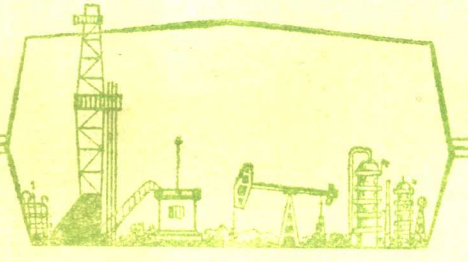
高等学校教学用书

石油数学地质

于志钧　赵旭东　编著



石油工业出版社

石油数学地质

于志钧 赵旭东 编著

石油工业出版社

内 容 简 介

本书按华东石油学院制定的《石油数学地质》教学大纲编写。全书包括：多元统计分析、计算机绘制地质图件、地质数据时间序列分析、石油资源预测四篇共十六章，供70至80学时类型课程讲授之用，其它类型酌减。本书除介绍基本原理和方法之外，一般附有FORTRAN程序并有典型算例可供上机实习之用。本书是石油高等院校石油地质专业的教材。此外，还可供高等院校普通地质矿产专业及中等专业学校石油地质专业教学参考用，也可供石油勘探科研与生产工程技术人员参考。

石 油 数 学 地 质

于志钧　赵旭东　编著

石油工业部教材编译室编辑（北京902信箱）

石油工业出版社出版

（北京安定门外外馆东后街甲36号）

地质印刷厂排版印刷

新华书店北京发行所发行

787×1092毫米 16开本 16$^{1}/_{2}$印张 3 插页 392千字　印1—3,000

1986年 9 月北京第 1 版 1986年 9 月北京第 1 次印刷

书号：15037·2483 定价：2.70 元

前　言

石油地质勘探工作进入二十世纪八十年代遇到三个方面的新问题，它们是：

寻找新油气田的难度越来越大；

使用众多的测量、化验、鉴定和测试的新技术和新方法；

积累的各种地质、物理和化学信息资料暴增。

这使得传统的人工处理方法变得更加困难，例如，目前石油地质勘探技术可以提供以千计的各种类型地质数据，然而，地质人员依靠人工处理就无法利用这样大量的数据，甚至连阅读都成为十分困难的事，因而大量的有用的找油信息被浪费掉。这样的事情在商业领域中尤为突出，为了解决商业信息问题，早在五十年代就引进了电子计算机技术，处理商业信息。进入六十年代，为了同样的目的，计算机技术被引进地质学领域，特别是石油地质勘探领域。现代计算机技术与古老的地质学相结合，形成一个新的边缘学科——数学地质。

数学地质是应用计算机处理大量地质信息（数据）的方法性的学问，它包含的主要内容有：

多元统计分析；

地质过程数学模拟；

地质绘图自动化；

矿产资源预测；

数据库。

多元统计分析包含：多元回归分析、聚类与分类分析、判别分析、因子分析。在此之前，研究地质学问题，由于人工计算的限制，对许多多变量的问题，仅研究二个变量间的关系，这无疑是很片面的。采用多元统计方法使得地质学定量研究提升到一个新高度。

地质过程是很复杂的，但了解它的真实过程对寻找石油与天然气，又是十分重要的。人工的定性了解已经不能满足日益困难的寻找油气藏的需要，因之许多地质过程的数学模型被建立起来，其中有地质构造发展模型、沉积模型、石油生成模型等等。目前已经利用这些种类的模型来寻找油气藏。

地质绘图占据了地质科技人员大量的时间，使得他们不能有更多的时间集中精力综合研究所进行的地质科研课题，所以地质绘图自动化一直是地质科技人员的一个追求目标。现代计算机系统已经为绘制地质图件自动化提供了可能性，目前已经能够绘制各种等值线图、剖面对比图、曲线图、座标图、立体曲面图、趋势面图等等，这无疑将使地质科技人员可以用更多的时间进行综合研究，同时大大提高图的质量和加快石油勘探速度。

地质数据库是现代计算机应用进入高级阶段的信息资源基础。没有一个强大的地质数据库支持的计算机系统，就脱离不了手工操作，就不能进行大规模的计算机侦察油气分布地区和石油勘探决策。所以说建立地质数据库是计算机应用水平从低级过渡到高级的标志。

我国数学地质的开拓者是现地震局地质研究所付研究员徐道一同志，他于1965年，在

美国开始把计算机技术引进地质学领域之后不久，在我国电子计算机应用尚处早期试验阶段，就把因子分析方法引进我国，研究四川盆地的碳酸盐岩地层剖面划分问题。之后，在1978年召开首届全国数学地质学术讨论会，成立了隶属中国地质学会的中国数学地质专业委员会。1981年召开了第二届数学地质学术讨论会。数学地质在我国发展非常迅速，目前已渗透到地质学的一切分支学科之中，获得了很好的效果。在建立、发展数学地质工作中，现数学地质专业委员会主任委员中国科学院地质研究所付研究员刘承祚同志做出了很大的贡献。

在石油地质勘探领域，数学地质工作在1970年起步，最早从事石油数学地质研究性工作的有现华东石油学院北京研究生部的于志钧付教授、现胜利油田地质科学研究院工程师韩玉笈、史彩云、王永福、北京石油勘探开发科学研究院计算中心工程师石广仁等。应该指出，胜利油田地质科学研究院数学地质室的同志们在研制数学地质应用软件、石油生成数学模型、地质绘图和石油地质勘探数据库的许多领域中作出了重大的贡献并成功地应用数学地质方法寻找出新的油藏。在研制石油资源评价方法和应用方面，北京石油勘探开发科学研究院工程师赵旭东同志作了很多工作。目前，我国石油地质勘探生产、科研和教学部门已经形成了一支实力雄厚，接近或赶上世界先进水平的数学地质队伍。1983年，召开了首届全国石油数学地质学术讨论会，检阅了石油数学地质队伍，交流推广和提高了数学地质在石油地质勘探中应用的方法和理论。

为了提高我国高等石油院校石油地质勘探专业学生的专业素质，适应生产、科研和新的技术革命的需要，我国各石油高等院校已经设置了《石油数学地质》课程。本书就是根据华东石油学院《石油数学地质》教学大纲要求编写的。全书共分四篇：第一篇多元统计分析方法；第二篇计算机绘制地质图件；第三篇地质数据的时间序列分析；第四篇石油资源预测方法。本书前三篇是由于志钧同志编写的，其中华东石油学院勘探系李汉林同志写了回归分析和聚类与分类分析二章；北京工业学院甘仞初同志编写了最大熵谱分析一章；第四篇是赵旭东同志编写的。

本书经国家地震局地质研究所徐道一同志及中国科学院地质研究所刘承祚同志审阅，提出了宝贵的修正意见，作者致以衷心的谢意。

本书为石油高等院校石油地质勘探专业教材，亦可供石油地质勘探或其它地质矿产的生产、科研人员参考之用。对非石油地质专业的高等地质院校各专业可做教学参考书。

最后，限于作者的工作经验和学术水平，本书缺点在所难免，望读者指正。

作者　1986年5月北京

目　录

第一篇　多元统计分析

第一章　基本统计概念的多元推广

大多数观测的单元素自然现象都可用正态分布描述，正态分布的概念可以推广到多元的情况。

假设我们在一个地区采集岩石标本，对每块标本测量了许多种性质。我们可把单独一块标本的一系列测定作为一个向量〔X〕=〔$X_1X_2\cdots X_m$〕，共有m个测定性质（变量）。若向量〔X〕代表的样品是从许多独立活动结果，即各元素的分布是相互无关的独立变量构成的总体中随机抽取的，那么所研究的向量将趋向于多元正态分布。分别考虑，每个变量是正态分布的并由平均值μ_k和方差σ_k^2确定分布性质。合并概率分布是一个相当于m维的正态分布，它具有向量平均值〔μ〕=〔$\mu_1\mu_2\cdots\mu_m$〕及归纳为对角矩阵形式方差：

$$[\Sigma^2]=\begin{bmatrix}\sigma_1^2 & 0 & 0\cdots\cdots 0\\ 0 & \sigma_2^2 & 0\cdots\cdots 0\\ \vdots & \vdots & \vdots \quad \vdots\\ 0 & 0 & 0\cdots\cdots\sigma_m^2\end{bmatrix}$$

第一节　多元正态分布

一、多元正态分布的形式

我们从一元正态分布推广，一元正态分布的密度函数形式为

$$f(x)=\frac{\sigma^{-1}}{(2\pi)^{1/2}}\exp\left[-\frac{1}{2}(x-\mu)\ (\sigma^{2-1}(x-\mu)\right]$$

把x具有上述分布律记作$x\sim N(\mu,\ \sigma^2)$。

推广到多元正态变量$X=[x_1x_2\cdots x_m]'$，密度函数具有下述形式：

$$f(x)=\frac{|\Sigma^{-1}|^{1/2}}{(2\pi)^{m/2}}\exp\left[-\frac{1}{2}(X-\mu)'\Sigma^{-1}(X-\mu)\right]$$

这里：μ均X的期望值，称均向量；Σ为X的协方差矩阵，即

$$\mu=[\mu_1\mu_2\cdots\mu_m]'$$

$$\Sigma=\begin{bmatrix}\sigma_{11} & \sigma_{12}\cdots\cdots\sigma_{1m}\\ \sigma_{21} & \sigma_{22}\cdots\cdots\sigma_{2m}\\ \vdots & \vdots \quad \vdots\\ \sigma_{m1} & \sigma_{m2}\cdots\cdots\sigma_{mm}\end{bmatrix}$$

$|\Sigma^{-1}|$为矩阵Σ的逆矩阵行列式。把X具有上述分布律记作$X\sim N(\mu, \Sigma)$。

二、多元正态分布的性质

这里我们仅列出多元正态分布的一些重要性质，不作证明。

1．*多元正态变量X的任意线性变换仍然是正态变量，即*

若$X=[x_1x_2\cdots x_m]'$服从正态分布$N(\mu, \Sigma)$，则任意线性变换

$$Y=CX+D$$

仍服从正态分布 $N(C\mu+D, C\Sigma C')$。这里：

$$C=\begin{bmatrix} C_{11} & C_{12} & \cdots\cdots & C_{1m} \\ C_{21} & C_{22} & \cdots\cdots & C_{2m} \\ \vdots & \vdots & & \vdots \\ C_{n1} & C_{n2} & \cdots\cdots & C_{nm} \end{bmatrix}$$

$$D=[d_1d_2\cdots d_m]'$$

2．*多元正态变量各分量的线性组合服从一元正态分布，即*

若$X\sim N(\mu, \Sigma)$

$C'=[C_1C_2\cdots C_m]$为任一数量向量则线性组合：

$C'X=C_1X_1+C_2X_2+\cdots+C_mX_m$

服从一元正态分布

$$N\left(\sum_{i=1}^{m}C_i\mu_i,\ \sum_{i=1}^{m}\sum_{j=1}^{m}C_iC_j\sigma_{ij}\right)$$

推论1：多元正态变量的每一个分量服从正态分布。

推论2：多元正态变量的任何部分分量之和仍服从正态分布。

3．*多元正态变量X的任何一个分量子集的分布(称为边缘分布)仍为正态分布。*

第二节　t检验的多元推广——检验统计量T^2

检验从具有表征值μ_0及未知的方差σ^2的总体取的随机样品。统计量：

$$t=\frac{(\bar{X}-\mu_0)\sqrt{n}}{\sqrt{S^2}}$$

推广这个检验到多变量的情况是以样品的向量平均值代替$\bar{X}$；总体向量平均值代替μ_0；方差—协方差矩阵代替S^2。

我们定义$[\mu]$为总体向量平均值，于是样品向量平均值为$[\bar{X}]$。类似地，$[\Sigma^2]$是总体方差—协方差矩阵，于是$[S^2]$代表样品方差—协方差矩阵。两向量相减得出样品与总体平均值之差的列向量：

$$[\bar{X}]-[\mu]=[\bar{X}-\mu]$$

代入上式，得

$$t=\frac{[\bar{X}-\mu]\sqrt{n}}{\sqrt{[S^2]}}$$

我们要应用这个检验，必须把向量及矩阵简化成单一数。如果我们给列向量$[X-\mu]$乘一个行向量，结果就是一个单一数。所以，我们定义一个任意行向量$[A]$，它的转置是

一个列向量〔A〕′。把〔$\bar{X}-\mu$〕乘以〔A〕，得出一个简单数，并且在$\sqrt{〔S^2〕}$前后各乘以〔A〕及〔A〕′也得一个简单数。这时，检验统计量：

$$t=\frac{〔A〕\cdot〔\bar{X}-\mu〕\sqrt{n}}{〔A〕\sqrt{〔S^2〕}〔A〕'}$$

从而把零假设

$$H_0:\ 〔\mu_1〕=〔\mu_0〕$$

变换为 $H_0^*:\ 〔A〕〔\mu_1〕=〔A〕〔\mu_0〕$

这样的变换无疑只有当新的假设H_0^*对〔A〕的所有可能值都成立时，原始假设H_0才为真。于是我们对统计量平方消去根式，新的检验统计量用T^2表示，有

$$T^2=n〔\bar{X}-\mu〕'\cdot〔S^2〕^{-1}\cdot〔\bar{X}-\mu〕$$

T^2的临界值可由下面的关系式确定：

$$F=\frac{n-m}{m(n-1)}T^2 \qquad (1-1-1)$$

这里n为样本容量；m为变量数。

检验统计量T^2是检验量t的多元引伸，称作郝泰灵(Hotelling)统计量。它可以由常规F分布表或专用T^2分布表查得临界统计量。

第三节　检验两个P维正态样本

这个检验，我们可以考虑为检验一个特殊总体平均向量的一个样本。我们用两个独立随机样本的集合来代替，检验它们的平均向量是等价的。

假设二个样本取自多变量正态分布总体，二者有相同的未知方差—协方差矩阵〔Σ^2〕。我们想检验零假设条件：

$$H_0:\ 〔\mu_1〕=〔\mu_0〕$$

相反

$$H_1:\ 〔\mu_1〕\neq〔\mu_0〕$$

零假设条件是，第一个样本的总体平均向量与第二个样本的总体平均向量相同。为此，我们把两个多变量样本合并计算其共同的方差—协方差矩阵：

$$〔S_p^2〕=\frac{1}{n_1+n_2-2}\ (〔SP_1〕+〔SP_2〕)$$

这里：矩阵〔SP_1〕、〔SP_2〕的元素为

$$SP_{jk}=\sum_{i=1}^{n}(A_{ij}A_{ik})-\frac{\sum_{i=1}^{n}A_{ij}\sum_{i=1}^{n}A_{ik}}{n}$$

这里：A_{ij}是第j个变量的第i次观测值；A_{ik}是第k个变量的第i次观测值；SP_{jk}是矩阵的第j行第k列元素。

接着计算二个平均向量的差〔$\bar{X}_1$〕−〔$\bar{X}_2$〕或〔$\bar{X}_1-\bar{X}_2$〕。统计量T^2有

$$T^2=\frac{n_1n_2}{n_1+n_2}〔\bar{X}_1-\bar{X}_2〕'\cdot〔S_p^2〕^{-1}\cdot〔\bar{X}_1-\bar{X}_2〕$$

或
$$T^2=\frac{n_1n_2}{n_1+n_2}D^2 \qquad (1-1-2)$$

D^2称为马哈拉诺毕斯(Mahalanobis)距离。

统计量T^2的显著性可由F变换确定：

$$F=\frac{n_1+n_2-m-1}{(n_1+n_2-2)m}T^2 \qquad (1-1-3)$$

这里：n_1为第一个样本的观测次数；n_2为第二个样本的观测次数；m为变量数；m及(n_1+n_2-m-1)分别为第一与第二自由度。

我们介绍这一章的目的是说明，以后的全部多元统计分析都是从正态分布总体取的样本出发的，以样本的方差协方差矩阵相等为基本假设的。

第二章　回归分析

第一节　多元线性回归分析

一、回归分析的概念和任务

在实际问题中，经常遇到一些相互制约而又相互依赖的事物。即然事物间存在着制约性和依赖性，那么用来描述这些事物内在联系的量之间也就必然具有一定的关系。我们把描述某一事物的量称做因变量y，而把描述另外m个事物的量称做自变量x_1、x_2…、x_m。这样，变量y和x_1、x_2…、x_m间的关系基本上可分为以下两类：

确定型关系，即函数关系。

例如，在已知圆柱体底圆半径R和圆柱体高H时，则圆柱体的体积V可由

$$V=\pi R^2H$$

确定。又如，曲边梯形的面积S是曲边纵坐标在其底边上的积分，即

$$S=\int_a^b f(x)dx$$

一般情况下，变量y与x_1、x_2…、x_m之间的函数关系记为

$$y=f(x_1, x_2, \cdots, x_m)$$

这一类函数关系是数学分析的研究对象。

非确定型关系，即相关关系。

仅就地质学中的问题而言，具相关关系的变量很多。众所周知，有机质随矿物质沉积后，转化成石油所需要的时间t，主要依赖于地层的温度T；另外，也与地层的压力P、有机质的类型及其它地质因素有关。简单地说，埋藏深、温度高，有机质成熟就快，转化成石油所需要的时间就短，反之就长。但是，它们之间的具体关系式是不知道的。实际上，由于地质因素的多样性和复杂性，也不可能确定出它们之间的严格表达式。又如，岩石的渗透率，它一方面受孔隙度的影响；另一方面，它还与孔道截面大小、形状、连通性等因素有关。一般说来，岩石的渗透率随着有效孔隙度的增加而变大。地质上诸如此类的例子很多，我们就不一一列举了。

我们把相互制约、相互依赖、关系式是属于非确定型的变量叫做具有相关关系的变量。

若变量y与x_1, x_2, …, x_m间有关系式

$$y=a_0+a_1x_1+a_2x_2+\cdots+a_mx_m+\varepsilon \qquad (1-2-1)$$

就说y与x_i(i=1,2,…,m)之间具有m元线性相关关系，简称m元线性关系。(1—2—1)式称为线性回归模型。式中a_0, a_1, …, a_m是待定常数，ε是误差项，且$\varepsilon\sim N(0,\sigma^2)$。确定了常数$a_0$, a_1, …a_m后，就确定了相关变量间的数学表达式。回归分析就是确定相关变量间数学表达式的一种统计方法。

回归分析的任务是：

1．根据变量y，x_1，x_2，…，x_m的n组观测值

$$(x_{1k}, x_{2k}, \cdots, x_{mk}, y_k)；\ k=1,2,\cdots,n。$$

求出a_0、a_1、…、a_m的最佳估计值b_0、b_1、…、b_m，得到一个方程

$$\hat{y}=b_0+b_1x_1+\cdots+b_mx_m \qquad (1-2-2)$$

（1—2—2）式称为y对$x_i(i=1,2,\cdots,m)$的回归方程，它代表的平面（超平面）叫做回归平面，b_0、b_1、…b_m叫做回归系数。在此，又可以把有关建立回归方程的理论和计算方法统称为回归分析。

2．把相关变量按线性模型（1—2—1)处理，得到回归方程（1—2—2)，它反映的相关程度如何？也就是（1—2—2）是否有实用价值？就要对回归方程进行检验。除了经过实践检验外，还可用统计的方法，对回归方程作显著性检验。

3．经检验，若（1—2—2）能代表y与$x_i(i=1,2,\cdots, m)$的线性关系，并且精度也满足研究问题的要求，就可利用回归方程（1—2—2）对变量y进行预测或控制。

二、用最小二乘法确定回归系数b_0、b_1、…b_m

设变量y与$x_i(i=1,2,\cdots,m)$间有线性关系

$$y=a_0+a_1x_1+\cdots+a_mx_m+\varepsilon \qquad (1-2-3)$$

如前所述，根据n组观测值

$$(x_{1k}, x_{2k}, \cdots, x_{mk}, y_k) \quad k=1,2,\cdots,n$$

我们只能求出a_0、a_1、…、a_m的最佳估计值b_0、b_1、…、b_m，得回归方程

$$\hat{y}=b_0+b_1x_1+\cdots+b_mx_m \qquad (1-2-4)$$

从数学上知，(1—2—4）是m+1维空间的一个平面。对于任意一个观测点$(x_{ik}, y_k)i=1,2,\cdots,m,k=1,2,\cdots n$，参看三维空间图（1-2-1)，则回归平面有一投影点$(x_{ik}, \hat{y}_k)$，$\delta_k=(y_k-\hat{y}_k)$称为剩余，建立回归方程的原则就是使剩余平方和

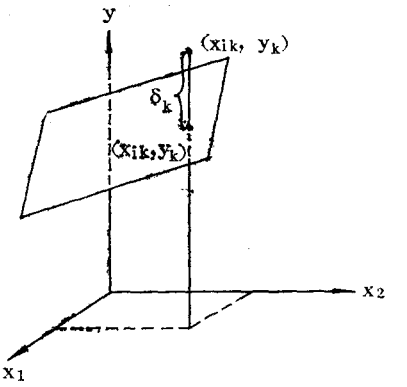


图 1-2-1　三维空间图

$$Q_1=\sum_{k=1}^{n}\delta_k^2=\sum_{k=1}^{n}(y_k-\hat{y}_k)^2$$

达到最小。Q_1是b_0、b_1、…、b_m的二次函数，且$Q_1>0$，它满足

$$\begin{cases}\dfrac{\partial Q_1}{\partial b_0}=0\\ \dfrac{\partial Q_1}{\partial b_1}=0\end{cases} \quad i=1,2,\cdots,m$$

即

$$\sum_{k=1}^{n}(y_k-b_0-b_1x_1-\cdots-b_mx_m)=0$$

$$\sum_{k=1}^{n}(y_k-b_0-b_1x_1-\cdots-b_mx_m)x_{1k}=0$$

$$\vdots \qquad\qquad \vdots \qquad\qquad (1-2-5)$$

$$\sum_{k=1}^{n}(y_k-b_0-b_1x_1-\cdots-b_mx_m)x_{mk}=0$$

从线性方程组（1—2—5）可以解出回归系数 b_0、b_1、…、b_m，得到回归方程（1—2—4）。

为了计算方便，化简线性方程组（1—2—5）。

从线性方程组（1—2—5）的第一个方程解出

$$b_0=\bar{y}-\sum_{i=1}^{m}b_i\bar{x}_i \qquad (1-2-6)$$

式中，$\bar{y}=\frac{1}{n}\sum_{k=1}^{n}y_k$；$\bar{x}_i=\frac{1}{n}\sum_{k=1}^{n}x_{ik}$，$i=1,2,\cdots,m$。

把（1—2—6）式代入线性方程组（1—2—5）的后m个方程，并由（1—2—5）的第一个方程等于零可得到

$$\sum_{k=1}^{n}[(y_k-\bar{y})-b_1(x_{1k}-\bar{x}_1)-\cdots-b_m(x_{mk}-\bar{x}_m)](x_{ik}-\bar{x}_i)=0 \qquad (1-2-7)$$

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

将（1—2—7）式展开整理得

$$\sum_{k=1}^{n}(y_k-\bar{y})(x_{ik}-\bar{x}_i)=\sum_{k=1}^{n}\sum_{j=1}^{m}(x_{ik}-\bar{x}_i)(x_{ik}-\bar{x}_i)$$

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

在上式中令

$$S_{iy}=\sum_{k=1}^{n}(y_k-\bar{y})(x_{ik}-\bar{x}_i),\ i=1,2,\cdots,m$$

$$S_{ij}=\sum_{k=1}^{n}(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j),\ i,j=1,2,\cdots,m$$

S_{iy}是变量y与x_i的协方差，S_{ij}是变量x_i和x_j的协方差。至此（1—2—5）式可写成

$$\sum_{j=1}^{m}S_{ij}b_j=S_{iy},\ i=1,2,\cdots,m \qquad (1-2-8)$$

方程组（1—2—8）称为正规方程组。从（1—2—8）可以解出系数 b_1、b_2、…、b_m，再将 b_1、b_2、…、b_m代入（1—2—6）式解出b_0，于是得回归方程（1—2—4）。

使剩余平方和Q_1最小，确定参数a_0、a_1、…、a_m的方法叫做最小二乘法。可以证明，用这种方法确定的回归方程是最佳方程。

三、回归方程的方差分析和显著性检验

假定变量y与x_i($i=1,2,\cdots,m$）间存在着线性关系，用最小二乘法确定了回归系数，建立了y对x_i($i=1,2,\cdots,m$）的回归方程。但是，在用最小二乘法确定回归系数时，只是

要求剩余平方和Q_1达到最小，而没有用到线性相关的假定条件。因此即便给出的n组观测值在m+1维空间中呈球状分布的情况下，按线性回归模型用最小二乘法同样可以求出一个所谓的"回归方程"。但是，这个回归方程根本就不能反映变量y与$x_i(i=1,2,\cdots,m)$之间的变化关系，也就不能用它来对y进行预测或控制。为此就要研究回归方程所反映的相关程度，即对回归方程作显著性检验。为了统计检验的需要，先对m元线性回归方程进行方差分析。

1．*m元线性回归方程的方差分析*

分解变量y的总偏差平方和Q，

$$\begin{aligned} Q &= \sum_{k=1}^{n}(y_k-\bar{y})^2 = \sum_{k=1}^{n}[(y_k-\hat{y}_k)+(\hat{y}_k-\bar{y})]^2 \\ &= \sum_{k=1}^{n}(y_k-\hat{y}_k)^2+\sum_{k=1}^{n}(\hat{y}_k-\bar{y})^2+\sum_{k=1}^{n}(y_k-\hat{y}_k)(\hat{y}_k-\bar{y}) \\ &= Q_1+Q_2 \end{aligned} \qquad (1-2-9)$$

式中：$Q_1=\sum_{k=1}^{n}(y_k-\hat{y}_k)^2$；$Q_2=\sum_{k=1}^{n}(\hat{y}_k-\bar{y})^2$；

$$\sum_{k=1}^{n}(y_k-\hat{y}_k)(\hat{y}_k-\bar{y})=0$$

从（1—2—9）式知，Q可以分解为Q_1和Q_2两部分。Q是变量y与其平均值之差的平方和，称为总偏差平方和，它的大小反映了观测值y_k的离散程度。Q_1主要是线性回归模型引起的偏差，另外也有其它随机因素的影响，称为剩余平方和。Q_2是$\hat{y}_k$与y的平均值$\bar{y}$之差的平方和，它反映了变量$x_i(i=1,2,\cdots,m)$的变化对y引起的波动，称为回归平方和。

用f表示自由度，则Q、Q_1、Q_2的自由度分别是

$$f_Q=n-1;\quad f_{Q_1}=n-m-1;\quad f_{Q_2}=m,$$

且满足等式

$$f_Q=f_{Q_1}+f_{Q_2}$$

2．*显著性检验*

（1）复相关系数法

从（1—2—9）式知，Q_1愈小，Q_2就愈接近于Q，说明变量y与$x_i(i=1,2,\cdots,m)$的线性关系越密切，即回归方程代表的变化关系就越好。由此可用比值Q_2/Q作为衡量回归方程显著性的一个指标。我们定义

$$R=(Q_2/Q)^{\frac{1}{2}}$$

为y与$x_i(i=1,2,\cdots,m)$的复相关系数。当R愈接近±1，表明回归方程的显著性越高；反之，当R愈接近0时，y与$x_i(i=1,2,\cdots,m)$的线性关系越差，甚至可以说它们之间毫无线性关系。

（2）F分布检验法

假设H_0：变量y与$x_i(i=1,2,\cdots,m)$间"无线性关系"。当假设H_0为真时，Q_1就比较大，Q_2就比较小，比值Q_2/Q_1就小，当小于一定界限时，就肯定原假设H_0；当大于一定界限时，就否定原假设H_0，即变量y与$x_i(i=1,2,\cdots,m)$间有着密切的线性关系。

统计量

$$F=\frac{Q_2/m}{Q_1/(n-m-1)}=\frac{Q_2(n-m-1)}{mQ_1}$$

遵从第一自由度为m、第二自由度为(n－m－1)的F(m,n－m－1)分布。根据给定的的信度（检验水平）α，在$F_\alpha(m,n-m-1)$分布表上查出临界值F_α。当$F>F_\alpha$时，否定原假设H_0，这时称回归方程是显著的，可以付诸应用；若$F\leqslant F_\alpha$，则肯定原假设H_0，即求出的回归方程没有实际意义。

四、利用回归方程进行预测或控制

显著性的回归方程，就可用于对y的预测或控制。

我们所说的预测，就是把定值$x_{i0}(i=1,2,\cdots,m)$代入回归方程计算出

$$\hat{y}_0=b_0+b_1x_{10}+\cdots+b_mx_{m0}$$

用$\hat{y}_0$作为y_0的估计值。估计值$\hat{y}_0$的误差取决于回归方程的显著性。剩余标准差，也叫剩余方差

$$\sigma=\sqrt{Q_1/(n-m-1)}$$

刻划出了$\hat{y}_0$误差的大小。当自变量取定值$x_{i0}(i=1,2\cdots,m)$，并且$\min\limits_{k=1,2\cdots n}X_{ik}\leqslant X_{i0}\leqslant\max\limits_{k=1,2\cdots n}X_{ik}$时，对应的$y_0$落在区间

$$\hat{y}_0-2\sigma<y_0<\hat{y}_0+2\sigma$$

内的概率为$(1-\alpha)$，α可取为0.01～0.10，即$(1-\alpha)=0.90\sim0.99$内插预测是可靠的。当自变量取定值$x_{i0}(i=1,2,\cdots,m)$，并且$x_{i0}<\min\limits_{k=1,2\cdots n}X_{ik}$或$x_{i0}>\max\limits_{k=1,2\cdots n}X_{ik}$时，$y_0$的可靠程度就难以估计了。因为这时回归方程可能已不能代表y与$x_i(i=1,2,\cdots,m)$的相关关系了。所以，外推预测只能供参考。

如何调整自变量$x_i(i=1,2,\cdots,m)$，使y值落在区间$y_1\leqslant y\leqslant y_2$内，这就是所谓的控制问题。只要取$x_{i0}(i=1,2,\cdots,m)$满足

$$\hat{y}_0-2\sigma\geqslant y_1$$
$$\hat{y}_0+2\sigma\leqslant y_2$$

即可实现对y的控制。

五、非线性回归问题

研究相关变量时，经常碰到变量之间的相关关系不是呈线性关系，而表现为某种曲线关系。例如，生油门限时间t与地层温度T及埋深H的方程式

$$\ln t=A\frac{1}{T+273}+B\frac{1}{H}+C$$

中，令$\ln t=y$、$x_1=\frac{1}{T+273}$、$x_2=\frac{1}{H}$就化为线性模型

$$y=Ax_1+Bx_2+C$$

另一种情况，就是根本不知道变量间的相关类型，也不可能通过作图的方法了解它们的类型时，一般采用较高次的函数式或一些初等函数的线性组合，假定它们的类型。然后，再采用变换的方法化为线性模型处理。

六、m元线性回归的计算步骤

首先给出计算Q、Q_1、Q_2及S_{ij}的简便公式：

$$Q=\sum_{k=1}^{n}(y_k-\bar{y})^2=\Sigma y_k^2-2\bar{y}\Sigma y_k+\Sigma\bar{y}^2$$

$$=\Sigma y_k^2-2\bar{y}\cdot n\cdot\frac{1}{n}\Sigma y_k+n\bar{y}^2=\Sigma y_k^2-n\bar{y}^2$$

$$Q_1=Q-Q_2$$

$$Q_2=\sum_{k=1}^{n}(\hat{y}_k-\bar{y})^2=\sum_{k=1}^{n}(b_0+b_1x_{1k}+\cdots+b_mx_{mk}-\bar{y})^2$$

$$=\sum_{k=1}^{n}(\bar{y}-\sum_{i=1}^{m}b_i\bar{x}_i+\sum_{i=1}^{m}b_ix_{ik}-\bar{y})^2=\sum_{k=1}^{n}\left[\sum_{i=1}^{m}b_i(x_{ik}-\bar{x}_i)\right]^2$$

$$=\sum_{k=1}^{n}\left[\sum_{i=1}^{m}\sum_{j=1}^{m}b_ib_j(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_i)\right]$$

$$=\sum_{i=1}^{m}\sum_{j=1}^{m}b_ib_j\sum_{k=1}^{n}(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_i)$$

$$=\sum_{i=1}^{m}\sum_{j=1}^{m}b_ib_jS_{ij}=\sum_{i=1}^{m}b_iS_{iy}$$

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

$$-\sum_{k=1}^{n}x_{jk}\bar{x}_i+\sum_{k=1}^{n}\bar{x}_i\bar{x}_j$$

$$=\sum_{k=1}^{n}x_{ik}x_{jk}-n\bar{x}_i\bar{x}_j$$

为了计算方便，令$y=x_{m+1}$，则$S_{iy}=S_{im+1}$；另外，正规方程组的系数矩阵关于主对角线是对称的，为此我们只计算主对角线在内的上三角元素。

第二节　逐步回归分析

一、逐步回归的提出

在m元线性回归分析中，我们拟定了m个与y有线性关系的变量$x_i(i=1,2,\cdots,m)$，在把它们对y的作用视为等同的条件下，建立了y对变量$x_i(i=1,2,\cdots,m)$的回归方程

$$\hat{y}=b_0+b_1x_1+\cdots+b_mx_m$$

并对该回归方程作了显著性检验。回归分析在油气资源预测、地震预报、工业自动控制及其他领域中都有着广泛的应用。但是m元线性回归分析中的复相关系数和F检验法，都是检验m个变量$x_i(i=1,2,\cdots,m)$共同对y起的作用，而没有考虑m个变量中哪些对y的影响更大。事实上，它们各自对y的作用是不同的。单独看其中某些变量，可能它们对y的相关程度都较高，但在m个变量建立的回归方程中，可能其中有一部分变量对y的作用就会显得无足轻重。原因在于变量间也存在着程度不同的相关关系，m个变量中对y作用更大