实用回归分析方法

周 纪 芗 编著

上海科学技术出版社

实用回归分析方法

周纪芗 编著

上海科学技术出版社

实用回归分析方法

周纪芗 编著

上海科学技术出版社出版

（上海瑞金二路450号）

新华书店上海发行所发行 商务印书馆上海印刷厂印刷

开本850×1168 1/32 印张5.5 字数141,000

1990年3月第1版 1990年3月第1次印刷

印数 1—3,300

ISBN 7-5323-1323-9/O·125

定价：2.95元

内 容 简 介

本书是专门论述数理统计中应用广泛的一个重要分支——回归分析的一本著作。主要内容包括：一元线性回归，可以化为一元线性回归的曲线回归，多元线性回归，多项式回归，自变量的选择和逐步回归，书中还辟出最后一章专门论述对最小二乘估计的一些改进方法。本书论述严谨，实例生动，并且还给出了若干用 Basic 语言编写的程序以及用微机算出的一些例题的结果，这些内容增强了本书的实用性。

本书可供大专院校数理统计、管理科学、经济学等专业的师生作教学参考书，也可供许多部门的科技工作者参阅。

序

本书介绍的是数理统计中的一种常用方法——回归分析方法，它是研究变量间相关关系的一种统计方法，在实践中有着广泛的应用.

本书主要通过实例介绍用最小二乘法求线性回归的基本方法和步骤，同时介绍了自变量选择的若干准则及对最小二乘估计的一些改进方法. 本书也适当给出了一些结论的证明. 阅读这些证明需要一定的概率统计和线性代数的知识. 为方便读者阅读本书，除第六章外，在其他有关各章中都将它们集中放在一节中(§1.7, §3.6, §5.4)，跳过这一节并不影响阅读下面的内容. 为方便应用，用 Basic 语言编写的若干程序附在有关章节之中(§1.8, §3.7, §5.5). 书中的例1.3、例4.3及第五、六章全部例题的结果都是在微机上算出的，如果用书中列出的数据运用计算器计算，其结果会与书中的结果有一些误差.

茆诗松教授阅读了全部书稿，提出了许多宝贵意见，在此表示衷心感谢. 由于编者水平有限，书中定有不少不妥之处，恳请读者批评指正.

编著者

引　言

我们在生产实践和科学研究中经常要和各种变量打交道．有些变量可能共处于一个统一体中，它们相互联系、相互制约．为了深入了解事物的本质，往往需要找出描述这些变量之间相互关系的数学表达式．

在实践中，人们发现变量之间的关系一般可分为两类．一类变量间有完全确定的关系，这种关系能用确定的函数形式表示出来．例如电路中著名的欧姆定律就可以用

$$V=IR$$

来表示三个变量——电流 I、电阻 R 与电压 V——之间的关系，只要其中两个已知，第三个变量的值也就完全确定了．

另一类变量间也有一定的关系，但这种关系无法用一个精确的数学式子来表达．例如合金的强度与合金中碳的含量有密切的关系，但是不能由碳的含量精确知道此合金的强度，这是因为合金的强度还受到许多其他因素以及一些无法控制的误差的影响．这种关系在实践中是大量存在的，我们称变量间的这种关系为*相关关系*。

当然，这两类关系是不同的，但是它们之间也并不存在一条严格的界限．随着对问题的研究的深入，变量间的相关关系可能转化为确定性的关系；另一方面，由于测量误差等原因，确定性关系也常以相关关系的形式表现出来．

回归分析就是通过试验和观测，去寻找隐藏在变量间的相关关系的一种数学方法．简单来讲，回归分析是研究变量之间相关关系的一种数学方法，它是数理统计的一个重要分支，在实践中有着广泛的应用．例如求经验公式，寻找产品的产量或质量指标与生产条件之间的关系，气象预报，病虫害预报，建立自动控制中的数学

模型，某些新标准的制订等等，都经常要用到回归分析这一工具.

在回归分析中，我们主要研究以下一些问题：

1° 从一组数据出发，确定变量间是否存在相关关系. 如果存在相关关系的话，需要确定它们之间的定量关系表达式，并对它的可信程度作统计检验.

2° 从共同影响一个变量的许多变量中，判断哪些变量的影响是显著的，哪些变量的影响是不显著的.

3° 利用所找到的定量关系表达式对变量进行预测或控制.

目 录

第一章　一元线性回归

§1.1　一元线性回归的数学模型

我们首先考虑最简单的情况，即两个变量的情况.

若已知变量 x 和 y 之间存在一定的相关关系，希望找出 y 的值是如何随 x 的值的变化而变化的规律，这时通常称 y 为因变量，称 x 为自变量. 那么如何来确定 y 与 x 之间相关关系的表达形式呢？为此我们通过一个例子来加以说明.

例 1.1　由专业知识知道，某合金的抗拉强度 $y(\mathrm{kg/mm^2})$ 与合金中含碳量 $x(\%)$ 之间有一定的相关关系. 为了了解其相关关系的表达形式，第一步是通过试验或从生产记录中去收集 n 组 y 与相应的 x 的值. 表 1.1 就是收集到的 12 个 y 与相应的 x 的值.

表 1.1　合金中含碳量 x 与抗拉强度 y 的数据

编号	x	y	编号	x	y
1	0.10	42.0	7	0.16	49.0
2	0.11	43.5	8	0.17	53.0
3	0.12	45.0	9	0.18	50.0
4	0.13	45.5	10	0.20	55.0
5	0.14	45.0	11	0.21	55.0
6	0.15	47.5	12	0.23	60.0

第二步是画一张散点图. 这是在研究两个变量间相关关系时常用的一种直观办法，以 x 为横坐标，以 y 为纵坐标，每一对数据 (x_α, y_α) 作为一个点在坐标纸上以“×”表示出来，$\alpha=1, 2, \cdots, n$. 本例的散点图见图 1.1.

第三步是观察这张散点图. 从图 1.1 我们发现，这 n 个点基本上在一条直线 l 附近，从而我们可以认为 y 与 x 的关系基本上

是线性的，而这些点与直线 l 的偏离是由其他一切随机因素的影响而造成的. 故我们可以假定表 1.1 中的数据有如下结构：

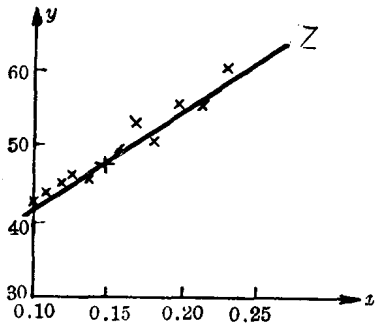

图1.1

$$y=\beta_0+\beta x+\varepsilon, \qquad (1.1)$$

其中 $\beta_0+\beta x$ 表示 y 随 x 的变化而线性变化的部分，ε 是一切随机因素影响的总和，有时也简称随机误差，它是不可观测其值的随机变量，并假定其数学期望 $E(\varepsilon)=0$，方差 $D(\varepsilon)=\sigma^2$. 在涉及到分布时，可进一步假定 ε 服从正态分布 $N(0,\ \sigma^2)$. x 可以是随机变量也可以是一般变量，以下讨论中我们总认为 x 是一般变量，即它是可以精确测量或严格控制的. 由以上假定可知 y 也是一个随机变量，但其值是可以观测的，其数学期望是 x 的线性函数：

$$E(y)=\beta_0+\beta x, \qquad (1.2)$$

这便是 y 与 x 的相关关系的形式.

对表 1.1 的 n 组观测，由(1.1)可认为有

$$\begin{cases} y_\alpha=\beta_0+\beta x_\alpha+\varepsilon_\alpha, \\ \text{各 } \varepsilon_\alpha \text{ 相互独立}, E(\varepsilon_\alpha)=0,\ D(\varepsilon_\alpha)=\sigma^2, \\ \qquad \alpha=1,\ 2,\ \cdots,\ n. \end{cases} \qquad (1.3)$$

(1.3)可以认为是一元线性回归的数学模型.

我们的首要任务就是要根据表 1.1 去求出(1.2)中未知参数 β_0 与 β 的估计 $\hat{\beta}_0$、$\hat{\beta}$，由此可得 $E(y)$ 的估计：

$$\hat{y}=\hat{\beta}_0+\hat{\beta}x. \qquad (1.4)$$

(1.4)称为 y 关于 x 的一元线性回归方程. 这便是我们要求的 y 与 x 之间的定量关系表达式，其图象便是图 1.1 中的直线 l. 称此直线为回归直线，$\hat{\beta}$ 也称为回归系数，它是回归直线的斜率，$\hat{\beta}_0$ 为回归常数，它是回归直线的截距.

§1.2 参数 β_0, β 的最小二乘估计

要求出回归方程(1.4)就是要求出 β_0 与 β 的估计，求此估计的一个直观想法便是希望对一切 x_α，观测值 y_α 与回归值 $\hat{y}_\alpha=\hat{\beta}_0+\hat{\beta}x_\alpha$ 的偏离达到最小，为此我们通常采用最小二乘法来求 β_0 与 β 的估计．令

$$Q(\beta_0, \beta)=\sum_{\alpha=1}^{n}(y_\alpha-\beta_0-\beta x_\alpha)^2. \tag{1.5}$$

所谓 β_0 与 β 的最小二乘估计是指使下式成立的 $\hat{\beta}_0$ 与 $\hat{\beta}$：

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

由于 $Q(\beta_0, \beta)$ 是 β_0, β 的非负二次函数，其最小值必存在，同时它是 β_0, β 的可微函数，故根据微积分学中的极值原理，$\hat{\beta}_0$, $\hat{\beta}$ 应是下列方程组的解：

$$\begin{cases}\left.\dfrac{\partial Q}{\partial \beta_0}\right|_{\beta_0=\hat{\beta}_0,\beta=\hat{\beta}}=-2\sum\limits_{\alpha=1}^{n}(y_\alpha-\hat{\beta}_0-\hat{\beta}x_\alpha)=0,\\ \left.\dfrac{\partial Q}{\partial \beta}\right|_{\beta_0=\hat{\beta}_0,\beta=\hat{\beta}}=-2\sum\limits_{\alpha=1}^{n}(y_\alpha-\hat{\beta}_0-\hat{\beta}x_\alpha)x_\alpha=0.\end{cases} \tag{1.6}$$

(1.6)可化简成

$$\begin{cases} n\hat{\beta}_0+(\sum\limits_{\alpha}x_\alpha)\hat{\beta}=\sum\limits_{\alpha}y_\alpha,\\ (\sum\limits_{\alpha}x_\alpha)\hat{\beta}_0+(\sum\limits_{\alpha}x_\alpha^2)\hat{\beta}=\sum\limits_{\alpha}x_\alpha y_\alpha.\end{cases} \tag{1.7}$$

(1.7) 称为正规方程组，其中"$\sum_\alpha$"表示"$\sum_{\alpha=1}^{n}$"，下面不再重复说明．

为求正规方程组(1.7)的解，我们可从第一式中求得

$$\hat{\beta}_0=\frac{1}{n}\sum_\alpha y_\alpha-\hat{\beta}\cdot\frac{1}{n}\sum_\alpha x_\alpha=\bar{y}-\hat{\beta}\bar{x},$$

将它代入第二式，得

$$(\sum_\alpha x_\alpha)(\bar{y}-\hat{\beta}\bar{x})+(\sum_\alpha x_\alpha^2)\hat{\beta}=\sum_\alpha x_\alpha y_\alpha,$$

经过整理，得

$$\left[\sum_{\alpha} x_{\alpha}^{2}-\frac{1}{n}\left(\sum_{\alpha} x_{\alpha}\right)^{2}\right]\hat{\beta}=\sum_{\alpha} x_{\alpha} y_{\alpha}-\frac{1}{n}\left(\sum_{\alpha} x_{\alpha}\right)\left(\sum_{\alpha} y_{\alpha}\right).$$

故求得其解为：

$$\begin{cases}\hat{\beta}=\dfrac{\sum\limits_{\alpha} x_{\alpha} y_{\alpha}-\dfrac{1}{n}\left(\sum\limits_{\alpha} x_{\alpha}\right)\left(\sum\limits_{\alpha} y_{\alpha}\right)}{\sum\limits_{\alpha} x_{\alpha}^{2}-\dfrac{1}{n}\left(\sum\limits_{\alpha} x_{\alpha}\right)^{2}}, \\ \hat{\beta}_{0}=\bar{y}-\hat{\beta}\bar{x}.\end{cases} \tag{1.8}$$

为简化记号，令

$$l_{xy}=\sum_{\alpha}\left(x_{\alpha}-\bar{x}\right)\left(y_{\alpha}-\bar{y}\right)=\sum_{\alpha} x_{\alpha} y_{\alpha}-\frac{1}{n}\left(\sum_{\alpha} x_{\alpha}\right)\left(\sum_{\alpha} y_{\alpha}\right),$$

$$l_{xx}=\sum_{\alpha}\left(x_{\alpha}-\bar{x}\right)^{2}=\sum_{\alpha} x_{\alpha}^{2}-\frac{1}{n}\left(\sum_{\alpha} x_{\alpha}\right)^{2},$$

其中，

$$\bar{x}=\frac{1}{n}\sum_{\alpha} x_{\alpha},\ \bar{y}=\frac{1}{n}\sum_{\alpha} y_{\alpha},$$

则最小二乘估计为

$$\begin{cases}\hat{\beta}_{0}=\bar{y}-\hat{\beta}\bar{x}, \\ \hat{\beta}=\dfrac{l_{xy}}{l_{xx}}.\end{cases} \tag{1.9}$$

下面我们来求例 1.1 的回归方程，计算过程通常用表格形式给出(表 1.2 与表 1.3)．为了下面进一步分析的需要，在此将 $l_{yy}=\sum\limits_{\alpha}\left(y_{\alpha}-\bar{y}\right)^{2}=\sum\limits_{\alpha} y_{\alpha}^{2}-\dfrac{1}{n}\left(\sum\limits_{\alpha} y_{\alpha}\right)^{2}$ 一起求出了，它在求回归方程时是不需要的．

若将 $\hat{\beta}_{0}$ 的表达式代入一元线性回归方程

$$\hat{y}=\hat{\beta}_{0}+\hat{\beta}x,$$

表 1.2　回归方程计算表(I)

编号	x	y	x^2	xy	y^2
1	0.10	42.0	0.0100	4.200	1764.00
2	0.11	43.5	0.0121	4.785	1892.25
3	0.12	45.0	0.0144	5.400	2025.00
4	0.13	45.5	0.0169	5.915	2070.25
5	0.14	45.0	0.0196	6.300	2025.00
6	0.15	47.5	0.0225	7.125	2256.25
7	0.16	49.0	0.0256	7.840	2401.00
8	0.17	53.0	0.0289	9.010	2809.00
9	0.18	50.0	0.0324	9.000	2500.00
10	0.20	55.0	0.0400	11.000	3025.00
11	0.21	55.0	0.0441	11.550	3025.00
12	0.23	60.0	0.0529	13.800	3600.00
Σ	1.90	590.5	0.3194	95.925	29392.75

表 1.3　回归方程计算表(II)

$\Sigma x_\alpha=1.90$	$\Sigma y_\alpha=590.5$	$n=12$
$\bar{x}=0.1583$	$\bar{y}=49.2083$	
$\Sigma x_\alpha^2=0.3194$	$\Sigma x_\alpha y_\alpha=95.925$	$\Sigma y_\alpha^2=29392.75$
$\frac{1}{n}(\Sigma x_\alpha)^2=0.3008$	$\frac{1}{n}(\Sigma x_\alpha)(\Sigma y_\alpha)=93.4958$	$\frac{1}{n}(\Sigma y_\alpha)^2=29057.5208$
$l_{xx}=0.0186$	$l_{xy}=2.4292$	$l_{yy}=335.2292$
	$\hat{\beta}=\frac{l_{xy}}{l_{xx}}=130.6022$	
	$\hat{\beta}_0=\bar{y}-\hat{\beta}\bar{x}=28.5340$	
	$\hat{y}=28.5340+130.6022x$	(1.10)

则可得另一种表达形式:

$$\hat{y}=\bar{y}+\hat{\beta}(x-\bar{x}), \tag{1.11}$$

由此可知回归直线过$(\bar{x}, \bar{y})$这一点,对作回归直线是很有帮助的.

当观测值较大(或过小)时,为简化计算,可以对数据作如下线性变换. 令

$$x'=\frac{x-c_x}{d_x}, \quad y'=\frac{y-c_y}{d_y}, \tag{1.12}$$

其中 c_x, c_y, d_x, d_y 都是适当选取的常数，且 $d_x\neq 0$, $d_y\neq 0$. 为建立 y 关于 x 的线性回归方程，可以先建立 y' 关于 x' 的线性回归方程，然后再将变换式(1.12)代入. 为此先对 (x_α, y_α), $\alpha=1,2,\cdots,n$, 作变换(1.12)，得到 (x'_α, y'_α), $\alpha=1, 2, \cdots, n$, 以这 n 组观测值，利用公式(1.9)求得一元线性回归方程：

$$\hat{y}'=\hat{\beta}'_0+\hat{\beta}'x', \tag{1.13}$$

再将变换(1.12)代入(1.13)，整理后得：

$$\begin{aligned}\hat{y}&=c_y+d_y\left(\hat{\beta}'_0+\hat{\beta}'\frac{x-c_x}{d_x}\right)\\&=c_y+d_y\hat{\beta}'_0-\frac{d_y}{d_x}\hat{\beta}'c_x+\frac{d_y}{d_x}\hat{\beta}'x\\&=\hat{\beta}_0+\hat{\beta}x.\end{aligned}$$

其中

$$\hat{\beta}_0=c_y+d_y\hat{\beta}'_0-\frac{d_y}{d_x}\hat{\beta}'c_x,$$

$$\hat{\beta}=\frac{d_y}{d_x}\hat{\beta}'.$$

例 1.2 为研究黄铜延性 y(%)关于退火温度 x(℃)之间的关系，现收集了如下数据：

x(℃)	300	400	500	600	700	800
y(%)	40	50	55	60	67	70

试求 y 关于 x 的一元线性回归方程.

解 由于 x 的取值为 100 的倍数，y 的取值大部分大于 50，故为了简化起见，可以令

$$c_x=500,\ d_x=100,\ c_y=50,\ d_y=1,$$

即令

$$x'=\frac{x-500}{100},\quad y'=y-50. \tag{1.14}$$

这样我们可以利用表 1.4 与表 1.5 求出 y' 关于 x' 的回归方程.

表 1.4

序号	x	y	x'	y'
1	300	40	−2	−10
2	400	50	−1	0
3	500	55	0	5
4	600	60	1	10
5	700	67	2	17
6	800	70	3	20
Σ			3	42

表 1.5

$$\Sigma x'_\alpha=3 \qquad \Sigma y'_\alpha=42$$

$$\bar{x}'=0.5 \qquad \bar{y}'=7$$

$$\Sigma x'^2_\alpha=19 \qquad \Sigma x'_\alpha y'_\alpha=124 \qquad \Sigma y'^2_\alpha=914$$

$$\frac{1}{n}(\Sigma x'_\alpha)^2=1.5 \qquad \frac{1}{n}(\Sigma x'_\alpha)(\Sigma y'_\alpha)=21 \qquad \frac{1}{n}(\Sigma y'_\alpha)^2=294$$

$$l_{x'x'}=17.5 \qquad l_{x'y'}=103 \qquad l_{y'y'}=620$$

$$\hat{\beta}'=5.8857$$

$$\hat{\beta}'_0=4.0571$$

$$\hat{y}'=4.0571+5.8857x' \tag{1.15}$$

再将(1.14)代入(1.15)并化简,得

$$\hat{y}-50=4.0571+5.8857\cdot\frac{x-500}{100},$$

故

$$\hat{y}=24.6286+0.058857x. \tag{1.16}$$

这与直接用 x 与 y 求得的回归方程是一致的,但这里计算较简单. 当然,对此问题,数据变换方式很多,不一定非用(1.14). 例如也可令 $y'=\frac{y-40}{10}$, $x'=\frac{x-300}{100}$,使 x', y' 中不出现负数. 总之,变换是以方便为原则的.

§1.3 回归方程的显著性检验

从上一节介绍的求 y 关于 x 的一元线性回归方程的过程可

知，在计算过程中并不需要假定 y 与 x 一定有相关关系．即使是对于 n 对杂乱无章的数据 (x_α, y_α)，同样可以按(1.8)求得回归系数的最小二乘估计，从而获得 y 关于 x 的一元线性回归方程．但是这种方程毫无实际意义．在§1.1我们曾经指出，只有在散点图上看到 n 个点落在一直线附近时才能认为 y 与 x 之间可配一元线性回归方程．为此我们必须定量地给出在什么情况下可以认为"n 个点落在一直线附近"，从统计上讲也就是 $E(y)$ 必须随 x 的变化而线性变化，即(1.2)中的 β 不能等于 0．所以问题就变成了去检验假设 $\beta=0$ 是否为真．若 $\beta=0$ 为真，说明不管 x 如何变化，$E(y)$ 并不是随 x 而线性变化的；反之若 $\beta\neq0$，则当 x 变化时，$E(y)$ 是随 x 而线性变化的，只有这时回归方程才是有意义的．

为检验假设 $\beta=0$ 是否为真，我们可以从分析引起各 y_α，$\alpha=1, 2, \cdots, n$ 的不同原因着手．n 个 y_α 的值之所以不同，可能有两个方面的原因：其一，若 $E(y)$ 确是随 x 线性变化的话，那么 x 的取值不同就是一个原因；其二是其他一切因素的影响．若前一方面的影响是主要的，那么 $\beta\neq0$，方程是有意义的，否则方程就没有意义．为此必须把这两个原因所引起的 y_α 的波动大小从 y_α 的总的波动中分解出来．

记

$$S_T=l_{yy}=\sum_\alpha(y_\alpha-\bar{y})^2, \tag{1.17}$$

称它为总的偏差平方和，它反映了各 y_α 的波动大小．由于

$$\begin{aligned}S_T&=\sum_\alpha(y_\alpha-\bar{y})^2\\&=\sum_\alpha(y_\alpha-\hat{y}_\alpha+\hat{y}_\alpha-\bar{y})^2\\&=\sum_\alpha(y_\alpha-\hat{y}_\alpha)^2+\sum_\alpha(\hat{y}_\alpha-\bar{y})^2+2\sum_\alpha(y_\alpha-\hat{y}_\alpha)(\hat{y}_\alpha-\bar{y}),\end{aligned}$$

由(1.6)可知

$$\begin{aligned}&\sum_\alpha(y_\alpha-\hat{y}_\alpha)(\hat{y}_\alpha-\bar{y})\\&=\sum_\alpha(y_\alpha-\hat{\beta}_0-\hat{\beta}x_\alpha)(\hat{\beta}_0+\hat{\beta}x_\alpha-\bar{y})=0,\end{aligned}$$

故