应用回归分析

张小蒂　编著

浙江大学出版社

应用回归分析

张 小 蒂 编著

浙江大学出版社

内 容 简 介

本书较全面地介绍了现代应用回归分析的基本理论和主要方法。全书共分十一章，除了系统地讨论了二元及多元回归模型、回归参数估计、回归假设检验和回归预测外，还深入浅出地论述了虚拟变量、选元、回归模型假定的违反及其处理等内容。阅读本书有助于理解和掌握回归分析这一重要的定量分析技术，增强研究、分析实际问题的能力。本书可作为高等院校试用教材，也可供各类科技工作者及各级经济管理人员学习参考。

应用回归分析

张小蒂　编著

责任编辑　傅百荣

*

浙江大学出版社出版

浙江上虞科技外文印刷厂印刷

浙江省新华书店发行

*

开本 1168×850 1/32　印张 10.125　字数 254 千字

1991 年 8 月第 1 版　1991 年 8 月第 1 次印刷

印数 0001—2000

ISBN 7—308—00797—9

F·054　定价：3.35 元

前　言

回归分析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，是现代应用统计学的一种重要分支，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。特别是近几年来计算机及有关统计软件的日渐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。因此，人们迫切要求学习回归分析的完整理论和方法。本书的编著正是为了适应这种客观需要而作的一种努力。

本书共分11章。第一章概括地介绍了学习回归分析需要具备的概率统计基础知识。第二章简要地论述了回归分析的基本概念。第三至第八章系统地讨论了二元及多元回归分析中的估计、检验及预测等主要内容，其中第八章运用线性代数中的矩阵方法讲述了多元回归模型。第九章通过阐明虚拟变量的性质和用途扩展了回归分述的应用范围。第十章针对回归分析实践中经常遇到又十分棘手的三个问题：异方差性、自相关性和多重共线性进行了深入详细的探讨并分别提出了处理的办法。第十一章通过举例具体介绍了选择自变量的四种方法。全书的内容具有一定的广泛性和较强的系统性。在内容介绍上，从易到难，由简入繁，力求深入浅出，简明扼要。对于回归分析中的大量数学公式一般都从其来源入手，循序渐进地导出，并借助于图、表及大量例题以便于读者的理

解和掌握。对其中的难点还作了专门的解释。本书在成稿过程中承研究生李晓钟协助，在此并致谢意。

限于水平，书中不足之处和错误在所难免，希望读者批评指正，以便进一步修改提高。

作　者

目　录

第四章　回归模型的考查

第五章　回归参数的检验

第六章　回归预测问题

第七章　回归分析中的几个专题

第八章　多元回归分析

第九章　虚拟变量及其应用

第十章　回归模型假定的违反及其处理

第十一章 自变量的选择

第一章　对有关数理统计学知识的回顾

本章仅就此书所涉及到的有关数理统计学基础知识作简要的回顾，而不作详细的证明。读者如在这方面有兴趣的话，可参阅数理统计学的专著或教材。

§1.1　随机变量及其分布

1．随机变量

简单地说，如果一个变量之值是由随机实验的结果确定的，则这种变量就称为随机变量(Random Variable)，简记为 RV。随机变量常用大写字母 X、Y 等来表示，而它们所取的值可用小写字母 x、y 等表示。

随机变量可以是离散型的，也可以是连续型的。离散型随机变量只取有限(或可数无限)个数值。而连续型随机变量则可在某数值区间中取任一数值。

设 Y 为一个取不同值 y_1、y_2、…、y_n 的离散型随机变量，则函数

$$\begin{aligned} f(y) &= P(Y=y_i) \quad i=1,2,\cdots,n \\ f(y) &= 0 \qquad\qquad Y \neq y_i \end{aligned} \tag{1.1.1}$$

称为随机变量 Y 的**概率密度函数** (Probability Density Func-

tion)，简记为 PDF。其中 $P(Y=y_i)$ 表示随机变量 Y 取值 y_i 的概率。

若 Y 是一个连续型随机变量，且满足下列条件：

$$f(y)\geqslant 0$$

$$\int_{-\infty}^{+\infty} f(y)\mathrm{d}y=1$$

$$\int_{b}^{a} f(y)\mathrm{d}y=P(a<y\leqslant b) \qquad (1.1.2)$$

则称 $f(y)$ 是 Y 的 PDF。$P(a<y\leqslant b)$ 表示 Y 位于区间 a 到 b 内的概率。如图 1.1.1 所示。

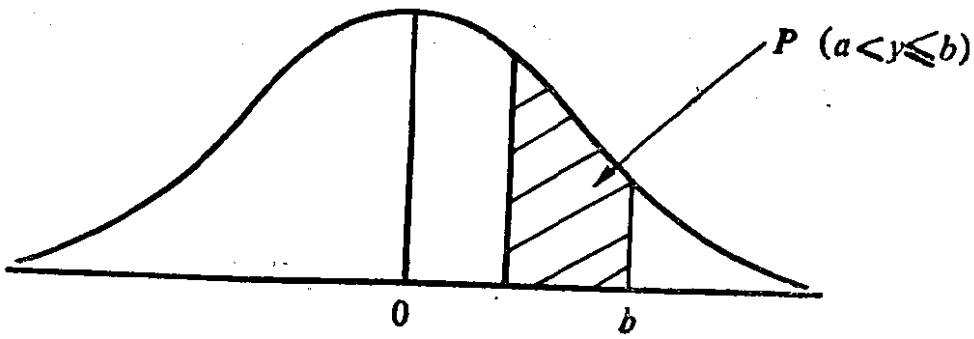

图1.1.1 连续型随机变量的PDF示意图

与离散型随机变量不同，连续型随机变量取某个特定值的概率等于零，这种变量的概率只有在某一区间内才是可以测度的。

2. 联合、边缘及条件概率密度函数

(Joint、Marginal、and Conditional PDF)

设 X 和 Y 为两个离散型随机变量，其函数

$$f(x,y)=P(X=x,\ Y=y) \qquad (1.1.3)$$

$$=0 \qquad \text{若 } X\neq x \text{ 和 } Y\neq y$$

称为离散型随机变量的联合 PDF。它给出了 X 取值 x、Y 取值 y 时的联合概率。

相对于 $f(x,y)$ 来说，$f(x)$ 和 $f(y)$ 称为边缘概率密度函数，并可由以下公式求得。

$$X\text{的边缘 PDF}=f(x)=\sum_{y} f(x,y) \qquad (1.1.4)$$

$$Y \text{ 的边缘 PDF} = f(y) = \sum_x f(x,y) \qquad (1.1.5)$$

式中 $\sum_y$ 和 $\sum_x$ 分别表示对 Y 所有的值和对 X 所有的值求和。

函数 $f(y|x) = P(Y = y | X = x)$ 称为 Y 的条件概率密度函数。类似地 X 的条件 PDF 为

$$f(x|y) = P(X = x | Y = y) \qquad (1.1.6)$$

X 和 Y 中的某个变量的条件 PDF 可用其联合 PDF 与另一变量的边缘 PDF 之比来表示。如：

$$X \text{ 的条件 PDF} = f(x|y) = \frac{f(x,y)}{f(y)} \qquad (1.1.7)$$

$$Y \text{ 的条件 PDF} = f(y|x) = \frac{f(x,y)}{f(x)} \qquad (1.1.8)$$

当 X 和 Y 为两个连续型随机变量时，其联合 PDF 具有以下性质：

$$f(x,y) \geqslant 0$$

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\mathrm{d}x\,\mathrm{d}y = 1$$

$$\int_c^d\int_a^b f(x,y)\mathrm{d}x\,\mathrm{d}y = P(a < x \leqslant b, c < y \leqslant d) \qquad (1.1.9)$$

从中可分别求出

$$X \text{ 的边缘 PDF} = f(x) = \int_{-\infty}^{+\infty} f(x,y)\mathrm{d}y \qquad (1.1.10)$$

$$Y \text{ 的边缘 PDF} = f(y) = \int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}x \qquad (1.1.11)$$

如果 X 和 Y（不论其是离散型还是连续型）的联合 PDF 可以表述为它们各自边缘 PDF 的乘积时，即

$$f(x,y) = f(x)f(y) \qquad (1.1.12)$$

则称这两个随机变量是统计独立的。

3．数学期望值和方差

随机变量的概率分布特征通常可以用它的数学期望值（亦称

为均值）及方差来概括。

若随机变量为 Y，则其数学期望值为

$$E(Y)=\sum_{y} yf(y) \qquad \text{若 } Y \text{ 为离散型}$$

$$=\int_{-\infty}^{+\infty} yf(y)\mathrm{d}y \qquad \text{若 } Y \text{ 为连续型} \qquad (1.1.13)$$

数学期望值具有以下性质：

(1) 常数的数学期望值就是该常数本身，即

$$E(a)=a \qquad a \text{ 为任一常数。} \qquad (1.1.14)$$

(2) $E(aY+b)=aE(Y)+b$，　a、b 为常数。　(1.1.15)

(3) 若 X 和 Y 为独立的随机变量，则有

$$E(XY)=E(X)E(Y) \qquad (1.1.16)$$

设随机变量 Y 的数学期望值为

$$E(Y)=\mu \qquad (1.1.17)$$

则 Y 所取值围绕期望值 $E(Y)$ 的散布程度可用其方差来度量，并定义为

$$\mathrm{Var}(Y)=\sigma^2=E(Y-\mu)^2=E\{[Y-E(Y)]^2\} \qquad (1.1.18)$$

$\mathrm{Var}(Y)$的计算公式为

$$\mathrm{Var}(Y)=\sum_{y}(Y-\mu)^2 f(y) \qquad \text{若 } Y \text{ 为离散型}$$

$$=\int_{-\infty}^{+\infty}(Y-\mu)^2 f(y)dy \qquad \text{若 } Y \text{ 为连续型} \qquad (1.1.19)$$

σ^2 的平方根 σ 为 Y 的标准差。

方差具有以下性质：

(1) $E(Y-\mu)^2=E(Y^2)-[E(Y)]^2$　(1.1.20)

(2) 常数的方差为零，即 $\mathrm{Var}(a)=0$。　(1.1.21)

(3) 若 a 和 b 为常数，则

$$\mathrm{Var}(aY+b)=a^2\mathrm{Var}(Y) \qquad (1.1.22)$$

(4) 若 X 和 Y 是两个相互独立的随机变量，则

$$\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y) \tag{1.1.23}$$

$$\mathrm{Var}(aX+bY)=a^2\mathrm{Var}(X)+b^2\mathrm{Var}(Y) \tag{1.1.24}$$

其中 a、b 为常数。

4. 协方差

设 X 和 Y 是两个数学期望值分别为 $E(X)$和 $E(Y)$的随机变量，则它们之间的协方差定义为

$$\mathrm{Cov}(X,Y)=E\{[Y-E(Y)][X-E(X)]\} \tag{1.1.25}$$

$$=E(XY)-E(X)E(Y)$$

类似地，$\mathrm{Cov}(Y,Y)=E(Y^2)-[E(Y)]^2=V_{ar}(Y)$ (1.1.26)

协方差的计算公式为

$$\mathrm{Cov}(X,Y)=\sum_y\sum_x[X-E(X)][Y-E(Y)]f(x,y)$$

[当变量为离散型时]

$$=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}[X-E(X)][Y-E(Y)]f(x,y)$$

[当变量为连续型时] (1.1.27)

协方差具有以下性质：

(1) $\mathrm{Cov}(X,Y)=\mathrm{Cov}(Y,X)$ (1.1.28)

(2) $\mathrm{Cov}(aX,bY)=ab\,\mathrm{Cov}(X,Y)$ a、b 为常数。 (1.1.29)

(3) $\mathrm{Cov}(a+X,b+Y)=\mathrm{Cov}(X,Y)$ a、b 为常数。 (1.1.30)

(4) $\mathrm{Cov}(a+bX,c+dY)=bd\,\mathrm{Cov}(X,Y)$

a、b、c、d 为常数。 (1.1.31)

(5) $\mathrm{Cov}(X_1+X_2,Y)=\mathrm{Cov}(X_1,Y)+(X_2,Y)$ (1.1.32)

(6) 若 X 和 Y 相互独立，则

$$\mathrm{Cov}(X,Y)=0 \tag{1.1.33}$$

设 $Y_1,Y_2,\cdots,Y_n$ 为 n 几个随机变量，则

$$E\left(\sum_{i=1}^{n} a_i Y_i\right) = \sum_{i=1}^{n} a_i E(Y_i) \quad \text{所有 } a_i \text{ 均为常数。} \quad (1.1.34)$$

$$\mathrm{Var}\left(\sum_{i=1}^{n} a_i Y_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \mathrm{Cov}(Y_i, Y_j) \quad (1.1.35)$$

所有 a_i 均为常数。

当 $n=2$ 时，

$$E(a_1Y_1 + a_2Y_2) = a_1E(Y_1) + a_2E(Y_2)$$

$$\mathrm{Var}(a_1Y_1 + a_2Y_2) = a_1^2\mathrm{Var}(Y_1) + a_2^2\mathrm{Var}(Y_2) + 2a_1a_2\mathrm{Cov}(Y_1, Y_2)$$

$$\mathrm{Var}(a_1Y_1 - a_2Y_2) = a_1^2\mathrm{Var}(Y_1) + a_2\mathrm{Var}(Y_2) - 2a_1a_2\mathrm{Cov}(Y_1, Y_2)$$

如果所有 Y_i 都为互相独立的随机变量，则

$$\mathrm{Var}\left(\sum_{i=1}^{n} a_i Y_i\right) = \sum_{i=1}^{n} a_i^2\mathrm{Var}(Y_i) \quad \text{所有 } a_i \text{ 均为常数。} \quad (1.1.36)$$

$$\mathrm{Cov}\left(\sum_{i=1}^{n} a_i Y_i,\ \sum_{i=1}^{n} b_i Y_i\right) = \sum_{i=1}^{n} a_i b_i\ \mathrm{Var}(Y_i) \quad (1.1.37)$$

所有 a_i 和 b_i 均为常数。

5．样本及总体相关系数

设 X 和 Y 为两个随机变量。(X_1, Y_1)、(X_2, Y_2)、……、(X_n, Y_n) 为一随机样本，度量 X 与 Y 相随变动的一个指标为样本协方差：

$$C_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} \quad (1.1.38)$$

下面是这个指标构成的原理。

如图所示，如果样本数据点落在 Ⅰ、Ⅲ象限，则 $(X_i - \bar{X})(Y_i - \bar{Y})$ 必同号，其积为正；如果样本数据点落在 Ⅱ、Ⅳ 象限，

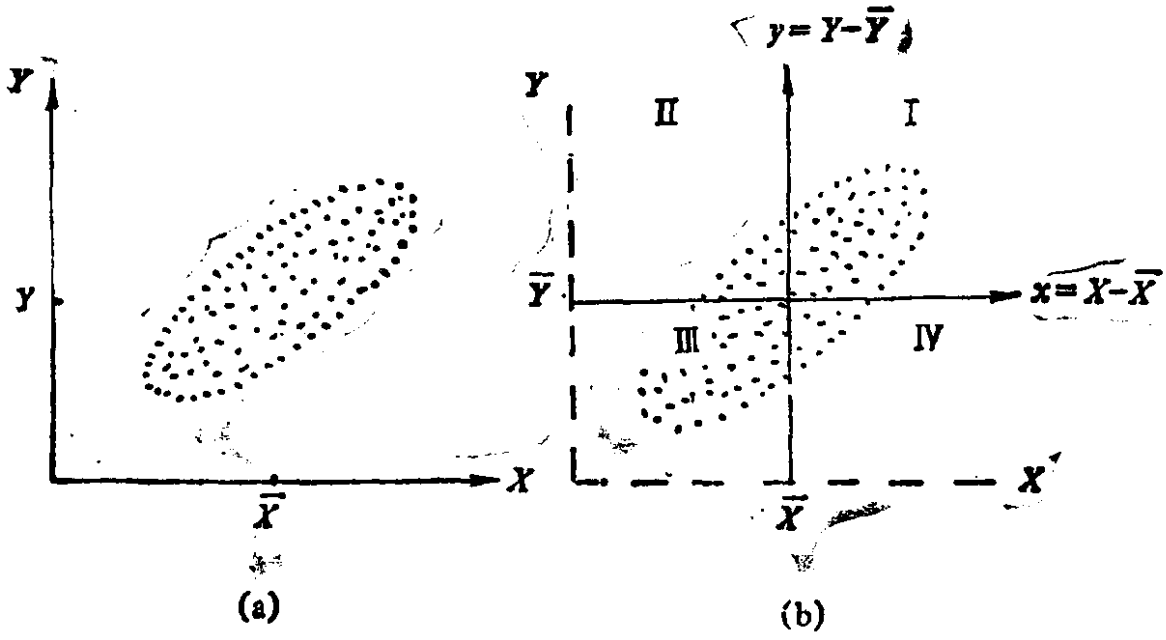

图1.1.2

则$(X_i-\bar{X})(Y_i-\bar{Y})$必异号，其积为负。可见，两离差之积的总和$\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})$可作为测量$X$和$Y$相随变动程度的一种指标，至少在正负号上是正确的。此外，当X和Y之间不存在相随变动关系时，它们的观测数据就会较均匀地分布在四个象限中，以致两离差之积的正负项在求和时会互相抵消而使$\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})$接近于零。但是，$\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})$这一项仍受样本数目$n$大小的影响，$n$大时，乘积较多，这一项就会变大，反之则会变小。消除这一影响的方法是把该项除以$(n-1)$，这就构成了样本协方差C_{XY}。应该指出，样本协方差C_{XY}还会受X和Y的度量单位的影响。消除这一缺陷的方法是将C_{XY}除以X和Y的样本标准差，使这个比值成为不再受度量单位影响的纯数，从而能比较真实地反映两变量之间相随变动的线性相关程度。它就叫样本相关系数，记为r。其计算公式为

$$r=\frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})\ (Y_i-\bar{Y})}{\sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}\cdot\frac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{n-1}}} \tag{1.1.39}$$

$$=\frac{\sum_{i=1}^{n}(X_i-\bar{X})\ (Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \tag{1.1.40}$$

其值域为 $-1\leqslant r\leqslant+1$ (1.1.41)

X 和 Y 之间的总体相关系数为 ρ，其定义为

$$\rho=\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \tag{1.1.42}$$

ρ是测定两个随机变量X和Y之间线性相关程度的一个度量。它的值域在 -1 与 $+1$ 之间。若 $\rho=0$，说明 X 和 Y 之间不存在线性相关关系；若 $\rho=-1$，说明 X 和 Y 之间存在着完全的线性负相关关系；若 $\rho=+1$，说明 X 和 Y 之间存在着完全的线性正相关关系。一般有

$$-1\leqslant\rho\leqslant+1 \tag{1.1.43}$$

6. 条件期望和条件方差

设 $f(x,y)$是随机变量 X 和 Y 的联合PDF。在给定 $X=x$ 的条件下，Y 的条件期望定义为

$$E(Y|X=x)=\sum_{y}yf(y|X=x) \qquad \text{若 } Y \text{ 为离散型}$$

$$=\int_{-\infty}^{+\infty}yf(y|X=x)\mathrm{d}y \qquad \text{若 } Y \text{ 为连续型} \tag{1.1.44}$$

Y 的条件方差定义为