TC

现代统计知识丛书

XIANDAI TONGJI ZHISHI CONGSHU

线性回归模型分析

俞大刚 著

中国统计出版社

线性回归模型分析

俞大刚　著

中国统计出版社

现代统计知识丛书

线性回归模型分析

XIANXING HUIGUI MOXING FENXI

俞大刚 著

*

中国统计出版社出版
新华书店北京发行所发行
房山先锋印刷厂印刷

*

787×1092毫米 32开本 5印张 9万字
1987年10月第1版 1987年10月北京第1次印刷
印数：1—4,000
ISBN 7—5037—0024—6/F·24
统一书号：4006.138 定价：0.95元

《现代统计知识丛书》序言

我们编写这套《现代统计知识丛书》的目的，一是为了弥补现有统计教材之不足，为统计教学增添新的内容；一是为了满足具有高中以上文化程度在职统计干部自学的需要，使他们的统计知识随着时间的推移而相应地得到更新。

在党的十一届三中全会前后，1978年12月国家统计局在四川峨眉召开“全国统计教学、科研规划座谈会”以来，已经出版的我国学者编写的统计教材的数量，大大超过了“文化大革命”前的十七年，在一定程度上，内容也有所更新。这些教材，在满足统计教学的亟需方面，起了重要的作用。但是，四化建设和经济体制改革正在不断地向前推进，统计科学也在继续发展。这些统计教材，已经落后于形势的发展，不能完全适应四个现代化的要求。统计教材有待进行全面的充实和更新。

在职统计干部进行有计划的自学，不断提高业务能力，是我国造就统计人才的一个重要途径。我们一直在努力探索具有中国特色的统计工作道路，为实现统计现代化的目标而努力。在职统计干部现有的统计知识，有的已经适应不了统计现代化的需要，而许多现代化的统计知识，他们还没有掌握起来。广大统计干部，正面临着新的挑战，他们的统计知识也亟需得到补充和更新。

为满足上述两方面的要求，需要以马列主义、毛泽东思

想为指针，从中国的实际情况出发，吸收国际上统计科学的新成果，编写一套具有中国特色的现代化的新的统计教材。但是，经济体制的改革还在深入进行，统计工作也在不断变化，要很快编写一套在较长时期内适用的新的统计教材，条件还不够成熟。至于先就教材中的某一侧面进行比较深入的剖析与论述，编写小册子以充实统计新知识，补充统计教材之不足，为逐步更新统计教材创造有利条件，则是必要的，也是不难做到的。这就是编写这套《现代统计知识丛书》的由来。

邓小平同志提出："教育要面向现代化，面向世界，面向未来。"这是教育工作的方针，也是我们编写《现代统计知识丛书》的方针。《丛书》选题，应当包括我国三十多年来统计工作经验的总结，重点应当放在党的十一届三中全会以来经验的总结。中国统计工作的改革要立足在自己创造的经验的基础上。另一方面，我们必须向国际上先进的统计理论和实践学习，要注意在统计工作中运用数学方法和电子计算机的新方法，还要探索在统计中对信息论、控制论和系统工程论的运用问题。这也是《丛书》选题的重点。介绍外国经验，是为了根据中国的国情加以运用。当然，把外国的经验同我国的情况结合起来，需要一个过程，有时需要较长的过程。作者在坚持四项基本原则的前提下，可以阐发自己的独立见解，可以介绍和评述不同的学派，通过百家争鸣，共同探求真理。《丛书》将根据我国统计工作现代化的长期目标和中期规划的需要，有计划地进行编写。每一本书都要求在现有水平的基础上提高一步，写出新意，向深度和广度发展。

我们的这一设想，希望得到广大统计实际工作者和理论、教学工作者的支持，为《现代统计知识丛书》写稿，并提供宝贵意见，共同为促进我国统计工作现代化的实现而努力。

《现代统计知识丛书》编辑委员会

1985年12月

目　录

一、绪　论

线性回归模型分析是处理变量之间关系的一种统计方法，利用这种方法有助于我们认识客观事物的定量关系及其内在的规律，它的应用几乎遍及所有科学技术和国民经济的各个部门，随着我国社会主义现代化建设的发展，它将起着越来越广泛的作用。

线性回归模型分析是数理统计学中的一个重要部分。我们知道，数理统计方法是以概率论为基础，通过样本来了解和判断总体的统计特性的方法。对总体的某些参数的估计，关于某些统计量所作假设的检验，便是其中的重要内容。在这里，我们先来谈谈在这本小册子中将涉及到的一些有关数理统计的几个重要概念。

（一）在数理统计中，正态分布占有特别重要的地位。在许多领域里，我们考察或研究的课题所遇到的许多总体，似乎都能很好地近似于正态分布。试考虑一个关于工业产品质量的例子来解释，一个工业产品的生产过程要受到许多因素影响，而且很明显的是所有随机因素都具有微小的效应，产品质量的好坏可看成所有这些因素总效应的反映。假定每个因素的效应来自某个总体的观察值，那么总效应本质上就是来自这个总体的一组观察值的平均数。在抽样调查中，我们知道用单纯随机抽样法抽取的样本，其样本平均数的分布，是基于概率论中以下两个定理：

1．总体的分布如果是正态分布，样本平均数的分布也是正态分布；

2．总体的分布无论如何分布，样本含量充分大，样本平均数的分布近似正态分布。

在客观世界中，虽然并非所有的分布都是正态的，但是接近正态的分布往往是经常碰到的。

从样本作出对总体的推断时，必须认识样本观察值的各种函数的分布，由于正态分布的抽样分布常常比来自别的总体的样本分布在分析上较易处理，所以正态分布在统计研究中扮演着重要角色。

对于任何正态分布，它的样本x落在任意区间(a, b)的概率记作$P(a<x<b)$，则有

$$P(a<x<b)=\frac{1}{\sigma\sqrt{2\pi}}\int_a^b e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx$$

式中μ是总体平均数，σ是总体标准差，$\pi\approx3.1416$。

由附表（二）可知：

$(\mu-\sigma, \mu+\sigma)$ 的概率为68.3%

$(\mu-1.96\sigma, \mu+1.96\sigma)$ 的概率为95%

$(\mu-2\sigma, \mu+2\sigma)$ 的概率为95.4%

$(\mu-3\sigma, \mu+3\sigma)$ 的概率为99.7%

（二）如果考虑$x_1, x_2, \cdots, x_n$为来自平均数为零、方差为1的正态分布的随机样本，记它们的平方和为χ^2（卡方），即

$$\chi^2=x_1^2+x_2^2+\cdots+x_n^2$$

它是服从参数为n的χ^2分布，其密度函数为：

$$f(\chi^2)=\begin{cases}\dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}(\chi^2)^{\frac{n}{2}-1}e^{-\frac{1}{2}\chi^2}, & \chi^2>0\\ 0, & \chi^2\leqslant 0\end{cases}$$

式中，$\Gamma(x)$是伽玛函数，参数n称为分布的自由度，e是自然对数的底。

χ^2分布的应用有直接和间接两方面，间接的应用是指在χ^2分布的基础上去推导出t分布和F分布，直接的应用，至少有两方面：

1．利用χ^2分布对正态总体的方差进行统计推断；

2．利用χ^2分布解决实测频数和理论频数是否符合的问题。

最近二十多年来，在人口统计中，人们发现所有规范化了的生育模式函数可以比较准确地用χ^2概率密度曲线来逼近，即是说，妇女生育模式函数大致上满足χ^2分布。近代许多人口统计学家利用它来对未来人口发展趋势作定量预测，比用传统的“年龄移算法”更符合发展过程中的客观实际，预测结果也具有更高的精确度。

为对正态总体的未知方差进行估计或假设检验，不论总体平均数为已知或未知，都可用它来作出判断。

（三）另一种在实践中具有重要意义的分布，是正态变量对与之独立的卡方变量的平方根之比的分布，这种分布是由W.S.Gosset在1908年用笔名Student发表而闻名的t分布，它的密度函数为：

$$\varphi(t)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}},\quad -\infty<t<+\infty$$

式中n为自由度，其它记号的含意同前。

t分布的重要价值在于对总体的方差未知，可用小样本对总体的未知平均数作出统计推断。

（四）还有一种有很大实用价值的分布，就是两个独立卡方变量的比的分布（F分布），它的密度函数为：

$$h(F)=\begin{cases}f_1^{\frac{f_1}{2}}f_2^{\frac{f_2}{2}}\dfrac{\Gamma\left(\frac{f_1+f_2}{2}\right)}{\Gamma\left(\frac{f_1}{2}\right)\Gamma\left(\frac{f_2}{2}\right)}\dfrac{F^{\frac{f_1}{2}-1}}{(f_2+f_1F)^{\frac{f_1+f_2}{2}}}, & F>0\\ 0, & F\leqslant 0\end{cases}$$

式中f_1称为第一自由度，f_2称为第二自由度。如果两总体的平均数和方差都未知，而要对两总体的方差作出估计或比较，则F分布为我们提供了方便。

以上几种分布，在本书中都会用到。

二、一元线性回归的几何意义和最小性质

我们在各种实践中，经常会遇到一些来自同一总体中的变量，这些变量是互相联系和互相依存的，它们之间存在着一定的关系。

变量之间的关系，大体可分为两类：一类是变量之间存在着完全确定的关系，即大家所熟悉的函数关系；另一类是变量之间的关系具有某种不确定性，称为相关关系，或简称为相关。

函数和相关是两种不同类型的变量关系，但它们之间并不存在着不可逾越的障碍，相关变量之间虽不具备完全确定的关系，但通过对现象的大量观察，可以探索它们之间的统计规律性，这类统计规律称为回归关系，有关回归关系的计算方法和理论称为回归分析，而对相关变量之间的关系用相关指标来表明它们之间关联的程度，其计算方法和理论则称为相关分析。实际上，回归分析和相关分析是既有联系又有区别的两种方法，在本书中都将涉及到这两种方法的内容，而我们的目的是利用数理统计方法侧重找出具有相关关系的回归方程式，对所找出的回归方程进行统计分析，统计检验，统计预测，并给出它们的精度估计。

下面谈一元线性回归的几何意义和最小性质。

假设我们有两个变量ξ和η，它们之间的变化具有相关关系，也就是说，变量ξ的变化会引起η作相应的变化，但它们

的变化关系是不确定的，即当$\xi=x_0$时，η有许多可取的值（见图2.1）。

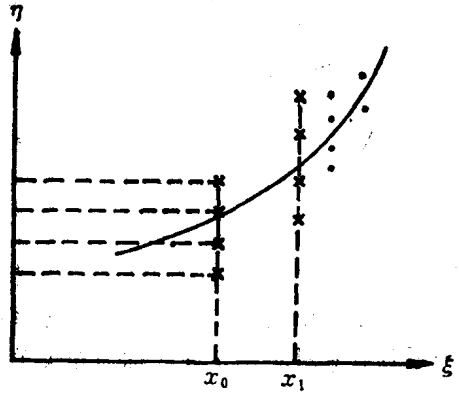

图 2.1

为了找出变量η与ξ之间的关系，一个很自然的想法是取$\xi=x_0$时所有η值的平均数作为对应$\xi=x_0$时η的代表值，亦即取

$$\hat{\eta}|_{\xi=x_0}=y_0=E(\eta|_{\xi=x_0})$$

式中$E(\eta|_{\xi=x_0})$表示在$\xi=x_0$条件下η的条件期望值。同样，对应于$\xi=x_1$，我们取

$$\hat{\eta}|_{\xi=x_1}=y_1=E(\eta|_{\xi=x_1})$$

作为对应于$\xi=x_1$时η的代表值，一般地，对于任何一个ξ的可取值x，我们都相应地取

$$\hat{\eta}|_{\xi=x}=y=E(\eta|_{\xi=x})$$

当x变化时，上式右端是x的一个确定的函数，记为

$$u(x)=E(\eta|_{\xi=x})$$

于是，我们就可以用一个确定的函数

$$y = u(x) \tag{2.1}$$

来大体上描述变量η与ξ之间的变化情况，并称公式(2.1)为η对ξ的回归方程，它反映了ξ在固定条件下η的平均状态的变化，或者说，η对ξ的回归就是η对ξ的条件期望函数。

回归函数$u(x)$具有一个重要的“最小”性质，即对于任意函数$\varphi(x)$，恒有

$$E[\eta-\varphi(x)]^2 \geqslant E[\eta-u(x)]^2 \tag{2.2}$$

当且仅当$\varphi(x)=u(x)$时，上式等号才成立。

证：事实上，对于任意固定的x，在$\xi=x$时有

$$\begin{aligned} E[\eta-u(x)]^2 &= E(\eta-E(\eta))^2 \\ &= E\{[\eta-\varphi(x)]-[E(\eta)-\varphi(x)]\}^2 \\ &= E[\eta-\varphi(x)]^2-2E[\eta-\varphi(x)][E(\eta)-\varphi(x)] \\ &\quad +[E(\eta)-\varphi(x)]^2 \\ &= E[\eta-\varphi(x)]^2-[E(\eta)-\varphi(x)]^2 \end{aligned}$$

显然，$[E(\eta)-\varphi(x)]^2 \geqslant 0$，且仅当$E(\eta)=\varphi(x)$时，等式才成立，所以

$$E[\eta-u(x)]^2 \leqslant E[\eta-\varphi(x)]^2$$

这说明，在一切x的函数中，只有用回归函数$u(x)$作为η的估计，才能使估计的偏差平方和达到最小，因此，回归方程(2.1)就成了研究变量间相关关系的一个重要工具。

一般说来，从任意函数$\varphi(x)$中找一个使估计偏差最小的回归函数$u(x)$是困难的，通常我们往往限制$\varphi(x)$为某一函数类，例如我们限制$\varphi(x)$为线性函数类，则形如

$$E(y)=\alpha+\beta x \tag{2.3}$$

就称它为η对ξ的一元线性回归方程，这里α、β是和x无关

的参数。由（2.3）可把随机变量y表示成

$$y = \alpha + \beta x + e \tag{2.4}$$

其中$e = y - E(y)$是随机变量，$E(e) = 0$。

推而广之，η对m个变量ξ_1，ξ_2，…，ξ_m的线性回归方程为：

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$$

或 $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + e$$

我们将要讨论的内容，主要是变量之间的线性回归问题。

三、用最小二乘法求一元线性回归方程

假设随机变量

$$y \sim N[E(y),\ \sigma^2]$$

其中$E(y)=u(x)=\alpha+\beta x$，α，β为待定常数，也就是说，随机变量y与x满足线性模型

$$y=\alpha+\beta x+e$$

式中的e有如下几个基本假定：

1．e是一个随机变量，对于每个观测值x_i来说，它可以分别以一定的概率取正值、负值或零值，不同观测值的误差是相互独立的；

2．e的期望值$E(e)=0$，即是说，如果考虑了e的所有可能取的值，它们的平均值为零；

3．e的方差$D(e)=\sigma^2$，即对所有x_i来说，每个e_i的方差都相同；

4．e表示众多因素微小影响的综合反映，由中心极限定理可知，它是正态变量。

以上几个关于e的基本假定，用记号表示，可以写成：

$$e \sim N(0,\ \sigma^2)$$

其图示见第10页图3.1。

今对y值作n次独立观测，得一组观测值：

$$(x_i,\ y_i),\ i=1,\ 2,\ \cdots,\ n$$

我们要根据这一组观测值来确定y对x的“最佳”拟合。如

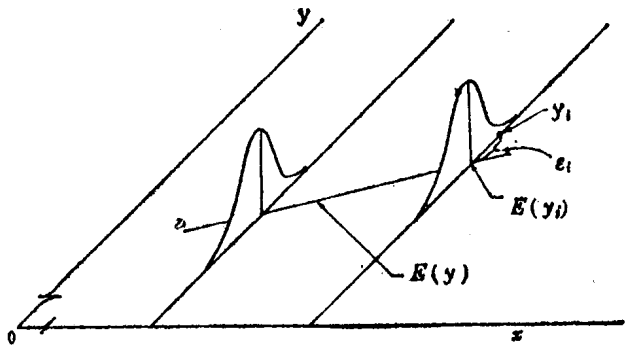

图 3.1

前所述，我们可用y的平均状态作为y的估计式：

$$\hat{y} = a + bx, \qquad (3.1)$$

式 (3.1) 称为经验回归方程，记号“∧”（读如hat）。因此我们的任务就是要确定经验回归方程 (3.1) 中的两个参数a和b的估计值，并要指出这样定出的估计所具有的精度。

一个常用的方法就是最小二乘法。

设y是变量x的函数，含有参数a及b，即

$$y = f\ (a,\ b;\ x)$$

今对y和x作n次观测得 $(x_i,\ y_i)$ $(i = 1,\ 2,\ \cdots,\ n)$。于是 y 的估计值$\hat{y}_i = f(a,\ b;\ x_i)$与观测值$y_i$的绝对误差为

$$|e_i| = |y_i - \hat{y}_i| \quad (i = 1,\ 2,\ \cdots,\ n)$$

所谓最小二乘法，就是要求上面n个误差的平方和，使得经验回归函数$\hat{y} = f(a,\ b;\ x)$与观测值 $y_1,\ y_2,\ \cdots,\ y_n$ 达到