高等学校现代统计学系列

实用回归分析（第二版）

何晓群　闵素芹　编著

Applied Regression Analysis

高等教育出版社

高等学校现代统计学系列教材

Applied Regression Analysis

实用回归分析

Shiyong Huigui Fenxi

（第二版）

何晓群　闵素芹　编著

高等教育出版社·北京

内容提要

回归分析是现代统计学中应用较为活跃的模型分析技术。本书旨在提高社会、经济、管理类本科生的量化分析水平，选择众多的回归分析方法中最为实用的基本模型分析技术，结合社会经济与管理中的实际问题，利用SPSS统计软件对回归建模分析方法作了系统介绍。

本书既可作为高等学校统计学类专业教材，也可作为人文社会科学、财经管理类专业工作者的参考书。

图书在版编目(CIP)数据

实用回归分析／何晓群，闵素芹编著. --2版. --北京：高等教育出版社，2014.5

ISBN 978-7-04-039553-2

Ⅰ.①实… Ⅱ.①何…②闵… Ⅲ.①回归分析-高等学校-教材 Ⅳ.①O212.1

中国版本图书馆CIP数据核字(2014)第090103号

策划编辑 张晓丽	责任编辑 张晓丽	特约编辑 张让让	封面设计 赵 阳
版式设计 余 杨	插图绘制 于 博	责任校对 胡美萍	责任印制 韩 刚

出版发行	高等教育出版社	网 址	http://www.hep.edu.cn
社 址	北京市西城区德外大街4号		http://www.hep.com.cn
邮政编码	100120	网上订购	http://www.landraco.com
印 刷	保定市中画美凯印刷有限公司		http://www.landraco.com.cn
开 本	787mm×960mm 1/16		
印 张	18.75	版 次	2008年5月第1版
字 数	330千字		2014年5月第2版
购书热线	010-58581118	印 次	2014年5月第1次印刷
咨询电话	400-810-0598	定 价	29.50元

物料号 39553-00

高等学校现代统计学系列教材编委会

（按姓氏笔画排序）

总　序

统计学是一门收集、整理和分析数据的科学和艺术。这里的“数据”通指“信息的载体”，涵盖了大千世界中的文本、图像、视频、时空数据、基因数据等。统计学是一个独立的学科，在历史上曾隶属于数学，但统计学与数学有着本质的区别，因此统计学教育有其自身的特点和要求，这些特点表现为：(1) 统计学研究的是随机现象，而数学研究的是确定性的规律；(2) 统计学是一门应用性很强的学科，许多概念和原理来自于实际的需要，不是数理逻辑的产物；(3) 数据在统计学中扮演了重要的角色。目前，统计学已被列为一级学科。

在过去的30年中，随着生命科学、信息科学、物质科学、资源环境、认知科学、工程技术、经济金融和人文科学等众多学科的发展，产生了许多新的统计学分支，如风险管理、数据挖掘、基因芯片分析等。此外，计算机及其有关软件在统计教育和应用中扮演了越来越重要的角色，它们提供了越来越多的图形表达和分析的方法，使得许多原来教科书中重要的内容，现在已变得无足轻重。统计教育必须要改革才能适应高速发展的形势。

大学的统计教育可分为两大类，一类是非统计学专业的课程，另一类是统计学专业的教学设计。非统计学专业的学生学习统计的目的是为了应用，在大学阶段，课程不多，主要是学习基础的统计概念和方法，学会使用统计软件，培养其解决实际问题的能力。统计学专业的课程设置十分重要，应向国际靠拢，对教师队伍的要求也较高。虽然这两类学生的教育有很多共同点，但在课程设置中必须加以区分。

我国的统计教育在过去受苏联的影响很深，把统计学作为数学的一个分支，在内容上偏理论，少应用，过于强调概率论在统计中的作用。统计学是一门应用性很强的学科，应从实际问题、从数据出发，通过统计的工具来揭示数据内部的规律。用“建模”的思路来教统计，使学生能更加容易理解统计的概念和方法，知道如何将实际问题抽象为统计模型，反过来又指导实践。对非统计学专业的学生，要强调统计的应用。学生要能熟练地使用至少一个统计软件包。对于统计学专业的学生，要培养学生对实际问题的建模能力。有些实际问题可直接应用现有的统计方法来解决，如问卷调查的统计分析。有些问题在初次接触时并不像一个统计问题，必须有坚实的统计基础和对实际问题的洞察力，才能从中发掘出统计模型。要培养学生的这种能力及统计思想(统计思想是统计文化的一

部分,是用统计学的逻辑思考问题)。教师在授课中要结合较多的应用例子,要求学生做案例研究,鼓励学生参加建模比赛,参加企业的实际项目。

为满足我国统计教育发展的需要,我们计划编写一套面向高校本科生、特别是一般院校,适用于统计学专业和非统计学专业的系列教材。系列教材的编写宗旨是:突出教学内容的现代化,重视统计思想的介绍,适应现代统计教育的特点及时代发展的新要求;以统计软件为支撑,注重统计知识的应用;内容简明扼要,生动活泼,通俗易懂。编写原则为:(1) 从数据出发,不是从假设、定理出发;(2) 从归纳出发,不是从演绎出发;(3) 强调案例分析;(4) 重统计思想的阐述,弱化数学证明的推导。系列教材分为两个方向,一个面对统计学专业,另一个面对非统计学专业和应用统计工作者。

高等学校现代统计学系列教材是适应形势的要求,由高等教育出版社邀请专家组成"高等学校现代统计学系列教材编委会"负责选题、审稿,由高等教育出版社出版。

以上是我们编写这套教材的背景和理念,希望得到读者的支持,特别是高校领导和教学一线的教师的支持。我们希望使用这套教材的师生和读者多提宝贵意见,使教材不断完善。

高等学校现代统计学系列教材编委会

第二版前言

《实用回归分析》一书自2008年5月出版以来，受到数万读者的关注，许多学校将其作为应用统计类本科生教材。一些老师在使用过程中也提出了许多宝贵意见和中肯建议，这些都促使我们尽快对本书做出修订。本次修订更加突出了回归模型的实际应用与统计软件的结合。比起第一版我们主要做了以下工作：

1. 例题增加了SPSS软件操作的详细步骤，以便读者更轻松地掌握数据分析方法，迅速运用到实际工作中。SPSS软件采用IBM SPSS Statistics 20.0英文版。

2. 将第9章中SAS软件涉及的内容修订为基于SPSS软件运行。

3. 教材中的例题与习题数据更新至2010年。

4. 第5章、第6章和第9章中的个别例题进行了更换，与相对应的理论更加契合。

5. 纠正了第一版中的个别错误。

6. 习题参考答案更加清晰与详尽。

在高等教育出版社李蕊、张晓丽女士的精心策划下，本书的再版得到中国现场统计研究会多元统计分析专业委员会许多同仁的支持和帮助。西京学院执行院长任芳博士为笔者提供了时间和资金方面的大力支持。加盟本书修订的中国传媒大学理学院副教授闵素芹博士做了大量工作，如果修订之后本书有了增色则大都是闵博士所为。在此我们谨对本书出版有帮助的朋友表示衷心的感谢！

由于我们的学术水平和教学实践还有待提高，书中难免仍有不妥之处，恳切期望读者提出批评意见。

何晓群

2013年3月1日

于西京学院应用统计科学研究中心

第一版前言

面对21世纪深刻的社会变革和迅猛的经济发展，我国的高等教育面临严峻的挑战和难得的机遇。时代呼唤我们的学生学习一些量化分析的研究方法，掌握定性与定量有机结合的研究技能。《实用回归分析》一书正是适应这一需要而编写的。

统计方法与技术是许多学科研究运用的基本方法。学习和运用统计方法已成为时代对我们的要求。以经济学为例，现代经济学一个很重要的标志就是模型技术的应用，而这里的模型技术更多的是指统计模型技术。诺贝尔经济学奖获得者萨缪尔森(Samuelson)曾说，第二次世界大战以后的经济学是计量经济学的时代。计量经济学的模型技术就是以统计学中的回归模型为基础的。自1969年诺贝尔经济学奖设立以来，已有70多位学者获奖，其中2/3以上获奖者是统计学家、计量经济学家和数学家。从大多数获奖者的经历和著作看，他们对统计方法的运用具有娴熟的技巧。瑞典皇家科学院爱立克·伦德伯教授在诺贝尔经济学奖的授奖仪式上精辟地指出："你们都是把经济学发展为数学和定量科学的先行者，你们借助于发展成熟的理论和统计分析来创造经济政策和经济规划的合理基础的贡献，涉及重大科学突破。"这足以说明统计量化分析研究已成为现代经济科学研究的重要手段。

笔者假定学生已具有线性代数、概率论与数理统计的基础知识，本着提高学生量化分析能力的宗旨，在众多统计方法中，仅选择部分最实用的回归分析方法。在不失理论严密性的前提下，力求将问题的背景、方法的思想与原理、具体的步骤、分析的技巧讲清楚。为重点突出方法的思想和应用，每种方法尽可能结合中国社会、经济、管理方面的实际问题，以案例研究为导向，运用现在流行的统计软件SPSS或Excel为进行量化分析起一定示范作用。

本书的编写主要是根据笔者过去编写的《回归分析与经济数据建模》、《应用回归分析》、《现代统计分析方法与应用》中的材料重新组合而成。根据笔者的经验，如有计算机配合，学生掌握这些基本方法和技能并不困难。选用本书的教师可有一定的灵活性，根据不同专业有选择地讲授该书内容。本书参考教学学时为54学时。

本书也可作为现代统计分析方法课的教材。此书还可作为从事社会、经济、管理等研究和实际工作的同志进行量化研究的参考书。

由于本人学识有限，书中谬误之处在所难免，恳请读者批评指正。

中国人民大学应用统计科学研究中心
中国人民大学六西格玛质量管理研究中心
何晓群

2007 年 10 月 1 日
于北京九鼎山庄长安仁居

目　录

第1章

统计学基础

为了更顺利地学习本课程的内容，本章将对统计学中的一些基本概念和术语作一简要回顾.

§1.1 统计数据的整理与描述

统计学是研究数据规律的方法论学科，统计数据是统计学研究的主要内容. 借助统计学方法研究任何实际问题，首先要做的工作就是收集数据，收集数据是一项很重要的基础工作. 收集数据的一般方法是查阅各种统计年鉴和报表，再就是运用某种调查方法获取欲研究问题的有关数据. 抽样调查获取数据的方式在我国方兴未艾，抽样调查的方法很多，专业性很强，现在已有不少抽样技术的专著. 需要利用抽样方法获取数据的研究者，还需很好地学习有关抽样技术的知识.

一、总体与样本

在一个统计问题中，通常把所要调查研究的事物或现象的全体称为总体，而把组成总体的每个元素(成员)称为个体，一个总体中所含的个体的数量称为总体的容量. 例如要研究某城市居民的家庭收入状况，那么这个城市所有家庭的收入状况就是我们研究的总体，而每个家庭的收入状况就是个体.

为了推断总体的某些特征，需要从总体中按一定的抽样技术抽取若干个体，将这一抽取过程称为抽样. 所抽取的部分个体称为样本，样本中所含个体的数量称为样本容量. 如在研究居民家庭收入时，随机抽取 1 000 户来进行调查，这 1 000户就是一个样本，样本容量就是 1 000.

二、统计量

通过抽样或查统计年鉴得到的原始数据，一般是杂乱无章的，很难从中直接看出有价值的东西. 因此，对获取的原始数据一般需要加以整理，以便把人们感兴趣的信息提取出来，并用简明醒目的方式加以表述. 画原始数据的散点图、饼

图、直方图等方法是直观表达数据的常见方式.统计学中最主要的提取信息方式就是对原始数据进行一定的运算,以算出某些代表性的数字,足以反映出数据某些方面的特征,这种数字被称为统计量.用统计学语言表述就是:统计量是样本的函数——它不依赖于任何未知参数.

例如样本均值和样本方差就是最重要的常用统计量.

均值是对数据集中特征的描述,方差是对数据波动特征的描述.

设 $x_1, x_2, \cdots, x_n$ 是一组独立的随机样本,则样本均值为

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

样本方差为

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2,$$

样本标准差为

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

例如有两组数据

$$(4,6,8,10,12), \quad (6,7,8,9,10),$$

它们的均值 $\bar{x}$ 都是8,这说明两组数据都以8为中心.读者可计算它们的方差,第一组数据的方差比第二组的要大,说明第一组数据相对均值8来说比较分散,第二组数据相对均值8来说比较集中.由这两组数据可以很直观地看出均值及方差的意义.

需要注意的是,方差带单位是没有实际意义的,只有标准差带上单位才有实际意义.

三、变异系数

如果两组数据的计量单位相同,且均值一样,可以利用标准差来比较两组数据的离散程度.但当两组数据的计量单位不同或均值不同时,就不能直接利用比较两组数据的标准差来分析两组数据的离散程度.由此引入变异系数 V:

$$V = \frac{S}{\bar{x}}.$$

例如,(4,5,6,7,8)与(40,50,60,70,80)两组数据的标准差分别是1.58和15.8,如果仅从标准差来看,显然第二组数据分散程度较大.但是由于两组数据的均值不同,分别为6和60,单纯由标准差来判断数据的分散程度就不合适.实际上,当我们算出两组数据的变异系数时,得到 V 都是0.26.比较而言,两组数

据的相对分散程度就是相同的了.

四、偏度与峰度

偏度和峰度是描述统计数据分布偏斜和陡峭程度的统计量.

偏度用偏度系数 V_1 来描述:

$$V_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{S^3(n-1)},$$

其中 S 为样本标准差.

偏度系数 V_1 的意义可由图 1.1 表示.

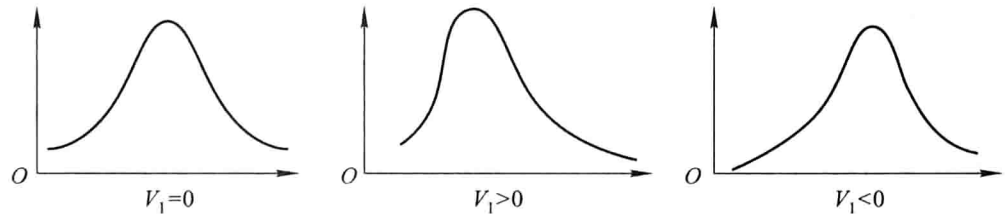

图 1.1　偏度系数不同时的数据分布图

峰度用峰度系数 V_2 表示:

$$V_2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{S^4(n-1)}.$$

当峰度系数 $V_2 = 3$ 时,一般为正态分布,见图 1.2.

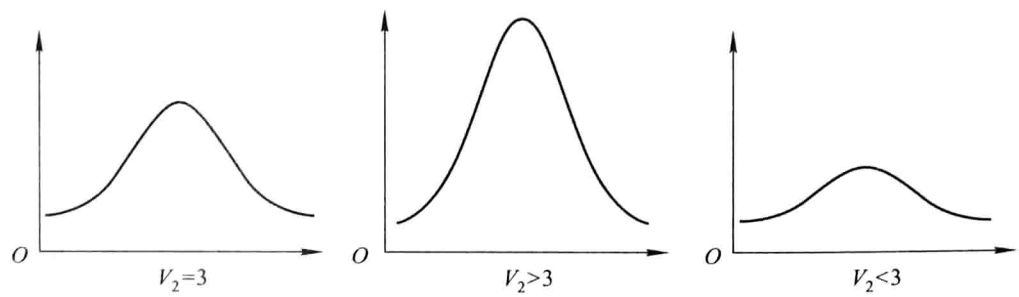

图 1.2　峰度系数不同时的数据分布图

五、累积频数分布

在社会经济调查中,经常会得到频数数据. 例如家庭月收入按等级划分时,就会得到每个等级的家庭数,通常将这些数据列在表中(如表 1.1)或画成直方图.

读者可依收入等级从低到高画出累积频数的直方图.

表 1.1　累积频数分布表

收入等级/元	家　庭　数	
	频数	累积频数
5 000 ~ 6 000	800	800
6 000 ~ 7 000	700	1 500
7 000 ~ 8 000	500	2 000
8 000 ~ 9 000	300	2 300

在社会经济研究中，M. E. Lorentz（洛伦兹）曲线是累积频数的典型应用. 如果按收入从低到高排列，各收入等级的家庭的累积数（百分比）为横坐标，与之相对应的收入的累计（百分比）为纵坐标，所得到的曲线就是西方经济学中著名的 Lorentz 曲线. 在宏观经济的收入与分配研究中，就可运用这一描述方法.

图 1.3 中对角线 OA 是均匀收入分布线. 图中 B 点表明在数量上占全体 40% 的家庭在收入上也占 40%. 收入分布不大可能绝对平均，所以 Lorentz 曲线一般并不是一条直线. 图中 C 点表示从最低收入开始的 40% 的家庭收入的合计还占不到总收入的 20%.

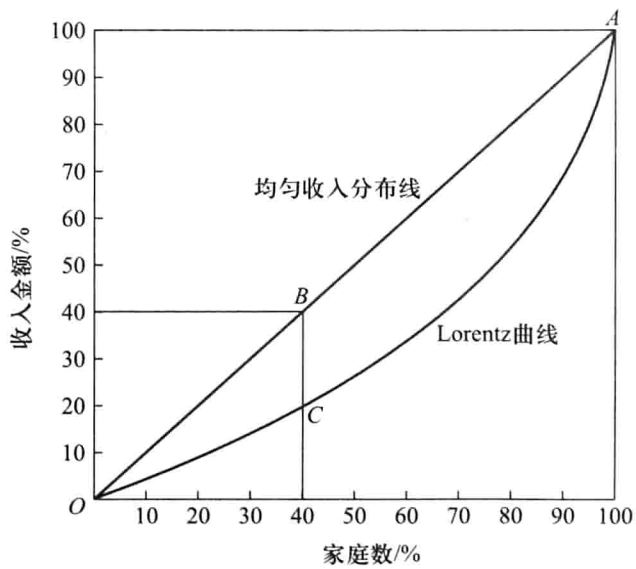

图 1.3　Lorentz 曲线图

关于累积频数的百分比曲线可拓宽到衡量贫富差距的 Gini（基尼）系数. Gini系数理论在中国当今的宏观经济研究中非常有用.

§1.2 几种重要的概率分布

一、正态分布

在经济研究和工商管理中,有许多随机变量的概率分布都可用正态分布来描述.例如一个城市居民的家庭收入、消费支出,某种股票月收益的百分比,某种产品的某质量特性指标都可近似用正态分布来描述.在实际问题的研究中,可以通过该随机变量的抽样数据的频数直方图与正态概率分布的钟形曲线相比较,来判断该随机变量是否为正态随机变量.

正态随机变量 X 的概率密度函数的形式如下:

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}},\qquad -\infty<x<+\infty,$$

其中,μ 为正态随机变量 X 的均值,σ^2 为正态随机变量 X 的方差.

通常对具有均值为 μ,方差为 σ^2 的正态概率分布,记为 $N(\mu,\sigma^2)$.于是有正态随机变量 $X\sim N(\mu,\sigma^2)$.

一般来说,正态分布的密度曲线是以 μ 为中心,在 μ 的两侧呈对称的形状,曲线的形状像一个钟的剖面,故称为钟形曲线.σ 越大,密度曲线的峰度越低;σ 越小,密度曲线的峰度越高.无论参数 μ 和 σ 取何值,密度曲线下所覆盖的面积均等于 1.正态分布的密度曲线见图 1.4.

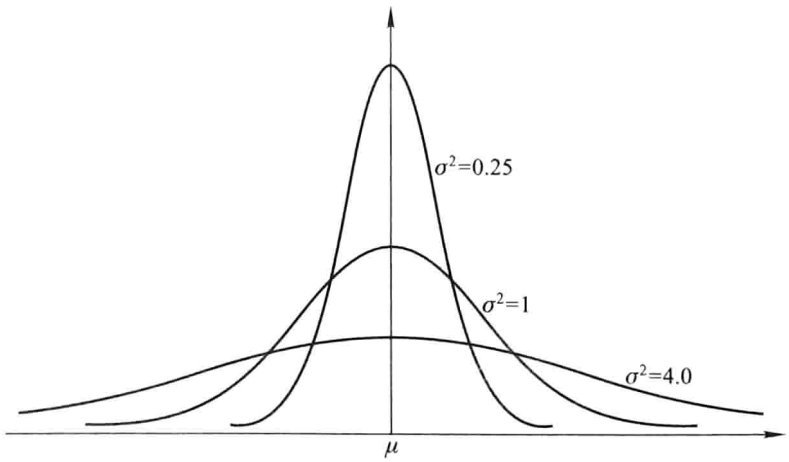

图 1.4 σ^2 不同时正态分布的密度曲线

正态分布曲线下,位于 $\mu\pm\sigma,\mu\pm2\sigma,\mu\pm3\sigma$ 之间的面积分别约占总面积的 68.26%,95.45%,99.73%,如图 1.5 所示.