应用线性回归分析

周复恭 黄运成 编著

应用线性回归分析

周复恭　黄运成　编著

中国人民大学出版社

应用线性回归分析

周复恭　黄运成　编著

中国人民大学出版社出版发行
（北京西郊海淀路39号）
中国人民大学出版社印刷厂印刷
（北京鼓楼西大石桥胡同61号）
新华书店经销

开本：850×1168毫米32开　印张：10
1989年8月第1版　1989年8月第1次印刷
字数：239 000　册数：1—3 000

ISBN 7-300-00616-7
O·20　定价：3.00元

前　言

回归分析和相关分析是应用数理统计学的重要组成部分，近年来又有了新的发展，已经成为应用数理统计学的一个新的分支。它在实际问题，特别是经济问题的研究中有着广泛的应用。

《应用线性回归分析》是作者在几年教学实践的基础上，参阅了有关的书籍，为本科学生学习这门课程编写的，也可以作为广大实际统计工作者在自学“应用线性回归分析”之用。

本书共计九章，除完整地介绍回归分析和相关分析的原理和方法外，还分章介绍了多重共线性、异方差性和自相关性及其处理方法，此外还介绍了虚拟变量和线性回归分析中自变量的选择问题。后者除介绍自变量选择的原则、标准外，还着重介绍了各种选择方法并对其进行了比较。虽然这些方面的介绍还是初步的，但就回归分析和相关分析讲，本书涉及的内容还是比较广泛的，并为读者进一步学习“应用线性回归分析”提供了一个良好的基础。

回归分析，特别是多元回归分析需要进行大量的运算工作。电子计算机的普及和应用，为进行回归分析和相关分析提供了快速的运算手段，以之为工具不仅可以解除应用者的繁重计算之苦，也可以做到计算准确和时间上的节约。鉴于此，本书的最后一章对 TSP 软件包的功能和使用做了简介，并以实例说明了 TSP 使用的方法和步骤。无疑这为还不太熟悉 TSP 软件包使用的读者提供了方便。

本书附有回归分析和相关分析中常用的五个分布表，这也为

读者学习本书提供了便利条件。

本书给出了回归分析和相关分析中进行计算的公式，并对其中某些公式进行了数理证明，以便读者更好地掌握其内容。鉴于读者在学习本书之前已经学过作为先行课程的矩阵代数，本书未专门介绍矩阵。如果有的读者还不太熟习矩阵代数，请读者自行阅读有关书籍，以便理解本书中的有关部分。

学以致用。把回归分析和相关分析应用于实际问题的研究时，除需熟练掌握回归分析和相关分析的原理和方法外，一如本书一再提到的，还需具有较为扎实的有关问题的理论分析能力、专业知识和统计方法应用的经验。在这里强调指出这一点，作者以为是有意义的。也请读者注意本书中提到的，某种方法怎样用是正确的，怎样用就有欠妥当。

在编写过程中，我们曾参考成晓梅编写的《Micro TSP 软件使用说明书》，摘引范福仁《生物统计学》一书的附表 9 作为本书的附表 5，在此对上述两位同志谨致谢忱。

本书前言、第五、第六和第七章由周复恭编写，其余各章由黄运成编写。全书由周复恭总纂定稿。

由于作者水平有限，书中不当和错误之处在所难免。恳请读者指正，在此谨致谢意。

作　者

1987年 8 月

目　录

第一章　简单线性回归分析…… 1
§1.1　引言 …… 1
§1.2　回归参数的估计 …… 4
§1.3　线性关系的检验 ……16
§1.4　回归参数的检验 ……33
§1.5　回归预测问题 ……38
第二章　简单线性相关分析……51
§2.1　相关系数的来源及其意义 ……51
§2.2　相关系数r的抽样分布 ……64
§2.3　总体相关系数ρ的假设检验……73
第三章　多元线性回归分析……88
§3.1　引言 ……88
§3.2　回归参数的估计 ……89
§3.3　回归分析中的假设检验 ……93
§3.4　回归预测问题 ……103
§3.5　复相关分析和偏相关分析 ……114
§3.6　一般线性回归模型简介 ……120
§3.7　曲线回归问题 ……130
第四章　多重共线性问题的处理 ……134
§4.1　多重共线性的含义及其后果 ……134
§4.2　多重共线性的检验 ……143
§4.3　补救多重共线性影响的方法 ……146
第五章　异方差性问题的处理 ……154

§5.1 异方差性问题的性质 ……154
§5.2 异方差性的后果 ……158
§5.3 异方差性的检验 ……162
§5.4 补救异方差性的方法 ……166
附录：加权最小平方法 ……169
第六章 自相关性问题的处理 ……171
§6.1 自相关现象的性质 ……171
§6.2 自相关现象的后果 ……180
§6.3 自相关现象的检验 ……185
§6.4 自相关现象的处理 ……195
第七章 虚拟变量 ……203
§7.1 虚拟变量的性质 ……203
§7.2 有一个数量变量和一个具有两种分类的品质变量的回归分析 ……205
§7.3 有一个数量变量和一个多于两类的品质变量的回归分析 ……208
§7.4 一个数量变量和两个品质变量的回归分析 ……210
§7.5 虚拟变量在季节分析中的应用 ……212
§7.6 举例 ……213
第八章 线性回归分析中的变量选择问题 ……216
§8.1 引言 ……216
§8.2 问题的公式化 ……216
§8.3 变量删除的后果 ……217
§8.4 变量选择的一般原则 ……220
§8.5 评价方程的标准 ……222
§8.6 变量选择的方法 ……227
§8.7 举例 ……233
附录一 方程的错误的技术要求对回归系数估计量和预测值的影响 ……240
附录二 江苏省连云港市纺织工业年度报表（1983） ……244

第九章　时间序列分析软件包（TSP）简介 ……245
§9.1　Micro　TSP软件概述 ……245
§9.2　Micro　TSP软件使用初步 ……253
§9.3　TSP软件应用的实例分析 ……260
附表1　标准正态累积概率表 ……288
附表2　t分布的百分位数 ……290
附表3　F分布表 ……292
附表4　D—W统计量临界值表 ……302
附表5　简单相关系数临界值表……304
参考书目 ……305

第一章　简单线性回归分析

§1.1　引　　言

事物间的发展变化往往表现为事物间的数量关系的发展变化，研究事物间的量变规律是认识其性质的一种重要方法。在自然界和人类社会中，普遍存在着两类数量关系，一类是确定性关系，即一种现象的数量变化完全确定了另一种现象的数量变化，这种精确的数量关系反映了事物间的因果联系，它可以用一种确定的函数表达式来反映，因此又称为函数关系或因果关系。如某种物品的销售收入 Y 在价格 p 确定的情况下完全由其销售量 X 决定，即

$$Y = pX$$

一般地说，对这种精确的数量关系可以用一种确定性的数学模型来描述，如

$$Y = f(X) \qquad (1.1.1)$$

其中：Y 代表因变量，X 代表自变量，f 代表 Y 和 X 之间的函数关系。因果关系是现实关系的抽象化，是进行了大量试验以后，通过对试验数据的收集、归纳、整理及分析等理论上的加工而提炼出来的事物内部存在的一种必然联系。这种联系往往并非那么一目了然，但又确是我们所期望得到的关系，因此，描述这种关系的（1.1.1）式所表示的模型又称为理论模型或因果模型。

因果关系是一种理想化关系，然而，人们首先发现的事物之

间的关系往往不是这种关系，而是另一类关系——统计关系。所谓统计关系是指在现实生活中表现出来的事物之间的相随变动的规律性，它是一种现实关系。可以说，因果关系是统计关系的深化和抽象化，是统计关系的科学化和理论化，而统计关系则是因果关系建立的前提和基础，是因果关系的现实来源。统计关系是一种经验关系，它有四大特性：（1）大量性。统计关系是通过对所研究事物进行了大量试验和观察以后而建立起来的一种关系。大量性表现了事物之间的客观性和现实性，不在大量性基础上建立起来的关系往往缺乏真实性。（2）平均性。统计关系是从平均的意义上表现了事物之间的关系，这种平均的数量联系不可能和事物间的个别联系一一吻合，个别例证甚至还可能与我们所建立的统计关系相违背，但是从平均的意义上说，这种关系是正确的，它反映了事物间的量变倾向和趋势。（3）随机性。统计关系不能和事物间的个别数量联系一一吻合的原因是因为存在着随机因素的干扰。实际上，一种事物的数量变化与多种因素有关，但是，由于客观条件的限制，我们只可能考察决定该事物量变的主要因素。统计关系反映的是这些主要因素和该事物之间的量变规律性，而那些大量的、非主要因素对事物量变的影响，我们无法一一考察，只有把它们看成大量的随机干扰，并假定由于随机因素的相互抵消，其平均干扰为零。（4）条件性。我们对事物的研究总是在一定的条件下进行的，任何科学试验也都是在一定的条件下做出的，因此所得到的试验结果及由之建立的统计关系也只是在一定的条件下是正确的和成立的。离开了具体的条件，这种关系就不复存在。而泛谈统计关系、一般地讲统计关系、离开了具体的条件谈统计关系是不科学和不正确的。综合起来说，统计关系是一种非确定性关系，对这种关系进行定量描述的数学模型为回归模型，如

$$Y = E(Y \mid X) + \varepsilon \qquad (1.1.2)$$

其中 Y为因变量，$E(Y|X)$为给定X的条件下Y的条件均值，ε 为随机变量，且

$$E(\varepsilon)=0 \tag{1.1.3}$$

令

$$f(X)=E(Y|X) \tag{1.1.4}$$

则（1.1.2）式又可以表示为

$$Y=f(X)+\varepsilon \tag{1.1.5}$$

我们称（1.1.4）式为一个回归方程，特别是，当

$$E(Y|X)=\beta_0+\beta_1 X \tag{1.1.6}$$

时，叫做简单线性回归方程。因此，在这种情况下，统计关系有时又称为回归关系或相关关系。

从广义上讲，回归和相关问题统称为相关分析；从狭义上讲，研究诸变量间联系的紧密程度的方法称为相关分析，研究一个或一组变量的变动对另一个变量的变动的影响程度的方法称为回归分析。

回归和相关问题最初来源于生物界，生物界中存在着许多相关现象，达尔文曾认为相关是生物有机体的一种有规律的特性。1880年英国生物学家兼统计学家高尔顿（Galton，1822—1911）在其所著《相关及其度量——以人体测定材料为根据》一文中首次使用了“相关”（correlation）一词，并提出用“相关指标”（index of correlation）度量变量间的相关程度。1889年，高尔顿的《自然遗传》一书出版，至此，高尔顿的相关论基本完成。在高尔顿的论文中，他发现同一种族中儿子的平均高度，介于其父亲的高度与种族平均高度之间。父亲矮的，其儿子的平均高度较父亲高，但比种族平均高度矮；父亲高的，其儿子的平均高度较父亲矮，但比种族平均高度高。儿子的高度有返归于种族平均高度的趋势，亦即回归于种族平均高度。在这里，高尔顿首次应用了“回归”（regression）一词，这就是“回归”一词在遗传学上的原始含义。现在，已经把“回归”一词的含义加以拓广，凡

是利用一个变量或一组变量的变异来估计或预测另一个变量的变异情况，都称之为回归。

在回归模型中，凡是变量之间的关系是线性关系的模型都称之为线性回归模型，否则，就称之为非线性回归模型。在线性回归模型中，若自变量X的个数只有一个，则称简单线性回归模型，或二元（这里的“元”指在回归模型中变量的个数，包括因变量Y）线性回归模型，如

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad (1.1.7)$$

其中 β_0、β_1 称为模型参数，或回归参数，β_0 也称为回归截距项，β_1 又称为回归斜率。若自变量的个数有两个或两个以上，则称为多元回归模型或复回归模型，如

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \qquad (1.1.8)$$

相关分析也可以作类似分类，在线性相关分析中，分析两个变量间联系的紧密程度的方法称为简单线性相关分析，分析一个变量和一组变量（两个及两个以上）间联系的紧密程度的方法称为多元相关分析或复相关分析。本章研究的重点是简单线性回归分析，多元线性回归分析的问题我们将在第三章研究。在本章中，我们重点研究回归参数的估计、回归方程的检验、回归系数的检验及回归预测等问题。

§1.2 回归参数的估计

在上一节中，我们已经给出了简单线性回归模型的一般形式，即

$$Y = E(Y|X) + \varepsilon \qquad (1.2.1)$$

其中

$$E(Y|X) = \beta_0 + \beta_1 X \qquad (1.2.2)$$

Y为因变量，X为自变量，β_0、β_1 为未知的回归参数，ε 为随机变量。我们称（1.2.1）式为理论回归模型，（1.2.2）式为理论

回归方程。若对 Y 和 X 分别进行了 n 次独立观测，得到以下 n 对观测值

$$(Y_i,\ X_i),\quad i=1,2,\cdots,n$$

则有

$$Y_i=\beta_0+\beta_1X_i+\varepsilon_i\qquad i=1,2,\cdots,n\qquad(1.2.3)$$

我们称（1.2.3）式为理论回归模型的有限样本模型，简称有限样本模型。令

$$E(Y_i|X_i)=\beta_0+\beta_1X_i\qquad i=1,2,\cdots,n\qquad(1.2.4)$$

（1.2.4）式为理论回归方程的有限样本回归方程，简称回归方程，$E(Y_i|X_i)$ 为 Y_i 的条件期望值。对随机变量 ε_i，通常有以下假定

$$\varepsilon_i\sim N(0,\ \sigma^2)\qquad i=1,2,\cdots,n\qquad(1.2.5)$$

即　（1）ε_i 是一个独立的正态随机变量；

（2）ε_i 的均值为零

$$E(\varepsilon_i)=0\qquad(1.2.6)$$

（3）ε_i 的方差为常数 σ^2

$$D(\varepsilon_i)=\sigma^2\qquad(1.2.7)$$

另外，在回归分析中，我们还假定 X_i 是一个非随机变量，且没有测量误差。这样，就使我们研究的问题大大简单化了，在（1.2.3）式中，由于 ε_i 的存在，所以 Y_i 也是一个随机变量，且

$$\left.\begin{aligned}&Y_i\sim N(E(Y_i|X_i),\ \sigma^2)\\&E(Y_i|X_i)=\beta_0+\beta_1X_i\\&D(Y_i)=\sigma^2\\&i=1,\ 2,\ \cdots,\ n\end{aligned}\right\}\qquad(1.2.8)$$

强调一下，（1.2.5）式、（1.2.8）式及 X_i 是非随机变量

且无观测误差的假定是回归分析中的重要假定，以后所给的结论都是建立在这些假定的基础上的。读者在以后使用本章所介绍的方法时，一定要注意这些基本假定是否满足，若这些基本假定遭到严重破坏，则通过回归分析所得到的结论可能和实际的情况不完全相符，甚至得出错误的结论。在专门的回归分析的教科书中都给出了考察这些基本假定是否满足的方法以及在假定被破坏的情况下的处理技术。然而，在本书中，限于篇幅，我们不能把这些问题都展开论述，请有兴趣的读者查阅本书附录中所列的参考书。

现在，我们把问题集中在对（1.2.3）式的研究上。在（1.2.3)式中，Y_i的变动由两部分构成，一部分是可观测因素X_i，另一部分是不可观测因素 e_i，即

$$Y_i = (\beta_0 + \beta_1 X_i) + e_i, \qquad i = 1,2,\cdots,n$$

可观测因素　　不可观测因素

我们的估计只能对可观测因素进行，即估计(1.2.4）式，在获得了样本（Y_i，X_i），$i=1,2,\cdots,n$以后，具体的任务就是估计未知参数 β_0、β_1。怎样获得β_0、β_1的样本估计量呢？通常的方法是最小平方法，即我们要寻找这样的一种估计量，使得随机误差项 e_i 的平方和达到最小。具体方法如下：

由（1.2.3）式得

$$e_i = [Y_i - (\beta_0 + \beta_1 X_i)], \qquad i = 1,2,\cdots,n$$

分别对上式两边先平方，然后求和，得

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_i)]^2 \qquad (1.2.9)$$

现在可以把我们的任务表述为：求 β_0、β_1的估计量β_0、β_1，使得

$$S(\beta_0,\ \beta_1)=\min\sum_{i=1}^{n}[Y_i-(\beta_0+\beta_1X_i)]^2 \tag{1.2.10}$$

分别对（1.2.10）式求 β_0、β_1 的偏导数并令其等于零，得

$$\left.\begin{aligned}\frac{\partial S(\beta_0,\ \beta_1)}{\partial\beta_0}&=0\\ \frac{\partial S(\beta_0,\ \beta_1)}{\partial\beta_1}&=0\end{aligned}\right\} \tag{1.2.11}$$

由（1.2.11）式，得

$$\left.\begin{aligned}\sum_{i=1}^{n}Y_i&=\beta_0 n+\beta_1\sum_{i=1}^{n}X_i\\ \sum_{i=1}^{n}X_iY_i&=\beta_0\sum_{i=1}^{n}X_i+\beta_1\sum_{i=1}^{n}X_i^2\end{aligned}\right\} \tag{1.2.12}$$

我们称(1.2.12)式为求回归参数 β_0、β_1 的正规方程组。由（1.2.12）式，得

$$\beta_1=\frac{\sum_{i=1}^{n}X_iY_i-n\bar{X}\bar{Y}}{\sum_{i=1}^{n}X_i^2-n\bar{X}^2} \tag{1.2.13}$$

$$\beta_0=\bar{Y}-\beta_1\bar{X} \tag{1.2.14}$$

其中

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$$

$$\bar{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i$$

(1.2.13)式、(1.2.14)式给出了回归参数 β_0、β_1 的最小平方估计量 $\hat{\beta}_0$、$\hat{\beta}_1$。现在，我们可以给出(1.2.4)式的估计关系式

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \qquad i = 1,2,\cdots,n \tag{1.2.15}$$

其中：$\hat{Y}_i$ 代表 $E(Y_i|X_i)$ 的估计量。(1.2.15)式的图形是一条回归直线，$\hat{\beta}_0$ 代表直线的截距，$\hat{\beta}_1$ 代表直线的斜率，如图1—1所示。

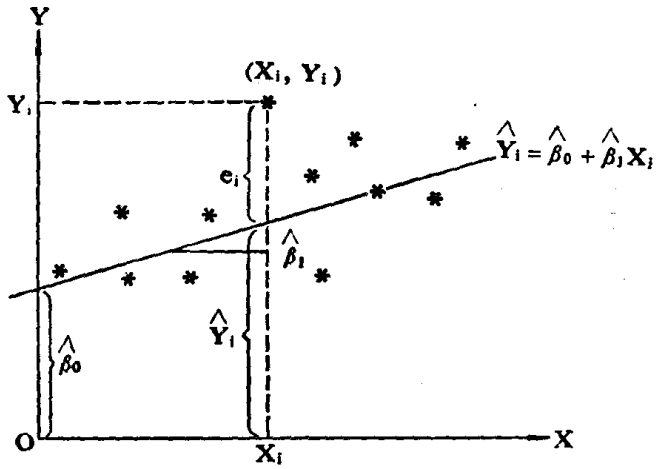

图1—1 回归直线

在图1—1中，e_i 是估计残差，它是 ε_i 的估计量。上述最小平方法估计参数 β_0、β_1 的过程也可以等价地表述为：我们要寻找 β_0、β_1 的这样一种估计量 $\hat{\beta}_0$、$\hat{\beta}_1$，使得残差平方和 $\sum_{i=1}^{n} e_i^2$ 达到最小，即

$$\sum_{i=1}^{n} e_i^2 = \min$$

有的教科书在叙述最小平方法时介绍的是后一种方法，但是，它来源于前一种方法。从图1—1易见

$$\begin{aligned} Y_i &= \hat{Y}_i + e_i \\ &= \beta_0 + \beta_1 X_i + e_i \qquad i = 1,2,\cdots,n \end{aligned} \tag{1.2.16}$$

即 $e_i = Y_i - \hat{Y}_i \qquad i = 1,2,\cdots,n$

可以证明（本书证明过程略）e_i 也是一个均值为零、方差为 σ^2 的正态随机变量

$$e_i \sim N(0,\ \sigma^2) \qquad i = 1,2,\cdots,n \tag{1.2.17}$$

令

$$\begin{aligned} x_i &= X_i - \bar{X} \\ y_i &= Y_i - \bar{Y} \\ i &= 1,2,\cdots,n \end{aligned} \tag{1.2.18}$$

x_i、y_i 分别代表变量 X_i、Y_i 的离差，（1.2.18）式代表对变量 X_i、Y_i 进行中心化的过程，这种形式我们以后要经常使用。对变量进行中心化以后，可以简化计算，使许多表达式更加简单明了。如对（1.2.13）式，我们可以表示为

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \tag{1.2.19}$$

其中

$$\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

$$= \sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}$$

$$\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2$$