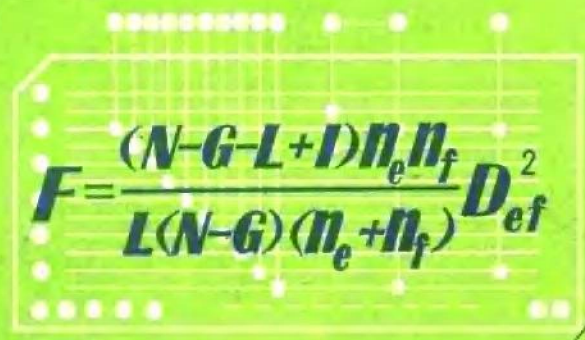


湖南医学院卫生系
卫生统计学教研组

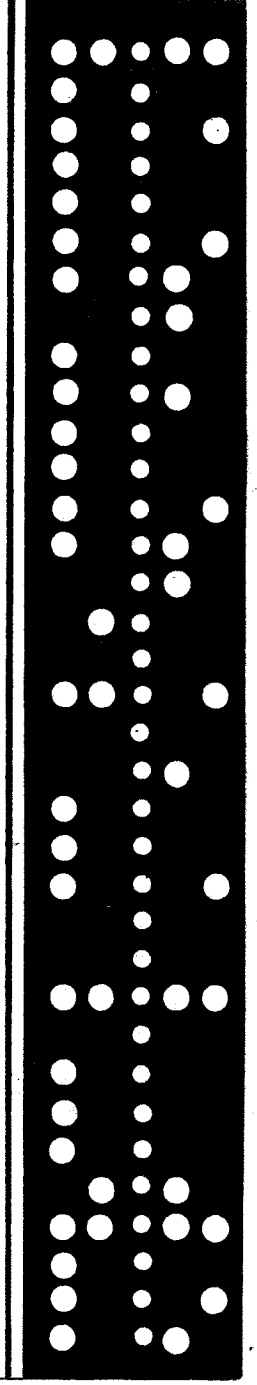
医用多因素分析

湖南科学技术出版社

医用多因素分析

主编：黄正南
编写：黄正南　刘树仁　孙振球
审阅：杜养志

湖南科学技术出版社
一九八〇年·长沙



医用多因素分析

湖南医学院卫生系卫生
统计学教研组

责任编辑：朱　杰

*

湖南科学技术出版社出版
（长沙市展览馆路14号）
湖南省新华书店发行　湖南省新华印刷二厂印刷

*

1980年9月第1版第1次印刷
字数：97,000　印张：4.75　印数：1——11,200
统一书号：14204·25　定价：0.52元

前　言

在医学科研和防治工作过程中，人们经常碰到多因素分析的问题。过去，人们常用单因素分析方法进行统计处理，要求固定研究单一因素以外的其他因素，结果很难从整体上对问题进行判断。比如，冠心病的发生，与年龄、饮食、运动、职业、精神状态等都有关系，用通常的单因素分析方法（如t检验、χ^2检验等），只能对各因素的影响个别地加以对比研究，而用多因素分析，则可以把上述各种因素的内在联系揭示出来。

本书分七讲，介绍多因素的回归分析、判别分析、聚类分析和正交试验，各讲内容有一定的独立性，读者可以结合实际工作和科研工作的需要而选读。每一讲的内容分成两部分：第一部分介绍各种多因素分析的基本概念；第二部分通过实例说明分析的方法和步骤。为了系统化和条理化，并用方框图作了总结。读者在未用到某讲内容时，也可先只看基本概念，真正需要用时再看方法和步骤。

本书的编写得到了省卫生局科技办公室、湖南医学院卫生系系主任王翔朴副教授和湖南医学院科研处杨世鞭副处长等的大力支持，在此特表谢意。

编　者

1980年6月

目　录

第一讲　多元线性回归

一　基本概念

客观世界的任何事物或现象都不是孤立的，而是与其他事物或现象相互联系着的。因此，在医学科研和防治工作中，也大量涉及到多因素的相互作用问题。通常生物机体出现某一现象或某一结果（例如肺活量），往往是很多因素（例如体重、胸围和胸围之呼吸差等）综合作用所致。前者称为因变量，一般以y表示；后者称为自变量，一般以x_1，x_2，…，x_m表示，简记为x_i($i=1, 2, \cdots, m$)。m为自变量的个数。

如果两个或两个以上的自变量x_i与因变量y的关系能用一个方程联系起来，就称为多元回归方程（一个自变量和一个因变量的回归方程为一般的一元回归方程，为多元回归方程的特殊形式），m个自变量和一个因变量的回归方程叫m元回归方程。多元回归方程中最简单而又一般的是多元线性回归方程，即

$$\hat{y} = a + b_1x_1 + b_2x_2 + \cdots + b_mx_m \qquad (1)$$

$\hat{y}$为总体y的均数的估计值，a为常数项，b_i ($i=1,2, \cdots,m$)为因变量$\hat{y}$对第i个自变量x_i的偏回归系数。多元线性回归是研究一种事物或现象与其他多种事物或现象在数量上相互联系和相互制约的统计方法。

多元线性回归分析的内容包括：1.由n个样品（n最好至少是m的5至10倍），即m＋1个变量（m个自变量和1个因变量）的n组观测数据x_{1k}，x_{2k}，……，x_{mk}，y_k $(k=1, 2, \cdots, n)$，求出b_i和a，从而建立多元线性回归方程。2.对多元线性回归方程的好坏进行检验或评价。

建立多元线性回归方程的用途主要有：1.确定每个自变量和因变量间的数量关系。2.从较易测得的x_i来推算较难测得的y。如根据就诊者的各项症状、体征及化验指标来推算是否患有某种难于直接诊断的疾病。3.从已发生的x_i来预测将发生的y。如流行病预报、存活期估计等。4.由于引入某些自变量能缩减因变量的变差，因此可比较精确地确定不同x_i值的y的正常值范围。

用多元线性回归可解决一些非线性回归的问题，常用的如：

1.多项式回归：在曲线回归分析中，相当广泛的一类曲线可用多项式逼近。设自变量为x，因变量的估计值为$\hat{y}$，多项式逼近的曲线回归方程为$\hat{y}=a+b_1x+b_2x^2+\cdots+b_mx^m$，这时如果设$x_1=x$，$x_2=x^2$，…，$x_m=x^m$，则m次多项式回归方程就可转换成m元线性回归方程。

2.趋势面分析：研究平面（也可扩展到多维空间）上某事物或某现象的分布趋势，如某疾病的地区分布，设平面坐标为(x, y)，相应的某事物或某现象的估计值为$\hat{z}$，则m阶趋势面即$\hat{z}$用x和y的m阶方程表示，如二阶趋势面为$\hat{Z}=a+b_1x+b_2y+b_3x^2+b_4xy+b_5y^2$，这时如果设$x_1=x$，$x_2=y$，$x_3=x^2$，$x_4=xy$，$x_5=y^2$，则二阶趋势面的方程就转换成五个自变量的

5元线性回归方程。

如果多元线性回归方程的自变量个数较少，可用电子计算器计算甚至手工计算；如果自变量个数较多，由于计算太繁，需用电子计算机处理，但原理是一样的。

二 方法和步骤

通过实例说明方法和步骤：10名女中学生测得体重 x_1（公斤）、胸围 x_2（厘米）、胸围之呼吸差 x_3（厘米）及肺活量 y（毫升），如表1所示，试作前三项指标与肺活量之多元线性回归分析。

表1 10名女中学生体重、胸围、胸围之呼吸差及肺活量数据

k(编号)	x_{1k}	x_{2k}	x_{3k}	$y_k(x_{4k})$
1	35	69	0.7	1600
2	40	74	2.5	2600
3	40	64	2.0	2100
4	42	74	3.0	2650
5	37	72	1.1	2400
6	45	68	1.5	2200
7	43	78	4.3	2750
8	37	66	2.0	1600
9	44	70	3.2	2750
10	42	65	3.0	2500
Σx_{ik}	405	700	23.3	23150
$\bar{x}_i$	40.5	70.0	2.33	2315

（一）多元线性回归方程的求法

1.列表计算相关系数

为表达方便，把 y记作x_{m+1}，本例把y记作x_4。在初始数据的表1的下栏，计算出各变量值的和Σx_{ik}和均数

$$\bar{x}_i = \frac{1}{n}\sum_{k=1}^{n} x_{ik} \qquad i = 1,2,\cdots,m+1 \qquad (2)$$

再用表2计算相关系数。

表2　　相关系数计算表

i	j	$\Sigma x_{ik}x_{jk}$	$(\Sigma x_{ik})(\Sigma x_{jk})$	l_{ij}	r_{ij}
1	1	16501	164025	98.5	1
1	2	28373	283500	23	0.1718
1	3	964.4	9436.5	20.75	0.6409
1	4	946550	9375750	8975	0.6945
2	2	49182	490000	182	1
2	3	1650.9	16310	19.9	0.4522
2	4	1630800	16205000	10300	0.5864
3	3	64.93	542.89	10.641	1
3	4	57035	539395	3095.5	0.7288
4	4	55287500	535922500	1695250	1

i和j表示变量的位次。如i＝1和j＝1，表示要计算x_1和x_1的相关系数；i＝1和j＝2，表示要计算x_1和x_2的相关系数；如此类推。必须计算出所有一个变量本身间和二个变量间的相关系数。

$\sum x_{ik}x_{jk}$当i＝j时表示变量值的平方和，例如对于i＝j＝1，则

$$\sum_{k=1}^{10} x_{1k}x_{1k} = \sum_{k=1}^{10} x_{1k}^2$$

$$= 35^2 + 40^2 + 40^2 + 42^2 + 37^2 + 45^2 + 43^2 + 37^2 + 44^2 + 42^2$$

$$= 16501$$

$\sum x_{ik}x_{jk}$ 当i不等于j时表示两变量值之积的和，例如对于i＝1和j＝2，则

$$\sum_{k=1}^{10} x_{1k}x_{2k} = 35\times 69 + 40\times 74 + 40\times 64 + 42\times 74 + 37\times 72$$

$$+45\times68+43\times78+37\times66+44\times70+42\times65$$

$$=28373$$

$(\sum x_{ik})(\sum x_{jk})$ 当i = j时表示变量值之和的平方，例如对于 i = j = 1，则

$$\left(\sum_{k=1}^{10}x_{1k}\right)\left(\sum_{k=1}^{10}x_{1k}\right)=\left(\sum_{k=1}^{10}x_{1k}\right)^2=405^2=164025$$

$(\sum x_{ik})(\sum x_{jk})$当 i 不等于 j 时表示两变量值之和的积，例如对于 i = 1和 j = 2，则

$$\left(\sum_{k=1}^{10}x_{1k}\right)\left(\sum_{k=1}^{10}x_{2k}\right)=405\times700=283500$$

l_{ij}表示变量值的离均差平方和或两变量值的离均差之积的和，计算公式为

$$l_{ij}=\sum_{k=1}^{n}\left(x_{ik}-\bar{x}_i\right)\left(x_{jk}-\bar{x}_j\right)$$

$$=\sum_{k=1}^{n}x_{ik}x_{jk}-\frac{1}{n}\left(\sum_{k=1}^{n}x_{ik}\right)\left(\sum_{k=1}^{n}x_{jk}\right)$$

$$i,\ j=1,\ 2,\ \cdots,\ m+1 \qquad (3)$$

l_{ij}当i = j时表示变量值的离均差平方和，例如对于i = j = 1，则

$$l_{11}=\sum_{k=1}^{10}\left(x_{1k}-\bar{x}_1\right)\left(x_{1k}-\bar{x}_1\right)$$

$$=\sum_{k=1}^{10}\left(x_{1k}-\bar{x}_1\right)^2$$

$$=\sum_{k=1}^{10}x_{1k}^2-\frac{1}{10}\left(\sum_{k=1}^{10}x_{1k}\right)^2$$

$$= 16501 - \frac{164025}{10}$$

$$= 98.5$$

l_{ij} 当 i 不等于 j 时表示两个变量值的离均差之积的和，例如对于 i = 1和 j = 2，则

$$l_{12} = \sum_{k=1}^{10}\left(x_{1k} - \bar{x}_1\right)\left(x_{2k} - \bar{x}_2\right)$$

$$= \sum_{k=1}^{10} x_{1k}x_{2k} - \frac{1}{10}\left(\sum_{k=1}^{10} x_{1k}\right)\left(\sum_{k=1}^{10} x_{2k}\right)$$

$$= 28373 - \frac{283500}{10}$$

$$= 23$$

r_{ij} 表示 x_i 和x_j 的相关系数，计算公式为

$$r_{ij} = \frac{\sum_{k=1}^{n}\left(x_{ik} - \bar{x}_i\right)\left(x_{jk} - \bar{x}_j\right)}{\sqrt{\sum_{k=1}^{n}\left(x_{ik} - \bar{x}_i\right)^2}\sqrt{\sum_{k=1}^{n}\left(x_{jk} - \bar{x}_j\right)^2}}$$

$$= \frac{l_{ij}}{\sqrt{l_{ii}}\sqrt{l_{jj}}} \qquad i,\ j = 1,2,\cdots,\ m+1 \qquad (4)$$

r_{ij} 当i = j 时等于1，当 i 不等于 j 时，例如对于 i = 1和j = 2，则

$$r_{12} = \frac{l_{12}}{\sqrt{l_{11}}\sqrt{l_{22}}} = \frac{23}{\sqrt{98.5}\sqrt{182}} = 0.1718$$

因为 $l_{ij} = l_{ji}$，所以 $r_{ij} = r_{ji}$。例如，因为 $l_{12} = l_{21}$，所以 $r_{12} = r_{21}$，r_{21}在表2中不予列出，其余类推。

2.建立标准回归系数的正规方程，求标准回归系数。如需

要，也可写出标准回归方程。

把一般回归方程（1）的变量x_i（包括y，即x_{m+1}）作标准化转换

$$x_i' = \frac{x_i - \bar{x}_i}{s_{x_i}}$$

其中$\bar{x}_i$和s_{x_i}是x_i的均数和标准差，于是原回归方程(1) 转换为标准回归方程

$$\hat{y}' = \tilde{b}_1 x_1' + \tilde{b}_2 x_2' + \cdots + \tilde{b}_m x_m' \qquad (5)$$

$\tilde{b}_i$ ($i=1, 2, \cdots, m$) 为标准回归系数。

标准回归系数$\tilde{b}_i$可由下列正规方程求出，其中$r_{i\,(m+1)}$ ($i=1, 2, \cdots, m$) 表示r_{iy}。

$$\begin{cases} r_{11}\tilde{b}_1 + r_{12}\tilde{b}_2 + \cdots + r_{1m}\tilde{b}_m = r_{1\,(m+1)} \\ r_{21}\tilde{b}_1 + r_{22}\tilde{b}_2 + \cdots + r_{2m}\tilde{b}_m = r_{2\,(m+1)} \\ \cdots\cdots\cdots\cdots \\ r_{m1}\tilde{b}_1 + r_{m2}\tilde{b}_2 + \cdots + r_{mm}\tilde{b}_m = r_{m\,(m+1)} \end{cases} \qquad (6)$$

在本例，把表 2 算得的相关系数代入(6)，其正规方程为

$$\begin{cases} 1\tilde{b}_1 + 0.1718\tilde{b}_2 + 0.6409\tilde{b}_3 = 0.6945 \\ 0.1718\tilde{b}_1 + 1\tilde{b}_2 + 0.4522\tilde{b}_3 = 0.5864 \\ 0.6409\tilde{b}_1 + 0.4522\tilde{b}_2 + 1\tilde{b}_3 = 0.7288 \end{cases}$$

解正规方程，即可求得标准回归系数$\tilde{b}_i$值。这里介绍用加减消元法列表逐步解正规方程的方法。

把正规方程两边的相关系数转录成表3。

表3　　相 关 系 数 表

1	0.1718	0.6409	0.6945
0.1718	1	0.4522	0.5864
0.6409	0.4522	1	0.7288

把表 3 的第 1 行、第 2 行和第 3 行分别看作方程①、②和③。

第 1 步：消第一个变量

②－①×0.1718；③－①×0.6409。得

1	0.1718	0.6409	0.6945
0	0.9705	0.3421	0.4671
0	0.3421	0.5892	0.2837

第 2 步：消第二个变量

首先②/0.9705＝②′，得

1	0.1718	0.6409	0.6945
0	1	0.3525	0.4813
0	0.3421	0.5892	0.2837

然后①－②′×0.1718；③－②′×0.3421。得（如果运算熟练，消变量的中间表可省掉，而直接得出下表。）

1	0	0.5803	0.6118
0	1	0.3525	0.4813
0	0	0.4686	0.1190

第 3 步：消第三个变量

首先③/0.4686 = ③′，得

1	0	0.5803	0.6118
0	1	0.3525	0.4813
0	0	1	0.2539

然后① − ③′ × 0.5803；② − ③′ × 0.3525。得

1	0	0	0.4645
0	1	0	0.3918
0	0	1	0.2539

于是最后一列数据即为 $\tilde{b}_i$ 值，即

$\tilde{b}_1 = 0.4645$　　$\tilde{b}_2 = 0.3918$　　$\tilde{b}_3 = 0.2539$

由 (5)式得标准回归方程为

$$\hat{y}' = 0.4645x_1' + 0.3918x_2' + 0.2539x_3'$$

标准回归方程中的标准回归系数，由于消除了单位，可用以比较自变量对因变量影响的大小。$\tilde{b}_i$ 的绝对值愈大，x_i 对 y 的影响愈大。在本例，由于 $\tilde{b}_1$ 最大，$\tilde{b}_2$ 次之，$\tilde{b}_3$ 最小，故体重（x_1）对肺活量的影响最大，胸围（x_2）的影响次之，胸围之呼吸差（x_3）对肺活量的影响最小（这在第二讲逐步回归中将更加得到证实）。

现在把用逐步消元法解正规方程的一般步骤总结如下：把原相关系数记为 $r_{ij}^{(0)}$，第 1 步的记为 $r_{ij}^{(1)}$，即

$r_{11}^{(l)}$	$r_{12}^{(l)}$	…	$r_{1m}^{(l)}$	$r_{1(m+1)}^{(l)}$
$r_{21}^{(l)}$	$r_{22}^{(l)}$	…	$r_{2m}^{(l)}$	$r_{2(m+1)}^{(l)}$
…	…	…	…	…
$r_{m1}^{(l)}$	$r_{m2}^{(l)}$	…	$r_{mm}^{(l)}$	$r_{m(m+1)}^{(l)}$

则第l+1步消去第k个变量的计算公式为

$$r_{ij}^{(l+1)}=\begin{cases} r_{kj}^{(l)}/r_{kk}^{(l)} & (i=k\text{即第k行}) \\ r_{ij}^{(l)}-r_{ik}^{(l)}r_{kj}^{(l)}/r_{kk}^{(l)} & (i\neq k\text{即其他行}) \end{cases} \tag{7}$$

如要消去第二个变量，则

$$r_{ij}^{(l+1)}=\begin{cases} r_{2j}^{(l)}/r_{22}^{(l)} & (i=2) \\ r_{ij}^{(l)}-r_{i2}^{(l)}r_{2j}^{(1)}/r_{22}^{(l)} & (i\neq 2) \end{cases}$$

3.求一般回归系数和常数项，从而建立回归方程

由标准回归系数$\tilde{b}_i(i=1,2,\cdots,m)$和离均差平方和$l_{(m+1)(m+1)}$（即$l_{yy}$）、$l_{ii}$（i=1，2，…,m）求一般回归系数的公式为

$$b_i=\tilde{b}_i\sqrt{\frac{l_{(m+1)(m+1)}}{l_{ii}}} \qquad i=1,2,\cdots,m \tag{8}$$

在本例m=3，由求得的$\tilde{b}_i$（i=1，2，3）值和表2中的l_{44}、l_{ii}（i=1，2，3）值，可得

$$b_1=\tilde{b}_1\sqrt{\frac{l_{44}}{l_{11}}}=0.4645\sqrt{\frac{1695250}{98.5}}=60.94$$

$$b_2=\tilde{b}_2\sqrt{\frac{l_{44}}{l_{22}}}=0.3918\sqrt{\frac{1695250}{182}}=37.81$$

$$b_3=\tilde{b}_3\sqrt{\frac{l_{44}}{l_{33}}}=0.2539\sqrt{\frac{1695250}{10.641}}=101.34$$

由回归系数b_i（i＝1，2，…，m）和均数$\bar{x}_{m+1}$（即$\bar{y}$）、$\bar{x}_i$（i＝1，2…，m）求回归方程中常数项的公式为

$$a = \bar{x}_{m+1} - \sum_{i=1}^{m} b_i \bar{x}_i \qquad (9)$$

在本例m＝3，由求得的b_i（i＝1，2，3）值和表1中的$\bar{x}_i$（i＝1，2，3，4）值，可得

$$a = \bar{x}_4 - \sum_{i=1}^{3} b_i \bar{x}_i$$

$$= 2315 - (60.94 \times 40.5 + 37.81 \times 70.0 + 101.34 \times 2.33)$$

$$= -3035.89$$

把求得的b_i值和a值代入（1），得回归方程为

$$\hat{y} = -3035.89 + 60.94x_1 + 37.81x_2 + 101.34x_3$$

注意的是，由于一般回归方程中的回归系数有单位，所以不能用回归系数的绝对值大小来比较自变量对因变量影响的大小。在本例不能由于b_3的绝对值最大，就认为胸围之呼吸差（x_3）对肺活量的影响最大。实际上，由前面计算的标准回归系数的比较，已知胸围之呼吸差对肺活量的影响最小。

（二）多元线性回归方程的方差分析

对拟合的多元线性回归方程的好坏进行检验或评价用方差分析。总平方和为$\sum(y_i - \bar{y})^2 = l_{(m+1)(m+1)}$（即$l_{yy}$），设回归平方和$\sum(\hat{y}_i - \bar{y})^2$为U；剩余平方和$\sum(y_i - \hat{y}_i)^2$为Q，则多元（m元）线性回归方程的方差分析如表4所示。

表4　　m元线性回归方程的方差分析表

来源	平方和	自由度	均方	F
总和	$l_{(m+1)(m+1)}$	$n-1$		
回归	$U=\sum_{i=1}^{m} b_i l_{i(m+1)}$	m	U/m	$F=\dfrac{U/m}{Q/n-m-1}$
剩余	$Q=l_{(m+1)(m+1)}-U$	$n-m-1$	$Q/(n-m-1)$	

在本例$m=3$，由表2中的l_{44}（即l_{yy}）、l_{i4}（$i=1, 2, 3$）值和求得的b_i（$i=1, 2, 3$）值，可得

$$l_{44}=1695250$$

$$U=\sum_{i=1}^{3} b_i l_{i4}$$

$$=60.94\times 8975+37.81\times 10300+101.34\times 3095.5$$

$$=1250077$$

$$Q=l_{44}-U$$

$$=1695250-1250077$$

$$=445173$$

方差分析如表5所示。

表5　　方差分析表

来源	平方和	自由度	均方	F
总和	1695250	9		
回归	1250077	3	416692.33	5.62
剩余	445173	6	74195.50	

据自由度$\nu_1=3$和$\nu_2=6$查F值表，$F_{0.05(3,6)}=4.76$，现