高等医药院校教材

孙尚拱　编著

实用多变量统计方法与计算程序

北京医科大学
中国协和医科大学　联合出版社

实用多变量统计方法
与
计算程序

孙尚拱　编著

北京医科大学
中国协和医科大学联合出版社

实用多变量统计方法与计算程序

编　　著　　孙尚拱

责任编辑　　暴海燕

*　　*　　*

北京医科大学
中国协和医科大学联合出版社出版

（社址：北京医科大学院内）

新华书店总店科技发行所发行　各地新华书店经销

北京密云华都印刷厂印刷

*　　*

开本：787×1092 1/16　印张：21.125　字数：498 千字

1990年12月第一版 1990年12月第一次印刷　印数：1—4000册

ISBN 7-81034- 037-9/R·38　　定价：4.20元

前　言

随着多变量统计在国际上的日益深入发展与普遍应用，国内的自然科学、社会科学及经济领域中的统计工作者也日益从多变量角度去收集和分析统计资料。本书是作者多年来从事多变量统计的应用、教学及科研的结果；它是为应用多变量统计方法的实际工作者，研究生及有关的高年级大学生中的非数理统计专业者而写。

本书的主要特点：

(1) 实用性。除第二章外，其它几乎所有公式都不给证明，强调模型、概念、方法及结果的解释。方法的选材，视其在实用上的重要性与普遍性；对有关方法及概念的介绍更多的是从实用角度而不是从数学出发。

(2) 把国内外很多的研究新成果，特别是国内（当然也包括作者本人）的成果反映在书内。作者希望能对来之不易的数据尽可能地分析得细致一些，信息提取得更多一些。因此很多旧有的标题，比如判别分析，典则相关分析等等，都充实进大量的新内容。

(3) 本书附有10个用*Fortran*语言编写的计算程序。这些程序中很多是作者所首创，其它则是在旧方法中或多或少充实进新内容的新编程序。

(4) 本书不是单纯的介绍统计方法，它也是多年解决实际问题及教学工作中的体会、看法与见解，也包含着其他同行们的很多经验在内。

学习本书应具备初步的统计学及简单的微积分知识，有关线性代数方面的基本知识见附录2。

魏宗舒先生曾对本书稿作过非常认真细致的审查并为本书的出版作了很多努力，特此致谢。

一些常用记号定义如下：

$E(y)$ 表示y的平均值（或称期望值），

$\mathrm{var}(y)$ 表示y的方差，

$\mathrm{Corr}(x, y)$ 表示x与y的相关系数，

$\exp(x) = e^x$

$\triangleq$ 表示定义或“记为”

本书一定会有很多不妥之处，请批评指正。

孙尚拱

1988年12月于

北京

目　　录

第一章　回归分析(Regression Analysis)

§1·1　基本概念

自然现象及社会活动中，变量之间不一定都存在有确定性的函数关系。但往往存在有一定的统计相关性。回归分析就是运用数学手段，在大量统计资料中找出这种相关性，并作定量分析。

回归分析中总有两类变量；一类是因变量，也称准则变量或反应变量；一类是自变量，也称回归变量、预测变量、独立变量或设计变量（当后者可以人为控制时）。当因变量只有一个（记为y），而自变量有多个时（记$x_1, x_2, \cdots, x_m$），则y与（$x_1, x_2, \cdots, x_m$）之间的回归，称之为多重回归。如因变量也有多个时，记为$y_1, y_2, \cdots, y_q$，则($y_1, y_2 \cdots, y_q$)与（$x_1, x_2, \cdots, x_m$)之间的回归关系，称之为多元（或多变量）回归，但这种多个因变量的回归问题也常化为多个多重回归法处理。一般习惯上也把多重回归称为多元回归。

表　1·1·1

镇序	Z_1 外来人口（人）	Z_2 常住人口（人）	W 工农业总产值（万元）
1.	28070	42208	4464.34
2.	7382	11479	929.89
3.	4320	23961	4338.00
4.	4161	15655	2687.25
5.	16435	17408	1860.21
6.	12381	7356	886.75
7.	12996	10052	1313.86
8.	11024	15806	2153.95
9.	19040	9739	3553.81
10.	33767	12175	6721.16
11.	20879	10217	3648.39
12.	29669	23718	3461.89
13.	10687	8148	2428.72
14.	8419	8373	1388.73
15.	4199	8148	300.42
16.	2903	6595	527.83
17.	908	6286	113.99
18.	4169	5580	245.73

例1·1·1，近几年来由于开放政策，深圳经济特区中，外来人口大幅度增加，为了考察特区中外来人口对本地经济发展的贡献，深圳特区的统计局收集宝安县在1987年末所属的18个镇的人口与工农业产值数据，见表1·1·1。

一般说来，人口的多少对经济的发展是有影响的。在此例中要考察人口对经济的影响，则把表1·1·1中工农业总产值当作因变量(W)，外地及本地人口看作两个自变量(Z_1, Z_2,)。

回归分析主要解决下面几个问题。

1. 确定因变量，(上例中工农业总产值)与各个自变量，(本地人口与外地人口数量）之间是否存在有相关关系。如有，指出相关性有多大，比如例1·1·1中工农业总产值(W）的变动信息中，有多大的一部分是被外地人口(Z_1)及本地人口(Z_2)的变动所决定的?进一步应找出它们之间相互关系的定量表示式。指出如何用自变量(Z_1, Z_2)去预测因变量(W）的变化及给出预报精度。

2. 找出每个自变量对因变量作用的大小及方向性。它与一般单变量统计分析不同点是：考察Z_1（外地人口）对因变量W（工农业总产值）的影响大小时是把其它所有自变量

(表1·1·1中仅Z_2)都固定不变时，考察Z_1对W的作用大小，这种考察法是与单元统计分析中考察Z_1与W关系时，把Z_2当作似乎不存在一样（即Z_2可以自由变动）的考察法是根本不同的。因此，当一个多变量问题，如把它们简单地拆成多个“单变量统计”法去考察，其结果也常与多变量统计分析的结果不一致。一般说来，当多个自变量之间存在有相关性时，把本质上是多变量相关问题拆成多个单变量去处理是不合理的。

3．对于前述问题的结论（比如W与(Z_1,Z_2)的相关性）可以给出概率值。即可以求出有多大的把握说(Z_1,Z_2)影响W，如用(Z_1,Z_2)预报W，有多大把握说W的理论值在给定的范围中？对于第3个问题的解答，往往对资料有一定的要求，比如正态分布，样本之间彼此独立同分布等等。

在实际问题中，实际工作者不一定对上述三个问题都有兴趣，比如在因素分析问题中，往往对问题1不感兴趣。还应指出，对于本质上是多变量相关的数据，用多变量统计分析，在理论上虽然比单变量统计法科学及合理，但这并不是说，该批数据经任一种多变量分析法后，其结果就都是“正确”或“合理”的。实际情形及大量实例表明，情况远非如此，多变量统计分析中结果的不合实际情形到处可见。产生错误结果的原因是多方面的。因此，实际工作者千万不要以为经过“多变量统计分析”后，其结果就一定是正确的。客观实际的检验及专业知识仍是最根本的标准。以后我们将详细讨论。

一、指标数量化

回归分析是寻找以数量表示的自变量与因变量间的统计规律。因此，一切变量（实际工作中常称为指标、因素等）都必须给以数量化。比如表1·1·1中每个变量本身就已经是数量了，不存在数量化问题。但其它实际问题中情况并未必都如此。比如寻找人的血压与性别的关系，男性与女性就不是数量指标，因此必须给以数量。最简单也是最常用的数量化方法是0—1法。具体做法如下：

如x_1表示性别，x_2表示“病人”与“非病人”，则可令

$$x_1=\begin{cases}1 & \text{男性}\\ 0 & \text{女性}\end{cases},\qquad x_2=\begin{cases}1 & \text{病人}\\ 0 & \text{非病人}\end{cases},$$

如果某个指标比如文化程度，本来分成四级：“文盲，小学，中学，大学及大学以上”，则可以把这个指标用4个自变量或3个自变量表示它。比如想用4个自变量表示上述的文化程度，则可令：

$$x_1=\begin{cases}1 & \text{文盲}\\ 0 & \text{其它(指非文盲者)}\end{cases},\quad x_2=\begin{cases}1 & \text{小学}\\ 0 & \text{其它}\end{cases},\quad x_3=\begin{cases}1 & \text{中学}\\ 0 & \text{其它}\end{cases},$$

$$x_4=\begin{cases}1 & \text{大学及大学以上}\\ 0 & \text{其它}\end{cases}\qquad(1\cdot1\cdot1)$$

这样在数据登记时，原来的文化程度只占用一列（或一行），而用回归分析处理时，该指标就必须变成4列（或4行），即用4个自变量代表文化程度。这种数量化法的优点是：把文化程度考察得很细致。比如考察文化程度$(x_1,\cdots,x_4)$与血压(y)的关系时，上法可以把每个级别的文化程度与血压的关系找出来。但文化程度的上述数量化方法有两个缺点：

(1) 上述（1·1·1)表示法不能用于做后面要叙述的后退法线性回归分析，因为该法要求自变量之间不能有完全的线性相关情形，而(1·1·1)表示法中显然有$x_1+x_2+x_3+x_4=1$，即其中一个文化等级可以用另外3个文化等级推算出来，比如$x_1=1-(x_2+x_3+x_4)$。因此，

我们可以只用3个自变量（比如$x_2, x_3, x_4,$）就可以表示上述的文化程度。这时数据登记中，文化程度所在的列（或行）即变成3列（或3行）。被取消了的等级x_1的作用如何估计？可以用下述方法估计：比如记z为另外的非文化程度的自变量，如果我们计算所得的回归方程（记$\hat{y}$为y的估计值）为：

$$\hat{y} = a + b_2x_2 + b_3x_3 + b_4x_4 + cz$$

这时$a + b_2$就是x_2（小学）对于y的作用大小；而常数a就是x_1（文盲）对于y的作用大小。所以四个文化程度等级（x_1……x_4）对于y的作用大小就是$(a,\ a+b_2,\ a+b_3,\ a+b_4)$，相互比较此四个数及符号，就可以得出4个文化程度级别对于y作用的大小及方向（方向性视符号而定）。

（2）在某些实际问题中，并不关心每一个文化程度等级对于y的作用大小。而只想用一个自变量x（而不是用3个或4个自变量）去表示文化程度。这时如用（1·1·1）式就不合要求（计算量也大了些），对于这种要求我们就可以用有序的四个数值（比如，0,1,2,3）去表示上述四个级别的文化程度："0"表示文盲"1"表示小学等等。这种表示法的最大优点是计算简单，缺点是解的不定性。因为我们没有理由不可以应用(10,12,18,30,)的4个数去表示4个级别的文化程度，但该结果与用(0,1,2,3)法结果就不大会相同。而在0—1法中，"0,1"的两个数是可以随便的。比如性别例子中如令

$$x_1 = \begin{cases} 10 & \text{男性} \\ 5 & \text{女性} \end{cases}$$

与用(1,0)表示男或女，可以证明是等价的。正因为如此，在数量化方法中，0—1法是使用得最多也是最方便的方法。

二、回归模型

记$\boldsymbol{x} = (x_1, x_2, \cdots\cdots x_m)'$，为$m$个自变量组成的自变量向量，$y$为因变量。如我们要找一个函数$f(\boldsymbol{x}) = f(x_1, x_2 \cdots x_m)$以使$f(\boldsymbol{x})$与$y$的误差的平方在平均意义下为最小，数学记号为：

$$E(f(\boldsymbol{x}) - y)^2 = \min \qquad (1\cdot1\cdot2)$$

则可以证明这个$f(\boldsymbol{x})$必是y的条件均值$E(y/\boldsymbol{x})$，即如果我们可以求出x值下y的均值$E(y/\boldsymbol{x})$，则这$E(y/\boldsymbol{x})$就是(1·1·2)最小二乘方意义下的理想回归函数$f(\boldsymbol{x})$。但实际上我们往往不知道$E(y/\boldsymbol{x})$，比如在例1·1·1中，对给定的每一组数据(Z_1, Z_2)值后，工农业生产总值W的均值是什么？不知道。一般只能利用专业知识或采用多项式逼近法求它。而在例1·1·1中我们可以利用经济学中有名的柯布－道格拉斯生产函数去取代条件均值，即我们采用下面的数学模型

$$\hat{W} = K \cdot Z_1^{b_1} \cdot Z_2^{b_2} \qquad (1\cdot1\cdot3)$$

拟合表1·1·1的数据，其中K为比例常数，它包含未控制的因素。W表示用W中可用(Z_1, Z_2)估计出来的部分。如令

$$y = \ln(W),\quad x_1 = \ln(Z_1),\quad x_2 = \ln(Z_2),$$

则(1·1·3)等价于

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 \qquad (1\cdot1\cdot4)$$

其中b_0是（1·1·3）中比例常数引出的结果，$\hat{y} = \ln(\hat{W})$是$\ln(W)$的估计值，b是要用样本估计的参数。可见（1·1·3）模型虽然不是一个线性函数，但它可以变成一个线性模型。由

于于线性模型统计上处理起来方便，且已有一套相当成熟的理论，所以实际工作中遇到的一般性模型，总是想方设法把它变成线性模型。因此，一般计算机软件中的回归模型也几乎都是线性模型。

下面介绍一般性的线性回归模型。

假设有m个自变量$(x_1, x_2, \cdots\cdots, x_m)$，它与因变量$y$存在有如下的线性关系：

$$y = \beta_0 + \beta_1 x_1 + \cdots\cdots + \beta_m x_m + \varepsilon \qquad (1\cdot1\cdot5)$$

其中$\beta_0 + \beta_1 x_1 + \cdots\cdots + \beta_m x_m$称为$y$的线性回归部分。$\varepsilon$表示$y$与线性回归部分间的误差，一般称残差。(1·1·5) 是y与$(x_1, x_2, \cdots\cdots, x_m)'$间的数学模型，其中$(\beta_1, \beta_2 \cdots\cdots \beta_m)' = \beta$为理论回归系数向量，$\beta_0$为常数项，它们都称为模型中的参数，是需要用样本去估计的。

下面考察 (1·1·5) 中参数的意义及所选模型正确性的简单分析法。

由 (1·1·5)可见，当固定其它变量时，x_1增加一个单位y的增加值为b_1；$b_2, \cdots\cdots b_m$的意义类似。因此称b_i为x_i的偏回归系数。b_i实际上就是y关于x_i的偏导数，也称b_i是y在x_i上的变化率。由 (1·1·5) 可见，当仅允许x_1变化而其它$(x_2, \cdots\cdots, x_m)$变量都固定不变时，则x_1与y呈直线关系。因此，如果由专业知识及由样本中看出，x_1与y不成直线关系，这时用(1·1·5)模型去代表样本就不合适了。

另外从模型 (1·1·5) 还可以分析出，y在x_1上的变化率b_1是与自变量x_2(及其它自变量) 无关的。即x_2及其它自变量无论取什么数值，都不会影响y在x_1上的变化率。比如在细菌繁殖问题中，如记y为一定容器中细菌总量。x_1, x_2分别表示容器中的温度及营养物数量。采用 (1·1·5) 模型就意味着，营养物质x_2不管是多或少，容器的温度不论是多少度，则容器内温度每升高一度时，细菌的总数永远是增加b_1个。由于细菌繁殖总有一个适宜温度，在不适合很好繁殖的温度与在适宜温度时，每增加一度，y的增加值一般不会相同的。因此，对这样的细菌繁殖问题，如简单的采用 (1·1·5) 模型是不合理的。

当使用很不合理的数学模型时很难想象回归分析最后的结果能正确地反映客观实际。而在实际工作者中间，滥用回归分析的情形相当普遍：找来一个回归分析软件包，对自己的数据也不加任何分析就套用任一个软件，对计算出来的结果，盲目的加以相信，有时该结果是明显地反常，也牵强附会的去作解释，这都是不应该的。可是对一批数量不很大的数据，若没有相当的专业知识，要正确地确定模型也确非易事。常用的两个简单方法是：(1) 法，把样本中每个自变量单独抽出来与因变量y作图。从图上大体看看，如每个自变量与因变量基本成直线，则采用 (1·1·5) 或许是好的。比如对例(1·1·1)，把$x_1 = \ln Z_1$与$y = \ln W$作图，$x_2 = \ln Z_2$与$\ln y$作图，可得图 (1·1·1)。从图1·1·1的(a)及(b)可见，两图基本上还是直线形状，因此采用模型 (1·1·3)还是可以的。(2) 法，是先不对(x_1, x_2, y)作任何变换，做一次线性回归，再对某些变量 (特别是经统计检验不显著的变量) 作人为地变换 (见后面 § 1·7)，再看看衡量线性回归的指标是否有改善，以决定该批数据应取什么样的模型。

三、参数的估计

(1·1·5) 中的模型为理论模型，参数$\beta_0, \beta_1 \cdots \beta_m$为理论值。要把理论值准确地找出来是不大可能的，一般都是抽取一批数据 (样本) 去估计$\beta_0, \beta_1, \cdots, \beta_m$。记参数的样本估计值为$b_0, b_1, \cdots b_m$。一般记$\boldsymbol{b} = (b_1, b_2, \cdots, b_m)'$为参数向量，$b_0$为常数项。有的统计书中也把$b_0$作为$\boldsymbol{b}$向量的一个分量。记

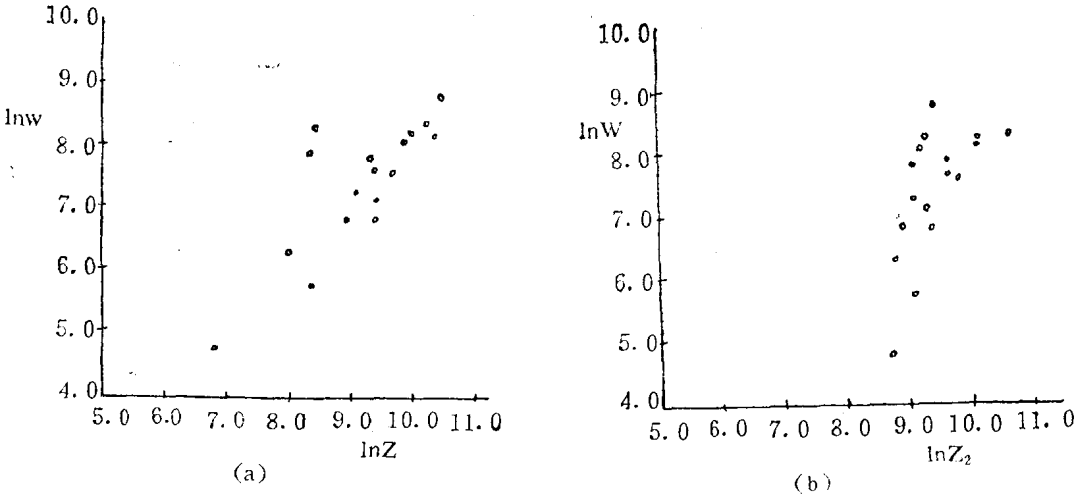

图1·1·1　表1·1·1的数据

$$\hat{y} = b_0 + b_1 x_1 + \cdots, + b_m x_m \qquad (1\cdot1\cdot6)$$

$\hat{y}$称为（1·1·5）中y的估计值，也称回归部分。

设对$(x_1, x_2, \cdots, x_m, y)$抽取几组样本，记第$i$组样本值为$(x_{i1}, x_{i2}, \cdots, x_{im}, y_i)$，$i = 1, 2, \cdots, n$.

记

$$X = \begin{pmatrix} x_{11}, & x_{12}, & \cdots, & x_{1m} \\ x_{21}, & x_{22}, & \cdots, & x_{2m} \\ \cdots & & \cdots & \\ x_{n1}, & x_{n2}, & \cdots, & x_{nm} \end{pmatrix}, \qquad Y = \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix}$$

则X，Y表示全部数据构成的数据矩阵（简称数据阵）。记$\bar{x}_j, \bar{y}$为x_j及y的洋本均数，其计算公式为：

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \bar{x} = (\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_m)'$$

$\bar{x}$为$(x_1, x_2, \cdots, x_m)'$的均数向量。求（1·1·6）中参数值的准则一般是用最小二乘方准则：它求参数值以使下式为最小

$$Q = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \qquad \hat{y}_i = b_0 + b_1 x_{i1} + \cdots + b_m x_{im} \qquad (1\cdot1\cdot7)$$

对于不同的b_0，b_1，$\cdots$，b_m值，(1·1·7)中可求出不同的Q值，能使Q达到最小值的参数值，称之为参数β_0，β_1，$\cdots$，β_m的最小二乘方估计值。

记：

$$l_{ij} = \sum_{t=1}^{n}(x_{ti} - \bar{x}_i)^2 = \sum_{t=1}^{n} x_{ti}x_{tj} - n\bar{x}_i\bar{x}_j, \quad i, j = 1, 2, \cdots, m$$

$$l_{iy} = \sum_{t=1}^{n}(x_{ti} - \bar{x}_i)(y_t - \bar{y}) = \sum_{t=1}^{n}(x_{ti} - \bar{x})y_t = \sum_{t=1}^{n} x_{ti}y_t - n\bar{x}_i\bar{y}$$

$$l_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}y_i^2-n\bar{y}^2,\quad s_y=\sqrt{\frac{l_{yy}}{n-1}},\quad s_i=\sqrt{\frac{l_{ii}}{n-1}}\qquad(1\cdot1\cdot8)$$

$$L_{xx}=\begin{pmatrix}l_{11}, & l_{12}, & \cdots, & l_{1m}\\ l_{21}, & l_{22}, & \cdots, & l_{2m}\\ \cdots & & & \cdots\\ l_{m1}, & l_{m2}, & \cdots, & l_{mm}\end{pmatrix},\quad l_{xy}=\begin{pmatrix}l_{1y}\\ l_{2y}\\ \vdots\\ l_{my}\end{pmatrix},\quad b=\begin{pmatrix}b_1\\ b_2\\ \vdots\\ b_m\end{pmatrix}\qquad(1\cdot1\cdot9)$$

L_{xx}称为$(x_1, x_2, \cdots, x_m)$的离差阵，l_{xy}为x向量与y的离差乘积阵。s_y称为y的标准差，s_i为x_i的标准差。利用样本，可以求出$\{l_{ij}\}$，$\{l_{iy}\}$及s_y，s_i。用数学方法中的极值方法，可以证明使（1·1·7）达最小值的b_i满足下面线性方程组：

$$\begin{cases}l_{11}b_1+l_{12}b_2+\cdots+l_{1m}b_m=l_{1y}\\ l_{21}b_1+l_{22}b_2+\cdots+l_{2m}b_m=l_{2y}\\ \cdots\qquad\qquad\cdots\\ l_{m1}b_1+l_{m2}b_2+\cdots+l_{mm}b_m=l_{my}\end{cases}\qquad(1\cdot1\cdot10)$$

求（1·1·10）线性方程组的解，即可得$\boldsymbol{b}$向量，如用矩阵表示，则为

$$\boldsymbol{b}=L_{xx}^{-1}\cdot l_{xy},\qquad b_0=\bar{y}-b_1\bar{x}_1-\cdots-b_m\bar{x}_m\qquad(1\cdot1\cdot11)$$

求出$\boldsymbol{b}$，b_0后代回（1·1·7），可得出使（1·1·7）中Q的最小值为

$$Q=l_{yy}-\boldsymbol{b}'l_{xy}=l_{yy}-\boldsymbol{b}'L\boldsymbol{b}\qquad(1\cdot1\cdot12)$$

Q称为回归方程（1·1·6）的残差平方和或失拟平方和。记

$$s=\sqrt{\frac{Q}{n-m-1}}\qquad(1\cdot1\cdot13)$$

把$\boldsymbol{b}$向量及b_0代回（1·1·6），即得出每一个样本点上对y_i的估计值$\hat{y}_i$，而s就描述了用$\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n$估计$y_1, y_2, \cdots, y_n$的一个平均性精确度指标，$s$称之为残差标准差，有的书也称之为回归标准误。

对表1·1·1数据，对每一个数据取对数后，可算得（$n=18$，$m=2$）；

$$\overline{\boldsymbol{X}}=\begin{pmatrix}9.1149\\ 9.3556\end{pmatrix},\quad L=\begin{pmatrix}15.2955, & 4.3345\\ 4.3345, & 4.9190\end{pmatrix},\quad l_{xy}=\begin{pmatrix}14.2705\\ 7.1380\end{pmatrix},\quad l_{yy}=21.6610,\qquad(1\cdot1\cdot14)$$

于是得

$$L^{-1}=\begin{pmatrix}0.08714, & -0.07678\\ -0.07678, & 0.27095\end{pmatrix},\quad \boldsymbol{b}=\begin{pmatrix}0.6954\\ 0.8383\end{pmatrix},\quad b_0=-6.8890$$

于是得y的回归方程为

$$\hat{y}=-6.8890+0.6954x_1+0.8383x_2\qquad(1\cdot1\cdot15)$$

其回归标准误$s=0.6193$

下面我们介绍回归系数公式的另外两种经常遇见的表示法：

（1）数据中心化的表示法

记$(\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_m, \bar{y})$为$(x_1, x_2, \cdots, x_m, y)$的样本均数，用$x_{ti}-\bar{x}_i$代替$x_{ti}$，用$y_t-\bar{y}$代替$y_t$，这样改造后的数据称为中心化了样本数据，显然，中心化的数据，每个变量的样本均数为零。记X为中心化了的数据阵，$\mathbf{Y}$为因变量中心化了的数据向量。即：

$$\mathbf{X}=\begin{pmatrix} x_{11}-\bar{x}_1, & x_{12}-\bar{x}_2, & \cdots, & x_{1m}-\bar{x}_m \\ x_{21}-\bar{x}_1, & x_{22}-\bar{x}_2, & \cdots, & x_{2m}-\bar{x}_m \\ \cdots & & & \cdots \\ x_{n1}-\bar{x}_1, & x_{n2}-\bar{x}_2, & \cdots, & x_{nm}-\bar{x}_m \end{pmatrix}, \quad \mathbf{Y}=\begin{pmatrix} y_1-\bar{y} \\ y_2-\bar{y} \\ \cdots \\ y_n-\bar{y} \end{pmatrix}$$

这时（1·1·9）中L_{xx}，L_{xy}，b即为：

$$\mathbf{L}_{xx}=\mathrm{X}'\mathrm{X},\ \mathbf{L}_{xy}=\mathrm{X}'\mathbf{Y},\ b=(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{Y} \tag{1·1·16}$$

常数项b_0仍用（1·1·11）中b_0公式。这时$\mathbf{Y}$的拟合值向量，记为$\hat{\mathbf{Y}}$，即为：

$$\hat{\mathbf{Y}}=\mathrm{X}\boldsymbol{b}=X(X'X)^{-1}X'Y=HY \tag{1·1·17}$$

其中，

$$\mathrm{H}=\mathrm{X}(X'X)^{-1}X' \tag{1·1·18}$$

（2）同时计算常数项的表示法

把常数项当作永远取值为1的自变量，于是数据阵（不要求中心化）X为：

$$\boldsymbol{X}=\begin{pmatrix} 1 & x_{11}, & x_{12}, & \cdots, & x_{1m} \\ 1 & x_{21}, & x_{22}, & \cdots, & x_{2m} \\ & \cdots & & & \cdots \\ 1 & x_{n1}, & x_{n2}, & \cdots, & x_{nm} \end{pmatrix} \tag{1·1·19}$$

$\mathbf{Y}$是$(y_1,\ y_2,\ \cdots,\ y_n)'$。这时回归系数公式为

$\boldsymbol{b}=(b_0,\ b_1,\ \cdots,\ b_m)'$，则

$$\boldsymbol{b}=(X'X)^{-1}X'Y,\ \hat{Y}=HY \tag{1·1·20}$$

H仍如（1·1·17），定义，但X用（1·1·19）式，Y未中心化。

四、决定系数，复相关系数

不管用何种方法表示回归系数，回归系数及下面一切有关量的数值不会改变。

下面我们考虑描述回归方程好坏的两个基本指标，我们称

$$U=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \tag{1·1·21}$$

为回归平方和。可以证明有等式

$$U=\sum_{i=1}^{m}l_{iy}b_i=L_{xy}'b \tag{1·1·22}$$

而且：

$$l_{yy}=U+Q \tag{1·1·23}$$

此式称之为离差平方和的分解公式。其中Q为残差平方和。显然，一个拟合得好的模型应使Q最小，也即U应最大（l_{yy}是被数据决定了的不变数），但U及Q均受量纲的影响，为了不受单位量纲的影响，我们称

$$D=U/l_{yy} \tag{1·1·24}$$

为回归方程（1·1·6）的决定系数，它表示在y的变动总信息（用l_{yy}表示）中，可以被回归方程拟合的百分比。如例1·1·1中，可算得：

$$U=14.2705\times0.6954+7.1380\times0.8383=15.9075$$

$$\therefore \quad D = U/l_{yy} = \frac{15.9075}{21.6610} = 0.7344 = 73.44\%$$

也就是说，在y的离差平方和中73.44%的变异可由（1·1·6）表示。显然如决定系数的值愈近似于1愈说明模型对资料的拟合能力愈好。除非随机干扰很少，一般说来决定系数73.44%是很高的了。反映模型对资料拟合程度的另一个指标为复相关系数ρ。ρ的严格定义是y与它的估计量$\hat{y}$之间的相关系数，此数可以证明永远是非负数。若用样本估计ρ（记为R），则R用下式计算：

$$R = \left(\frac{U}{l_{yy}}\right)^{1/2} = D^{1/2} \tag{1·1·25}$$

例1·1·1中$R = 0.8570$，显然R愈接近于1则说明模型拟合程度愈好。永远有$0 \leqslant R \leqslant 1$，由统计理论可以证明用（1·1·25）计算的样本复相关系数并不是理论ρ值的无偏性估计，而且总是过高地估计了ρ。因此一些近似的无偏性估计被发展起来，一个常用公式，称之为修正公式：

$$R_c^2 = R^2 - \frac{(m-1)(1-R^2)}{n-m} \tag{1·1·26}$$

在例1·1·1中，有 $R_c^2 = 0.8570^2 - \dfrac{(1-0.8570^2)}{18-2} = 0.71785$ 所以近似无偏的复相关系数ρ的估计值为

$$R_c = 0.8473$$

附带说明一下，使残差平方和Q为最小的参数解b及b_0也就是使决定系数D为最大的参数解。

五、标准化回归系数

建立在（1·1·9）及由（1·1·11）确定的回归系数是有量纲的，不同变量由于所使用的单位不同而使得它们的回归系数无法进行比较。为了克服这一缺点，引进标准化回归系数的概念。称下面的数据为标准化数据：

$$\begin{aligned} \tilde{x}_{ti} &= (x_{ti} - \bar{x}_i)/s_i \quad , \quad \text{其中 } s_i = \sqrt{l_{ii}/(n-1)} \\ \tilde{y}_t &= (y_t - \bar{y})/s_y \quad , \quad \text{其中 } s_y = \sqrt{l_{yy}/(n-1)} \end{aligned} \tag{1·1·27}$$

$t = 1, 2, \cdots, n \quad i = 1, 2, m$

其中s_i为x_i的样本标准差，s_y为y的样本标准差。经（1·1·27）处理后的数据，可以发现，每一个变量的样本均数为0，样本标准差为1。称下面的量

$$r_{ij} = \frac{l_{ij}}{\sqrt{l_{ii}}\sqrt{l_{jj}}} \quad r_{iy} = \frac{l_{iy}}{\sqrt{l_{ii}}\sqrt{l_{yy}}} \quad i, j = 1, 2, \cdots, m \tag{1·1·28}$$

为x_i与x_j的样本相关系数及x_i与y的样本相关系数。称

$$R_{xx} = \begin{pmatrix} r_{11}, & r_{12}, & \cdots, & r_{1m} \\ r_{21}, & r_{22}, & \cdots, & r_{2m} \\ \cdots & & & \cdots \\ r_{m1}, & r_{m2}, & \cdots, & r_{mm} \end{pmatrix}, \qquad r_y = \begin{pmatrix} r_{1y} \\ r_{2y} \\ \vdots \\ r_{my} \end{pmatrix} \tag{1·1·29}$$

为x的样本相关阵及x与y的样本相关系数向量。在公式（1·1·9）中如用$\tilde{x}_{ti}$代替x_{ti}，且改

记l_{ij}为 $\tilde{l}_{ij}$，l_{iy}为$\tilde{l}_{iy}$，则可以发现有

$$\tilde{l}_{ij}=r_{ij}(n-1)\ ,\quad \tilde{l}_{iy}=r_{iy}(n-1)$$

把$\tilde{l}_{ij}$及$\tilde{l}_{iy}$代入（1·1·11）中$\boldsymbol{b}$公式（改记为$\tilde{b}$）显然有

$$\tilde{\boldsymbol{b}}=R_{xx}^{-1}\cdot\mathbf{r}_y \qquad (1\cdot1\cdot30)$$

这里的回归系数$\tilde{\boldsymbol{b}}\triangleq(\tilde{b}_1,\ \tilde{b}_2,\cdots,\ \tilde{b}_m)'$显然是没有量纲的，它称之为回归方程的标准化回归系数。可以找出标准化回归系数$\tilde{\boldsymbol{b}}=(\tilde{b}_1,\ \tilde{b}_2,\ \cdots,\ \tilde{b}_m)$与一般回归系数 $\boldsymbol{b}=(b_1,\ b_2,\ \cdots,\ b_m)'$之间的关系式：

$$b_k=\frac{s_y}{s_k}\tilde{b}_k,\qquad \tilde{b}_k=\frac{s_k}{s_y}b_k\ ,\quad k=1,2,\cdots,m \qquad (1\cdot1\cdot31)$$

由于例1·1·1中已算得$b_1=0.6954, b_2=0.8383$，利用$s_1=(15.2955/(18-1))^{1/2}=0.9485$，$s_2=0.5379$，$s_y=1.1288$，代入（1·1·31）得$\tilde{b}_1=0.5844$，$\tilde{b}_2=0.3995$。比较可见，似乎可以认为“外来人口”$x_1$的变化比“本地人口”$x_2$的变化对于宝安县的工农业总产值的变化有更大的影响。但在本例中对$(\tilde{b}_1,\ \tilde{b}_2)$的解释似乎与对$(b_1,\ b_2)$的解释不一致。由于此例恰可以比较$b_1$，$b_2$的大小，其结论是：本地人口$x_2$每增加一个人时工农业总产值的增加比外地人口$x_1$每增加一个人时的产值为大。但在比较$(\tilde{b}_1,\ \tilde{b}_2)$时结论似乎相反，这个相反的结论来源于$x_1$与$x_2$有不同的标准差：$s_1=0.9485$比$s_2=0.5379$大很多。由于$\{\tilde{x}_{ti}\}$数据中每增加一个单位即是增加一个标准差，所以比较$\tilde{b}_1$，$\tilde{b}_2$，严格地讲是：当$x_1$，$x_2$都增加一个标准差时，由$x_1$引起的工农业总产值的增加比由$x_2$引起的总产值的增加要大。所以，这种比较法不同于“每增加一个人”时b_1与b_2的比较法。显然，如果x_1与x_2有相同的量纲，则比较标准化回归系数是不合适的；而x_1与x_2在绝大多数情形下是有不同的量纲。因此在无法直接比较b_1与b_2时，只好采用标准化数据法。

下面来看看例1·1·1中回归方程（1·1·15）的现实意义：把$\hat{y}=\ln\hat{W}$，$x_i=\ln Z_i$代回（1·1·15），取微分，得

$$\frac{d\hat{W}}{\hat{W}}=0.6954\frac{\mathrm{d}Z_1}{Z_1}+0.8383\frac{\mathrm{d}Z_2}{Z_2} \qquad (1\cdot1\cdot32)$$

这里$\mathrm{d}Z_1/Z_1$与$\mathrm{d}Z_2/Z_2$都是没有量纲的。他们与$\mathrm{d}W/W$一样都是增加量的相对数。比如，外来人口（Z_1）及本地人口（Z_2）都增加一倍，则工农业总产值将增加$(0.6954+0.8383)=1.5337$倍。而其中外地人口的贡献占$0.6954/1.5337=45.3\%$本地人口的贡献占54.7%。所以（1·1·32）公式可用于去估计人口变动时工农业总产值增加的百分数。

§1·2 回归模型的统计检验与预报

前一节中所述的回归方程及有关公式都未涉及统计检验，因此对因变量y的分布没有具体要求。在本节中，我们假定m个自变量x_1，x_2，…，x_m，是非随机性的变量。如果它们是随机变量，我们就规定它们是属于已给定具体值了的量，因此可以被当作非随机性变量处理。由于实际工作中往往用$(x_1,\ x_2,\ \cdots,\ x_m)$值去预报$y$值，所以，因变量总是随机变量，即具有不确定性。假定模型的理论公式为

$$y=\beta_0+\beta_1x_1+\cdots+\beta_mx_m+\varepsilon \qquad (1\cdot2\cdot1)$$

$\beta_0,\beta_1,\cdots,\beta_m$为理论系数，它们在最小二乘方准则下的样本估计为§1·1的$b_0,b_1,\cdots,b_m$，上

式中ε为理论残差，我们在本节中假定

$$\varepsilon \sim N(0,\ \sigma^2) \tag{1·2·2}$$

其中$N\ (0,\ \sigma^2)$，表示正态分布，其理论均值为0，方差为σ^2。以后将指出，(1·2·2) 的正态性条件是可以大大放松，甚至可以取消的。在回归分析理论中可以证明，在(1·2·2)条件下，几个基本的理论结果为：

(1) 残差平方和Q与回归系数向量b是独立的，且

$$E(Q) = (n-m-1)\sigma^2 \tag{1·2·3}$$

(2) 回归系数向量b及b_0为正态分布：记$\boldsymbol{b} = (b_1,\ b_2,\ \cdots b_m)'$，则

$$\boldsymbol{b} \sim N_m(\boldsymbol{\beta},\ \sigma^2 L_{xx}^{-1}) \tag{1·2·4}$$

具体：

$$E(b_i) = \beta_i,\ \ \mathrm{var}(b_i) = \sigma^2 l^{(i,i)},\ \ \mathrm{cov}(b_i,\ b_j) = \sigma^2 l^{(i,j)}$$

$$E(b_0) = \beta_0,\ \ \mathrm{var}(b_0) = \sigma^2\left(\frac{1}{n} + \bar{\boldsymbol{x}}' L_{xx}^{-1}\bar{\boldsymbol{x}}\right) \tag{1·2·5}$$

其中记$L_{xx}^{-1} = (l^{(i,j)})$。

下面介绍基本的假设检验：

一、回归模型中对线性假定的检验

如果y与$(x_1,\ x_2,\ \cdots,\ x_m)$没有线性关系，此话用数学表示法为：

$$H_0:\ \beta_1 = \beta_2 = \cdots = \beta_m = 0 \tag{1·2·6}$$

在H_0假定下，可以证明下面的统计量为F统计量

$$F(m, n-m-1) = \frac{n-m-1}{m} \cdot \frac{U}{Q} \tag{1·2·7}$$

其中F统计量的第一自由度为m，第二自由度为$n-m-1$，(1·2·7) 也可写成

$$F(m, n-m-1) = \frac{n-m-1}{m} \cdot \frac{R^2}{1-R^2} \tag{1·2·8}$$

如果上述F值大于显著性水平α（一般为0.10，0.05，0.01）下的临界值F_α，则应否定H_0，即认为y与$(x_1,\ x_2,\ \cdots,\ x_m)$间是存在有线性关系，也就是说 (1·2·6) 中的理论系数中至少有一个不应为零。对例1·1·1用 (1·2·8)，

$$F(2, 16-2-1) = \frac{0.8570^2}{1-0.8570^2} \times \frac{18-2-1}{2} = 20.7382$$

取$\alpha = 0.01$，则$F_{0.01}(2, 15) = 6.36$，F值远大于$F_{0.01}$，所以，否定假设H_0；也就是说，认为理论回归系数$(\beta_1,\ \beta_2,\ \cdots,\ \beta_p)$全为零（即$y$与$(x_1,\ x_2)$不存在线性关系）的说法是不对的，我们至少有99%以上的把握认为，例1·1·1的数据中y与$(x_1,\ x_2)$存在线性关系。

实际工作中一个容易犯的错误是：有的人误以为由 (1·2·7) 或 (1·2·8) 计算出来的F值愈大，愈说明y与$(x_1, x_2,\ \cdots,\ x_m)$的线性关系愈密切，因此也认为线性模型的假定也就愈合理，这种看法是错误的。原因是(1·2·8)的基本假定为$(\beta_1,\ \beta_2,\ \cdots,\ \beta_p)$全为零，否定了这个假定，并不等于$y$与$(x_1,\ x_2,\ \cdots,\ x_p)$的线性关系就密切。

如记ρ为y与$(x_1,\ x_2,\ \cdots,\ x_m)$的理论复相关系数，则上述基本假定$H_0$等价于$\rho = 0$，所以否定$H_0$，仅说明$\rho$不应为零。比如理论上如果$\rho = 0.001$，显然，它不为零，但实际上这么小的复相关系数没有什么意义。而从 (1·2·8) 可见，当样本数n无限增加时，R^2则

趋向于理论值ρ^2。明显可见（1·2·8）中F值正比于n。也就是说，如果理论复相关系数$\rho\neq 0$（不管多么小），只要n足够大，我们总可以使(1·2·8)中的F要多大就可以有多大。因此，把（1·2·8）中的F值（或由其计算出的概率值p）的大小作为线性模型好坏的标准是不正确的。正确地说，(1·2·8）的检验仅说明y与(x_1，x_2，…，x_p）是否存在有线性关系，绝不能说明这个关系的程度。当然，一般说来，具有较大理论值的ρ，公式(1·2·8）也容易出现较大的F值（样本数n不变时），这也是显然的。

我们根据文献[1]的结果，介绍如果y的正态分布不成立但等方差性（$\mathrm{var}(\varepsilon)=\sigma^2$)仍然成立时如何检验（1·2·6）中$H_0$。记

$$\boldsymbol{Z}_r=\left(\frac{x_{r1}-\bar{x}_1}{\sqrt{l_{11}}},\ \frac{x_{r2}-\bar{x}_2}{\sqrt{l_{22}}},\cdots,\frac{x_{rm}-\bar{x}_m}{\sqrt{l_{mm}}}\right)',\quad r=1,2,\cdots n$$

$$A=\sum_{r=1}^{n}(\boldsymbol{Z}_r'R_{xx}^{-1}\boldsymbol{Z}_r)^2,\quad B=\frac{1}{l_{yy}^2}\sum_{r=1}^{n}(y_r-\bar{y})^4$$

$$C_x=\frac{n-1}{m(n-3)(n-m-1)}\{n(n+1)A-(n-1)m(m+2)\}$$

$$C_y=\frac{n-1}{(n-3)(n-2)}\{n(n+1)B-3(n-1)\}$$

$$C=\frac{(n-3)C_xC_y}{2n(n-1)},\qquad \delta=1-\frac{C(n+1)}{(C+1)(n-1)} \tag{1·2·9}$$

其中R_{xx}为（x_1，x_2，…，x_m)的相关阵，则可以证明，即使（1·2·2）中正态分布条件不成立，则（1·2·7）或（1·2·8）的右边值仍然可用统计量F去近似，这时自由度要略作改变，即（1·2·7）或（1·2·8）改为

$$F(\delta m,\delta(n-m-1))\doteq\frac{n-m-1}{m}\cdot\frac{R^2}{1-R^2} \tag{1·2·10}$$

可以称δ为非正态性的修正因子。可以证明，如果（1·2·6）成立，则$E(C_y)=0$，于是δ理论值为1。这时（1·2·10）自然即（1·2·8)。C_x，C_y是有上下限的，可以证明有下面不等式：

$$-2\leqslant C_x\frac{(n-3)}{(n-1)}\leqslant n-1,\qquad -2\leqslant C_y\frac{(n-3)}{(n-1)}\leqslant n-1 \tag{1·2·11}$$

所以，当C_x与C_y并不是同时达到上限，又n足够大时从（1·2·9）可见，δ很易近似于1。也就是说，既使正态分布条件不成立，但如果是大样本，则上述关于H_0的检验是可以成立的。这时也可以不必对自由度作修正。但如果是小样本，则还是应对自由度作修正。

本节中方法的计算程序见程序1。

例1·1·1中，$n=18$，可算得：

$A=2.9326$，$B=0.1509$，$C=0.0346$，$\delta=0.9627$，由于$\delta\doteq 1$，所以，可不考虑修正F检验中的自由度。

二、回归模型拟合的优良性及实际误差

我们的目的是要考察用（1·2·1）的模型是否可以相当好地拟合已给的样本。首先考察拟合的残差。记（1·2·1）中y的估计值为$\hat{y}$，记估计残差为e，即每点上残差为：