应用数学丛书

多元统计分析及计算程序

林少宫　袁蒲佳　申鼎煊　编著

华中工学院出版社

多元统计分析及计算程序

林少宫　袁蒲佳　申鼎煊　编著

华中工学院出版社

多元统计分析及计算程序

林少宫　袁蒲佳　申鼎煊　编著

责任编辑　龙纯曼

华中工学院出版社出版发行

（武昌喻家山）

新华书店湖北发行所经销

华中工学院出版社沔阳印刷厂印刷

开本：787×1092　1/32　印张 9.625　字数：197,000

1987年1月第1版　1987年1月第1次印刷

印数：1—1,500

ISBN 7-5609-0023-2/O·3

统一书号：13255—044　　定价：1.65 元

内 容 简 介

本书较系统地介绍了多元统计分析和它的应用，并配有用标准 BASIC 语言编写的计算程序。内容包括有多元回归分析、方差分析、判别分析、聚类分析、主成分分析、因子分析、典型相关分析、对应分析、非线性映射及时间序列分析等。各章都有程序使用说明，框图和实例。

本书可供地质、气象、经济、医学、生物、农业、体育等方面的科技工作者和研究人员及学校师生参考。

作者，提供一套便于在微型电算机上进行多元分析的计算程序。诚然，在早先出版过的一些多元统计计算的书籍中，这些算法程序并不完全缺乏，但大多是用ALGOL语言编写的，特别是针对国产DJS—21机来实现的。在微型电算机日趋普及的今天，这些程序虽然仍有可借鉴的意义，但因微型机上均不配备ALGOL语言，故而不能直接使用。为此，作者尝试用BASIC语言重新编写这些算法。BASIC语言最为普及，容易学会，又具有一定的会话功能，可以说任何微型机上都可以使用它。现在，为了使这些程序更具有普遍适用性，作者有意把所编程序限制在BASIC语言的最基本语法内，而舍弃了一些较为特殊的部分如扩展BASIC的内容。这样看起来程序的某些方面如输出格式方面显得不够完美，但它能在已拥有的任何微型机上运行而不必去更改程序，从而对用户提供了极大便利。这也正是作者的至诚愿望。

本书取名为多元统计分析及计算程序，而不取名为多元统计分析的计算程序，是因为作者还抱着另一种愿望，即希望读者通过本书的阅读，能对多元统计分析的原理、模型、问题和方法有一个基本的了解。也就是说，撇开算法程序不谈，对于具备单元统计分析和矩阵代数基础知识的读者，本书也可作为应用多元统计分析的一种手册式的入门读物。

作者在编写本书的过程中得到了华中工学院数学系陆传务教授的热情支持和具体帮助，计算机系刘健教授在百忙中为本书审稿，在此一并表示衷心的感谢。

由于作者的水平所限，书中难免有不完善甚至错误的地方，尚祈读者不吝指正。

编　者

1985年8月

目　录

第十章　时间序列分析

第一章　多元回归分析

§1 引　　言

回归分析是研究随机现象中变量之间相关关系的一种数理统计方法。通过对变量实际观测值的分析、计算，建立一个变量与另一个（或一组）变量的所谓回归方程，经统计检验认为回归效果显著后，即可用于预测和控制。

回归分析在许多科学领域中都有广泛的应用，如经济中的产品销售量预测，农业中的病虫害或收成预测，自然现象中的地震、洪水预测等等。本章主要讨论多元线性回归和逐步回归。对非线性回归只作扼要的介绍。

设随机变量y与N个自变量x_1，…，x_N有线性关系，于是作y的N元线性回归模型

$$y=\beta_0+\beta_1x_1+\cdots+\beta_Nx_N+\varepsilon, \qquad (1.1.1)$$

其中ε是服从正态分布$N(0,\ \sigma^2)$的随机变量（误差），β_0，…，β_N为回归系数。

回归分析的主要步骤是：

（1） 由观测值确定参数（回归系数）β_0，…，β_N的估计值b_0，…，b_N，从而得y对x_1，…，x_N的线性回归方程

$$\hat{y}=b_0+b_1x_1+\cdots+b_Nx_N, \qquad (1.1.2)$$

其中$\hat{y}$表示对y的估计。

（2） 对线性关系、自变量的显著性进行统计检验。

（3） 利用回归方程进行预报或控制。

（4） 对预报精度作出估计。

§2 多元线性回归

（一）回归系数的最小二乘估计

设对y及x_1，…，x_N作M次观测得

$$(y_\alpha, x_{1\alpha}, \cdots, x_{N\alpha}), \alpha = 1, \cdots, M.$$

将观测值代入（1.1.1）式，得

$$y_\alpha = \beta_0 + \beta_1 x_{1\alpha} + \cdots + \beta_N x_{N\alpha} + \varepsilon_\alpha \quad (\alpha = 1, \cdots, M), \quad (1.2.1)$$

其中ε_α相互独立且服从正态分布$N(0, \sigma^2)$。

若b_i分别为$\beta_i (i = 0, 1, \cdots, N)$的估计值，则由(1.1.2)式得

$$\hat{y}_\alpha = b_0 + b_1 x_{1\alpha} + \cdots + b_N x_{N\alpha}. \quad (1.2.2)$$

称观测值y_α与回归值$\hat{y}_\alpha$之差$y_\alpha - \hat{y}_\alpha$为**残差（或剩余）**，记作$e_\alpha$，它是误差$\varepsilon_\alpha$的估计值。

用最小二乘法求b_i，即是选取b_i使

$$Q = \sum_{\alpha=1}^{M} e_\alpha = \sum_{\alpha=1}^{M} (y_\alpha - b_0 - b_1 x_{1\alpha} - \cdots - b_N x_{N\alpha})^2 = \min.$$

据微分学知识，由$\dfrac{\partial Q}{\partial b_i} = 0 \quad (i = 0, 1, \cdots, N)$，经整理可得

$$\begin{cases} S_{11} b_1 + \cdots + S_{N1} b_N = S_{y1}, \\ \cdots\cdots \\ S_{1N} b_1 + \cdots + S_{NN} b_N = S_{yN}, \end{cases} \quad (1.2.3)$$

其中

$$\begin{cases} S_{ij}=\sum_{\alpha=1}^{M}(x_{i\alpha}-\bar{x}_i)(x_{j\alpha}-\bar{x}_j) \quad (i,j=1,\cdots,N), \\ S_{yj}=\sum_{\alpha=1}^{M}(y_\alpha-\bar{y})(x_{j\alpha}-\bar{x}_j) \quad (j=1,\cdots,N), \\ \bar{x}_i=\frac{1}{M}\sum_{\alpha=1}^{M}x_{i\alpha},\ \bar{y}=\frac{1}{M}\sum_{\alpha=1}^{M}y_\alpha. \end{cases} \tag{1.2.4}$$

(1.2.3) 式称为**正规方程组**，b_i的系数组成的矩阵$\boldsymbol{S}=(S_{ij})$称为**系数矩阵**。由观测值算出离差 S_{ij}和S_{yj}后，即可解得b_i $(i=1,\cdots,N)$。再计算 $b_0=\bar{y}-b_1\bar{x}_1-\cdots-b_N\bar{x}_N$（见多元线性回归程序）。

为了讨论方便起见，记

$$\boldsymbol{X}=\begin{pmatrix} x_{11}-\bar{x}_1 & \cdots & x_{N1}-\bar{x}_N \\ \vdots & & \vdots \\ x_{1M}-\bar{x}_1 & \cdots & x_{NM}-\bar{x}_N \end{pmatrix},$$

$$\boldsymbol{b}=(b_1,\cdots,b_N)',$$

$$\boldsymbol{Y}=(y_1-\bar{y},\cdots,y_M-\bar{y})',$$

则（1.2.3）式写成矩阵式

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}=\boldsymbol{X}'\boldsymbol{Y}\text{或}\boldsymbol{S}\boldsymbol{b}=\boldsymbol{B},$$

其中$\boldsymbol{S}=\boldsymbol{X}'\boldsymbol{X},\boldsymbol{B}=\boldsymbol{X}'\boldsymbol{Y}$.当$\boldsymbol{S}$满秩时，有逆矩阵$\boldsymbol{S}^{-1}=\boldsymbol{C}=(C_{ij})$，于是

$$\begin{cases} \boldsymbol{b}=\boldsymbol{S}^{-1}\boldsymbol{B}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}, \\ b_0=\bar{y}-\sum_{i=1}^{N}b_i\bar{x}_i. \end{cases}$$

在逐步回归中，常将（1.2.3）式的离差S_{ij}改为相关系数

$$r_{ij}=\frac{S_{ij}}{\sqrt{S_{ii}}\sqrt{S_{jj}}} \quad (i,j=1,\cdots,N,y).$$

显然$r_{ii}=1$ $(i=1,\cdots,N,y)$.此时（1.2.3）式写成

$$\begin{cases} r_{11}b_1^* + \cdots + r_{N1}b_N^* = r_{y1}, \\ \cdots \qquad\qquad\qquad \cdots \\ r_{1N}b_1^* + \cdots + r_{NN}b_N^* = r_{yN}. \end{cases} \tag{1.2.5}$$

(1.2.5)式的解b_i^*与(1.2.3)式的解b_i有如下关系：

$$b_i = b_i^* \sqrt{S_{yy}} / \sqrt{S_{ii}} \qquad (i = 1, \cdots, N). \tag{1.2.6}$$

从而 $$b_0 = \overline{y} - \left(\frac{\sqrt{S_{yy}}}{\sqrt{S_{11}}} b_1^* \overline{x_1} + \cdots + \frac{\sqrt{S_{yy}}}{\sqrt{S_{NN}}} b_N^* \overline{x_N} \right). \tag{1.2.7}$$

相关系数矩阵$\boldsymbol{R} = (r_{ij})$的逆阵$(C_{ij}^*)$与离差矩阵$\boldsymbol{S} = (S_{ij})$的逆阵$\boldsymbol{C} = (C_{ij})$有如下关系：

$$C_{ij} = C_{ij}^* / \sqrt{S_{ii}} \sqrt{S_{jj}} \ .$$

在回归计算中，常把y作为第$N+1$个变量的(1.2.5)式的系数增广矩阵

$$\boldsymbol{R}^{(0)} = (r_{ij}) = \begin{pmatrix} r_{11} & \cdots & r_{N1} & r_{N+11} \\ & \cdots & \cdots & \\ r_{1N} & \cdots & r_{NN} & r_{N+1N} \\ r_{1N+1} & \cdots & r_{NN+1} & r_{N+1N+1} \end{pmatrix}. \tag{1.2.8}$$

称为相关系数矩阵。

当$N = 1$时，(1.2.1)式变成

$$y_\alpha = \beta_0 + \beta_1 x_\alpha + \varepsilon_\alpha \qquad (\alpha = 1, \cdots, M), \tag{1.2.9}$$

(1.2.2)式变成

$$\hat{y}_\alpha = b_0 + b_1 x_\alpha, \tag{1.2.10}$$

(1.2.3)式变成

$$\begin{cases} \dfrac{\partial Q}{\partial b_0} = -2\sum_{\alpha=1}^{M} (y_\alpha - b_0 - b_1 x_\alpha) = 0, \\ \dfrac{\partial Q}{\partial b_0} = -2\sum_{\alpha=1}^{M} (y_\alpha - b_0 - b_1 x_\alpha) x_\alpha = 0. \end{cases} \tag{1.2.11}$$

由(1.2.11)式解得

$$\begin{cases} b_1 = \dfrac{\sum_{\alpha=1}^{M} x_\alpha y_\alpha - M\overline{x}\ \overline{y}}{\sum_{\alpha=1}^{M} x_\alpha^2 - M\overline{x}^2}, \\ b_0 = \overline{y} - b_1\overline{x}, \end{cases} \tag{1.2.12}$$

其中 $\overline{x} = \dfrac{1}{M}\sum_{\alpha=1}^{M} x_\alpha$. $\overline{y} = \dfrac{1}{M}\sum_{\alpha=1}^{M} y_\alpha$.

（二）回归问题的统计检验

前面是在假定y与$\boldsymbol{x}=(x_1,\cdots,x_N)$具有线性关系的条件下建立线性回归方程的。但$y$与$\boldsymbol{x}$的关系是否为线性关系？所有$x_i(i=1,\cdots,N)$对$y$是否都有影响？需要作统计检验.在不同问题中，回归分析的重点各异。如果是要建立经验公式或作曲线拟合，则人们最关心的是算出回归系数估计值b_i；如果是为了预报，则最关心的是求预报值及其精度，即求$\hat{y}$及方差$\hat{\sigma}^2$；如果是要找影响y的主要因素，则须检验b_i的显著性，从而决定x_i的取舍。下面我们以经济计量中的统计准则为例，介绍统计检验的具体内容，其中包括：方差分析，F检验，t检验，拟合优度检验及标准差。

（1）方差分析及F检验

类似方差分析法，可将y的总离差平方和S_{yy}分解为

$$S_{yy} = U + Q,$$

其中

$$\begin{cases} S_{yy} = \sum_{\alpha=1}^{M} (y_\alpha - \overline{y})^2, \\ U = \sum_{\alpha=1}^{M} (\hat{y}_\alpha - \overline{y})^2 = \sum_{i=1}^{N} b_i S_{yi}, \\ Q = \sum_{\alpha=1}^{M} (y_\alpha - \overline{y}_\alpha)^2 = \sum_{\alpha=1}^{M} e_\alpha^2 = S_{yy} - U. \end{cases} \tag{1.2.13}$$

U,Q分别称为回归离差平方和和线差平方和.

它们的自由度依次是$f_{总}=M-1$,$f_U=N$,$f_Q=M-N-1$.

对于给定的观测值,S_{yy}也就给定了,于是,可用U和Q的相对大小来衡量回归效果.作统计量

$$F=\frac{U/N}{Q/(M-N-1)}. \tag{1.2.14}$$

为了检验线性回归方程是否有意义,可作假设

$$H_0:\ \beta_1=\beta_2=\cdots=\beta_N=0.$$

如果H_0成立,就表示回归方程无意义.可以证明,当H_0为真时,统计量F服从自由度为N和$M-N-1$的F分布.对于给定的显著性水平α,若计算值$F>F_\alpha(N,M-N-1)$,则在α显著水平上拒绝假设H_0,意谓线性回归方程有显著意义;反之,若$F\leqslant F_\alpha(N,M-N-1)$,则认为线性回归方程无显著意义.

(2)各自变量的重要性——β_i的显著性检验

线性回归方程的显著性,只说明回归方程中全部自变量的总体效果,并不意味着每个自变量都有效果.为了考察各个自变量对y的影响是否显著,必须逐一检验假设

$$H_0:\ \beta_i=0\ (i=1,\cdots,N).$$

为此,可采用t检验.由统计理论知,统计量

$$t_i=\frac{b_i-\beta_i}{\sqrt{C_{ii}}S_y}\ (i=1,\cdots,N)$$

服从自由度为$M-N-1$的t分布.其中C_{ii}为逆阵$\boldsymbol{S}^{-1}=\boldsymbol{C}$的对角线上的元素,$S_y$为$y_\alpha$的**剩余标准差**的无偏估计,$S_y=\sqrt{Q/(M-N-1)}$.当$H_0$为真时,有

$$t_i=\frac{b_i}{\sqrt{C_{ii}}S_y}\ (i=1,\cdots,N). \tag{1.2.15}$$

对于给定的显著性水平α,若计算值$t>t_\alpha(M-N-1)$,则拒

绝假设H_0，认为β_i显著不为0，即相应的x_i对y有影响；否则β_i与0无显著差别，即x_i对y无关重要，应该从回归方程中剔除之。然后再对剩下的自变量建立回归方程。

顺便指出，在下面的逐步回归中，检验变量x_i贡献的显著性是采用与t检验等价的F检验。设从N个变量中去掉x_i，回归平方和由U降到U'，则$W_i=U-U'$称为x_i在这N个变量的回归方程中的贡献。可以证明：$W_i=b_i^2/C_{ii}$，且

$$F_i=\frac{W_i/1}{Q/(M-N-1)}=\frac{b_i^2}{C_{ii}S_y^2}$$

在假设H_0下服从自由度为1和$M-N-1$的F分布。若算得$F_i>F_\alpha(1,M-N-1)$，则x_i的贡献显著；反之可从回归方程中剔除x_i.

（3） $D.W$检验

在回归分析中，有一个重要的假定，认为ε_α是独立随机变量，即在任一时刻的取值与任何前期值无关。当ε取值与它的前一期的值相关时，称这种自相关为**序列相关**。在经济变量中，常常出现这种情形，如销售量并非只受独立变量的影响，也部分地受到前期销售量的影响，因此误差项ε存在自相关。自相关性会影响t检验和F检验的效果，以致整个回归计算和分析需要重新考虑。为了检验是否存在ε的自相关，常采用Durbin-Watson统计量

$$D=\frac{\sum_{i=2}^{M}(e_i-e_{i-1})^2}{\sum_{i=1}^{M}e_i^2} \qquad (1.2.16)$$

进行所谓$D.W$检验，这里e_i是残差$y_i-\hat{y}_i$，即ε_i的估计值。$D.W$检验的步骤是：

假设H_0：ε_t与ε_{t-1}无自相关.

对于给定的显著水平α，自变量个数N，样本容量M，查D分布表得上临界值d_u和下临界值d_l，由(1.2.16)式算得D，并按表1.1处理假设H_0.

表 1.1

D 值	结 论
$(4-d_l)<D<4$	拒绝H_0，有负自相关
$(4-d_u)<D<(4-d_l)$	不作结论
$d_u<D<(4-d_u)$	接受H_0
$d_l<D<d_u$	不作结论
$0<D<d_l$	拒绝H_0，有正自相关

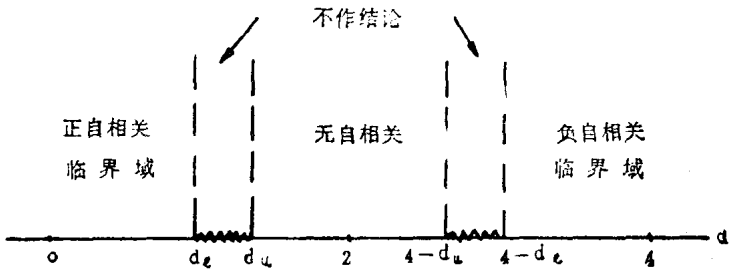

图 1.1

（4） 拟合优度检验

为了说明回归方程中全部自变量的总效果，除前面讲的F检验外，又可采用下面的无量纲值

$$R^2=\frac{U}{S_{yy}}=1-\frac{Q}{S_{yy}}=1-\frac{\sum_{t=1}^{M}e_t^2}{\sum_{t=1}^{M}(y_t-\bar{y})^2}$$

或

$$R=\sqrt{1-\frac{Q}{S_{yy}}} \qquad (1.2.17)$$

实质上U是回归方程中全部自变量的“方差贡献”，故R^2即为这种贡献在总离差中所占的比例.容易看出，$0 \leqq R^2 \leqq 1$；当$R^2 = 1$时，y_i与$\hat{y}_i$完全拟合。R^2愈接近1，y_i与$\hat{y}_i$的拟合愈好，所以R^2是检验回归方程拟合优度的指标。R称为**复相关系数**。

（三） 利用回归方程进行预测和控制

对于给定的样本值$x_{10}, x_{20}, \cdots, x_{N0}$，代入回归方程算得的

$$\hat{y}_0 = b_0 + b_1 x_{10} + \cdots + b_N x_{N0}$$

是y_0的点估计值。实用中常需给出y_0的预测区间。当M较大且x_{i0}较接近$\overline{x_i}$时，可以近似地认为

$$y_0 - \hat{y}_0 \sim N(0, \sigma^2),$$

于是对给定显著性水平$\alpha = 0.05$和$\alpha = 0.01$，可分别用下面两式进行预测和控制：

$$P(\hat{y}_0 - 1.96\sigma < y_0 < \hat{y}_0 + 1.96\sigma) = 0.95$$

和

$$P(\hat{y}_0 - 2.58\sigma < y_0 < \hat{y}_0 + 2.58\sigma) = 0.99.$$

由于σ未知，用无偏估计量

$$S_y = \sqrt{\frac{Q}{M-N-1}} = \sqrt{\frac{\sum_{i=1}^{M} e_i^2}{M-N-1}}$$

代替，故y_0的95%和99%预测区间分别为

$$(\hat{y}_0 - 1.96S_y, \hat{y}_0 + 1.96S_y)$$

和

$$(\hat{y}_0 - 2.58S_y, \hat{y}_0 + 2.58S_y).$$

在各项经济指标的预测中，往往M不可能很大，x_{i0}也不一定接近$\overline{x_i}$，这时对给定的显著性水平α，其预测区间为