现关系式	定量的	定量的	回归分析、典型相关分析
		可兼有定性和定量的	数量化理论Ⅰ
样品的分类	定性的	定量的	判别分析
		可兼有定性和定量的	数量化理论Ⅱ
变量或样品的分类	无	定量的	主成分分析、因子分析
		可兼有定性和定量的	数量化理论Ⅲ、Ⅳ、对应分析

数量化理论是正在发展着的理论。将它置于样本分布的严密理论基础之上，还是近年才开始的。这一理论中还有不少问题值得研究，例如项目的如何选取、类目的如何划分、定量数据转化为定性数据时对结果有怎样的影响等等，这些问题都有待于从理论与实践两方面来解决。正因为这样，它像多元分析中其它分支一样，也有褒贬的不同看法，这种正常的争论，对于一种理论、方法的健康发展是有益的。

目　　录

概　论 …… 1
第一章　数量化理论Ⅰ …… 1
§1.1　项目、类目及其反应 …… 1
§1.2　数量化理论Ⅰ的数学模型与解法 …… 4
§1.3　正规方程的解的性质 …… 9
§1.4　最大相关准则 …… 14
§1.5　予测精度及各项目对予测的贡献 …… 15
§1.6　选择项目的方法 …… 20
§1.7　兼有定性和定量说明变量的情形 …… 22
§1.8　关于数量化理论Ⅰ的数学模型 …… 24
§1.9　应用例 …… 32
第二章　数量化理论Ⅱ …… 49
§2.1　判别函数的求法 …… 49
§2.2　判据的确定 …… 61
§2.3　方程组$Cb = \lambda Db$的解的性质 …… 69
§2.4　多维判别问题 …… 72
§2.5　二组判别和数量化理论Ⅰ的一致性 …… 81
§2.6　数量化理论Ⅱ的数学模型和选择项目（或类目）的方法 …… 85
§2.7　应用例 …… 87
第三章　数量化理论Ⅲ …… 108
§3.1　数量化理论Ⅲ的提出和解法 …… 108

§3.2 最大相关准则 ……116
§3.3 兼有定性和定量说明变量的情形 ……120
§3.4 数量化理论Ⅲ和对应分析的等价性 ……124
§3.5 应用例 ……130
第四章 数量化理论Ⅳ ……137
§4.1 亲近度 ……137
§4.2 数量化理论Ⅳ的提出和解法 ……142
§4.3 多维空间情形 ……146
§4.4 应用例 ……150
附录1 阵矩……153
附录2 TQ-16 机 BCY 语言程序 ……168

第一章　数量化理论 I

数量化理论 I 与回归分析都是用于定量基准变量的予测问题，前者着重考虑说明变量为定性变量的情形，后者通常限于考虑定量的说明变量，但实际上两者是可以得到统一的。

§1.1　项目、类目及其反应

在数量化理论中，常把定性变量叫做**项目** (item)，而把定性变量的各种不同的取“值”叫做**类目** (category)。譬如，职业是项目，而工人、技术员、职员、教员等是这个项目的类目。

今考虑依据一些项目 $x_1, x_2, \cdots, x_m$ 对定量的基准变量 y 进行予测。设第 1 个项目 x_1 有 r_1 个类目 $c_{11}, c_{12}, \cdots, c_{1r_1}$，第 2 个项目 x_2 有 r_2 个类目 $c_{21}, c_{22}, \cdots, c_{2r_2}, \cdots$，第 m 个项目 x_m 有 r_m 个类目 $c_{m1}, c_{m2}, \cdots, c_{mr_m}$，总共有 $\sum_{j=1}^{m} r_j = p$ 个类目。假定我们观测了 n 个样品，测定的结果记入下列的项目、类目反应表（表1—1）。

表中 y_i 是基准变量 y 在第 i 个样品中的测定值，$\delta_i(j,k)$ $(i=1,\cdots 2,\cdots,n;\ j=1,2,\cdots,m;\ k=1,2,\cdots,r_j)$ 称为 j 项目之 k 类目在第 i 样品中的反应，并按下式确定

$$\delta_i(j,k)=\begin{cases}1 & \text{当第 } i \text{ 样品中 } j \text{ 项目的定性数据为 } k \text{ 类目时，}\\ 0 & \text{否则。}\end{cases} \tag{1.1.1}$$

表1—1 项目、类目反应表

样品号	项目 / 类目 / 基准变量	x_1	x_2	…	x_m
		$c_{11}, c_{12}, \cdots, c_1r_1$	$c_{21}, c_{22}, \cdots, c_2r_2$	…	$c_{m1}, c_{m2}, \cdots, c_mr_m$
1	y_1	$\delta_1(1,1), \delta_1(1,2), \cdots, \delta_1(1,r_1)$	$\delta_1(2,1), \delta_1(2,2), \cdots, \delta_1(2,r_2)$	…	$\delta_1(m,1), \delta_1(m,2), \cdots, \delta_1(m,r_m)$
2	y_2	$\delta_2(1,1), \delta_2(1,2), \cdots, \delta_2(1_1r_1)$	$\delta_2(2,1), \delta_2(2,2), \cdots, \delta_2(2,r_2)$	…	$\delta_2(m,1), \delta_2(m,2), \cdots, \delta_2(m,r_m)$
3	y_3	$\delta_3(1,1), \delta_3(1,2), \cdots, \delta_3(1,r_1)$	$\delta_3(2,1), \delta_3(2,2), \cdots, \delta_3(2,r_2)$	…	$\delta_3(m,1), \delta_3(m,2), \cdots, \delta_3(m,r_m)$
⋮	⋮	⋮ ⋮ ⋮	⋮ ⋮ ⋮	⋮	⋮ ⋮ ⋮
n	y_n	$\delta_n(1,1), \delta_n(1,2), \cdots, \delta_n(1,r_1)$	$\delta_n(2,1), \delta_n(2,2), \cdots, \delta_n(2,r_2)$	…	$\delta_n(m,1), \delta_n(m,2), \cdots, \delta_n(m,r_m)$

由所有 $\delta_i(j,k)$ 构成的 $n \times p$ 阶矩阵记以

$$X=\begin{pmatrix} \delta_1(1,1)\cdots\delta_1(1,r_1)\delta_1(2,1)\cdots\delta_1(2,r_2)\cdots\delta_1(m,1)\cdots\delta_1(m,r_m) \\ \delta_2(1,1)\cdots\delta_2(1,r_1)\delta_2(2,1)\cdots\delta_2(2,r_2)\cdots\delta_2(m,1)\cdots\delta_2(m,r_m) \\ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ \delta_n(1,1)\cdots\delta_n(1,r_1)\delta_n(2,1)\cdots\delta_n(2,r_2)\cdots\delta_n(m,1)\cdots\delta_n(m,r_m) \end{pmatrix} \quad (1.1.2)$$

称为**反应矩阵**。

为了便于理解，举一简单例子。根据经验，负重能力与体重和性别有关，负重能力是定量变量（譬如以负重来测定），体重是定量变量，但可按以下三种等级转化为定性变量：

轻：小于 100 斤，

中：大于等于 100 斤但小于 130 斤，

重：大于等于 130 斤。

性别显然是定性变量。这样，体重有 3 个类目，性别有 2 个类目。为了用体重和性别对负重能力进行予测，假定我们测定了10个样品，测定结果如表1—2。

表 1—2

样品号	基准变量	项目 / 类目	体重 x_1			性别 x_2	
			轻 c_{11}	中 c_{12}	重 c_{13}	女 c_{21}	男 c_{22}
1	3		1	0	0	1	0
2	5		0	1	0	1	0
3	6		0	0	1	1	0
4	7		1	0	0	0	1
5	9		0	1	0	0	1
6	11		0	0	1	0	1
7	9		1	0	0	0	1
8	7		0	1	0	0	1
9	7		0	0	1	1	0
10	6		0	1	0	1	0

表中第一行表明样品号为 1 的人，负重能力是 3，体重属于轻量级，性别是女．第六行表明样品号为 6 的人，负重能力是11，体重属于重量级，性别是男．余类推．

反应 $\delta_i(j,k)$ 有个重要的性质，即对每个固定的 i 和 j 有

$$\sum_{k=1}^{r_j}\delta_i(j,k)=1. \tag{1.1.3}$$

这是由于任一样品在每个项目中只有一个类目的反应是 1，其余类目的反应皆为 0，这一简单事实在以后是非常有用的．

在进行予测时，样品的数目不能取得太少，而且在取样上尽可能做到避免有所偏重．根据经验，一般应取样品数 $n\geqslant 2p$，当然多些更好．

§1.2 数量化理论 I 的数学模型与解法

在数量化理论 I 中，假定基准变量与各项目、类目的反应间遵从下列线性模型：

$$y_i=\sum_{j=1}^{m}\sum_{k=1}^{r_j}\delta_i(j,k)b_{jk}+\varepsilon_i,\ i=1,2,\cdots,n. \tag{1.2.1}$$

这里 b_{jk} 是仅依赖于 j 项目之 k 类目的常系数，ε_i 是第 i 次抽样中的随机误差。

现在我们要根据最小二乘原理寻求系数 b_{jk} 的最小二乘估计，换言之即寻求 b_{jk} 使得

$$q=\sum_{i=1}^{n}\varepsilon_i^2=\sum_{j=1}^{n}\Big[y_i-\sum_{j=1}^{m}\sum_{k=1}^{r_j}\delta_i(j,k)b_{jk}\Big]^2 \tag{1.2.2}$$

达到最小值．为此，求 q 关于 $b_{\alpha\beta}$ 的偏导数并令其等于 0，得到

$$\frac{\partial q}{\partial b_{uv}} \equiv -2\sum_{i=1}^{n}\left[y_i - \sum_{j=1}^{m}\sum_{k=1}^{r_j}\delta_i(j,k)b_{jk}\right]\delta_i(u,v) = 0,$$

$$u = 1, 2, \cdots, m; \quad v = 1, 2, \cdots, r_u.$$

因为这是极小值点的必要条件，故如记使 q 达到最小值的 b_{jk} 以 $\hat{b}_{jk}$，则 $\hat{b}_{jk}$ 应满足上式，亦即满足

$$\sum_{j=1}^{m}\sum_{k=1}^{r_j}\left[\sum_{i=1}^{n}\delta_i(j,k)\delta_i(u,v)\right]\hat{b}_{jk} = \sum_{i=1}^{n}\delta_i(u,v)y_i,$$

$$u = 1, 2, \cdots, m; \quad v = 1, 2, \cdots, r_u. \tag{1.2.3}$$

如果用矩阵形式来表示，(1.2.3)可写成

$$\mathbf{X}'\mathbf{X}\hat{\mathbf{b}} = \mathbf{X}'\mathbf{y}, \tag{1.2.4}$$

其中

$$\mathbf{y}' = (y_1, y_2, \cdots, y_n),$$
$$\hat{\mathbf{b}}' = (\hat{b}_{11}, \cdots, \hat{b}_{1r_1}, \hat{b}_{21}, \cdots, \hat{b}_{2r_2}, \cdots, \hat{b}_{m1}, \cdots, \hat{b}_{mr_m}).$$

我们称方程组(1.2.3)或(1.2.4)为**正规方程**（组）。

矩阵是处理多元问题的有力工具，以后将更多地利用这一工具。这里将正规方程的推导再用矩阵工具重复一遍。记

$$\mathbf{b}' = (b_{11}, \cdots, b_{1r_1}, b_{21}, \cdots, b_{2r_2}, \cdots, b_{m1}, \cdots, b_{mr_m}),$$
$$\mathbf{e}' = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n).$$

则(1.2.1)可写成

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}, \tag{1.2.5}$$

并且

$$q = \mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}). \tag{1.2.6}$$

为使 q 达到最小值，必要条件是

$$\frac{\partial q}{\partial \mathbf{b}} \equiv -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{0}.$$

（关于矩阵求偏导数见附录）.因此所求的最小二乘估计 $\hat{\mathbf{b}}$ 应满足

$$\mathbf{X}'\mathbf{X}\hat{\mathbf{b}} = \mathbf{X}'\mathbf{y}.$$

对于§1.1中所举的例子，正规方程是

$$\begin{aligned}
3\hat{b}_{11} \qquad\qquad &+ \hat{b}_{21} + 2\hat{b}_{22} = 19 \\
4\hat{b}_{12} \qquad &+ 2\hat{b}_{21} + 2\hat{b}_{22} = 27 \\
3\hat{b}_{13} &+ 2\hat{b}_{21} + \hat{b}_{22} = 24 \\
\hat{b}_{11} + 2\hat{b}_{12} + 2\hat{b}_{13} &+ 5\hat{b}_{21} \qquad = 27 \\
2\hat{b}_{11} + 2\hat{b}_{12} + \hat{b}_{13} &\qquad + 5\hat{b}_{22} = 43
\end{aligned} \tag{1.2.7}$$

从这个具体例子我们看到，正规方程组的系数间有很强的规律性。首先，系数矩阵是对称的。第一个方程左端诸系数是矩阵 $\mathbf{X}'$ 的第一行分别与 $\mathbf{X}$ 的各列相乘而得，而右端是 $\mathbf{X}'$ 的第一行与 $\mathbf{y}$ 相乘而得。而 $\mathbf{X}'$ 的第一行恰由反应矩阵中第一项目的第一类目诸反应所构成。为了方便，我们称第一方程为第一项目第一类目的方程。依此类推，称第二方程为第一项目第二类目的方程，等等。由于系数矩阵的对称性，我们也把第一列称为第一项目第一类目的列，把第二列称为第一项目第二类目的列，等等。方程组的系数间另一个明显的规律是，在每一列（包括右端常数项的列）中，对应第一项目的前三个元素之和与对应第二项目的两个元素之和总是相等的。

现在来证明一般正规方程（1.2.4）都具有上述规律性。

1）正规方程（1.2.4）的系数矩阵是对称的。

这是由于系数矩阵是 $\mathbf{X}'\mathbf{X}$，而 $(\mathbf{X}'\mathbf{X})' = \mathbf{X}'\mathbf{X}$ 之故。

2）在正规方程（1.2.4）系数矩阵的各列（包括右端常数项的列）中，对应各项目的诸类目的元素之和皆相等。

证：考虑 j 项目的 k 类目之列，对应第 u 项目各元素之和为

$$\sum_{v=1}^{r_u}\sum_{i=1}^{n}\delta_i(j,k)\delta_i(u,v)=\sum_{i=1}^{n}\delta_i(j,k)\sum_{v=1}^{r_u}\delta_i(u,v)$$

$$=\sum_{i=1}^{n}\delta_i(j,k).$$

由于这个值与 u 无关，可见对每个 $u=1,\cdots,m$，它都是一样的。再考虑右端常数项的列，对应第 u 项目各元素之和为

$$\sum_{v=1}^{r_u}\sum_{i=1}^{n}\delta_i(u,v)y_i=\sum_{i=1}^{n}y_i\sum_{v=1}^{r_u}\delta_i(u,v)=\sum_{i=1}^{n}y_i.$$

它也与 u 无关，于是性质得证。

3）正规方程（1.2.4）中最多有 $\sum_{j=1}^{m}r_j-m+1$ 个方程是线性无关的。

证：由性质 2）可知对应每个项目的诸方程之和皆相同，故除某一个项目（例如第一项目）外，其余各项目的方程中如不去掉一个，放在一起必定线性相关，于是性质得证。

性质 3）表明正规方程（1.2.4）的系数矩阵 $\mathbf{X}'\mathbf{X}$ 是不满秩的，其秩 $R(\mathbf{X}'\mathbf{X})$ 最多是 $\sum_{j=1}^{m}r_j-m+1$，因此方程的解是无穷多的。在以后的讨论中，假定 $\mathbf{X}'\mathbf{X}$ 的秩是 $\sum_{j=1}^{m}r_j-m+1$（在实际问题中，只要样品数 n 足够大，一般总可保证这一点）。这时，我们可以对每个 $j=2,\cdots,m$，删去第 j 项目第一类目的方程并取 $\hat{b}_{j1}=0$，以使删除后的方程组的系数矩阵成为满秩的，故可唯一地解出其余的 $\hat{b}_{jk}$，这样去解正规方程不失一般性，并且确使 q 达到最小，关于这一点将在下一节中阐述。

在解出 $\hat{b}_{jk}$ 之后，便得到以下的予测方程

$$\hat{y}=\sum_{j=1}^{m}\sum_{k=1}^{r_j}\delta(j,k)\hat{b}_{jk},\qquad(1.2.8)$$

这里 $\delta(j,k)$ 表示任一样品在 j 项目的 k 类目的反应，当我们取得一样品时，便可由其反应 $\delta(j,k)$ 按 (1.2.8) 算出 $\hat{y}$，作为对基准变量 y 的予测值. 有时称 $\hat{b}_{jk}$ 为 j 项目 k 类目的得分，并用表格的形式表达予测方程，称之为项目类目得分表 (参看§1.9).

对于原给的 n 个样品，也可按 (1.2.8) 算出对应的予测值

$$\hat{y}_i = \sum_{j=1}^{m} \sum_{k=1}^{r_j} \delta_i(j,k)\hat{b}_{jk}, \quad i = 1, 2, \cdots, n. \tag{1.2.9}$$

并用它与已知的 $y_i (i = 1, 2, \cdots, n)$ 对比，以检验予测效果. (1.2.9)也可用矩阵形式写成

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}}, \tag{1.2.10}$$

其中

$$\hat{\mathbf{y}}' = (\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n).$$

对于前面的例子，解方程组(1.2.7)可得：

$\hat{b}_{11} = 3.81$, $\hat{b}_{12} = 4.86$, $\hat{b}_{13} = 6.74$, $\hat{b}_{21} = 0$, $\hat{b}_{22} = 3.79$,

因而得到予测方程为

$$\hat{y} = 3.81\delta(1,1) + 4.86\delta(1,2) + 6.74\delta(1,3) + 3.79\delta(2,2).$$

将原样品的予测值与实测值对比如表1—3所示.

表1—3

i	1	2	3	4	5	6	7	8	9	10	和
y_i	3	5	6	7	9	11	9	7	7	6	70
$\hat{y}_i$	3.81	4.86	6.74	7.59	8.64	10.52	7.59	8.64	6.74	4.86	70
$y_i - \hat{y}_i$	−0.81	0.14	−0.74	−0.60	0.36	0.48	1.40	−1.64	0.26	1.14	0