·数理统计丛书·

周纪芗 编

回归分析

华东师范大学出版社

·数理统计丛书·

回归分析

周纪芗 编

华东师范大学出版社

(沪)新登字 201 号

周纪芗 编

华东师范大学出版社出版发行.
(上海中山北路 3663 号)
新华书店上海发行所经销 江苏句容排印厂印刷
开本 850×1168 1/32 印张: 7.625 字数 190 千字
1993年3月第一版 1993年4月第一次印刷
印数 001—3000 本

ISBN7-5617-0890-4/O·021 定价: 4.85元

总　序

数理统计是一门应用性很强的学科。它是研究如何有效地收集、整理和分析受随机影响的数据，并对所考虑的问题作出推断或预测，直至为采取决策和行动提供依据和建议的一门学科。凡是有大量数据出现的地方，都要用到数理统计。人口调查、税收预算、测量误差、出生与死亡统计、保险业中赔款额和保险金的确定等，这些数理统计早期主要研究的问题，直到现在仍值得认真研究。建立在现代数学和概率论基础上的数理统计，在近半个世纪以来，在理论、方法、应用上都有较大的发展。抽样调查、试验设计、回归分析与回归诊断、多元分析、时间序列分析、非参数统计、统计决策函数、统计计算、随机模拟、探索性数据分析等统计方法相继产生并在实践中普遍使用，把以描述为主的统计发展到以推断为主的统计。今天，数理统计的内容已异常丰富，应用面广量大，成为当前最活跃的学科之一。

我国科技和经济的发展，需要大量的经过系统训练的数理统计专业人才。最近数年中，国家教育委员会已在一些高校先后设立各种统计专业，这将为我国数理统计发展开创新的局面。为了促进数理统计人才的培养，国家教育委员会于1984年召开了“数理统计教学座谈会”，会上交流了各校培养数理统计人才的经验，同时还指出，组织国内专家编写和出版一套数理统计专业的教材是当务之急。

我们认为，一本好的数理统计教材，首先应讲清统计思想与统计方法，所讲的理论应是在实践中有用的统计方法所必需的理论，在阐明各种统计方法时，应给出足够的问题背景和有关的数据，能使学生对数理统计有一个系统、全面的认识，并培养学生对

统计实践的兴趣。另外，文字流畅，带有趣味性，适合大学生阅读，当然也是一本好教材的必要条件。

鉴于国内对数理统计教材的急需，我们约请国内一些颇有造诣的专家，按照上述对教材的特定要求，编写了这套“数理统计丛书”，作为高等院校数理统计专业的基本教材，并将由华东师范大学出版社陆续出版。现在奉献给读者的《回归分析》即为该丛书之一。我们希望，这套丛书能对我们数理统计事业的发展起到一定的促进作用，能成为我国年轻一代统计学家的引路之石。

茆 诗 松

1986 年 11 月于华东师范大学

序　言

回归分析是数理统计应用最广泛的分支之一。本书叙述了经典的最小二乘方法与理论，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。

本书曾以讲义形式在华东师范大学数理统计专业中试用过六届，不少学生对讲义提出了许多宝贵意见，作者据此对讲义作了多次修改。这次出版之前，茆诗松教授又仔细阅读了全文，提出了许多宝贵的修改意见与建议。在此我们表示衷心的感谢。

由于编者水平有限，错误之处在所难免，恳请读者批评指正。

周　纪　芗

1991.1 于华东师范大学

引　言

在现实世界中，我们经常要与各种变量打交道，有些变量存于一个共同体中，人们就需要研究这些变量间的关系。

变量间的关系，常见的有两类。

一类称为“确定性关系”。例如，一个圆的半径 R 与周长 C 可以看成两个变量，当 R 的值确定后，C 可以作为 R 的函数

$$C=2\pi R$$

来确定其值。我们讲变量 C 与 R 间有确定性关系，其关系可用函数形式表达。这是一般非随机性的数学、物理、化学、工程等学科中研究的问题。

另一类变量间有一定的关系，但其关系不能用函数形式来表达。例如，高度相同的人其体重不完全相同，但一般讲较高的人其体重相应也重一点，较矮的人其体重轻一点，因而身高与体重间有一定的关系，然而它们间的关系不能用一个函数表达。我们称这种变量间的关系为“相关关系”，它是一种不确定性的关系。

回归分析就是要研究具有相关关系的变量间的统计规律性。

回归分析在工农业生产及科学研究中有着广泛的应用，在试验数据的处理、经验公式的寻找、产品的统计质量管理、市场预测、某些新标准的制订、自动控制中数学模型的建立、气象预报、地质勘探、医学卫生等许多领域中都经常应用回归分析。

学习这门课程需要有一定的线性代数、微积分及概率统计基础知识。

目　录

第一章　一元线性回归

§1.1　模　　型

最简单的回归分析是一元线性回归分析．为了弄清它解决什么问题，我们先看一个例子．

例 1.1　合金钢的强度 y 与钢材中碳的含量 x 有密切关系．为了冶炼出符合要求强度的钢常常通过控制钢水中的碳含量来达到目的，为此需要了解 y 与 x 之间的关系．

由于种种因素的影响，即使钢水中碳的含量相同，合金钢的强度也不会完全相同，因而它们间的关系是一种非确定性的关系．那么它们之间有没有关系？有什么样的关系？为此首先就要收集 n 组不同的碳含量 $x(\%)$ 对应的钢的强度 $y(\mathrm{kg/mm^2})$ 数据 (x_i, y_i)，$i=1,2,\cdots,n$．本例所收集的数据见表 1.1.1，这里 $n=10$．

表　1.1.1

i	$x(\%)$	$y(\mathrm{kg/mm^2})$
1	0.03	40.5
2	0.04	39.5
3	0.05	41.0
4	0.07	41.5
5	0.09	43.0
6	0.10	42.0
7	0.12	45.0
8	0.15	47.5
9	0.17	53.0
10	0.20	56.0

为了看清其规律，把(x_i,y_i)看成是平面直角坐标系中的点，画出“散点图”(见图 1.1.1)。

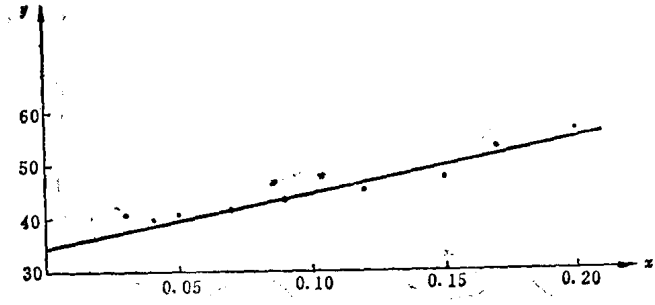

图 1.1.1

接着，我们可以观察散点图上点的分布规律。在本例中，这些点散布在一直线附近，但又不全在一条直线上，那么我们可以认为 y 与 x 之间的关系由两部分组成，一部分是由于 x 的变化引起的 y 线性变化部分，记为 $\beta_0+\beta_1x$，另一部分是由其它一切随机因素引起的，记为 ε：

$$y=\beta_0+\beta_1x+\varepsilon. \tag{1.1.1}$$

在(1.1.1)中，我们总假定 x 是一般变量，其值是可以精确测量或严格控制的，β_0、β_1 为未知参数，ε 是不可观测的随机误差，通常假定 $E(\varepsilon)=0$，$D(\varepsilon)=\sigma^2$，为了对参数作区间估计与假设检验的需要，通常还假定其服从正态分布，从而在上述假定下 $y\sim N(\beta_0+\beta_1x,\sigma^2)$。

对我们所获得的观测数据(x_i,y_i)，$i=1,2,\cdots,n$ 来讲，通常还假定各 $y_1,y_2,\cdots,y_n$ 间相互独立，从而可得到一元线性回归的数学模型：

$$\begin{cases} y_i=\beta_0+\beta_1x_i+\varepsilon_i,\ \ i=1,2,\cdots,n, \\ \text{各 } \varepsilon_i \text{ 独立同分布，其分布为 } N(0,\sigma^2) \end{cases} \tag{1.1.2}$$

我们今后称

$$E(y)=\beta_0+\beta_1x \tag{1.1.3}$$

为 y 关于 x 的回归函数，它在平均意义上表明了 y 与 x 之间的一种统计规律性。

我们所要研究的问题有：

(1) 如何根据样本(x_i, y_i)，$i=1,2,\cdots,n$ 求出 β_0、β_1 的估计。

若用 $\hat{\beta}_0$、$\hat{\beta}_1$ 分别记 β_0、β_1 的点估计，则称

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x \tag{1.1.4}$$

为 y 关于 x 的一元线性回归方程，为表明这是 x 的函数，必要时也记 $\hat{y}$ 为 $\hat{y}(x)$。

(2) 如何检验回归方程的可信度？

(3) 如果回归方程是可信的话，如何用它进行预测和控制。

§1.2 参数 β_0、β_1 的最小二乘估计

为估计模型(1.1.2)中的参数β_0、β_1，常采用最小二乘法，即要求观测值 y_i 与其拟合值 $\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1 x_i$ 之间的偏差平方和达到最小。若记

$$Q(\beta_0,\beta_1)=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2,$$

则 $\hat{\beta}_0$，$\hat{\beta}_1$ 应满足下列要求：

$$Q(\hat{\beta}_0,\hat{\beta}_1)=\min_{\beta_0,\beta_1} Q(\beta_0,\beta_1).$$

由于 $Q(\beta_0,\beta_1)$是 β_0、β_1 的非负函数，且关于 β_0、β_1 可微，故根据微积分原理有

$$\begin{cases}\dfrac{\partial Q(\beta_0,\beta_1)}{\partial \beta_0}\Big|_{\beta_0=\hat{\beta}_0,\beta_1=\hat{\beta}_1}=0,\\[2ex] \dfrac{\partial Q(\beta_0,\beta_1)}{\partial \beta_1}\Big|_{\beta_0=\hat{\beta}_0,\beta_1=\hat{\beta}_1}=0,\end{cases}$$

即

$$\begin{cases} \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0, \\ \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0. \end{cases} \tag{1.2.1}$$

称(1.2.1)为正规方程组，将其简化一下得

$$\begin{cases} \hat{\beta}_0 + \bar{x}\cdot\hat{\beta}_1 = \bar{y}, \\ \bar{x}\cdot\hat{\beta}_0 + \frac{1}{n}\Sigma x_i^2\cdot\hat{\beta}_1 = \frac{1}{n}\Sigma x_i y_i. \end{cases}$$

(以下凡不加说明处，"Σ"表示"$\sum_{i=1}^{n}$")解得

$$\begin{cases} \hat{\beta}_1 = \dfrac{l_{xy}}{l_{xx}}, \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}, \end{cases} \tag{1.2.2}$$

其中，

$$\bar{x} = \frac{1}{n}\Sigma x_i, \qquad \bar{y} = \frac{1}{n}\Sigma y_i,$$

$$l_{xx} = \Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 - \frac{1}{n}(\Sigma x_i)^2,$$

$$l_{xy} = \Sigma(x_i - \bar{x})(y_i - \bar{y}) = \Sigma x_i y_i - \frac{1}{n}(\Sigma x_i)(\Sigma y_i).$$

称(1.2.2)为 β_0、β_1 的最小二乘估计，简记为 β_0、β_1 的 LSE。

回归方程

$$\begin{aligned} \hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x \\ &= \bar{y} + \hat{\beta}_1(x - \bar{x}) \end{aligned}$$

在平面直角坐标系中的图象称为回归直线，它必过$(0, \hat{\beta}_0)$与$(\bar{x}, \bar{y})$两点。

下面我们就来求例 1.1 中钢的强度 y 关于碳含量 x 的一元线性回归方程，为使计算过程明了，常将中间结果列成如下表格形式(表1 .2.1)。

表 1.2.1

$\Sigma x=1.02$ $\bar{x}=0.102$ $\Sigma x^2=0.1338$	$n=10$ $\Sigma xy=48.555$	$\Sigma y=449.0$ $\bar{y}=44.9$ $\Sigma y^2=20443$
$\frac{1}{n}(\Sigma x)^2=0.10404$ $l_{xx}=0.02976$	$\frac{1}{n}(\Sigma x)(\Sigma y)=45.798$ $l_{xy}=2.757$	$\frac{1}{n}(\Sigma y)^2=20160.1$ $l_{yy}=282.9$
	$\beta_1=\frac{l_{xy}}{l_{xx}}=92.64$	
	$\beta_0=\bar{y}-\hat{\beta}_1\bar{x}=35.45$	

(其中 l_{yy} 是为后面需要求的) 由此得 y 关于 x 的一元线性回归方程为:

$$\hat{y}=35.45+92.64x. \tag{1.2.3}$$

为了后面讨论的需要，我们来研究一下 $\hat{\beta}_0$、$\hat{\beta}_1$ 的若干性质。

性质 1 $\quad \hat{\beta}_1\sim N(\beta_1,\sigma^2/l_{xx})$. (1.2.4)

证：由于

$$\hat{\beta}_1=l_{xy}/l_{xx}=\Sigma\frac{x_i-\bar{x}}{l_{xx}}y_i$$

是 n 个独立正态随机变量的线性组合，因而它也服从正态分布，而正态分布由其期望与方差唯一决定，下面就来求 $\hat{\beta}_1$ 的期望与方差：

$$\begin{aligned}E(\hat{\beta}_1)&=\Sigma\frac{x_i-\bar{x}}{l_{xx}}E(y_i)\\&=\Sigma\frac{x_i-\bar{x}}{l_{xx}}(\beta_0+\beta_1x_i)\\&=\beta_1,\end{aligned}$$

$$\begin{aligned}D(\hat{\beta}_1)&=\Sigma\left(\frac{x_i-\bar{x}}{l_{xx}}\right)^2D(y_i)\\&=\sigma^2/l_{xx},\end{aligned}$$

因此有(1.2.4)。 [证毕]

性质 2 $\quad \hat{\beta}_0\sim N\left(\beta_0,\left(\frac{1}{n}+\frac{\bar{x}^2}{l_{xx}}\right)\sigma^2\right)$. (1.2.5)

证：由于

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$= \sum \left(\frac{1}{n} - \frac{x_i - \bar{x}}{l_{xx}} \bar{x} \right) y_i,$$

同性质 1 的理由知 $\hat{\beta}_0$ 服从正态分布，又

$$E(\hat{\beta}_0) = E(\bar{y}) - E(\hat{\beta}_1)\bar{x}$$

$$= \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x}$$

$$= \beta_0,$$

$$D(\hat{\beta}_0) = \sum \left(\frac{1}{n} - \frac{x_i - \bar{x}}{l_{xx}} \bar{x} \right)^2 D(y_i)$$

$$= \left(\frac{1}{n} - 2 \sum \frac{\bar{x}(x_i - \bar{x})}{n \cdot l_{xx}} + \sum \bar{x}^2 \frac{(x_i - \bar{x})^2}{l_{xx}^2} \right) \sigma^2$$

$$= \left(\frac{1}{n} + \frac{\bar{x}^2}{l_{xx}} \right) \sigma^2,$$

由此得(1.2.5)。 [证毕]

性质 3 $\operatorname{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}}{l_{xx}} \sigma^2$。 (1.2.6)

证：

$$\operatorname{cov}(\hat{\beta}_0, \hat{\beta}_1) = \operatorname{cov}(\bar{y} - \hat{\beta}_1 \bar{x}, \hat{\beta}_1)$$

$$= \operatorname{cov} \left(\frac{1}{n} \sum y_i, \sum \frac{x_i - \bar{x}}{l_{xx}} y_i \right) - \bar{x} \cdot D(\hat{\beta}_1)$$

$$= \sum \frac{x_i - \bar{x}}{n l_{xx}} \sigma^2 - \bar{x} \cdot \frac{\sigma^2}{l_{xx}}$$

$$= -\frac{\bar{x}}{l_{xx}} \sigma^2.$$ [证毕]

由以上三个性质可知，$\hat{\beta}_0$、$\hat{\beta}_1$ 分别是 β_0、β_1 的无偏估计，除了 $\bar{x} = 0$ 外，$\hat{\beta}_0$ 与 $\hat{\beta}_1$ 不独立。

由以上性质还可知，对固定的 x 来讲，

$$\hat{y} = \hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

也是 $y_1, y_2, \cdots, y_n$ 的线性组合，且

$$E(\hat{y}) = E(\hat{\beta}_0) + E(\hat{\beta}_1)x = \beta_0 + \beta_1 x = E(y),$$

$$D(\hat{y}) = D(\hat{\beta}_0) + D(\hat{\beta}_1) \cdot x^2 + 2\operatorname{cov}(\hat{\beta}_0, \hat{\beta}_1)x$$

$$=\left(\frac{1}{n}+\frac{\bar{x}^2}{l_{xx}}\right)\sigma^2+\frac{x^2}{l_{xx}}\sigma^2-2\frac{\bar{x}x}{l_{xx}}\sigma^2$$
$$=\left[\frac{1}{n}+\frac{(x-\bar{x})^2}{l_{xx}}\right]\sigma^2,$$

故

$$\hat{y}\sim N\left(\beta_0+\beta_1 x,\ \left(\frac{1}{n}+\frac{(x-\bar{x})^2}{l_{xx}}\right)\sigma^2\right), \qquad (1.2.7)$$

由此可知 $\hat{y}$ 是 $E(y)$的无偏估计，而 $D(\hat{y})$随 x 与 $\bar{x}$ 的距离$|x-\bar{x}|$的增大而增大.

从上面的讨论中，我们还可以注意到一点，在用 LSE 求 y 关于 x 的一元线性回归方程的系数 $\hat{\beta}_0,\hat{\beta}_1$ 及在讨论 $\hat{\beta}_0$、$\hat{\beta}_1$ 的期望、方差、协方差时只涉及到 $E(\varepsilon_i)=0,D(\varepsilon_i)=\sigma^2$ 及各$\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n$不相关，而在讨论 $\hat{\beta}_0,\hat{\beta}_1$ 的分布时要用到 ε 的正态性. 但在下面几节讨论假设检验与区间估计问题时都是在 $\varepsilon\sim N(0,\sigma^2)$ 的假定下进行的.

§1.3 回归方程的显著性检验

从求一元线性回归方程系数的最小二乘估计公式(1.2.2)可知，不管 y 与 x 之间是否有线性相关关系，只要给出了 n 对数据 $(x_i,y_i),i=1,2,\cdots,n$，总可由(1.2.2)求出 $\hat{\beta}_1$ 与 $\hat{\beta}_0$，从而写出回归方程 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x$，然而这方程不一定有意义. 那么，什么是一个有意义的回归方程呢？我们研究回归方程目的是寻找 y 与 x 之间的统计规律性，即要找出 $E(y)$随 x 变化的规律. 在一元线性回归中，β_1 反映了 $E(y)$随 x 线性变化的变化率，若 $\beta_1=0$，就意味着 $E(y)$不随 x 作线性变化，那么我们给出的一元线性回归方程就没有意义，若 $\beta_1\neq 0$，那么方程才有意义. 因而对回归方程作显著性检验就是要检验假设

$$H_0:\ \beta_1=0 \qquad (1.3.1)$$

是否为真. 为此我们需要寻找一个检验的统计量，它在(1.3.1)为