数量化理论及其应用

吉林大学数学系概率统计教研室
吉林省林业科学研究所经营研究室　合编

吉林省林业勘测第二大队　翻印

一九七六年二月

目　　录

概　论

数量化理论始于五十年代，起初它的应用　于“计量社会学”。随着电子计算机的广泛应用，六十年代以后，它在自然科学中的应用日益增多。

毛主席教导我们，“大家明白，不论做什么事，不懂得那件事的情形，它的性质，它和它以外事情的关联，就不知道那件事的规律，就不知道如何去做，就不能做好那件事。”数量化理论是多元分析的一支。因此必须弄清楚它与多元分析中其它主要方法的联系与区别。

多元分析中的某些方法是把许多变量分成两个组，并研究一变量组与另一变量组的关联。当某变量x被视为变化的原因，变量y视为变化的结果时，则把x叫做说明变量(explanatory variable)，或予测变量(predictor variable)而把y叫做基准变量(criterion variable)。由说明变量构成的变量组叫做说明变量组，而由基准变量构成的变量组叫做基准变量组。

此外，多元分析中的方法还因调查或实验而得到的数据资料的性质而异。从数据的性质上分，有“量的数据”以及“质的数据”。譬如身长、体重、人口、产量等都是“量的数据”。而性别(男女)、职业、品种、试验设计中质的因子之水准等都是“质的数据”。

根据上述两点，我们可以把多元分析中的主要方法归纳如下：

主要目的	基准变数		说明变量	主要方法
予测	有	量的数据	仅有量的数据	多元回归分析、正准相关分析
发现关系式		量的数据	有质的数据	数量化理论 I
样本的分类		质的数据	仅有量的数据	判别分析
		质的数据	有质的数据	数量化理论 II
变量或样本的分类	无		仅有量的数据	主成分分析、因子分析
			有质的数据	数量化理论 III

“量的数据”与“质的数据”之间不是不可转换的。当我们将数轴划分为不相交的区间时，这些区间就是“质的数据”，从而“量的数据”就转换为“质的数据”了。反之，从某一原则出发，将“质的数据”转化为“量的数据”，以便用于予测或分类的理论，就是数量化理论。

数量化理论对于变量的分布不作任何要求，并且变量间的关系为非线性时，也可以应用。加之“量的数据”可以转化为“质的数据”，于是它的应用的广泛性就更为明显了。

第一章 数量化理论 I

数量化理论 I 与多元回归分析都是用于基准变量是“量的数据”时的予测问题，不同之处在于前者的说明变量中有“质的数据”，而后者仅有“量的数据”。

在数量化理论中，为了方便起见，常把“质的数据”的变量叫做项目（Item）。而这个变量的“质的数据”叫做这个项目的类目（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。

假定为了予测某现象，而该现象可用“量的数据”来表示，为此除了测定基准变量之外，还测定了 m 个项目。第一个项目有 r_1 个类目，第二个项目有 r 个类目……第 m 个项目有 r_m 个类目，总共有 $\sum_{e=1}^{} r_e = P$ 个类目。为了达到予测之目的，测定的样本的数目

不能太少，根据经验，样本数n应满足

$$n \geq 2p \qquad （当然n愈大愈好）$$

通常把每个样本的测定结果记入下列的项目类目反应表中。

样本的项目、类目反应表

样本号 \ 基准变量 \ 类目 \ 项目		x_1	x_2	……	x_j	……	x_m
		$c_{11}\ c_{12}\cdots\cdots c_{1r_1}$	$c_{21}\ c_{22}\cdots\cdots c_{2r_2}$	……	$c_{j1}\ c_{j2}\cdots\cdots c_{jr_j}$	……	$c_{m1}\ c_{m2}\cdots\cdots c_{mr_m}$
1	y_1	$\delta_1(1,1)\cdots\cdots\delta_1(1,r_1)$	$\delta_1(2,1)\cdots\cdots\delta_1(2,r_2)$	……	$\delta_1(j,1)\cdots\delta_1(j,r_j)$		$\delta_1(m,1)\cdots\delta_1(m,r_m)$
2	y_2	$\delta_2(1,1)\cdots\cdots\delta_2(1,r_1)$	$\delta_2(2,1)\cdots\cdots\delta_2(2,r_2)$	……	$\delta_2(j,1)\cdots\delta_2(j,r_j)$		$\delta_2(m,1)\cdots\delta_2(m,r_m)$
3	y_3	$\delta_3(1,1)\cdots\cdots\delta_3(1,r_1)$	$\delta_3(2,1)\cdots\cdots\delta_3(2,r_2)$	……	$\delta_3(j,1)\cdots\delta_3(j,r_j)$		$\delta_3(m,1)\cdots\delta_3(m,r_m)$
⋮	⋮	⋮ ⋮	⋮ ⋮		⋮ ⋮		⋮ ⋮
n	y_n	$\delta_n(1,1)\cdots\cdots\delta_n(1,r_n)$	$\delta_n(2,1)\cdots\cdots\delta_n(2,r_2)$	……	$\delta_n(j,1)\cdots\delta_n(j,r_j)$	……	$\delta_n(m,1)\cdots\cdots\delta_n(m,r_n)$

表中y_i是测得的第i个样本的基准变量，

而$\delta_i=(j\cdot K)=1$：当第i个样本的第j个说明变量的“质的数据”

是第k个类目时；

0：否则

为了便于理解，我们举一个虚拟的例子：

根据经验负重能力与体重与性别有关。负重能力是“量的数据”（譬如以负重来测定），体重是“量的数据”，但可按小于１００斤，大于等于１００斤而小于１３０斤，大于等于１３０斤而转换成“质的数据”，性别显然是“质的数据”。由于体重有3个类目，性别有2个类目，于是测定了１０个样本，以便予测负重能力。并将测定结果记入下列项目、类目反应表中：

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

为了方便起见，今后把反应表中由$\delta_i(j\cdot K)$组成的矩阵$(\delta_i(j\cdot K))$称为反应矩阵，并记以X。

反应表

样本号 \ 项目	基准变量 \ 类目	体重 (x_1) 轻 C_{11}	中 C_{12}	重 C_{13}	性别 (x_2) 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

从这个例子的反应表中可以看出，对于每个样本来说，其每一个项目中只有一个类目的反应是1，余者皆为0。换言之反应矩阵中隶属于一个项目元素之行和为1。于是如果反应矩阵X有m个项目，第一个项目有r_1个类目，第二个项目有r_2个类目，……，第m个项目有r_m个类目，则反应矩阵X的秩$\le \sum_{j=1}^{m} r_j - m + 1$。如果样本数n充分大，则实际上可以有反应矩阵X的秩$= \sum_{j=1}^{m} r_j - m + 1 < n$。以后的讨论都在上述假定下进行的。

§1.1 数量化理论Ⅰ的导出

在数量化理论Ⅰ中，假定基准变量y_i与各项目、类目的反应$\delta_i(j \cdot k)$之间有下列线性关系成立：

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

其中b_{jk}是仅仅依赖于j项目的k类目的常数。y_i、ε_i、$\delta_i(j \cdot k)$都依赖于第i个样本，$\delta_i(j \cdot k)$也还依赖j项目的k类目。

下面我们用最小二乘法原理，去求b_{jk}的估计值$\hat{b}_{jk}$。当我们用$\hat{b}_{jk}$代替b_{jk}时，由(1.1.1)可得

$$y_i = \sum_{j=1}^{m} \sum_{k=1}^{r_j} \delta_i(j \cdot k)\, \hat{b}_{jk} + \hat{\varepsilon}_i \qquad i = 1, \cdots\cdots n \quad (1,1,2)$$

欲使 $q = \Sigma \varepsilon_a^2 \longrightarrow min$ （极小化），其必要条件令

$$\frac{\partial q}{\partial \hat{b}_{uv}} = 0$$

即可得为

$$\frac{\partial q}{\partial \hat{b}_{uv}} = -2\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{m}\sum_{k=1}^{r_j}\delta_i(j\cdot k)\hat{b}_{jk}\right)\delta_i(u\cdot v) = 0$$

$$u = 1, \cdots m$$
$$v = 1 \cdots r_u$$

于是有

$$\sum_{j=1}^{m}\sum_{k=1}^{r_j}\left(\sum_{i=1}^{n}\delta_i(j\cdot k)\delta_i(u\cdot v)\right)\hat{b}_{j\cdot k} = \sum_{i=1}^{n}\delta_i(u\cdot v)y_i \qquad \begin{matrix} u = 1, \cdots m \\ v = 1, \cdots r_u \end{matrix} \quad (1,1,3)$$

故方程组（1．1．3）就是 $\hat{b}$ j k应满足的方程。

如果令

$$Y^T = (y_1 y_2 \cdots\cdots y_n) \quad \hat{B}^T = (\hat{b}_{1\cdot1}, \hat{b}_{1\cdot2}, \hat{b}_{1r_1}, \hat{b}_{21}, \cdots \hat{b}_{m_1} \cdots \hat{b}_{m\cdots})$$

其中 Y^T 是矩障Y（实际上是向量）的转置矩障，B T是B的转置，则方程组（1．1．3）可以写成

$$X^T \times \hat{B} = X^T Y \qquad (1.1.4)$$

为了使一些同志熟悉有关矩障运算，我们把推导方程组(1．1．3）的过程，使用矩障这一工具冉重复一遍，以便于以后使用矩障。

如果令 $\hat{D}^T=(\hat{\varepsilon}_1,\hat{\varepsilon}_2,\cdots,\hat{\varepsilon}_n)$ 则（1．1．2）可以写成

$$Y=X\hat{B}+\hat{D} \qquad (1.1.5)$$

且

$$
\begin{aligned}
Q &= \sum_{i=1}^{n}\hat{\varepsilon}_i^2=\hat{\varepsilon}_1^2+\hat{\varepsilon}_2^2+\cdots\cdots+\hat{\varepsilon}_n^2 \\
&= \hat{D}^T\hat{D} \\
&= (Y-X\hat{B})^T(Y-X\hat{B}) \quad (\text{由}1.1.5) \\
&= Y^TY-2\hat{B}^TX^TY+\hat{B}^TX^TX\hat{B}
\end{aligned}
$$

使 $Q=D^TD$ 为最小的必要条件，由 D T D 对 B 求偏导数 *），并令其为 0（零向量）即可得：

$$\frac{\partial}{\partial B}(\hat{D}^T\hat{D})=-2X^TY+2X^TX\hat{B}=0$$

于是有

$$X^TX\hat{B}=X^TY \qquad (1.1.4)$$

前面所举例子的相应于（1．1．3）或（1，1，4）的方程组是：

$$
\begin{aligned}
3\hat{b}_{11} \qquad\qquad\qquad\qquad\quad \hat{b}_{21}+2\hat{b}_{22} &= 19 \\
4\hat{b}_{12}+2\hat{b}_{21}+2\hat{b}_{22} &= 27 \\
3\hat{b}_{13}+2\hat{b}_{21}+\hat{b}_{22} &= 24 \qquad (1.1.6) \\
\hat{b}_{11}+2\hat{b}_{12}+2\hat{b}_{13}+5\hat{b}_{21} \qquad\quad &= 27 \\
2\hat{b}_{11}+2\hat{b}_{12}+\hat{b}_{13}+ \qquad 5\hat{b}_{22} &= 43
\end{aligned}
$$

*）关于矩阵求偏导数见附录。

这个方程组的系数有很强的规律性。首先系数矩障是对称矩障。第一个方程是矩障 X^T 的第一行分别与 X 的各列相乘而得，而右端是 X^T 的第一行与 Y 相乘而得，而 X^T 的第一行恰好是反应矩障第一项目第一类目的反应构成的，为了方便起见有时把第一个方程称为对应于第一项目第一类目的方程。类似地把最后一个方程叫做对应于第二项目第二类目的方程，也分别把系数矩障的第一行叫做对应于第一项目第一类目的行……。由于方程组的系数矩障是对称的，有时我们也把系数矩障的第一列叫做对应于第一项目第一类目的列，………最后一列叫做对应于第二项目第二类目的列。方程组（1．1．6）系数矩障的规律性还表现在每一列中前三个元素之和等于后两者元素之和。方程右端的列向量也是如此。换言之，对于每个列来说，该列的对应于一个项目的元素之和皆相等（不同列的自然可以不等）。（1．1．6）的这些性质具有普遍性，可以证明如下：

先看（1．1．4）的右端，对应第 u 项目的元素之和是

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

$u=1,2,\cdots\cdots m$

其中 r_u 是第 u 项目中的类目数。于是对于右端这个性质已证毕。再看系数矩障对应第 j 项目第 k 类目之列，其对应于第 u 项目的行的诸元素之和是

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

其中u—1，2，……m，证毕。

这个性质说明了方程（1．1．4）中最多只有

$\sum_{j=1}^{m} r_j - m + 1$ 个方程是线性无关的。

在以后的讨论中，我们假定方程组（1．1．4）的系数矩阵$X^T X$的秩是$\sum_{j=1}^{m} r_j - m + 1$。在解方程组（1．1．3）或（1．1．4）时，为了方便起见，我们先删去第j项目第一类目（j—2……m）所对应的方程，然后令$\hat{b}_{j1}=0$（j—2，…m）。此时由于方程组的系数是满秩的，故可唯一地解出其余的$\hat{b}_{jk}$。这样去解不失一般性，关于这一点将在下一节阐述。

算出$\hat{b}_{jk}$后，我们便可以得到各样本的基准变量的予测值，它由下式给出：

$$\hat{y}_i=\sum_{j=1}^{m}\sum_{k=1}^{r_j} \delta_i(j,k)\hat{b}_{jk} \qquad i=1,\cdots\cdots,n \quad (1.1.7)$$

由（1．1．7）算出的$\hat{y}_i$，有时称为第i个样本的得分。

如果令Y^T—（y_1，……y_n）则（1．1．7）可以写成

$$\hat{Y}=X\hat{B} \qquad (1.1.8)$$

下面我们来证明$\sum_{i=1}^{n} y_i=\sum_{i=1}^{n}\hat{y}_i$

为此将（1．1．8）代入（1．1．4）则有

$$X^T\hat{Y}=X^TY$$

对　对于我们前面的例子，（1．1．9）可写成

$$\begin{pmatrix} 1001001000 \\ 0100100101 \\ 0010010010 \\ 1110000011 \\ 0001111100 \end{pmatrix} \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \hat{y}_4 \\ \hat{y}_5 \end{pmatrix} = \begin{pmatrix} 1001001000 \\ 0100100101 \\ 0010010010 \\ \hdashline 1110000011 \\ 0001111100 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix}$$

从而有

$$\begin{pmatrix} \hat{y}_1+\hat{y}_4+\hat{y}_7 \\ \hat{y}_2+\hat{y}_5+\hat{y}_8+\hat{y}_{10} \\ \hat{y}_3+\hat{y}_7+\hat{y}_9 \\ \hat{y}_1+\hat{y}_2+\hat{y}_3+\hat{y}_9+\hat{y}_{10} \\ \hat{y}_4+\hat{y}_5+\hat{y}_6+\hat{y}_7+\hat{y}_9 \end{pmatrix} = \begin{pmatrix} y_1+y_4+y_7 \\ y_2+y_5+y_8+y_{10} \\ y_3+y_7+y_9 \\ y_1+y_2+y_3+y_9+y_{10} \\ y_4+y_5+y_6+y_7+y_8 \end{pmatrix}$$

比较前三个分量就有

$$\sum_{i=1}^{10} \hat{y}_i = \sum_{i=1}^{10} y_i$$

对于一般的情形，只要比较（1．1．9）的对应于一个项目一p

分分量就可以得到 $\sum\limits_{i=1}^{n} y_i = \sum\limits_{i=1}^{n} \hat{y}_i$ 。由于

$$\frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i ,$$

常把这个性质记成 $\bar{y} = \bar{\hat{y}}$

§1.2 方程组$X^TXB=X^TY$的解的性质

为了方便起见，我们称$\hat{b}_{jK}$为对应于j项目K类目的得分或称j项目K类目的得分。今设b^o_{11}，b^o_{12}，……$b^o_{1r_1}$，0，b^o_{21}，……$b^o_{2r_2}$，0，b^o_{32}，……$b^o_{3r_3}$，……，0，b^o_{m2}，…$b^o_{mr_m}$ 是按上一节的方法得到的方程（1。1。4）的解，且设 $\hat{b}'_{11}, \hat{b}'_{12}, \cdots \hat{b}'_{1r}, \hat{b}'_{21}, \hat{b}'_{22} \cdots\cdots \hat{b}'_{2r_2}, \hat{b}'_{31} \cdots\cdots$

$$\cdots\cdots \hat{b}_{3r_3} \cdots\cdots, \hat{b}'_{m1} \cdots\cdots \hat{b}'_{mr_m}$$

是（1。1。4）的任一解，我们来讨论这二解之间有什么关系，以及对予测有什么影响。显然

$$\hat{b}'_{11}-\hat{b}^o_{11}, \hat{b}'_{12}-\hat{b}^o_{12}, \cdots\cdots \hat{b}'_{1r_1}-\hat{b}^o_{1r_1}, \hat{b}'_{21}-0, \hat{b}^o_{22}, \cdots\cdots,$$

$$\hat{b}'_{2r_2}-\hat{b}^o_{2r_2}, \cdots\cdots \hat{b}'_{m1}-0, \hat{b}'_{m2}-\hat{b}^o_{m2} \cdots\cdots \hat{b}'_{mr_m}-\hat{b}^o_{mr_m}$$

是$X^TXB=0$的解。根据上一节的讨论，方程中对应一项目的b_{jK}的系数之和皆相等这一性质可知

$$\underbrace{-\hat{b}'_{21}-\hat{b}'_{31}\cdots\cdots\hat{b}'_{m1}, -\hat{b}'_{21}-\hat{b}'_{31}\cdots\cdots\hat{b}'_{m1}, \cdots\cdots \hat{b}'_{21}\cdots\cdots\hat{b}'_{m1}}_{r_1\text{个}}$$

$$\underbrace{\hat{b}'_{21}\cdots\cdots\hat{b}'_{21}}_{r_2\text{个}}, \cdots\cdots \underbrace{\hat{b}'_{m1} \cdots\cdots \hat{b}'_{m1}}_{r_m\text{个}}$$

也是XTXB＝0的解。在该方程组中删去对应于j

（$j=2,\cdots m$）项目1类目的方程，并把其余每个方程对应于j

（$j=2,\cdots\cdots m$） 项目1类目的项移到右端，这个新方程组是满满秩

的，有唯一解。于是

$\hat{b}_{11}-\hat{b}_{11}=-\hat{b}'_{21}-\hat{b}'_{31}\cdots\cdots,-\hat{b}'_{m1}$，亦即$\hat{b}'_{11}=\hat{b}^{o}_{11}-\hat{b}'_{21}\cdots\cdots-\hat{b}'_{m1}$，

同理有：$\hat{b}'_{12}=\hat{b}^{o}_{12}-\hat{b}'_{21}\cdots\cdots-\hat{b}'_{m1},\cdots\cdots,\hat{b}'_{1r_1}=\hat{b}^{o}_{1r_1}-\hat{b}'_{21}\cdots\cdots-\hat{b}'_{m1}$。

$$\hat{b}'_{21}=\hat{b}'_{21},\hat{b}'_{22}=\hat{b}^{o}_{22}+\hat{b}'_{21},\cdots\cdots,\hat{b}'_{2r_2}=\hat{b}^{o}_{2r_2}+\hat{b}'_{21},\cdots\cdots$$

$$\hat{b}'_{m1}=\hat{b}'_{m1},\hat{b}'_{m2}=\hat{b}^{o}_{m2}+\hat{b}'_{m1},\cdots\cdots,\hat{b}'_{mr_m}=\hat{b}^{o}_{mr_m}+\hat{b}'_{m1}.$$

由于用（1。1。7）进行予测时，是在每个项目中取一个类目且只取一个，可见此二解的予测结果完全相同。

下面讨论的性质将用于第四章。在上一节中我们假定的模型是

$$y_i=\sum_{j=1}^{m}\sum_{K=1}^{r_j}\delta_i(j,K)b_{jK}+\varepsilon_i \qquad i=1,\cdots\cdots n \quad (1.2.1)$$

现在我们考虑一新模型

$$y_i=\sum_{j=1}^{m}\sum_{K=1}^{r_j}[\delta_i(j,K)-\bar{\delta}(j,K)]b'_{jK}+\varepsilon_i \qquad i=1,\cdots\cdots n \quad (1.2.2)$$

其中 $\bar{\delta}(j,K)=\frac{1}{n}\sum_{i=1}^{n}\delta_i(j,K)$