应用回归分析

盛承懋　李慧芬　钱君燕编译

上海科学技术文献出版社

应用回归分析

盛承燃　李慧芬　钱君燕　编译

上海科学技术文献出版社

应用回归分析

盛承懋　李慧芬　钱君燕　编译

*

上海科学技术文献出版社出版发行

（上海市武康路2号）

新华书店经销

昆山亭林印刷厂印刷

开本 787×1092 1/32　印张 7.625　字数 184,000

1989年3月第1版　1989年3月第1次印刷

印数：1—2,500

ISBN 7-80513-266-6/Z·69

定价：3.95元

《科技新书目》178-599

前　言

回归分析已经成为广大科研人员（自然科学与社会科学），经济管理工程技术和生态环境工作人员的最有用的统计分析工具之一。近十多年来人们在大量应用的基础上，发展并形成了近代回归分析的理论与方法。

本书的目的是深入浅出地介绍近代回归分析的应用思路与方法；如何利用残差图简捷又直观地分析数据；应用残差图揭示各种因素间的内在关系，检查与纠正回归模型的错误。

为了实现上述目的，本书介绍了大量的经济管理、工业生产、科学研究及生态环境的分析实例，结果均通过计算机给出。这些方法的使用将大大推动现代统计工具在生产科研中的应用。

本书所介绍的内容对研究数据分析的读者来说是十分直观有效的。本书可以作为大专院校统计与非统计专业学生的教科书或教学参考书。由于本书较突出应用的思路和方法，也便于实际工作者作为自学读物，使通过阅读本书后，能独立、有条不紊且有把握地进行数据分析。

本书对于各种技巧与方法的介绍并不作严格理论推导，为了满足某些想了解这一内容的读者，我们在各处都注明了可供查阅的原著。

本书的第一章到第四章由盛承懋编写；第五章到第七章由李慧芬编写；第八章和第九章由钱君燕编写；黄树颜总纂。在编写本书的过程中，得到中国纺织大学吴让泉教授的指导。上海

交通大学王亨时教授对稿件进行了认真细致地审阅，同济大学王福保教授也对本书十分关心。我们谨对他们表示衷心感谢。

编　者

1988年1月

目　录

第一章　一元线性回归

1.1 引　言

回归分析可以广义地定义为对一些变量之间的关系的分析。由于它为建立变量之间的函数关系提供了一种简单的方法，已经成为最有用的统计工具之一。这个关系可用因变量 y 与一个或多个自变量 $x_1, x_2, \cdots, x_p$ 之间的方程来表示。这个方程称为回归方程，一般取

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

的形式。这里 $b_0, b_1, b_2, \cdots, b_p$ 称为回归系数，是由数据确定的。只包含一个自变量的上述方程，称为一元线性回归方程，包含多于一个自变量的上述方程，称为多元线性回归方程。本书对这两类问题都作了介绍，其中一元回归的一个典型的实例是关于计算机的维修时间，与所维修元件的数目之间的关系的分析。这里有一个因变量（计算机的维修时间），与一个自变量（所维修元件的数目）。一个较复杂的多元回归的实例是用大量的环境与社会经济因素（自变量），去解释在不同的地理区域中由所有原因造成的人类死亡的总体年龄（因变量）。这两类问题在书中都作了介绍。

明确地求出回归方程，在某种意义上就是分析的最终结果，它是 y（因变量）与一组自变量 x 之间的关系的一种概括。方程可以用于各种不同的目的，它可以用来估价每个 x 的重要性，用来分析那些含有 x 的改变值的因素的作用，或对于给定的一组 x 预测 y 的取值。虽然回归方程是最终结果，然而还有许多重要的问

接结果。回归分析作为分析数据的方法，常被用来帮助弄清在一个特定的范围内变量之间的相互关系（假定在这个范围内所取的数据是有效的）。有时数据是在控制的条件下收集的，于是主要感兴趣的不是因子，甚至可以认为因子是不变的；而更多的情形，数据在非实验的条件下收集，很少能被研究者所控制。回归分析的任务就是要尽可能多地去获悉数据所表达的有关情景。我们的重点是要弄清，遵循建立方程的途径，揭示了哪些与回归方程一样重要且有益的问题。

我们从一元线性回归模型开始研究，在这一章中建立模型，叙述假定与标准的理论结果。书中不作理论上的推导，而是通过实例来熟悉这些标准结果，所表示出来的公式，仅作参考。并且始终假定所需的统计量，可以用现有的回归软件由计算机算出。熟悉回归分析基本概念的读者可以从"残差分析"这一节开始学习，然后做1.8节上的例题。对数学推导感兴趣的读者，可以参阅这一章最后的评注，那里所列举的书目，包括了回归问题的一系列重要的进展。

1.2 数据与模型的说明

数据由因变量（或称响应变量）y与一个自变量（或称解释变量）x的n次观测值组成，通常如表1.1所示。

表1.1 数据的组成

观测次数	y	x_1
1	y_1	x_{11}
2	y_2	x_{12}
3	y_3	x_{13}
⋮	⋮	⋮
n	y_n	x_{1n}

y 与 x_1 之间的关系确定为一个线性模型：

$$y_i=\beta_0+\beta_1x_{1i}+u_i,\quad i=1,2,\cdots,n \tag{1.1}$$

其中，β_0 与 β_1 是常数，称为模型的回归参数，而 u_i 是随机扰动。假定在所研究的观测值的范围中，线性方程 (1.1) 提供了 y 与 x_1 之间的真实关系的一个合理的近似，即 y 近似地是 x_1 的线性函数。u 度量了在这种近似下的偏差。假定对于 x_1 的每一个固定的值，u 的取值是随机的、相互独立的，且服从均值为 0、方差同为 σ^2 的分布。系数 β_1 也可以解释为 x_1 增加一个单位时，相应的 y 的增量。

1.3 估计与假设检验

用最小二乘法估计参数 β_0 与 β_1，要求使残差平方和 $S(\beta_0,\beta_1)$ 为最小，这里

$$S(\beta_0,\beta_1)=\sum_{i=1}^{n}u_i^2=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_{1i})^2$$

使 $S(\beta_0,\beta_1)$ 最小的 β_0 与 β_1 的值，即 b_0 与 b_1 由

$$b_1=\frac{\sum(y_i-\bar{y})(x_{1i}-\bar{x}_1)}{\sum(x_{1i}-\bar{x}_1)^2} \tag{1.2}$$

与

$$b_0=\bar{y}-b_1\bar{x} \tag{1.3}$$

给出，其中

$$\bar{y}=\frac{1}{n}\sum y_i$$

与

$$\bar{x}_1=\frac{1}{n}\sum x_{1i}$$

根据上述关于 u 的假定，所得的量 b_0 与 b_1 是 β_0 与 β_1 的无偏估计，它们的方差是

$$\mathrm{Var}(b_1)=\frac{\sigma^2}{\sum(x_{1i}-\bar{x}_1)^2} \tag{1.4}$$

$$\mathrm{Var}(b_0)=\sigma^2\left[\frac{1}{n}+\frac{\bar{x}_1^2}{\sum(x_{1i}-\bar{x}_1)^2}\right] \tag{1.5}$$

σ^2 的无偏估计 S^2 由

$$S^2=\frac{1}{n-2}\sum(y_i-b_0-b_1x_{1i})^2 \tag{1.6}$$

给出，在式(1.4)与(1.5)中用 S^2 代替 σ^2，得到 b_0 与 b_1 的方差的无偏估计。

对应于第 i 次观测值，由模型预测的响应值为

$$\hat{y}_i=b_0+b_1x_{1i} \tag{1.7}$$

对应于第 i 次观测值的残差是

$$e_i=y_i-\hat{y}_i \tag{1.8}$$

而第 i 个标准化残差由

$$e_{i,s}=\frac{e_i}{S} \tag{1.9}$$

确定，这里 S 由式(1.6)可得。

为了对回归模型中的参数构造置信区间与作假设检验，进一步假定 u 服从正态分布。

对于正态的假定，β_1 的最小二乘估计 b_1 服从均值为 β_1、方差由式(1.4)给出的正态分布。为了检验原假设 $H_0(\beta_1=\beta_1^0)$，这里 β_1^0 是研究者所选的一个常数，合适的检验统计量是

$$t=\frac{b_1-\beta_1^0}{\mathrm{s.e.}(b_1)} \tag{1.10}$$

其中，s.e.(b_1) 是 b_1 的标准误差，它由

$$\mathrm{s.e.}(b_1)=\frac{S}{[\sum(x_{1i}-\bar{x}_1)^2]^{1/2}} \tag{1.11}$$

给出。式(1.10)中的统计量 t 服从自由度为 $n-2$ 的 t 分布。通过比较观测值与选取适当的 t 临界值，即可完成检验。通常的检验是对 $\beta_1^0=0$ 所作的，这时 t 简化为 b_1 与它的标准误差之比。

β_1的置信系数为$1-\alpha$的置信限由

$$\left\{b_1 \pm t\left(n-2, \frac{\alpha}{2}\right)[\text{s.e.}(b_1)]\right\} \tag{1.12}$$

给出，这里$t\left(n-2, \frac{\alpha}{2}\right)$是自由度为$n-2$的$t$分布的$(1-\alpha)$百分位数。回归直线的截距$b_0$服从均值为$\beta_0$、方差由式(1.5)给出的正态分布。关于检验$H_0(\beta_0=\beta_0')$的统计量是

$$t=\frac{b_0-\beta_0'}{\text{s.e.}(b_0)} \tag{1.13}$$

这里β_0'是指定的值，而

$$\text{s.e.}(b_0)=S\left[\frac{1}{n}+\frac{\bar{x}_1^2}{\sum(x_{1i}-\bar{x}_1)^2}\right]^{1/2} \tag{1.14}$$

统计量t服从自由度为$n-2$的t分布。β_0的置信系数为$1-\alpha$的置信限为

$$\left\{b_0 \pm t\left(n-2, \frac{\alpha}{2}\right)[\text{s.e.}(b_0)]\right\} \tag{1.15}$$

1.4 拟 合 度

在得到β_0，β_1与σ^2的估计量之后，自然希望去测定方程(1.1)中的模型与观测数据的拟合度。为此目的而被广泛采用的指标是计算关于y与$\hat{y}$的样本相关系数，它定义为

$$R=\frac{\sum(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})}{[\sum(y_i-\bar{y})^2\sum(\hat{y}_i-\bar{\hat{y}})^2]^{1/2}} \tag{1.16}$$

其中，$\bar{\hat{y}}$是$\hat{y}$的平均，R的数值介于-1与1之间。这个拟合度指标可以看成是y与x_1之间的线性关系强度的一个度量。相关系数的平方R^2可以写成

$$R^2=1-\frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar{y})^2} \tag{1.17}$$

式(1.16)与(1.17)中所给的 R 的定义是等价的。式(1.17)所给的指标 R^2 可以解释为 y 中被 x_1 所解释的总变差的比例。如果 R^2 接近于1，则 x_1 解释了 y 中变差的大部分。为了检查 x_1 是否解释了 y 中的大部分变差，检验原假设 $H_0(\rho=0)$ 对于对立假设 $H_1(\rho\neq 0)$，这里 ρ 是总体相关系数。检验这个假设的合适的统计量为

$$t=\frac{|R|\sqrt{n-2}}{\sqrt{1-R^2}} \tag{1.18}$$

它服从自由度为 $n-2$ 的 t 分布。通过比较所得的 t 值与选取适当的 t 的临界值，即可完成这个检验。

显然，如果 y 与 x_1 之间不存在线性关系，则回归系数 $\beta_1=0$。因此，关于 $H_0(\beta_1=0)$ 与 $H_0(\rho=0)$ 的统计检验是一致的。虽然在式(1.10)与(1.18)中给出的检验统计量看上去不同，而实际可以证明它们是等价的。

1.5 预测值与标准误差

所拟合的回归方程，可以用来对任意选取的自变量值 x_1^0，预测因变量 y 的值，预测值 $\hat{y}_0$ 是

$$\hat{y}_0=b_0+b_1x_1^0 \tag{1.19}$$

且方差为

$$\mathrm{Var}(\hat{y}_0)=\sigma^2\left[1+\frac{1}{n}+\frac{(x_1^0-\bar{x}_1)^2}{\sum(x_{1i}-\bar{x})^2}\right] \tag{1.20}$$

$\hat{y}_0$ 的方差的估计量可以由在式(1.20)中以 S^2 代替 σ^2 得到。置信系数为 $1-\alpha$ 的预测值的置信限是

$$\left\{\hat{y}_0\pm t\left(n-2,\frac{\alpha}{2}\right)\mathrm{s.e.}(\hat{y}_0)\right\}$$

其中

$$\mathrm{s.e.}(\hat{y}_0)=S\left[1+\frac{1}{n}+\frac{(x_1^0-\bar{x}_1)^2}{\sum(x_{1i}-\bar{x}_1)^2}\right]^{1/2} \tag{1.21}$$

所预测的响应变量服从均值为 $\mu_0=\beta_0+\beta_1 x_1^0$ 的正态分布。如果我们对平均响应感兴趣，则 μ_0 的估计量为

$$\hat{\mu}_0=b_0+b_1x_1^0 \tag{1.22}$$

它的方差为

$$\mathrm{Var}(\hat{\mu}_0)=\sigma^2\left[\frac{1}{n}+\frac{(x_1^0-\bar{x}_1)^2}{\sum(x_{1i}-\bar{x}_1)^2}\right] \tag{1.23}$$

注意 μ_0 的点估计与预测响应 $\hat{y}_0$ 恒等，在解释中的差异，是与两个量各自的方差有关的。

1.6 拟合的测定

我们已经指出上下文中用来作推断的一元线性回归模型的基本结果。这些结果根据数据所计算的综合统计量可得。只有当与模型中的残差项有关的假定满足时，这些结果才成立。因此，研究残差的结构以及反映到图像上的数据的式样是非常重要的。一个大的 R^2 值或一个显著的 t 统计量，并不能保证数据拟合得很好。为了强调这一点，Anscombe(1973)[11] 构造了四组数据，每组的式样不同，但具有相同的综合统计量[1]。数据与图像表示于表 1.2 与图 1.1 中。仅按检查综合统计量进行分析，不能发现数据式样中的差异，因而得出一种错误的分析结果。

表 1.2 具有相同综合统计量值的四组数据[1]

	x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4
001	10	8.04	10	9.14	10	7.46	8	6.58
002	8	6.95	8	8.14	8	6.77	8	5.76
003	13	7.58	13	8.74	13	12.74	8	7.71
004	9	8.81	9	8.77	9	7.11	8	8.84
005	11	8.33	11	9.26	11	7.81	8	8.47
006	14	9.96	14	8.10	14	8.84	8	7.04
007	6	7.24	6	6.13	6	6.08	8	5.25
008	4	4.26	4	3.10	4	5.39	19	12.50
009	12	10.84	12	9.13	12	8.15	8	5.56
010	7	4.82	7	7.26	7	6.42	8	7.91
011	5	5.68	5	4.74	5	5.73	8	6.89