陈希孺　王松桂　编著

近代实用回归分析

广西人民出版社

$$=X\beta+e$$

近代实用回归分析

陈希孺　王松桂　编著

☆

广西人民出版社出版

（南宁市河堤路14号）

广西新华书店发行　　贵县印刷厂印刷

开本 787×1092 1/32　10.5 印张　231 千字

1984 年 2 月第 1 版　1984 年 2 月第 1 次印刷

印数 1—4,600 册

书号：13113·30　定价 1.30 元

序言

回归分析是数理统计中应用最广泛的分支之一。直到约六十年代中期为止，应用工作者所使用的回归分析方法基本上限于以最小二乘法为主体的线性回归。近十余年来，回归分析作为数理统计的一个分支有了多方面的发展。本书是介绍这些新发展中较富实用意义的部分，所介绍的新方法，大多已在国外见诸实用，有些在国内也已开始受到应用工作者的注意和使用。但是，由于大量的材料都散见在一些杂志文献中，查找和学习都不方便，本书的写作就是想以不大的篇幅对有关材料作比较系统的收集和整理，以方便想了解这方面情况的同志。

本书的对象主要是具有一定数学和概率统计知识的应用统计工作者和高等学校有关专业师生，也希望对某些对这方面发展感兴趣的理论工作者能有所帮助。由于这个原因，虽则我们编写的重点在于实用方法，但对理论问题也给予一定的注意。

本书要求读者具备微积分、线性代数和矩阵的初步知识，以及初等概率统计知识。要阅读§4·4，需要一些测度论的知识。

本书一共分四章。第一章是关于古典线性回归的预备知识。第二章是关于回归自变量的选择，即在一个包含很大数目的可供选择的自变量的问题中，怎样去挑选出为数不多的重要的自变量。第三章是关于线性回归系数的估计问题，介绍了近十多年来发展的一系列重要的估计方法，这些方法的提出是企图改进目前常用的最小二乘估计。第四章有两个内容，一个是非参数回归，这里不需要假定回

归为线性或其它任何特定的形式，因而可以适用于更广的范围；另一个是所谓“稳健回归”，它与第三章一样，也是讨论线性回归系数的估计，是企图从另一个角度来改进最小二乘估计。

要在实际问题中恰当而有效地使用回归分析这个工具，除了掌握其方法和理论外，很重要的是需要以分析的态度来对待它，需要积累经验，以找出在特定领域内有效的方法，和恰当处理一些要凭经验（当然，还有对问题所涉及的专业知识的了解）酌定的问题。本书所介绍的方法，其历史都不长，应用上积累的经验还不很多，所以当我们把这些方法推荐给读者的时候，特别期望能够在实际工作中使用和考验这些方法，以积累自己的经验，并进一步探索新问题、发展新方法。这也是本书书名中特别标出“实用”二字的用意。当然，由于这些方法的历史不长，加上我们所掌握的资料的限制，书中引述的具体应用实例还不够多，我们决定采用现在这个书名，除了在选材上它确实主要从实用的角度着眼外，另一层用意，也许是更重要的，是为了表明我们希望能引起实用工作者对这方面发展的注意。

本书的编写工作可以说是从1977年开始的，从那时起，我们开始陆续收集关于这方面的文献，积累了一些资料。以后的两年中，作者曾先后在北京、开封、武汉、广州等地举办的统计讲习班上讲授了这些内容，通过这些学术活动，对初稿作了几次修改。一些同志曾提出了不少有益的意见，作者谨借这个机会表示衷心地感谢。当然，由于作者水

平所限，书中肯定还会有不少欠妥的地方，恳切希望同行专家和广大读者不吝赐教。

张尧庭同志对原稿作了仔细的审阅。孙庭恒同志为本书绘制了插图。对以上提到的同志，作者也谨借这个机会表示衷心的感谢。

作者　1981年8月

目 录

第一章　线性回归的若干预备知识

本书是叙述近十多年来回归分析的一些较有实用意义的新发展。这些新发展，多与在最小二乘法和Gauss—Markov或正态假定的基础上建立的线性回归的理论和方法，有密切的联系。因此，了解这些内容，对于弄清楚近代回归分析中一些问题产生的根源，以及处理这些问题的思想和方法，是必要的。根据本书的性质，可以要求读者具备这个基础。但是，特别是为了方便那些主要兴趣在于应用的同志，并使本书内容大体上成为自封的，我们打算在这开头的一章中，对线性回归的基本知识，以适应本书需要的形式作一概括性的叙述。在编写这一章时，我们特别注意不要使它成为一些公式和记号的罗列，而着重于对基本概念、问题提法和处理方法，作一种富于实际背景的论述，这对于理解这门学科的实质是至为重要的。然而，限于篇幅和本书的性质，我们并不打算处处从一种入门读物的要求去写，也不讲求完全的系统性。希望获得更为全面和系统知识的读者，可参看有关文献，例如文献〔1〕和〔2〕。

向量和矩阵是讨论回归分析的重要工具，下面将本书中一贯采用的若干记号列举出来，以便查阅。

1．向量用一个字母，例如U来记，必要时指出其分量：

$$U=\begin{pmatrix}U_1\\U_2\\\vdots\\U_n\end{pmatrix} \qquad (1\cdot1)$$

U_1, U_2, …, U_n都是实数，U_j称为U的第 j 个分量，而n称为U的维数。值得注意的是，当提到向量时总是指列向量而言。

2. 矩阵用一个字母，如 A 来记，必要时写出其元素，如

$$A=(a_{ij})_{m\times n}=\begin{pmatrix}a_{11}&a_{12}&\cdots&a_{1n}\\a_{21}&a_{22}&\cdots&a_{2n}\\\cdots&\cdots&&\\a_{m1}&a_{m2}&\cdots&a_{mn}\end{pmatrix},$$

a_{ij}称为A的 (i,j) 元。当m = n时称为方阵，而a_{11}, …, a_{nn}称为其主对角线元。又n阶单位阵记为I_n，或简记为I。

3. 矩阵 A的转置记为 A′，即若 $A=(a_{ij})_{m\times n}$，则 $A'=(b_{ij})_{n\times m}$，其中 $b_{ij}=a_{ji}$。特别，列向量(1·1)的转置为行向量$U'=(U_1, U_2, \cdots U_n)$。因此，有时为书写方便，把列向量 (1·1) 写为$U=(U_1, U_2, \cdots, U_n)'$的形式。

4. 向量$U=(U_1,U_2,\cdots,U_n)'$和$V=(V_1,V_2,\cdots,V_n)'$的内积 $U'V=V'U=\sum_{i=1}^{n}U_iV_i$。而$\sqrt{U'U}$，即向量U的欧氏长度，常记为$\|U\|$。

5. n阶方阵 $A=(a_{ij})_{n\times n}$的迹 (trace)，定义为

$$\mathrm{tr}(A) = \sum_{i=1}^{n} a_{ii}$$

6. 矩阵A的秩记为R(A)。若A为n阶方阵而R(A) = n，则称A为满秩的，这时以A^{-1}记其逆。

7. 若A为正定（半正定）方阵，则记为$A>0$，$(A\geqslant 0)$。若A－B为正定（半正定）方阵，则记为$A>B$ $(A\geqslant B)$。

§1·1 线性回归模型

（一）相关关系与回归模型

自然科学、工程技术以至社会科学中许多问题的研究，往往归结为弄清楚一些有关变量的联系。在许多问题中，这种联系具体反映为寻求其中一个变量Y（为方便计以下称之为因变量）通过其它的变量X_1，…，X_p（以下称之为自变量）表达出来。在这方面可分出两种基本类型：第一类的特征是：只要知道了自变量X_1，…，X_p所取的值，因变量Y所取之值就唯一地确定了。这种关系叫做确定性关系，即数学上的所谓函数关系。另一类关系叫做非确定性关系，有时也称为相关关系。它的特征是：因变量Y所取的值与自变量X_1，…，X_p所取的值有关系，但这种关系没有密切到可以唯一决定的程度。例如，在一项农业研究中，可能要考虑每亩播种量X_1、每亩施肥量X_2对每亩产量Y的影响，当然，Y的值与X_1，X_2有很大关系，但由X_1，X_2并不能唯一地决定Y。

非确定性关系产生的原因，一则是由于在不少问题中，

对特定因变量Y有影响的因素（即变量）往往为数极多。而由于我们的认识水平和客观条件的限制，在问题中考虑进来的自变量 X_1，…，X_p，一般只是全部有关变量的一部分。其它未被考虑进来的变量，由于在研究工作中未加控制，往往以一种随机的方式影响因变量Y的取值，从而造成不确定性。再则，即使在某种稀少的情况下，我们已把全部有关变量都考虑进来了，但由于观测仪器、外界环境、参加试验工作人员等等方面的偶然性原因，也会影响Y的取值。

不论产生非确定性的原因如何，从数学的角度看，它使我们面临这样一种情况：在给出了自变量 X_1，…，X_p 的值 x_1，…，x_p 后，Y的取值仍带有随机性。就是说，Y是一个随机变量，其分布由 X_1，…，X_p 所取的值 x_1，…，x_p 所定。在这里我们抓到了函数关系与相关关系的差别的实质：一个是由 x_1，…，x_p 决定Y的值，一个是由 x_1，…，x_p 决定Y的概率分布。因此，从数学上研究相关关系，就归结为弄清Y在给定 $X_1=x_1$，…，$X_p=x_p$ 时的（条件）概率分布随 x_1，…，x_p 而变化的规律。回归分析是进行这种研究一种有力的方法。这个方法的第一步是引进回归函数 $f(x_1, \cdots, x_p)$，它是在给定 $X_1=x_1$，… $X_p=x_p$ 的条件下，Y的条件期望值，即

$$f(x_1, \cdots, x_p) = E(Y|X_1=x_1, \cdots, X_p=x_p) = E(Y|x_1, \cdots, x_p)$$

而方程

$$y = f(x_1, \cdots, x_p) \qquad (1\cdot2)$$

则称为（理论）回归方程。回归方程可以解释为是描述了因变量Y随自变量 X_1，…，X_p 变化的平均情况。在实用上，它可以被解释为用一种确定的函数关系来近似地代替复杂的

相关关系，或者说，回归方程（1·2）反映了Y与$X_1,\cdots,X_p$的（相关）关系的“主要方面”。

确定了回归函数$f(x_1,\cdots,x_p)$以后，我们就可以把Y对$X_1,\cdots,X_p$的依赖关系分解为两部分：

$$Y=f(X_1,\cdots,X_p)+e \qquad (1\cdot3)$$

e是一随机变量，由回归函数的定义，它显然满足条件：不论$X_1,\cdots,X_p$取怎样的值，有

$$E(e)=0 \qquad (1\cdot4)$$

（1·3）和（1·4）一起，构成一个理论上的回归模型。（1·3）式可解释为对Y的一种分解：第一项，即回归函数f，反映了其“平均趋势”或“主要部分”，第二项e可以看作为某种随机干扰，通常称为随机误差，也叫模型误差，这一项的存在使Y与$X_1,\cdots,X_p$的关系成为非确定性的。可以说，不确定性的程度，取决于这一项的影响的大小。

现在我们来较仔细地分析一下回归函数这一项。有两种情况：一是对函数f毫无所知，这时回归模型称为非参数性的，这种模型将在本书第四章讨论。另一种是回归函数f的总的数学形式已知，但包含若干个未知参数，最简单且最重要的情况是f为$x_1,\cdots,x_p$的线性函数：

$$f(x_1,\cdots,x_p)=\beta_0+\beta_1x_1+\cdots+\beta_px_p \qquad (1\cdot5)$$

这里$\beta_0,\cdots,\beta_p$为未知参数，β_0称为常数项，而$\beta_1,\cdots,\beta_p$称为回归系数。这时称回归模型为线性的。有些表面上不是线性的回归模型，例如

$$f(x_1,x_2)=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2$$

可以通过人为地引进某些新的“自变量”而转化为线性的。例

如，引进新“自变量”$x_3=x_1x_2$，即化为线性形式。然而，这时三个“自变量”x_1，x_2，x_3中，独立起作用的只有两个。本章及以下两章所处理的都是线性回归模型。回归分析的一个主要问题，就是估计回归函数中所包含的未知参数，以及检验有关这些参数的假设。在这方面，随机误差e的性质有着重要的作用，如果假定e服从正态分布，则回归模型称为正态的。正态线性模型是回归分析中研究最深入的模型。

（二）在有了观察或试验数据时，线性回归模型的表述

设有了线性回归模型

$$Y=\beta_0+\beta_1X_1+\cdots+\beta_pX_p+e \qquad (1\cdot6)$$

要估计未知数β_0，…，β_p之值，或讨论其它有关这些参数的问题，需要进行观察或试验，以得到样本。在此首先要区分两个基本的情况：一是X_1，…，X_p的取值可以由人们控制，例如在前面所举的关于播种量、施肥量和产量的关系的例中，播种量X_1和施肥量X_2都可以由试验者加以控制，这时，我们可以在一些经过选定的(X_1, X_2)的值处进行试验，以观察其对应的Y值。在这种情况下。自变量自然看成是非随机的。另一种情况是自变量X_1，…，X_p的取值也是由随机观察所得，而不能由人任意安排。例如，考虑人的身高（自变量X）与体重（因变量Y）的关系，在进行观察时，我们随机地抽出一些人，测出他们的身高和体重。这些值，包括X值在内，都非事先可任意选定的。在这种情况下，往往较自然的是把自变量也看成是随机变量。在第四章我们将考虑这种情况。然而，特别是在线性模型的情况下，人们往往不采取这种看法，而宁愿把自变量仍看成是非随机的。这

在理论上有某些方便，且在若干重要方面，并不影响问题的实质。在本书的前三章我们就采取这个观点。

这样，我们假定进行了n次试验或观察，在第i次观察中，X_1，…，X_p 和Y分别取值X_{1i}，…，X_{pi}和Y_i，$i=1$，…，n；又设在第i次观察中随机误差e取值e_i，$i=1,\cdots,n$。应当注意的是，e_1，…，e_n的值是不能观察到的。根据模型（1·6），我们得到n个方程：

$$Y_i=\beta_0+\beta_1X_{1i}+\cdots+\beta_pX_{pi}+e_i,\quad i=1,\cdots,n \qquad (1·7)$$

模型（1·6）描述了相关关系的理论结构，而统计工作的出发点则是（1·7），因此可以把（1·7）说成是有了样本以后的回归模型。为方便起见，我们以后在提到回归线性模型时，常是指（1·7）而言，这不会引起任何混淆。

各次观察的随机误差e_1，…，e_n，本身都是随机变量。回归模型（1·7）的概率性质取决于它们的概率性质。首先，依（1·4），有

$$E(e_i)=0,\quad i=1,2,\cdots,n \qquad (1·8)$$

关于进一步的假定，最常用的有以下两种：

1。Gauss—Markov假定（或条件），它包含以下两点：

等方差性：

$$Var(e_i)=\sigma^2,\quad i=1,\cdots,n,\quad 0<\sigma^2<\infty \qquad (1·9)$$

不相关性：

$$Cov(e_i,e_j)=0,\quad 当i\neq j,\ i,j=1,2,\cdots,n \qquad (1·10)$$

这两条结合起来，大致可解释为：各次观测（不论在其中自变量取什么值）所受的随机影响程度相同，且任意两次观察的误差大小并无关联。

2。正态假定（或条件）：$e_1,\cdots,e_n$独立，$e_i\sim N(0,\sigma^2)$，$i=1,\cdots,n$。

显然，正态假定包含了Gauss—Markov假定。又需注意，不论在Gauss—Markov假定或正态假定下，误差方差σ^2都看作是未知的，这是模型的一个重要未知参数。

引用矩阵记号，可以把（1·7）写成比较简洁的形式。为此，令

$$\mathbf{1}=(1,1\cdots,1)' \qquad (n维) \tag{1·11}$$

$$\mathbf{X}=\begin{Bmatrix} x_{11} & x_{21} & \cdots & x_{p1} \\ x_{12} & x_{22} & \cdots & x_{p2} \\ \cdots & & \cdots & \\ x_{1n} & x_{2n} & \cdots & x_{pn} \end{Bmatrix} \tag{1·12}$$

$$\mathbf{B}=\begin{Bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{Bmatrix}, \quad \mathbf{e}=\begin{Bmatrix} e_1 \\ \vdots \\ e_n \end{Bmatrix}, \quad \mathbf{Y}=\begin{Bmatrix} Y_1 \\ \vdots \\ Y_n \end{Bmatrix} \tag{1·13}$$

则有

$$\mathbf{Y}=\mathbf{1}\beta_0+\mathbf{X}\beta+\mathbf{e} \tag{1·14}$$

我们知道，一个随机向量$\xi=(\xi_1,\cdots,\xi_n)'$的均值向量，定义为$E(\xi)=(E(\xi_1),\cdots,E(\xi_n))'$，其方差阵记为$VAR(\xi)$，定义为

$$VAR(\xi)=\{Cov(\xi_i,\xi_j)\}_{n\times n}$$

在这些记号下，（1·8）可写为

$$E(\mathbf{e})=\mathbf{0} \tag{1·15}$$

而Gauss—Markov假定可表为

$$VAR(\mathbf{e})=\sigma^2\mathbf{I}_n, \qquad 0<\sigma^2<\infty \tag{1·16}$$

正态假定可表为

$$\mathbf{e}\sim N(0,\sigma^2\mathbf{I}_n) \tag{1·17}$$

模型（1·14）中包含自变量$(X_1,\cdots,X_p)$的n组值$(x_{1i},\cdots,x_{pi})$，$i=1,2,\cdots,n$，它们常被称为n个“试验点”。在

$X_1,\cdots,X_p$ 的值确可由试验者任意选定时，这个称呼是名副其实的。在这种情况下，人们就可以利用试验点的可自由选择性，去找出某种比较好的选择，以使当利用由之产生的模型（1·14）去进行统计推断时，能得到好的效果，这就是回归分析的设计问题。由于这个原因，由（1·12）确定的矩阵 $\mathbf{X}$，常被称为设计矩阵。即使在自变量 $x_1,\cdots,x_p$ 的值不能自由选定的情况，也常这样称呼 $\mathbf{X}$。当然，这时 $\mathbf{X}$ 实际上并不具备设计的意义。

模型（1·14）的写法，突出了常数项 β_0 的地位。这不仅是一个形式上的写法问题，在实际意义上，β_0 只反映度量原点的选定，而回归系数 β_i 则反映 Y 随 X_i 的变化，二者的性质不同。由于这个原因，在回归分析中处理一些问题时，往往有必要把常数项 β_0 与回归系数 $\boldsymbol{\beta}$ 分别对待。读者在本书中将会看到若干这样的例子。§1·2中所讨论的关于设计矩阵 $\mathbf{X}$ 的中心化、标准化问题也与此有关。

但是，在某些主要是纯理论性的问题中，并无必要把 β_0 单独分开。引进

$$\tilde{\boldsymbol{\beta}} = (\beta_0,\beta_1,\cdots\beta_p)' \tag{1·18}$$

$$\tilde{\mathbf{X}} = (\mathbf{1} \vdots \mathbf{X}) \tag{1·19}$$

可将（1·14）写为

$$\mathbf{Y} = \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} + \mathbf{e} \tag{1·20}$$

这个简洁形式对讨论许多理论问题更为方便。

在本章以下几节中，我们要讨论线性回归模型的一些重要的统计问题。有关的理论证明放到本章最后一节，即§1·5，只对应用感兴趣的读者可暂时跳过这一节。

§1·2　最小二乘估计

（一）常数项和回归系数的最小二乘估计

设有了线性模型（1·14），我们把它写为（1·20）的形式。为了记号上的方便，在本节前三段中，我们把$\widetilde{X}$和$\widetilde{\beta}$分别简记为X和β，在以后我们还会多次这样做。这不致引起混淆，因为在每次这样做时，我们都会加以明确的声明。这样，我们有模型

$$Y = X\beta + e \tag{1·21}$$

若记$X_i = (1, X_{1i}, \cdots, X_{pi})$，$i = 1, \cdots, n$，可将（1·21）写为

$$Y_i = X_i' \beta + e_i \qquad i = 1, \cdots, n \tag{1·22}$$

本段讨论β的估计问题。估计的方法基于下面的想法：设想β的真值为β_0，则$X_i' \beta_0$表示在第i个试验点$(X_{1i}, \cdots, X_{pi})$处Y的平均值，而Y_i为Y的实际观测值。容易想象，当β_0确为β的真值或与之很接近时，差

$$\sigma_i(\beta_0) = Y_i - X_i' \beta_0, \qquad i = 1, \cdots, n \tag{1·23}$$

的绝对值应倾向于较小。反之，若β_0与β的真值差距较大，则上述绝对值倾向于较大。这个考虑引导到如下的估计法：作函数

$$L(\beta) = \sum_{i=1}^{n} (Y_i - X_i' \beta)^2 = \|Y - X\beta\|^2 \tag{1·24}$$

然后定出$\hat{\beta}$（$\hat{\beta}$与X，Y有关），使