实用回归分析

方开泰 全 辉 陈庆云 编著

科学出版社

实 用 回 归 分 析

方开泰　全　辉　陈庆云 编著

科 学 出 版 社

1 9 8 8

内 容 简 介

回归分析是数理统计中很重要的方法．本书重点介绍回归分析的基本思想和常用方法，有些方法是近十几年才发展起来的，并且很有实用价值（如岭回归、压缩估计、最优回归子集等）．书中介绍的方法均用例子来说明，以便读者理解和使用．

本书可供工程技术人员、管理干部、科学工作者阅读，也可供大专院校有关专业师生参考．

实 用 回 归 分 析

方开泰　全　辉　陈庆云　编著

责任编辑　毕　颖

科学出版社出版

北京朝阳门内大街137号

中国科学院印刷厂印刷

新华书店北京发行所发行　各地新华书店经售

*

1988年10月第一版　开本：787×1092　1/32

1988年10月第一次印刷　印张：11 1/2

印数：0001—5,040　字数：260,000

ISBN 7-03-000544-9/O·140

定价：4.30元

实用回归分析

方开泰 全 辉 陈庆云 编著

科学出版社

1988

内容简介

回归分析是数理统计中很重要的方法．本书重点介绍回归分析的基本思想和常用方法，有些方法是近十几年才发展起来的，并且很有实用价值（如岭回归、压缩估计、最优回归子集等）．书中介绍的方法均用例子来说明，以便读者理解和使用．

本书可供工程技术人员、管理干部、科学工作者阅读，也可供大专院校有关专业师生参考．

实用回归分析

方开泰　全　辉　陈庆云　编著

责任编辑　毕　颖

科学出版社出版

北京朝阳门内大街137号

中国科学院印刷厂印刷

新华书店北京发行所发行　各地新华书店经售

*

1988年10月第一版　开本：787×1092　1/32

1988年10月第一次印刷　印张：11 1/2

印数：0001—5,040　字数：260,000

ISBN 7-03-000544-9/O·140

定价：4.30元

证 $$\begin{aligned} E(\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}) &= E[\operatorname{tr}(\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x})] \\ &= E[\operatorname{tr}(\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}')] \\ &= \operatorname{tr}[\boldsymbol{A}E(\boldsymbol{x}\boldsymbol{x}')] \\ &= \operatorname{tr}[\boldsymbol{A}D(\boldsymbol{x}) + \boldsymbol{A}\boldsymbol{\mu}\boldsymbol{\mu}'] \\ &= \operatorname{tr}(\boldsymbol{A}\Sigma) + \operatorname{tr}(\boldsymbol{A}\boldsymbol{\mu}\boldsymbol{\mu}') \\ &= \operatorname{tr}(\boldsymbol{A}\Sigma) + \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu}. \end{aligned}$$

第二章　一元线性回归

第一节　回归方程的建立

一、问题提出

一元线性回归是处理两个变量之间关系的最简单的模型. 本章将详细讨论这个模型. 一元线性回归虽较简单，但从中可以了解回归分析方法的基本思想、方法和应用.

我们首先通过一个例子来说明如何建立一元线性回归方程.

例 2.1　为了估计山上积雪融化后对下游灌溉的影响,在山上建立了一个观察站,测量了最大积雪深度(x)与当年灌溉面积(y),得到连续 10 年的数据如表 2.1.

为了研究这些数据中所蕴含的规律性，我们把各年最大积雪深度作横坐标,相应的灌溉面积作纵坐标,将这些数据点标在平面直角坐标图上,如图 2.1，这个图称为散点图.

从图 2.1 看到，数据点大致落在一条直线附近. 这告诉我们变量 x 与 y 之间的关系大致可看作是线性关系. 从图 2.1 还看到,这些点又不都在一条直线上，这表明 x 与 y 的关系并没有确切到给定 x 就可以唯一地确定 y 的程度. 事实上,还有许多其它因素对 y 产生影响,如当年的平均气温，当年的降雨量等等. 这些都是影响 y 取什么值的随机因素. 如果我们只研究 x 与 y 的关系，可以假定有如下结构式:

$$y = a + bx + \varepsilon \tag{2.1}$$

表 2.1　最大积雪深度与灌溉面积观测数据

年　序	最大积雪深度 x（尺）	灌溉面积 y（千亩）
1	15.2	28.6
2	10.4	19.3
3	21.2	40.5
4	18.6	35.6
5	26.4	48.9
6	23.4	45.0
7	13.5	29.2
8	16.7	34.1
9	24.0	46.7
10	19.1	37.4

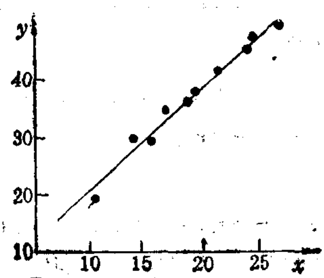

图　2.1

式中 α 和 β 是未知常数，称为回归系数，ε 表示其它随机因素对灌溉面积的影响．(2.1)式通常称为一元线性回归模型．在这个模型中一般假定 ε 是随机干扰或随机误差，它是一个随机变量（有关随机变量等基本概念请参看文献[10]），且满足

$$\begin{cases} E(\varepsilon) = 0 \\ \mathrm{Var}(\varepsilon) = \sigma^2 \end{cases} \tag{2.2}$$

这里 $E(\varepsilon)$ 表示 ε 的数学期望，$\mathrm{Var}(\varepsilon)$ 表示 ε 的方差，且 $\sigma^2 > 0$．在实际问题中经常假定 ε 遵从正态分布．

用 n 表示观察值的组数. $(x_1, y_1),(x_2, y_2),\cdots,(x_n, y_n)$ 表示 n 组观察值. 如果它们符合模型(2.1), 则

$$y_i = \alpha + \beta x_i + \varepsilon_i, \ i = 1,2,\cdots,n \tag{2.3}$$

由(2.2)应有

$$E(\varepsilon_i) = 0, \ \mathrm{Var}(\varepsilon_i) = \sigma^2, \ i = 1,2,\cdots,n$$

通常还假定 n 组数据是独立观察的, 因而 $\varepsilon_1, \varepsilon_2,\cdots,\varepsilon_n$ 是相互独立的随机变量.

对(2.1)两边求数学期望得

$$E(y) = \alpha + \beta x \tag{2.4}$$

该式表示当 x 已知时,可以精确地算出 $E(y)$. 由于 ε 是个不可控制的随机因素, 通常就用 $E(y)$作为 y 的估计, 故得

$$\hat{y} = \alpha + \beta x \tag{2.5}$$

式中 $\hat{y}$ 表示 y 的估计.

类似地对(2.3)式两边求数学期望, 得

$$E(y_i) = \alpha + \beta x_i, \ i = 1, 2,\cdots,n \tag{2.6}$$

或

$$\hat{y}_i = \alpha + \beta x_i, \ i = 1,2,\cdots,n \tag{2.7}$$

回归分析的首要任务是通过 n 组观察值来估计 α 与 β. 对 α 与 β 常用两种方法进行估计, 即最小二乘法和极大似然法. 我们只介绍前者. 为了记号的简单, 今后 α 与 β 的估计一般不用 $\hat{\alpha}$ 与 $\hat{\beta}$ 表示,而用 a 与 b 来表示. 我们称

$$\hat{y} = a + bx \tag{2.8}$$

为**回归方程**(或回归直线), a 与 b 称为**回归系数**. 在实际问题中用 $\hat{y} = a + bx$ 代替 $E(y) = \alpha + \beta x$ 作为 y 的估计.

二、最小二乘原理

如果 x 与 y 有精确的线性关系, 则应有

$$y_i = \hat{y}_i, \quad i = 1, 2, \cdots, n$$

但是，由于测量误差以及其它随机因素的干扰，一般 $y_i \neq \hat{y}_i$，因此，我们要确定一条直线，也就是要确定 α 与 β 的估计值 a 与 b，使回归直线(2.8)与所有数据点都比较"接近"．为了刻划这种"接近"的程度，我们引进**残差**的概念，所谓**残差**是指观察值 y_i 与回归值 $\hat{y}_i = a + bx_i$ 的偏差．用 e_i 表示残差（$i = 1,2,\cdots,n$）．很自然，可以用绝对残差和，即

$$\sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

来度量观察值与回归直线的接近程度．绝对残差和越小，回归直线就与所有数据点越接近．但是人们考虑到绝对残差和的数学处理有一定的麻烦，所以在古典回归中，一般用残差平方和

$$\begin{aligned} Q \equiv Q(a,b) &= \sum_{i=1}^{n} e_i^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - a - bx_i)^2 \end{aligned} \tag{2.9}$$

来刻划所有观察值与回归直线的偏离程度．

所谓**最小二乘法**，就是选择 a,b 使 $Q(a,b)$ 达最小，以这样的 a，b 作为 α，β 的估计值，所得的回归直线与所有观察值最接近．因而，用最小二乘法配出的直线 $\hat{y} = a + bx$ 就是在所有直线中残差平方和 Q 最小的一条．Q 是关于 a,b 的二次函数，所以它的最小值总是存在的．根据微积分中求极值的方法，a，b 应满足下列方程组

$$\left\{ \frac{\partial Q}{\partial a} = -2 \sum_{i=1}^{n} (y_i - a - bx_i) = 0 \right.$$

$$\left\{\frac{\partial Q}{\partial b} = -2\sum_{i=1}^{n}(y_i - a - bx_i)x_i = 0\right.$$

或等价于

$$\begin{cases}\sum_{i=1}^{n} y_i - na - b\sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2 = 0\end{cases} \tag{2.10}$$

记

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.11}$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{2.12}$$

$$\begin{aligned} L_{xx} &= \sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \end{aligned} \tag{2.13}$$

$$\begin{aligned} L_{xy} &= \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\ &= \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) \end{aligned} \tag{2.14}$$

则(2.10)的第一式可化为

$$a = \bar{y} - b\bar{x} \tag{2.15}$$

代入(2.10)的第二式，经整理得

$$L_{xx}b = L_{xy}$$

即

$$b = \frac{L_{xy}}{L_{xx}} \tag{2.16}$$

将(2.16)代入到(2.15)式就可求出 a 的值，于是回归方程为

$$\hat{y} = a + bx \tag{2.17}$$

由(2.15)和(2.17)式得

$$\hat{y} - \bar{y} = b(x - \bar{x}) \tag{2.18}$$

从(2.18)式可见，回归直线(2.17)是通过点$(\bar{x}, \bar{y})$的，这对回归直线的作图很有帮助.

现在利用这些公式来计算例 2.1 的回归方程.

根据表 2.1 的数据计算得

$$\bar{x} = \frac{1}{10}(15.2 + 10.4 + \cdots + 19.1) = 18.85$$

$$\bar{y} = \frac{1}{10}(28.6 + 19.3 + \cdots + 37.4) = 36.53$$

$$L_{xx} = (15.2^2 + 10.4^2 + \cdots + 19.1^2) - \frac{1}{10}(18.85)^2$$

$$= 227.845$$

$$L_{xy} = (15.2 \times 28.6 + 10.4 \times 19.3 + \cdots + 19.1 \times 37.4) - \frac{1}{10} \times 188.5 \times 365.3$$

$$= 413.065$$

由(2.16)和(2.15)式得

$$b = \frac{L_{xy}}{L_{xx}} = \frac{413.065}{227.845} = 1.813$$

$$a = \bar{y} - b\bar{x} = 36.53 - 1.813 \times 18.85 = 2.355$$

回归方程为

$$\hat{y} = 2.355 + 1.813x$$

如果在图 2.1 上画出这个回归方程的图像，可以看到它与所有数据点都很接近.

三、最小二乘估计的性质

得到一个估计并不难，但我们希望得到一个好的估计. 考

察一个估计的好坏有许多原则,如无偏性、相合性(或相容性)、允许性、稳健性等. 鉴于本书的性质,我们仅讨论最小二乘估计的部分基本性质.

1. 无偏性

若 $\hat{T}$ 是参数 T 的一个估计,且满足 $E(\hat{T})=T$,则称 $\hat{T}$ 为 T 的无偏估计.

我们关心的是 a 与 b 是否是 α 与 β 的无偏估计.

由(2.6)式知

$$E(y_i)=\alpha+\beta x_i,\ i=1,2,\cdots,n$$

由此易得

$$E(\bar{y})=\frac{1}{n}\sum_{i=1}^{n}E(y_i)=\alpha+\beta\bar{x}$$

所以

$$\begin{aligned}E(y_i-\bar{y})&=E(y_i)-E(\bar{y})\\&=\beta(x_i-\bar{x}),\ i=1,2,\cdots,n.\end{aligned}$$

从而

$$\begin{aligned}E(b)&=\frac{\sum_{i=1}^{n}(x_i-\bar{x})E(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\\&=\frac{\sum_{i=1}^{n}(x_i-\bar{x})\beta(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\\&=\beta\end{aligned}\tag{2.19}$$

$$\begin{aligned}E(a)&=E(\bar{y}-b\bar{x})\\&=E(\bar{y})-\bar{x}E(b)\\&=\alpha+\beta\bar{x}-\beta\bar{x}\\&=\alpha\end{aligned}\tag{2.20}$$

(2.19)和(2.20)式表示 a,b 分别是 α,β 的无偏估计，这说明，若用同样方法对 α，β 作多次估计，它们的平均值将趋于 α，β. 这是最小二乘估计的一个重要性质.

显然有

$$
\begin{aligned}
E(\hat{y}) &= E(a + bx) \\
&= \alpha + \beta x \\
&= E(y)
\end{aligned}
\tag{2.21}
$$

这表明 $\hat{y}$ 是 $E(y)$的无偏估计，即回归值 $\hat{y}$ 可看作是实际观察值的平均值.

2. 关于 a 和 b 的方差

在实际应用中，仅知道估计是无偏的是不够的，还需要知道估计量本身的波动状况，这就要研究它的方差.

注意到

$$
\begin{aligned}
b &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i - \bar{y}\sum_{i=1}^{n}(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \sum_{i=1}^{n}\frac{x_i - \bar{x}}{\sum_{j=1}^{n}(x_j - \bar{x})^2}\, y_i
\end{aligned}
$$

由$\{y_i\}$相互独立及$\mathrm{Var}(y_i)=\sigma^2$，得

$$\mathrm{Var}(b)=\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sum_{j=1}^{n}(x_j-\bar{x})^2}\right)^2\mathrm{Var}(y_i)$$

$$=\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \tag{2.22}$$

大家知道，方差的大小表示随机变量取值波动的大小．(2.22)式表明，回归系数b的波动大小不仅与误差的方差σ^2有关，而且还取决于观察数据中自变量x的波动程度．如果x值波动较大(即取值比较分散)，则b的波动就较小，也就是估计值比较稳定．反之，如果原始数据x是在一个较小范围内取得的，则β的估计值就稳定性差，当然也就很难说精确了．这对我们如何安排试验，收集数据有一定的指导意义．

类似地，可以求得α的估计量a的方差．因为

$$a=\bar{y}-b\bar{x}$$

$$=\sum_{i=1}^{n}\left[\frac{1}{n}-\frac{\bar{x}(x_i-\bar{x})}{\sum_{j=1}^{n}(x_j-\bar{x})^2}\right]y_i$$

所以有

$$\mathrm{Var}(a)=\sum_{i=1}^{n}\left[\frac{1}{n}-\frac{\bar{x}(x_i-\bar{x})}{\sum_{j=1}^{n}(x_j-\bar{x})^2}\right]^2\mathrm{Var}(y_i)$$

$$=\left[\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]\sigma^2 \tag{2.23}$$

由此可知，回归系数a的方差不仅与σ和x的波动大小有关，而且还同观察数据的个数n有关．数据越多，且x的观察值

越分散，估计量 a 就越稳定。

关于 $\hat{y}_i$ 的方差，本章将要进行讨论。

3. 在正态总体的条件下，a 和 b 是 α 和 β 的极大似然估计。

假定 $\varepsilon_i \sim N(0,\sigma^2)(i=1,2,\cdots,n)$，由于 $\varepsilon_1, \varepsilon_2,\cdots,\varepsilon_n$ 相互独立，所以它们的联合密度（称为似然函数）为

$$L(\alpha,\beta)=\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2\right]$$

所谓 α 和 β 的极大似然估计 $\hat{\alpha}$ 和 $\hat{\beta}$ 是指

$$L(\hat{\alpha},\hat{\beta})=\operatorname*{Max}_{\alpha,\beta} L(\alpha,\beta)$$

可以证明，$\hat{\alpha}$ 和 $\hat{\beta}$ 正好是 a 和 b。

4. 将数据进行线性变换后的回归系数

若令

$$x_i'=c+dx_i,\ y_i'=c^*+d^*y_i,\ i=1,2,\cdots,n.$$

式中 $d>0$，$d^*>0$。

用 a^*，b^* 表示由 x_i'，y_i' 算得的 α，β 的最小二乘估计，则 a^*，b^* 与未经变换的数据算得的 α，β 的最小二乘估计 a，b 有如下关系

$$\begin{cases} b^*=\dfrac{d^*}{d}b \\ a^*=c^*-\dfrac{d^*}{d}bc+d^*a \end{cases} \tag{2.24}$$

这表明，如果变化 x 和 y 的量纲或基准线（例如，温度由华氏变换为摄氏），一般 a 和 b 要变化。特别当 $d=d^*$ 时（即 x 和 y 的量纲变化比例相同），则有 $b^*=b$，即 β 的估计不变。

第二节 回归方程的显著性检验

在一些场合下，试验点不那么接近于一条直线，这时也可

用最小二乘法得到一条回归直线。但这条直线并没有很好地反映变量 x 和 y 的实际关系，没有应用价值．因此，一方面我们要建立从经验上认为有意义的方程，另一方面我们要用数学方法对方程的显著性进行检验．本节我们介绍两种检验方法。

一、相关系数的显著性检验

设 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$ 是 (x,y) 的 n 组观察值，称

$$r=\frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}} \tag{2.25}$$

为 x 与 y 的**相关系数**．式中

$$\begin{aligned} L_{yy}&=\sum_{i=1}^{n}(y_i-\bar{y})^2 \\ &=\sum_{i=1}^{n}y_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}y_i\right)^2 \end{aligned} \tag{2.26}$$

称为 y 观察值的离差平方和．L_{xx}，L_{xy} 见(2.13)、(2.14)式．

相关系数 r 表示 x 和 y 的线性关系的密切程度．可以证明 $|r|\leqslant 1$.

比较回归系数 b 与相关系数 r 可得

$$\begin{aligned} r&=\frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}} \\ &=\frac{L_{xy}}{L_{xx}}\sqrt{\frac{L_{xx}}{L_{yy}}} \\ &=b\sqrt{\frac{L_{xx}}{L_{yy}}} \end{aligned}$$

所以 r 与 b 有相同的符号．当 $r>0$ 时，称 x 与 y 正相关．