高等学校统计学类系列教材

实用回归分析

何晓群 编著

高等教育出版社
HIGHER EDUCATION PRESS

高等学校统计学类系列教材

实用回归分析

何晓群　编著

高等教育出版社

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/何晓群编著. —北京：高等教育出版社，2008.5

ISBN 978-7-04-023894-5

Ⅰ. 实… Ⅱ. 何… Ⅲ. 回归分析-高等学校-教材 Ⅳ. O212.1

中国版本图书馆CIP数据核字(2008)第030214号

策划编辑	李 蕊	责任编辑	蒋 青	封面设计	王凌波
责任绘图	杜晓丹	版式设计	王艳红	责任校对	俞声佳
责任印制	宋克学				

出版发行	高等教育出版社	购书热线	010-58581118
社 址	北京市西城区德外大街4号	免费咨询	800-810-0598
邮政编码	100120	网 址	http://www.hep.edu.cn http://www.hep.com.cn
总 机	010-58581000	网上订购	http://www.landraco.com http://www.landraco.com.cn
经 销	蓝色畅想图书发行有限公司	畅想教育	http://www.widedu.com
印 刷	北京地质印刷厂		
开 本	787×960 1/16	版 次	2008年5月第1版
印 张	17.75	印 次	2008年5月第1次印刷
字 数	330 000	定 价	24.20元

物料号 23894-00

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50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：

$$(4,6,8,10,12),$$

$$(6,7,8,9,10),$$

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.

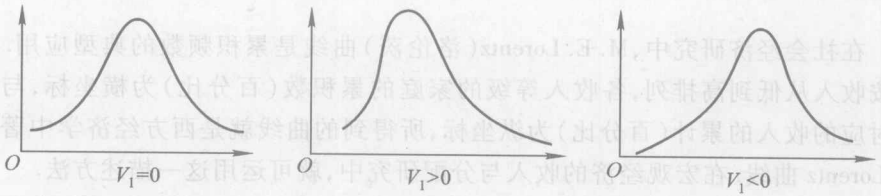

图 1.1

峰度用峰度系数 V_2 表示:

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

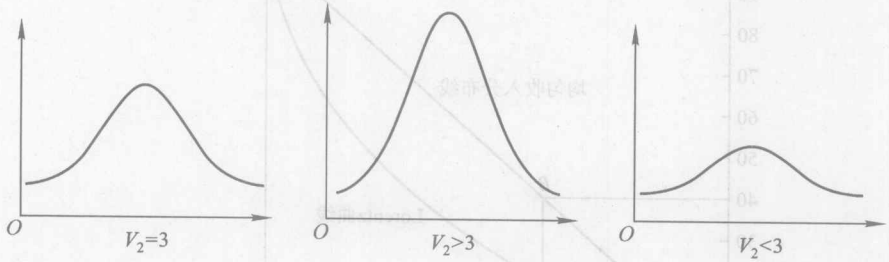

图 1.2

当峰度系数 $V_2 = 3$ 时,一般为正态分布,见图 1.2.

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
6 001 ~ 7 000	700	1 500
7 001 ~ 8 000	500	2 000
8 001 ~ 9 000	300	2 300

在社会经济研究中,M. E. Lorentz(洛伦茨)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%.

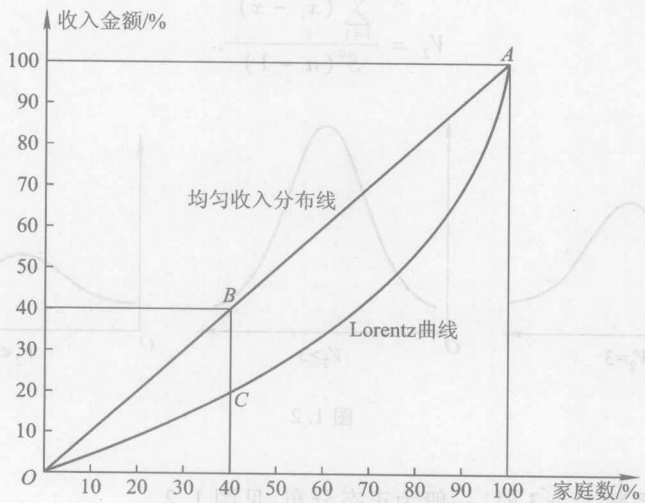

图 1.3

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.

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.

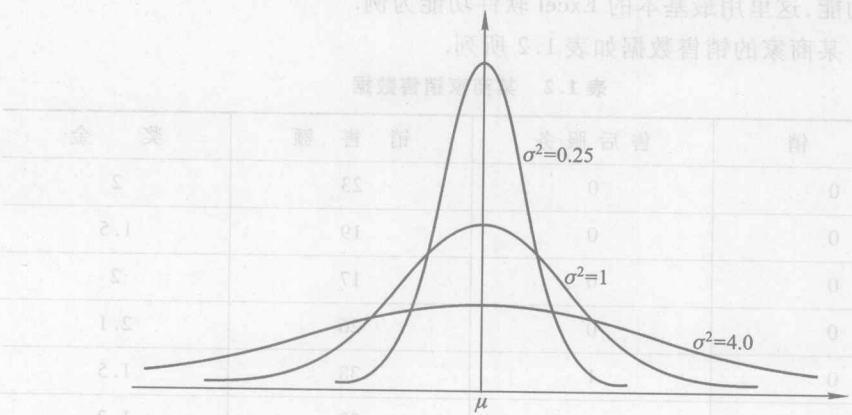

图 1.4 正态分布的密度曲线

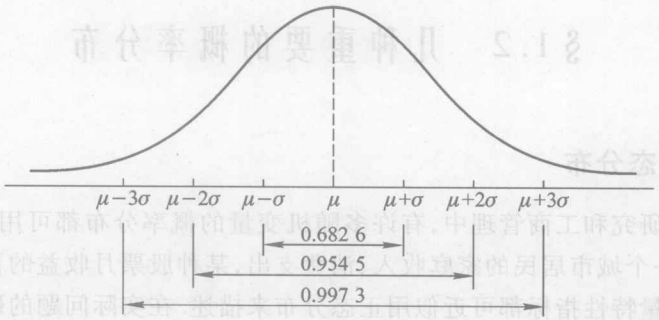

图 1.5

在正态分布的概率密度中，当 $\mu=0,\sigma=1$ 时，称随机变量 X 遵从标准正态分布，记为 $X\sim N(0,1)$.

关于正态分布的理论已很完善，数学上也易于处理. 此外，当一个经济问题的模型误差是由很多因素构成的时候，总体的分布与正态分布近似. 所以，在计量经济学和一些经济问题的建模研究中常假定为正态分布. 另外，当总体概率分布为正态分布时，作为从中抽出的样本，其统计量的样本概率分布有 χ^2 分布、t 分布、F 分布等. 因此，正态分布成为计量经济学乃至统计学中最重要的概念之一.

二、用 Excel 绘制正态概率图进行数据正态性的评估

在实际应用中，经常需要验证数据的正态性，许多统计软件都有数据正态性评估的功能，这里用最基本的 Excel 软件功能为例.

例 某商家的销售数据如表 1.2 所列.

表 1.2 某商家销售数据

促　销	售后服务	销售额	奖　金
0	0	23	2
0	0	19	1.5
0	0	17	2
0	0	26	2.1
0	1	28	1.5
0	1	23	1.2
0	1	24	1.6

续表

促　销	售后服务	销售额	奖　金
0	1	30	1.8
1	0	26	1.8
1	0	22	1.1
1	0	20	0.9
1	0	30	2.1
1	1	36	2.1
1	1	28	1.21
1	1	30	1.91
1	1	32	2.15
2	0	30	1.8
2	0	23	1.2
2	0	25	1.3
2	0	32	1.92
2	1	48	1.7
2	1	40	1.3
2	1	41	1.2
2	1	46	1.81

在进行方差分析之前，必须检验数据的正态性. 选取销售额这一列数据进行正态性评估. 其评估步骤为：

第 1 步：对销售额按照升序排列，记为 $X(j)$.

第 2 步：计算 $(j-0.5)/24$.

第 3 步：根据公式 $(j-0.5)/24 = P(Z \leqslant z_i) = \Phi(z_i)$，求出正态分位数 z_i.

单击 D2 单元格，选择“插入”→“函数”选项，在出现对话框中：

从“选择类别”窗口中选择“统计”.

从“选择函数”窗口中选择“NORMSINV”选项，选择“确定”，其界面如图 1.6.

再单击 D2 单元格，鼠标指向单元格右下角填充控点，按住鼠标左键往下拖至 D25 单元格，这样，计算出 C2:C24 区域中概率值对应的标准正态分位数如图 1.7 所示.

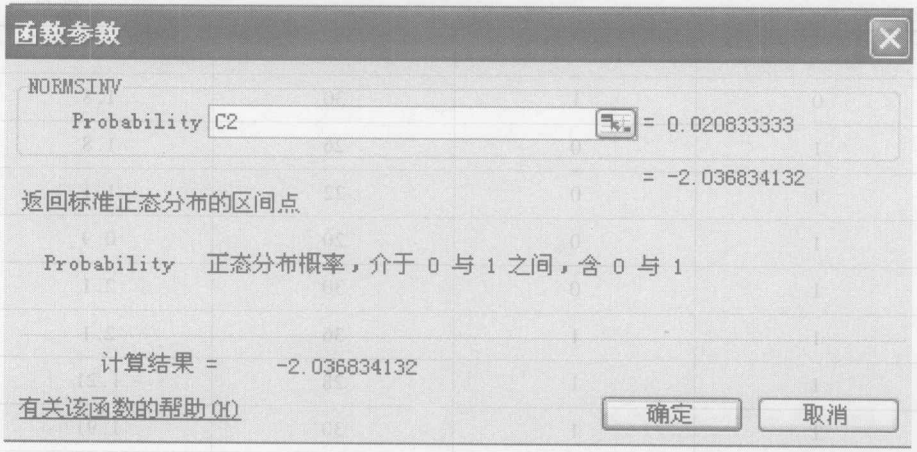

图 1.6

	A	B	C	D
1	j	销售额X(j)	(j-0.5)/24	Zi
2	1	17	0.020833333	-2.036834132
3	2	19	0.0625	-1.534120544
4	3	20	0.104166667	-1.258161561
5	4	22	0.145833333	-1.054472452
6	5	23	0.1875	-0.887146559
7	6	23	0.229166667	-0.741594044
8	7	23	0.270833333	-0.61029461
9	8	24	0.3125	-0.488776411
10	9	25	0.354166667	-0.37409541
11	10	26	0.395833333	-0.264146977
12	11	26	0.4375	-0.157310685
13	12	28	0.479166667	-0.05224518
14	13	28	0.520833333	0.05224518
15	14	30	0.5625	0.157310685

图 1.7

第 4 步：选择“工具”→“数据分析”，在“分析工具”中选择“回归”，单击“确定”. 以 z_i 为纵轴，$X(j)$ 为横轴，绘制标准正态概率图. 界面如图 1.8 所示.

单击“确定”，得到如图 1.9 所示的标准正态概率图.

其中，$X(j)$ 转化为其对应的百分比排位. 从图 1.9 可看出，由 $(X(j), z_i)$ 形成的点基本上在一条直线附近，可以说该组数据基本上遵从正态分布.

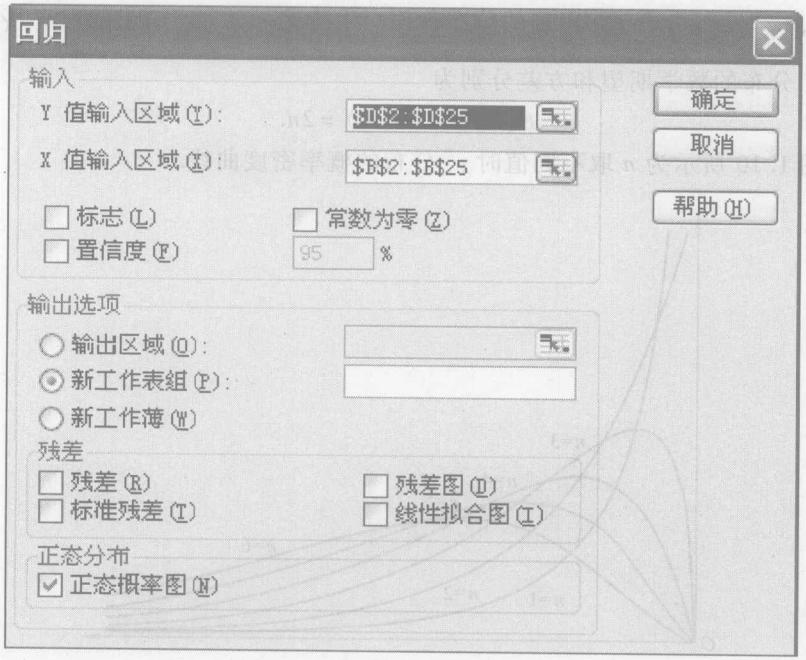

图 1.8

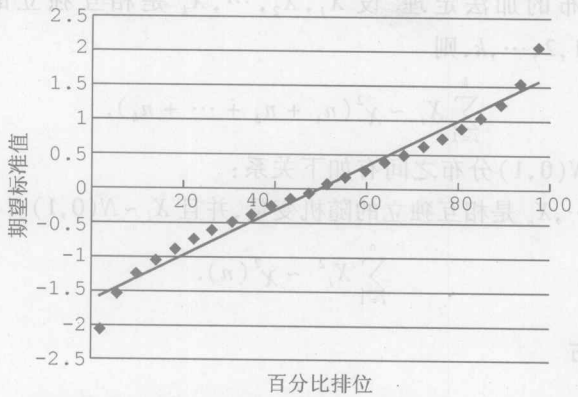

图 1.9 标准正态概率图

三、χ^2 分布

如果从标准正态分布 $N(0,1)$ 的总体中得到的 n 个随机变量分别为 X_1,