线性回归及其应用研究

XIANXING HUIGUI JIQI YINGYONG YANJU

荆庆林 著

吉林大学出版社

线性回归及其应用研究

XIANXING HUIGUI JIQI YINGYONG YANJU

荆庆林 著

吉林大学出版社

图书在版编目(CIP)数据

线性回归及其应用研究/荆庆林著.--长春:吉林大学出版社,2012.12

ISBN 978-7-5601-9473-8

Ⅰ.①线… Ⅱ.①荆… Ⅲ.①线性回归－研究 Ⅳ.①O212.1

中国版本图书馆 CIP 数据核字(2012)第 297612 号

书　名:线性回归及其应用研究

作　者:荆庆林　著

责任编辑:孟亚黎　责任校对:刘守秀　封面设计:王菊红

吉林大学出版社出版、发行　北京市登峰印刷厂　印刷

开本:850×1168 毫米　1/32　2012 年 12 月第 1 版

印张:5.375　字数:145 千字　2012 年 12 月第 1 次印刷

ISBN 978-7-5601-9473-8　定价:28.00 元

社址:长春市明德路 501 号　邮编:130021

发行部电话:0431－89580026/28/29

网址:http://www.jlup.com.cn

E－mail:jlup@mail.jlu.edu.cn

前　言

线性回归是根据数理统计中的回归分析，从而确定两种或者两种以上变量间相互依赖的定量关系的一种统计分析方法之一，随着高速电子计算机的日益普及，其在医学、生物、经济、工业、农业、管理、工程技术等领域都得到了长足的发展.

本书在撰写方面力求突出以下特点：

1. 力求做到难易适当、深入浅出，融会贯通；

2. 讲解理论重点、层次分明、通俗易懂；

3. 案例丰富，有利于读者掌握所学理论知识.

全书共分五章，第一章为回归分析的简介，本章分为三小节分别为回归分析的基本概念、一元线性回归及多元线性回归，主要讲述线性回归的基本理论知识，从而使读者对线性回归有初步认识，继而有助于后面理论知识的讲解；第二章讲述参数估计的相关知识；第三章讲述假设检验与预测；第四章讲述线性回归模型；第五章为应用案例与实验，本章列举了大量的应用案例，有助于对前几章所学理论的理解和巩固.

本书在撰写过程中得到了许多同行、专家的支持和帮助，在此表示衷心的感谢；撰写时参考了大量的相关著作和文献资料，选用了其中的部分内容和习题，在此向有关作者表示感谢.

由于作者水平有限，书中错误或不当之处在所难免，热忱欢迎同行和广大读者朋友批评指正.

作者

2012 年 8 月

目　录

第一章　回归分析的简介

1.1　回归分析的基本概念

在19世纪，生物统计学家高尔顿在研究父子身高的遗传规律时，对1 078对父子的身高进行了观察，用x表示父亲的身高，y表示成年儿子的身高，把这1 078对数据放到直角坐标系中发现，这些点基本在一条直线附近，该直线方程为(单位：in，1in = 2.54cm)：

$$y = 33.73 + 0.516x.$$

其观察结果表明：

(1) 父亲的身高每增加一个单位，则儿子的身高平均增加0.516个单位；

(2) 矮个子父亲的儿子们的平均身高要比父辈们平均身高要高些；

(3) 高个子父亲有生高个子儿子的趋势，然而高个子父亲的儿子们的平均身高要低于父辈们的平均身高.

这便是子代的平均身高有向中心回归的趋势，从而使得在一段时间内人的身高相对稳定. 在这之后回归分析的应用越来越广泛.

在现实生活中我们经常可以看到多个变量处于同一个过程，且这些变量相互关系、相互制约，其关系大致可分为两类：一类为确定性关系，即我们通常所说的函数关系；另一类为非确定性关系.

确定关系是指当一些变量的值确定后另一些变量的值也会随之

而确定的关系，这些变量之间的关系是完全已知的，可用函数 $y = f(x)$ 表示. 例如，电路中电阻值 R、电压 U 与电流 I 之间的关系为 $U = IR$ 等.

非确定性关系是指变量之间有着一定的依赖关系，然而当其中一些变量确定以后，另有一些变量的值虽然随之发生变化，但是却不能完全确定，此时变量间的关系则不能精确地用函数来表示. 例如，农作物的单位面积产量与施肥量之间有着密切的关系，但是这两变量之间的关系我们却不能用确定的函数关系来表示. 该例子中变量间的关系为不确定关系，通常我们将这样的关系称之为相关关系.

确定性关系与相关关系我们往往无法截然区分. 一方面，因为受到测量误差等随机因素的影响，确定关系在实际中往往会通过相关关系表现出来；另一方面，当我们对客观事物的内部规律有了深入了解时，相关关系则又可能转化为确定关系.

回归分析为数理统计中处理变量之间非确定关系的一种方法. “回归”是指随机变量的取值回归到其平均值. 回归分析是数理统计中研究一个响应变量与若干个预报变量之间相关关系的一种十分有效的方法，利用估计结果做预测和控制. 在大量重复试验中，由不确定性的变量所呈现出的统计规律性，称之为统计相关. 两个变量之间的相关关系称之为单相关，三个以上的变量的相关关系称之为复关系，在复相关条件下仅研究两个变量间的相关关系称之为偏相关.

设有两个变量 X 和 Y，其中 X 为可精确测量或者控制的非随机变量，Y 为随机变量，X 的变化会引起 Y 相应变化，但是它们之间的变化关系是不确定的. 若当 X 取得任一可能值 x 时，Y 相应地服从一定的概率分布，那么称随机变量 Y 与变量 X 之间存在相关关系.

若进行 n 次相互独立的试验从而测得试验数据如表 1-1-1 所示.

其中 x_i 表示变量 X 在第 i 次试验中观察到的值，y_i 表示随机变量 Y 相应的观察值. 一般将点 $(x_i, y_i)(i = 1, 2, \cdots, n)$ 画在直角坐标平面上，可得到散点图(图 1-1-1).

表 1-1-1　试验数据

X	x_1	x_2	…	x_n
Y	y_1	y_2	…	y_n

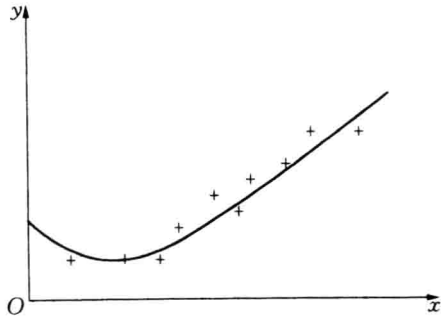

图 1-1-1

怎样根据这些观察值使用“最佳的、确定的”函数 $\hat{y}=f(x)$ 来表达变量 X 和 Y 之间的相关关系?称 $f(x)$ 为 Y 对 X 的回归函数,$\hat{y}=f(x)$ 为 Y 对 X 的回归方程. 由于找出回归函数 $f(x)$ 比较困难,所以一般限制函数 $f(x)$ 为某一类型的函数. 而函数 $f(x)$ 的类型一般由被研究问题的假设来确定. 若没有任何理由可确定 $f(x)$ 的类型,那么我们可根据在试验结果中得到的散点图来确定.

函数 $f(x)$ 的类型确定后,那么可设

$$f(x)=f(x;\alpha_1,\alpha_2,\cdots,\alpha_k),$$

其中 $\alpha_1,\alpha_2,\cdots,\alpha_k$ 为未知参数,那么如何根据试验所得数据合理选择参数 $\alpha_1,\alpha_2,\cdots,\alpha_k$ 的值,使得

$$\hat{y}=f(x;\alpha_1,\alpha_2,\cdots,\alpha_k)$$

在一定条件下“最佳地”表现变量 Y 与 X 之间的相关关系?

解决这个问题我们通常采用最小二乘法,即要求选取 $f(x;\alpha_1,\alpha_2,\cdots,\alpha_k)$ 中的参数,使其观察值 y_i 和相应的函数值 $\hat{y}=f(x;\alpha_1,\alpha_2,\cdots,\alpha_k)(i=1,2,\cdots,n)$ 的偏差平方和 $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ 达到最小值.

下面我们总结一下回归函数的获取步骤：

(1) 确定相关变量间的函数类型，通常从实际问题中获得，因此该种数学表达式也称之为经验公式；

(2) 根据回归标准，确定待定参数继而得到回归函数；

(3) 检验和判定建立的回归函数的有效性.

接下来我们对回归函数作几点说明：

(1) 回归函数并不是完全表达了相关变量间的全部统计规律，只是代表了其主要方面，也可以说是它的一个拟合；

(2) 经验公式的选取要以有效性为第一判断标准，在有效程度相差不大时，应选择函数形式简单，含待定参数少的数学表达式；

(3) 尽量选择与实际意义相符的函数类型.

回归分析的主要任务：

(1) 确定变量之间是否存在相关关系，若存在，则找出它们之间的数学表达式，将其称为回归方程，并且判断回归方程是否有效，判断哪些预报变量对响应变量为显著的，哪些预报变量为不显著的；

(2) 利用回归方程，依据一个或者多个变量的值，预测或者控制另一个的变量的取值，并估计该种预测可达到什么样的精度；

(3) 进行因素分析，对共同影响一个变量的诸变量之间，找出哪些因素为重要因素，哪些因素为次要因素以及它们之间的关系；

(4) 依据预测或者控制所提出的要求，选择实验，并且对实验进行设计.

1.2 一元线性回归

1.2.1 一元线性回归模型

首先我们看一个例题.

例 1-2-1　在某个化工生产过程中，为了研究温度 x(℃) 对产品的收率 y(%) 的影响，测得一组数据，如表 1-2-1 所示.

表 1-2-1

温度 x/℃	100	110	120	130	140	150	160	170	180	190
收率 y/%	45	51	54	61	66	70	74	78	85	89

将表 1-2-1 中各对数据在坐标平面上描出，如图 1-2-1 所示. 根据图 1-2-1 可知，随着温度 x 的升高，收率 y 也随之不断增加，数据点大致分布在某条直线的周围，我们则可认为在 y 和 x 之间存在着线性关系，由于存在着不可知或者不可观测的随机因素的影响，所以散点没有完全落在直线上.

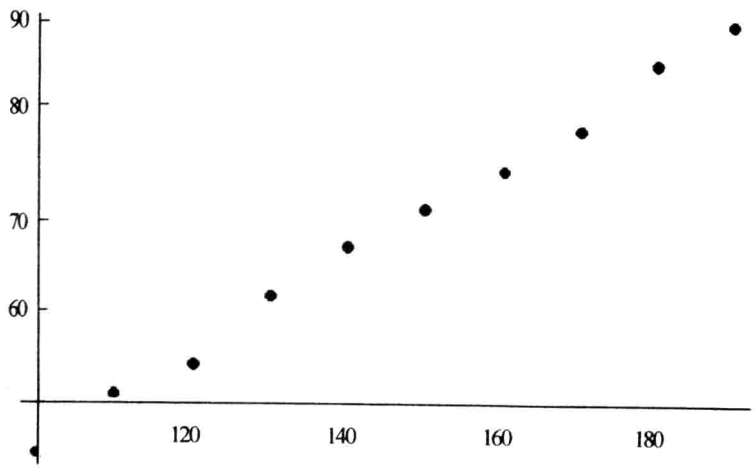

图 1-2-1

一般情况下，若给定一组数据

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n).$$

在直角坐标系中描出相应的点，从而得到散点图. 散点近似落在一条直接上，提出以下假设：

$$\begin{cases} y = \alpha + \beta x + \varepsilon \\ \varepsilon \sim N(0, \sigma^2) \end{cases}. \tag{1-2-1}$$

其中 α, β, σ^2 为待定的参数. ε 表示随机因素的影响. 因为这种影响常常为大量的不可知因素引起的，并且每个因素都起微小的作用，认为 ε 服从正态分布.

将式(1-2-1) 称为一元线性回归模型. 易知，y 的数学期望为

$$E(y) = \alpha + \beta x,$$

回归函数为

$$\hat{y} = \alpha + \beta x.$$

1.2.2 α, β 的最小二乘估计

在散点图上可画出直线，使得直线两边的散点分布比较均匀，从而可得到一条回归直线，那么哪一条直线最好，需要给出一个标准，最小二乘法为其标准之一.

取 x 的 n 个不全相同的值 $x_1, x_2, \cdots, x_n$ 作独立试验，从而得到样本 $(x_1, Y_1), (x_2, Y_2), \cdots, (x_n, Y_n)$. 由模型(1-2-1) 我们不难得出

$$Y \sim N(\alpha + \beta x, \sigma^2),$$

从而相应地有

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2), i = 1, 2, \cdots, n.$$

因为 $Y_1, Y_2, \cdots, Y_n$ 相互独立，所以 n 维随机向量 $(Y_1, Y_2, \cdots, Y_n)$ 的联合分布密度为

$$\begin{aligned} L &= \prod_{i=1}^{n} \frac{1}{\sigma \sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2} (y_i - \alpha - \beta x_i)^2 \right] \\ &= \left(\frac{1}{\sigma \sqrt{2\pi}} \right)^n \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \right]. \end{aligned} \tag{1-2-2}$$

使用最大似然估计法来估计未知参数 α, β. 对于任意一组观察值 $y_1, y_2, \cdots, y_n$，则式(1-2-2) 为其样本的似然函数. 易知，若要使 L 取最大

值，则式(1-2-2)右端方括弧中的平方和部分取得最小，即函数

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

取得最小值，则 $Q(\alpha,\beta)$ 可看作为 α,β 的二元函数，其表示为 y_i 偏离直线 $\alpha+\beta x_i$ 的偏差平方和. 选取 α,β 使得偏差平方和取得最小，称其为最小二乘原理.

取 $Q(\alpha,\beta)$ 分别关于 α,β 的偏导数，并且令它们等于零：

$$\begin{cases}\dfrac{\partial Q}{\partial\alpha}=-2\sum\limits_{i=1}^{n}(y_i-\alpha-\beta x_i)=0\\[2ex]\dfrac{\partial Q}{\partial\beta}=-2\sum\limits_{i=1}^{n}(y_i-\alpha-\beta x_i)x_i=0\end{cases},$$

解方程组可得

$$\begin{cases}n\alpha+(\sum\limits_{i=1}^{n}x_i)\beta=\sum\limits_{i=1}^{n}y_i\\[2ex](\sum\limits_{i=1}^{n}x_i)\alpha+(\sum\limits_{i=1}^{n}x_i^2)\beta=\sum\limits_{i=1}^{n}x_iy_i\end{cases}.\tag{1-2-3}$$

式(1-2-3)称为正规方程组.

因为 x_i 不全相同，则正规方程组的系数行列式

$$\begin{vmatrix}n & \sum\limits_{i=1}^{n}x_i\\[2ex]\sum\limits_{i=1}^{n}x_i & \sum\limits_{i=1}^{n}x_i^2\end{vmatrix}=n\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2=n\sum_{i=1}^{n}(x_i-\bar{x})^2\neq 0.$$

所以式(1-2-3)有唯一一组解. 从而解得 α,β 的最大似然估计值为

$$\beta\Big[n\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2\Big]=\sum_{i=1}^{n}x_iy_i-(\sum_{i=1}^{n}x_i)(\sum_{i=1}^{n}y_i)$$

可得

$$\hat{\beta}=\frac{n\sum_{i=1}^{n}x_iy_i-\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}$$

$$=\frac{\sum_{i-1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2},$$

或者

$$\hat{\beta}=\frac{L_{xy}}{L_{xx}}.$$

易知

$$\hat{\alpha}=\frac{1}{n}\sum_{i=1}^{n}y_i-\frac{\hat{\beta}}{n}\sum_{i=1}^{n}x_i=\overline{y}-\hat{\beta}\overline{x}. \tag{1-2-4}$$

其中$\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$，$\overline{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$.

在得到α，β的估计$\hat{\alpha}$，$\hat{\beta}$后，对于给定的x，则可取$\hat{\alpha}+\hat{\beta}x$作为回归函数

$$\mu(x)=\alpha+\beta x$$

的估计，即有$\widehat{\mu(x)}=\hat{\alpha}+\hat{\beta}x$，称其为$Y$关于$x$的经验回归函数. 记作

$$\hat{\alpha}+\hat{\beta}x=\hat{y},$$

方程

$$\hat{y}=\hat{\alpha}+\hat{\beta}x$$

称其为Y关于x的经验回归方程，简称为回归方程，图形称为回归直线.

根据式(1-2-3)中

$$n\alpha+\left(\sum_{i-1}^{n}x_i\right)\beta=\sum_{i=1}^{n}y_i$$

整理可得

$$\alpha+\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)\beta=\frac{1}{n}\sum_{i=1}^{n}y_i,$$

即有

$$\overline{y}=\alpha+\beta\overline{x}.$$

点$(\overline{x},\overline{y})$可看作为数据点$(x_i,y_i)(i=1,2,\cdots,n)$的几何重心，根据式(1-2-4)可知，其恰好落在回归直线上. 在x_i处的观察值y_i与$\hat{y}_i=\hat{\alpha}+\hat{\beta}x_i$的差$y_i-\hat{y}_i$称之为残差.

接下来我们计算例1-2-1中所给数据的回归直线方程.

此处$n=10$.

$$\sum_{i=1}^{10}x_i=1\ 450,\overline{x}=\frac{1}{10}\sum_{i=1}^{10}x_i=145.$$

$$\sum_{i=1}^{10}y_i=673,\overline{y}=\frac{1}{10}\sum_{i=1}^{10}y_i=67.3.$$

$$\sum_{i=1}^{10}x_i^2=218\ 500,\sum_{i=1}^{10}x_iy_i=101\ 570,\sum_{i=1}^{10}y_i^2=47\ 225.$$

$$L_{xx}=\sum_{i=1}^{10}x_i^2-n\overline{x}^2=8\ 250,$$

$$L_{yy}=\sum_{i=1}^{10}y_i^2-n\overline{y}^2=1\ 932,$$

$$L_{xy}=\sum_{i=1}^{10}x_iy_i-n\overline{x}\,\overline{y}=3\ 985.$$

$$\hat{\beta}=\frac{L_{xy}}{L_{xx}}=0.483\ 030,$$

$$\hat{\alpha}=\overline{y}-\hat{\beta}\overline{x}=-2.739\ 3.$$

回归直线方程为

$$\hat{y}=-2.739\ 3+0.483\ 030x.$$

代入x_i的值计算出$\hat{y}_i$，与y_i进行比较，见表1-2-2.

表 1-2-2　例 1-2-1 的计算表

x_i	y_i	$\hat{y}_i$	$y_i-\hat{y}_i$	$(y_i-\hat{y}_i)^2$
100	45	42.833	−0.568	0.256 0
110	51	47.663 3	0.601 7	0.362 0
120	54	52.493 6	−1.228 6	1.509 5
130	61	57.323 9	0.941 6	0.886 6
140	66	62.154 2	1.110 8	1.233 9
150	70	66.985 4	0.280 5	0.078 7
160	74	71.814 8	−0.549 8	0.302 3
170	78	76.645 1	−1.380 1	1.904 7
180	85	81.475 4	0.789 6	0.623 5
190	89	86.305 7	−0.040 7	0.001 7
$\sum$				7.158 9

画出回归直线的图形并且和散点图叠加在一起，如图 1-2-2 所示.

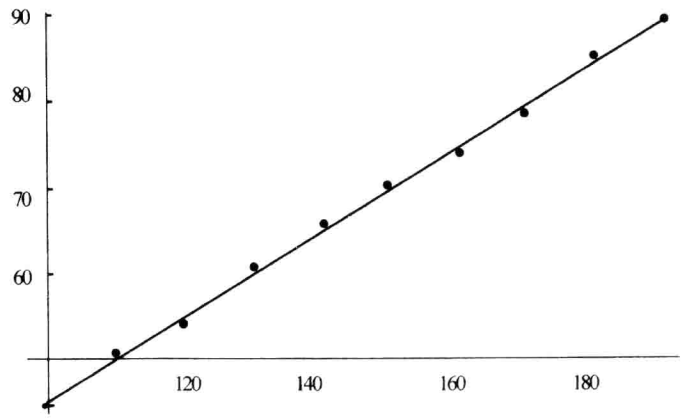

图 1-2-2

从图中我们可看出，数据点分布在直线的附近，其残差平方和仅为 7.158 9.

1.2.3　σ^2 的估计

根据中心极限定理可知，随机误差 ε 近似地服从正态分布 $N(0,\sigma^2)$，为了估计 σ^2 的值，需要考察残差平方和

$$Q_e=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}x_i)^2$$

其图形如图 1-2-3 所示.

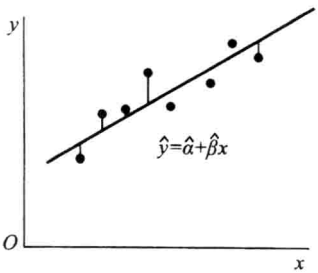

图 1-2-3

其为经验回归函数在 x_i 处的函数值 $\widehat{\mu(x_i)}=\hat{\alpha}+\hat{\beta}x_i$ 与 x_i 处的观察值 y_i 的偏差的平方和. 将 Q_e 作如下分解：

$$\begin{aligned}Q_e&=\sum_{i=1}^{n}[(y_i-\overline{y})-\hat{\beta}(x_i-\overline{x})]^2\\&=\sum_{i=1}^{n}(y_i-\overline{y})^2-2\hat{\beta}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})+\hat{\beta}^2\sum_{i=1}^{n}(x_i-\overline{x})^2\\&=L_{yy}-2\hat{\beta}L_{xy}+\hat{\beta}^2L_{xx}\\&=L_{yy}-\hat{\beta}^2L_{xx}.\end{aligned}$$

Q_e 的数学期望为