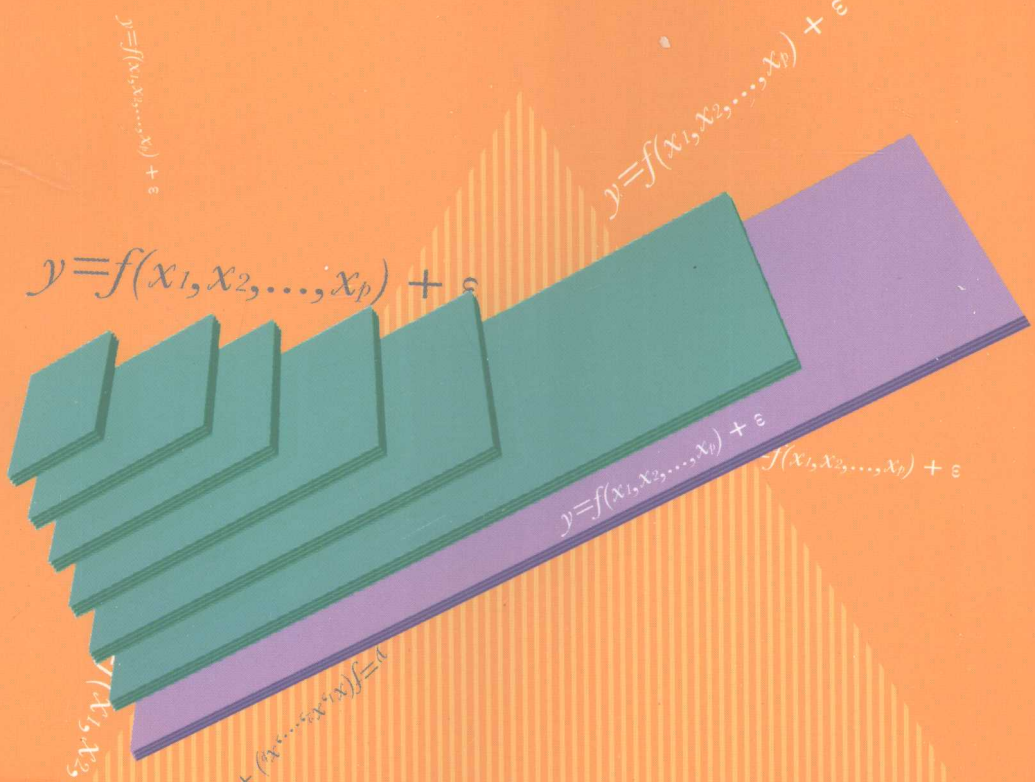
顾问 王静龙 艾春荣 徐国祥 周 勇

21世纪
高校统计学专业教材系列

应用回归分析

王黎明 陈 颖 杨 楠 编著



復旦大學出版社

21世纪

高校统计学专业教材系列

应用回归分析

王黎明 陈 颖 杨 楠 编著

復旦大學出版社

图书在版编目(CIP)数据

应用回归分析/王黎明,陈颖,杨楠编著. —上海:复旦大学出版社,2008.6
(博学·21世纪高校统计学专业教材系列)
ISBN 978-7-309-06058-4

Ⅰ.应… Ⅱ.①王…②陈…③杨… Ⅲ.回归分析-高等学校-教材
Ⅳ.0212.1

中国版本图书馆CIP数据核字(2008)第067278号

应用回归分析

王黎明 陈 颖 杨 楠 编著

出版发行 复旦大学出版社 上海市国权路579号 邮编200433
86-21-65642857(门市零售)
86-21-65100562(团体订购) 86-21-65109143(外埠邮购)
fupnet@fudanpress.com http://www.fudanpress.com

责任编辑 王联合
出 品 人 贺圣遂

印 刷 上海肖华印务有限公司
开 本 787×960 1/16
印 张 19.25
字 数 315千
版 次 2008年6月第一版第一次印刷
印 数 1—4 100

书 号 ISBN 978-7-309-06058-4/O·412
定 价 32.00元

preface

总 序

回顾新中国成立以来我国统计学科的发展道路，可以说是充满着曲折，也充满着希望。20世纪50年代照搬苏联模式，政府统计实务代替了统计学，否定了统计学作为一门方法论科学的存在。统计学的科学性和普遍应用性被曲解。这种情况直到1978年科学的春天来了之后才有了根本的改变。党的改革开放的政策使得统计学在中国得到了迅猛的发展。20世纪末国家教育部新颁的大学本科专业目录中统计学被确立为一级学科。这就要求我们做好统计学专业课程建设的工作。为此上海财经大学统计学系与复旦大学出版社携手，会同国内相关院校及港台相关名校专家，策划出版一套主要针对统计学专业本科生使用的、适应新时期需要的系列教材——博学·21世纪高校统计学专业教材系列。

本教材系列力求体现以下特点：

第一，教材主要考虑面向财经类统计学专业，同时也要考虑“大统计学”专业的需求，力求选材做到“精”和“新”。

第二，内容选择将广泛吸收国内外优秀教材的成果，在系统介绍基本理论和基本方法的同时，注意介绍新的、成熟的内容，以及统计学在实际问题中的应用。

第三，教材编写注重计算机的应用，根据教材的具体内容选讲相应的统计

软件，提高学生熟练运用统计方法和计算机技术解决实际的能力。

博学·21世纪高校统计学专业教材系列前期规划教材包括《统计学》、《数理统计学》、《应用回归分析》、《国民经济核算原理》、《应用时间序列分析》、《统计计算》、《SAS数据分析系统教程》、《非参数统计学》、《变点统计分析及其应用》、《多元统计分析》、《金融计量统计学》；后期还将与境内外知名高校专家合作，陆续出版《抽样调查技术》、《贝叶斯统计学》、《实验设计与质量控制》、《统计预测与决策》、《高等数理统计学》和《金融时间序列分析》等。本教材系列的编写大纲和书稿经过教材编写委员会的多次反复论证、认真讨论。感谢参与论证和编写的各位同行，希望他们的辛勤的劳动成果能够得到统计学界同行们的认可，获得同学的欢迎。这套系列教材的不当之处，恳请读者批评指正。为完善财经类统计学专业的教材建设，我们大家一起努力。

王静龙

2008年5月

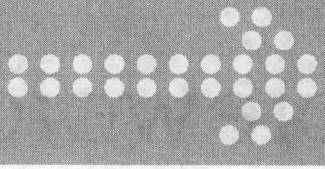
于华东师范大学

foreword

前言

回归分析是统计学中的一个非常重要的分支，是以概率论与数理统计为基础迅速发展起来的一种应用性较强的科学方法。它是由一组探求变量之间关系的技术组成，作为统计学应用最广泛的分支之一，在社会经济各部门以及各个学科领域都能得到广泛的应用。随着我国社会主义现代化建设的发展，人们越来越认识到应用定量分析技术研究问题的重要意义。特别是近些年来计算机及有关统计软件的日益普及，为在实际问题中进行大规模、快速、准确的回归分析运算提供了有力手段。

随着统计学在中国被确立为一级学科，统计学专业的课程设置已有了较大的变化，加强推断统计内容的学习和应用已成为中国统计界的共识。为了适应新的统计学学科体系和财经类统计学专业教学的需要，我们决定编写一套适应新时期需要的系列教材——博学·21世纪高校统计学专业教材。作为系列教材之一，应用回归分析是其中较为重要的一本教材。本书写作的指导思想是：既要保持较为严谨的统计理论体系，又要努力突出实际案例的应用和统计思想的渗透，结合统计软件较全面地系统介绍回归分析的实用方法。为了贯彻这一指导思想，本书将系统介绍回归分析基本理论和方法，在理论上，本书叙述了经典的最小二乘理论，同时又结合应用中出现的一些问题给出了对最小二乘估计的改进方法。中心主题是建立线性回归模型，评价拟合效果，并且作出结论。与此同时，本书也尽力结合中国社会、经济、自然科学等领域的研究实例，把回归分析方法与实际应用结合起来，注意定性分析与定量分析的紧密结合，努力



前言

把同行们以及我们在实践中应用回归分析的经验和体会融入其中。全书分为九章。第一章介绍了一般回归模型的定义，讨论了回归模型的主要任务和回归模型的建模过程。第二章详细地介绍了一元线性回归模型，给出了未知参数的最小二乘估计以及极大似然估计，还讨论了一元线性回归模型的预测问题以及数据变换问题。第三章系统讨论了多元线性回归模型。详细地讨论了最小二乘估计的优良性。对于假设检验，讨论了多元回归模型的显著性检验，以及其回归系数的显著性检验。第四章以残差为重要工具，讨论了回归模型的诊断问题。第五章和第六章讨论了多项式回归模型和含有定性变量的回归模型。第七章讨论了多元线性回归模型的有偏估计。重点介绍较常用的岭估计和主成分估计，同时也介绍其他的估计方法。第八章简单介绍了非线性回归模型，主要讨论了 Logistic 回归模型、Poisson 回归和广义线性模型。本书的最后一章介绍 SAS 统计软件在回归分析中的应用。本书可以作为统计学、数学以及经济学等专业的教材，学习本课程的学生需要熟悉随机变量、参数估计、区间估计、假设检验等思想，也要熟悉正态分布及其由它导出的分布，当然，学生也要具备微积分和线性代数知识。由于本书的内容较多，教师在选用此书作教材时可以灵活选讲。本书也可以作为非统计学专业研究生回归技术的教材。根据我们多年的教学实践，本书讲授 51 课时较为合适，若有计算机和投影设备的配合，教学将会更为方便和有效。在本书的写作过程中，始终得到了博学·21 世纪高校统计学专业教材编委会和复旦大学出版社的支持，编写大纲和书稿都经过教材编写委员会的多次认真讨论。

本书是我们多年教学和科研工作的积累，其中部分案例为体现其典型性引用了他人著作。在此，我们谨向对本书出版给予帮助的同行和朋友表示衷

心的感谢。本书的完成也是我们多年友好合作的结果，研究生苏艳和万里同学也参加了部分习题和案例的编写和整理工作，同时也参加了最后的统稿和校对工作。由于编者的水平有限，在取材及其结构上，难免会存在不够妥当的地方，错误之处也在所难免，恳请同行专家和广大读者能给我们宝贵的批评和建议。

王黎明　陈颖　杨楠

2008年2月

于上海财经大学

contents

目 录

第一章

回归分析的一般介绍

§1.1 变量间的统计关系

社会经济领域与自然科学等诸多现象之间始终存在着相互联系和相互制约的普遍规律。比如社会经济的发展与一定的经济变量的数量变化密切联系，社会经济现象不仅同与它有关的现象构成一个普遍联系的整体，同时在其内部也存在着彼此关联的因素，在一定的社会环境等诸多条件的影响下，一些因素推动或制约另外一些与之关联的因素发生变化。也就是说，社会经济现象的内部和外部联系中存在一定的相关性，要认识和掌握客观经济规律就必须探求经济现象间经济变量的变化规律，变量间的统计关系是经济变量变化规律的重要内容。

这些互相联系的经济现象和经济变量，其联系的紧密程度也是互不相同的。这中间极端的关系就是确定性关系，即一个变量的变化完全确定另外一个变量的变化。例如一个保险公司承保汽车5万辆，每辆保费收入是1 000元，则该公司汽车承保总额为5 000万元。即承保总收入为 y，承保汽车数为 x，则变量 y 和 x 的关系可以表示为：$y=1\,000\cdot x$。见图1.1。

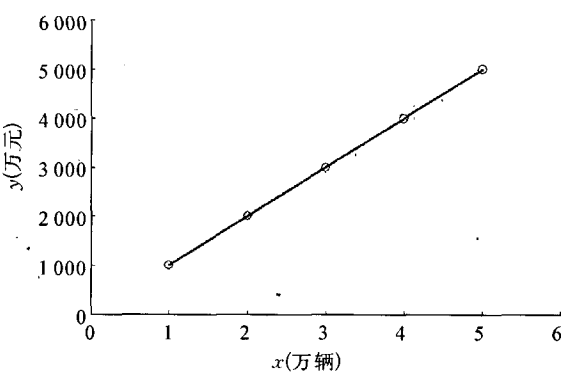


图1.1 函数关系图

从这个例子可以看出，每给定一个 x，就一定可以得到一个 y，即变量 y 与 x 之间完

全表现为一种确定性的关系——函数关系。在实际生活中，这样的例子还有很多。比如银行的一年存款利率为年息2.75%，存入的本金用x表示，到期的本息用y表示，则y与x有函数关系$y=(1+0.0275)x$，这里y与x仍具有线性函数关系，对于任意两个变量y与x的函数关系，可以表示为数学形式：$y=f(x)$。

一般而言，给定p个变量x_1，…，x_p，就可以确定变量y，称这种变量之间的关系为确定性关系。它往往可以用某一函数关系$y=f(x_1, \cdots, x_p)$来表示。

可是，在实际问题中，变量之间存在大量非确定的关系，它们之间虽存在着密切联系，但是其密切程度不是由确切关系能够刻画的。

为此，我们再看一个例子：

由日常生活，我们知道某种高档品的消费量(y)与城镇居民的收入(x)有密切关系。居民收入高了，这种消费品的销售量就大；居民收入低了，这种消费品的销售量就小。但是居民的收入并不能完全确定该种高档品的消费量。因为，商品的消费量还受着人们的消费习惯、心理因素、其他可替代商品的吸引程度以及价格的高低等诸因素的影响。也就是说，城镇居民的收入与该种高档品的消费量有着密切关系，且城镇居民的收入对该种高档品的消费量的多少起着主要作用，但是它并不能完全确定该种高档品的消费量。

在日常生活中，变量与变量之间表现为这种关系的有很多，比如粮食产量与施肥量之间的关系，银行储蓄额与居民收入之间的关系。

把以上概括为：变量x与变量y有密切关系，但是又没有密切到可以通过一个变量可以确定另一个变量的程度。它们之间是一种非确定性的关系，我们称这种关系为**统计关系**或**相关关系**。

应该指出的是，变量之间的函数关系和相关关系，在一定条件下是可以互相转化的。本来具有函数关系的经济变量，当存在观测误差时，其函数关系往往以相关的形式表现出来。而具有相关关系的变量之间的联系，如果我们对它们有了深刻的规律性认识，并且能够把影响因变量变动的因素全部纳入方程，这时的相关关系也可能转化为函数关系。另外，相关关系也具有某种变动规律性，所以，相关关系经常可以用一定的函数形式去近似地描述。经济现象的函数关系可以用数学分析的方法去研究，而研究社会经济现象的相关关系必须借助于统计学中的相关与回归分析方法。

回归分析就是讨论变量与变量之间的统计关系的一种统计方法。

§1.2 回归模型的一般形式

假设因变量 y 与一个或多个自变量 x_1，x_2，…，x_p 之间具有统计关系，我们把 y 称为因变量、响应变量或被解释变量，x_1，x_2，…，x_p 称为自变量、预报变量或解释变量。我们可以设想 y 由两部分组成，一部分由 x_1，x_2，…，x_p 能够决定，记为 $f(x_1, x_2, \cdots, x_p)$，另一部分由众多未加考虑的因素（包括随机因素）所产生的影响，它被看成随机误差，记为 ϵ。于是得到了如下统计模型：

$$y = f(x_1, x_2, \cdots, x_p) + \epsilon \tag{1.1}$$

其中，ϵ 称为随机误差，一般要求它的数学期望为 0，它的出现使得变量间关系的相关性得以恰当体现；$f(x_1, x_2, \cdots, x_p)$ 称为 y 对 x_1，x_2，…，x_p 的回归函数，或称为 y 对 x_1，x_2，…，x_p 的均值回归函数；(1.1)式称为回归模型的一般形式。

模型(1.1)清楚地表达了变量 x_1，x_2，…，x_p 与随机变量 y 的相关关系，数理统计学中的“回归”通常指散点分布在一直线（或曲线）附近，并且越靠近该直线（或曲线），点的分布越密集的情况。它也称为**直线（或曲线）的拟合**。

当概率模型(1.1)中的回归函数为线性时，(1.1)式变成为

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \tag{1.2}$$

其中 β_0，β_1，…，β_p 为未知参数。常称 β_0 为回归常数，β_1，…，β_p 为回归系数。这时我们称(1.2)式为**线性回归模型**。

在实际应用中，β_0，β_1，…，β_p 一般皆是未知的，为了应用需要将它们估计出来。估计就需要数据，假设样本观测值为 x_{i1}，x_{i2}，…，x_{ip}；y_i，$i = 1, 2, \cdots, n$，则线性回归模型可表示为

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i,\ i = 1, 2, \cdots, n \tag{1.3}$$

假设由这些数据给出了 β_0，β_1，…，β_p 的估计值，分别记为 $\hat{\beta}_0$，$\hat{\beta}_1$，…，$\hat{\beta}_p$。称

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p \tag{1.4}$$

为经验回归方程。

如果给定一组 x_1，x_2，…，x_p，由(1.4)式可以得到一个 y，记为 $\hat{y}$，$\hat{y}$ 称为 y 的一个**预测值**。

对模型(1.4),通常规定满足的基本假设有:

(1) 变量 x_1, x_2, …, x_p 是非随机变量,观测值 x_{i1}, x_{i2}, …, x_{ip} 是常数。

(2) 高斯-马尔可夫(Gauss-Markov)条件:G-M 条件(等方差及不相关的假定)

$$\begin{cases} E(\varepsilon_i) = 0, \ i = 1, 2, 3, \cdots, n \\ \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} 0, & i \neq j \\ \sigma^2, & i = j \end{cases} \end{cases}$$

(3) 正态分布的假定条件为

$$\begin{cases} \varepsilon_i \sim N(0, \sigma^2) \\ \varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n \text{相互独立} \end{cases}$$

对线性回归模型,通常要研究的问题有:

1. 如何根据样本 x_{i1}, x_{i2}, …, x_{ip}; y_i, $i = 1, 2, \cdots, n$, 求出 β_1, β_2, …, β_p 及方差 σ^2 的估计。

2. 对回归方程及回归系数的种种假设进行检验。

3. 如何根据回归方程进行预测和控制,以便进行实际问题的结构分析。

回归分析方法在生产实践中的广泛应用是它发展和完善的根本动力。如果从19世纪初(1809年)高斯(Gauss)提出最小二乘法算起,回归分析已有近200年的历史。从经典的回归分析方法到近代的回归分析方法,它们所研究的内容已非常丰富。如果按研究的方法来划分,回归分析研究的范围大致如下,见图1.2。

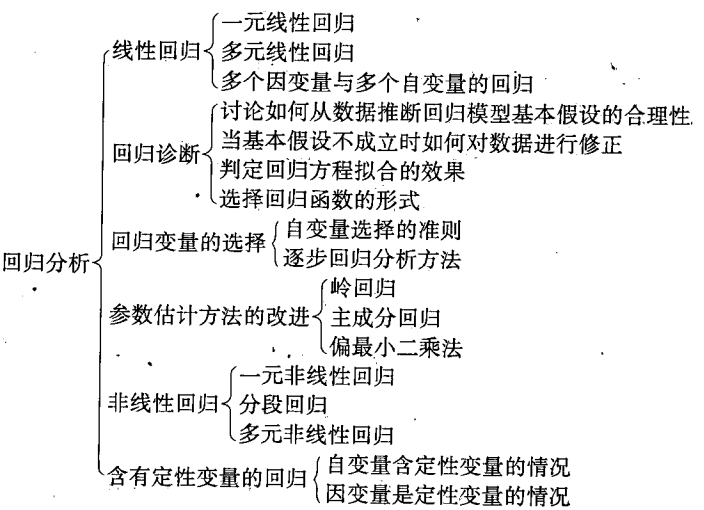


图 1.2 回归分析研究的范围

§1.3 回归方程与回归名称的由来

回归分析是处理变量 x 与 y 的关系的一种统计方法和技术。这里所研究的变量之间的关系是当给定 x 的值，y 的值不能确定，只能通过一定的概率分布来描述。于是，我们称给定 x 时 y 的条件数学期望

$$f(x)=E(y|x)$$

为随机变量 y 对 x 的回归函数，或称为随机变量 y 对 x 的均值回归函数。上式从平均意义上刻画了变量 x 与 y 之间的统计规律。在实际问题中，我们把 x 称为自变量，y 称为因变量。如果要由 x 预测 y，就是要利用 x，y 的观察值，即由样本观测值 (x_1, y_1)，(x_2, y_2)，…，(x_n, y_n) 来建立一个公式，当给定 x 值后，就代入此公式中算出一个 y 值，这个值就称为 y 的预测值。

“回归”一词的英文是“regression”，其基本思想和方法都是由英国著名生物学家、统计学家 F·高尔顿(F. Galton，1822—1911)在研究人类遗传问题时提出的。为了研究父代与子代身高的关系，高尔顿和他的学生、现代统计学的奠基者之一 K·皮尔逊(K. Pearson，1856—1936)在研究父母身高与其子女身高的遗传问题时，观察了 1 078 对夫妇，以每对夫妇的平均身高作为 x，而取他们的一个成年儿子的身高作为 y，将结果在平面直角坐标系上绘成散点图，发现趋势近乎一条直线。计算出的回归直线方程为

$$\hat{y}=33.73+0.561x$$

这种趋势及回归方程总的表明父母平均身高 x 每增加一个单位时，其成年儿子的身高 y 也平均增加 0.516 个单位。人们自然会这样想：若父亲身高为 x 英寸，其儿子身高应为 $x+1$ 英寸。但是所得的结论与此大相径庭。高尔顿发现：$x=72$ 英寸(大于平均身高 68 英寸)时，他们的儿子平均身高为 71 英寸，不但达不到 $72+1=73$ 英寸，反而比父亲低了 1 英寸；反过来，$x=64$ 英寸(小于平均身高 68 英寸)时，他们的儿子平均身高为 67 英寸，竟比预期的 $64+1=65$ 英寸高出 2 英寸。这个结果表明，虽然高个子父辈确有生高个子儿子的趋势，但父辈身高增加一个单位，儿子身高仅增加半个单位左右。反之，矮个子父辈确有生矮个子儿子的趋势，但父辈身高减少一个单位，儿子身高仅减少半个单位左右。通俗地说，一群特高个子父辈(例如排球运动员)的儿子们在同龄人中平均仅为高个子；一群高个子父辈的儿子们在同龄人中平均