YINGYONG HUIGUI FENXI

应用回归分析

王黎明 张日权 景英川 编著

中国海洋大学出版社

应用回归分析

王黎明　张日权　景英川　编著

中国海洋大学出版社
·青岛·

图书在版编目(CIP)数据

应用回归分析/王黎明,张日权,景英川编著.—青岛:中国海洋大学出版社,2005.6

ISBN 7-81067-697-0

Ⅰ.应…　Ⅱ.①王…　②张…　③景…　Ⅲ.回归分析

Ⅳ.O212.1

中国版本图书馆CIP数据核字(2005)第051428号

中国海洋大学出版社出版发行

(青岛市鱼山路5号　邮政编码:266003)

出版人:王曙光

编辑室电话:0532—88661615,82032122

网址:www.ouc.edu.cn

淄博恒业印务有限公司印刷

新华书店经销

*

开本:850mm×1 168mm　1/32　印张:7.375　字数:390千字

2005年7月第1版　2005年7月第1次印刷

印数:1～1 100　定价:29.00元

内容简介

本书以经典的最小二乘理论为基础，叙述了回归分析. 全书共分为 9 章. 第一章讨论了回归模型的主要任务及其建模过程；第二、三章详细地介绍了线性回归模型；第四章以残差为重要工具，讨论了回归模型的诊断问题；第五、六章讨论了多项式回归模型和含有定性变量的回归模型；第七章讨论了多元线性回归模型的有偏估计；第八章简单介绍了非线性回归模型；本书的最后一章介绍 SPSS 统计软件在回归分析中的应用.

本书可以作为统计学、数学以及经济学等专业的教材. 学习本课程的学生需要熟悉概率论与数理统计的基础知识，也要具备微积分和线性代数的相关知识.

前　言

回归分析由一组探求变量之间关系的技术组成，是数理统计应用最广泛的分支之一．在理论上，本书叙述了经典的最小二乘理论，同时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．中心主题是建立线性回归模型，评价拟合效果，并且作出结论．

全书分为 9 章．第一章介绍了一般回归模型的定义，讨论了回归模型的主要任务和回归模型的建模过程．第二章详细地介绍了一元线性回归模型，给出了未知参数的最小二乘估计以及极大似然估计，还讨论了一元线性回归模型的预测问题以及数据变换问题．第三章系统讨论了多元线性回归模型，最小二乘估计的优良性，多元回归模型的显著性检验以及其回归系数的显著性检验．第四章以残差为重要工具，讨论了回归模型的诊断问题．第五章和第六章讨论了多项式回归模型和含有定性变量的回归模型．第七章讨论了多元线性回归模型的有偏估计，重点介绍较常用的岭估计和主成分估计，同时介绍其他的估计方法．第八章简单介绍了非线性回归模型，主要讨论了 Logistic 回归模型和广义线性模型．本书的最后一章介绍 SPSS 统计软件在回归分析中的应用．

本书可以作为统计学、数学以及经济学等专业的教材，学习本课程的学生需要熟悉随机变量、参数估计、区间估计、假设检验等思想，也要熟悉正态分布及由其导出的分布，当然，学生也要具备微积分和线性代数的相关知识．

由于编者的水平有限，在取材及结构上，本书难免会存在不够

妥当的地方,错误之处也在所难免,恳请同行专家和广大读者能给我们宝贵的批评和建议.

编者

2004 年 12 月

目　录

第一章　引言 …………………………………………………… (1)

§1.1　变量间的统计关系 ………………………………… (1)

§1.2　回归模型的一般形式 ……………………………… (2)

§1.3　“回归”一词的由来 ………………………………… (4)

§1.4　建立实际回归模型的过程 ………………………… (4)

小结 ………………………………………………………… (7)

第二章　一元线性回归分析 ………………………………… (8)

§2.1　一元线性回归模型 ………………………………… (8)

§2.2　一元线性回归模型的假设 ………………………… (9)

§2.3　参数的最小二乘估计 ……………………………… (10)

§2.4　参数的极大似然估计 ……………………………… (12)

§2.5　最小二乘法估计的性质 …………………………… (14)

§2.6　一元线性回归模型的显著性检验 ………………… (16)

§2.7　一元线性回归模型的回归预测与区间估计 ……… (21)

§2.8　数据交换后的线性拟合 …………………………… (24)

小结 ………………………………………………………… (27)

习题 ………………………………………………………… (28)

第三章　多元线性回归分析 ………………………………… (30)

§3.1　多元线性回归模型 ………………………………… (30)

§3.2　多元线性回归模型的参数估计 …………………… (34)

§3.3　带约束条件的多元线性回归模型的参数估计 …… (40)

§3.4　多元线性回归模型的广义最小二乘估计 ………… (44)

§3.5 多元线性回归模型的假设检验 …………………… (46)
§3.6 多元线性回归模型的预测及区间估计 ………… (56)
§3.7 逐步回归与多元线性回归模型选择 …………… (59)
§3.8 多元数据变换后的线性拟合 ……………………… (70)
小结 ……………………………………………………… (77)
习题 ……………………………………………………… (79)
第四章 回归诊断 ………………………………………… (83)
§4.1 残差及其性质 ………………………………………… (83)
§4.2 回归函数线性的诊断 ………………………………… (85)
§4.3 误差方差齐性的诊断 ………………………………… (89)
§4.4 误差的独立性诊断 …………………………………… (92)
§4.5 异常点与强影响点 …………………………………… (98)
小结 ……………………………………………………… (100)
习题 ……………………………………………………… (101)
第五章 多项式回归 ……………………………………… (103)
§5.1 多项式回归 …………………………………………… (103)
§5.2 正交多项式回归 ……………………………………… (106)
§5.3 多项式对曲线的分段拟合 …………………………… (116)
小结 ……………………………………………………… (123)
习题 ……………………………………………………… (124)
第六章 含定性变量的数量化方法 ……………………… (125)
§6.1 自变量中含有定性变量的回归模型 ……………… (125)
§6.2 协方差分析 …………………………………………… (130)
小结 ……………………………………………………… (136)
习题 ……………………………………………………… (137)
第七章 多元线性回归模型的有偏估计 ………………… (139)
§7.1 引言 …………………………………………………… (139)
§7.2 岭估计 ………………………………………………… (148)

§7.3 主成分估计 …………………………………… (158)
§7.4 广义岭估计 …………………………………… (164)
§7.5 Stein 估计 …………………………………… (166)
小结 …………………………………… (168)
习题 …………………………………… (169)
第八章 非线性回归模型 …………………………………… (171)
§8.1 Logistic 回归模型 …………………………………… (171)
§8.2 广义线性模型 …………………………………… (175)
第九章 SPSS 统计软件在回归分析中的应用 …………………………………… (177)
§9.1 线性回归方程的建立 …………………………………… (178)
§9.2 SPSS 在线性回归模型中的应用例子 …………………………………… (186)
附表 1 F 检验的临界值 …………………………………… (192)
附表 2 t 分布的分位数 …………………………………… (204)
附表 3 检验相关系数 $\rho=0$ 的临界值 …………………………………… (208)
附表 4 F_{max} 的分位点 …………………………………… (209)
附表 5 G_{max} 的分位点 …………………………………… (211)
附表 6 $D-W$ 检验临界值 …………………………………… (213)
附表 7 正交多项式 …………………………………… (217)
参考文献 …………………………………… (224)

第一章　引言

§1.1　变量间的统计关系

我们先看一个例子.

例 1.1　一个保险公司承保汽车 5 万辆,每辆保费收入是 1 000元,则该公司汽车承保总额为5 000万元.即承保总收入为 y,承保汽车数为 x,则变量 y 和 x 的关系可以表示为 $y=1\,000x$.

从这个例子可以看出,每给定一个 x,就一定可以得到一个 y,即变量 y 与 x 之间完全表现为一种确定性的关系——函数关系.

一般而言,给定 p 个变量 $x_1,\cdots,x_p$ 就可以确定变量 y,称这种变量之间的关系为确定性关系.它往往可以用某一函数关系 $y=f(x_1,\cdots,x_p)$来表示.

下面再看一个例子.

例 1.2　日常生活中,我们知道某种高档品的消费量(y)与城镇居民的收入(x)有密切关系.居民收入高了,这种消费品的销售量就大;居民收入低了,这种消费品的销售量就小;但是居民的收入并不能完全确定该高档品的消费量.因为商品的消费量还受着人们的消费习惯、心理因素、其他可替代商品的吸引程度以及价格的高低等因素的影响.也就是说,城镇居民的收入与该高档品的消费量有着密切关系,且城镇居民的收入对该种高档品的消费量起着主要作用,但是,它并不能完全确定该高档品的消费量.

在日常生活中,变量与变量之间表现为上述关系的有很多.比如,粮食产量与施肥量之间的关系,银行储蓄额与居民收入之间的

关系.

综上所述,变量 x 与变量 y 有密切关系,但是又没有密切到可以通过一个变量去确定另一个变量的程度.它们之间的这种非确定性的关系,我们称之为统计关系或相关关系.

回归分析就是讨论变量与变量之间的统计关系的一种统计方法.

§1.2 回归模型的一般形式

假设因变量 y 与一个或多个自变量 $x_1,x_2,\cdots,x_p$ 之间具有统计关系,我们把 y 称为因变量、响应变量或被解释变量,$x_1,x_2,\cdots,x_p$ 称为自变量、预报变量或解释变量.我们可以设想 y 由两部分组成:一部分是由 $x_1,x_2,\cdots,x_p$ 能够决定的部分,记为 $f(x_1,x_2,\cdots,x_p)$,另一部分是由众多未加考虑的因素(包括随机因素)所产生的影响,它被看成随机误差,记为 ε.于是得到了如下统计模型:

$$y=f(x_1,x_2,\cdots,x_p)+\varepsilon, \tag{1.1}$$

式中,ε 称为随机误差,一般要求它的数学期望为 0,它的出现使得变量间关系的相关性得以恰当体现;$f(x_1,x_2,\cdots,x_p)$ 称为 y 对 $x_1,x_2,\cdots,x_p$ 的回归函数,或称为 y 对 $x_1,x_2,\cdots,x_p$ 的均值回归函数;模型(1.1)称为回归模型的一般形式.

模型(1.1)清楚地表达了变量 $x_1,x_2,\cdots,x_p$ 与变量 y 的相关关系.数理统计学中的"回归"通常是指散点分布在一直线(或曲线)附近,并且越靠近该直线(或曲线)则点的分布越密集的情况.它也称为直线(或曲线)的拟合.

当模型(1.1)中的回归函数为线性时,式(1.1)变为

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\varepsilon, \tag{1.2}$$

式中,$\beta_0,\beta_1,\cdots,\beta_p$ 为未知参数,常称 β_0 为回归常数,$\beta_1,\cdots,\beta_p$ 为回归系数.这时我们称式(1.2)为线性回归模型.

在实际应用中，$\beta_0,\beta_1,\cdots,\beta_p$ 一般皆是未知的. 为了应用，需要将它们估计出来. 估计就需要数据，假设样本观测值为 $x_{i1},x_{i2},\cdots,x_{ip};y_i,i=1,2,\cdots,n$，则线性回归模型可表示为

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_px_{ip}+\varepsilon_i,i=1,2,\cdots,n. \quad (1.3)$$

假设由这些数据给出了 $\beta_0,\beta_1,\cdots,\beta_p$ 的估计值，分别记为 $\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_p$. 称

$$y=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\cdots+\hat{\beta}_px_p \quad (1.4)$$

为经验回归方程.

如果给定一组 $x_1,x_2,\cdots,x_p$，由式(1.4)可以得到一个 y，记为 $\hat{y}$. $\hat{y}$ 称为 y 的一个预测值.

对模型(1.4)，通常规定满足它的基本假设有：

(1)变量 $x_1,x_2,\cdots,x_p$ 是非随机变量，观测值 $x_{i1},x_{i2},\cdots,x_{ip}$ 是常数.

(2)高斯-马尔可夫(Gauss-Markor)条件：G-M 条件(等方差及不相关的假定)

$$\begin{cases}E(\varepsilon_i)=0,i=1,2,3,\cdots,n,\\ Cov(\varepsilon_i,\varepsilon_j)=\begin{cases}0,i\neq j,\\ \sigma^2,i=j.\end{cases}\end{cases}$$

(3)正态分布的假定条件为

$$\begin{cases}\varepsilon_i\sim N(0,\sigma^2),\\ \varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n \text{ 相互独立.}\end{cases}$$

对线性回归模型，通常要研究的问题有：

(1)如何根据样本 $x_{i1},x_{i2},\cdots,x_{ip};y,i=1,2,\cdots,n$ 求出 $\beta_1,\beta_2,\cdots,\beta_p$ 及方差 σ^2 的估计.

(2)对回归方程及回归系数的种种假设进行检验.

(3)如何根据回归方程进行预测和控制，以便进行实际问题的结构分析.

§1.3 "回归"一词的由来

"回归"一词的英文是"regression". 它是由英国著名生物学家兼统计学家 Galton 在研究人类遗传问题时提出来的. 为了研究父代与子代身高间的关系,他收集了1 078对父子的身高. 用 x 代表父亲的身高,y 代表儿子的身高,将这1 078对数据(x_i,y_i)描绘在一坐标系中,发现它们大致在一条直线附近,即父亲的身高 x 增加时,儿子的身高 y 也倾向于增加;父亲比较矮时,儿子也倾向于比较矮. 这与我们的常识是一致的. Galton 得到的回归方程是

$$\hat{y}=33.73+0.561x.$$

这1 078个 x_i 的算术平均数 $\bar{x}=68$英寸,y_i 的算术平均数 $\bar{y}=69$英寸. 这说明子代身高平均增加了 1 英寸. 人们自然会这样想:若父亲的身高为 x 英寸,其儿子的身高应为 $x+1$ 英寸. 但是所得的结论与此大相径庭. Galton 发现:$x=72$ 英寸(大于平均身高 68 英寸)时,儿子的平均身高为 71 英寸,不但达不到$72+1=73$英寸,反而比父亲低了 1 英寸;反过来,$x=64$ 英寸(小于平均身高 68 英寸)时,儿子的平均身高为 67 英寸,竟比预期的 $64+1=65$ 英寸高出 2 英寸.

这种现象不是个别的,而是呈现一个一般规律:身高超过平均值的父亲,他们的儿子的平均身高将低于父亲的平均身高;反之,身高低于平均值的父亲,他们的儿子的平均身高将高于父亲的平均身高. Galton 对这个一般结论的解释是:大自然具有一种约束力,使人类的身高分布在一定时期内相对稳定而不产生两极分化,这就是所谓的回归效应. 从此引进了回归一词. 对于后面将要讲到的回归模型,回归效应不一定具有.

§1.4 建立实际回归模型的过程

我们先用逻辑框架图表示回归模型的建模过程,如图1.1所示.

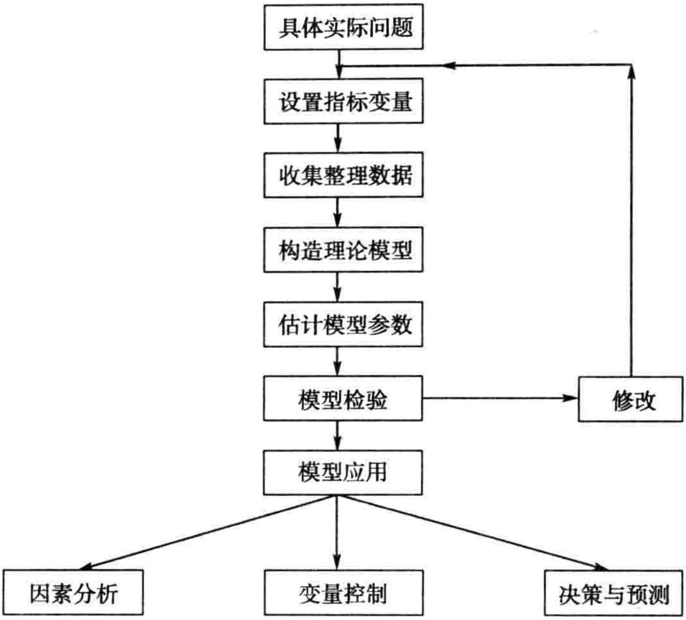

图 1.1　回归模型的建模过程

第一步,根据研究的目的设置指标变量.

回归分析模型主要是揭示事物之间相关变量的数量关系.首先要根据所研究的目的设置因变量 y,然后再选取与因变量有统计关系的一些变量作为自变量.

通常情况下,我们希望因变量与自变量之间具有因果关系.一般先定“果”,再寻找“因”.例如,要研究我国的通货膨胀问题,在金融理论的指导下,通常把全国零售物价总指数作为衡量通货膨胀的重要指标,那么,全国零售物价总指数作为因变量,影响全国零售物价总指数的有关因素就作为自变量.它包括国民收入、居民存

款、工农业总产值、货币流通量、职工平均工资、社会商品零售总额等 18 个变量.

第二步,收集整理统计数据.

常用的数据可分为时间序列数据和横截面数据. 时间序列数据就是按时间先后顺序排列的统计数据. 比如,历年来的国民收入、居民存款、工农业总产值等. 横截面数据就是在同一时间截面上收集的统计数据,如同一年在不同地块上测得的施肥量与小麦产量、同一年全国各大中城市的物价指数.

在实际收集数据时应该收集多少数据? 一般而言,收集的数据越多越好. 但是,在实际操作过程中,由于人力、物力等因素的限制,收集一个比较合理的数据量就可以了. 我们面临的另一个问题是如何收集数据. 关于这两个问题的讨论可以参考有关抽样调查方面的书籍.

收集到数据以后,有时这些数据并不是直接可以使用的,需要对它们进行一些处理,比如拆算、差分、取对数、标准化、补缺、处理异常数据等.

第三步,构造理论模型.

首先,研究所讨论问题的机理,根据其机理确定理论模型. 比如,要研究资本 k 以及劳动 l 对产出 y 的影响. 由数理经济学的理论可知它们存在如下的关系:

$$y=ak^{\alpha}l^{\beta},$$

式中,α,β 分别为资本和劳动对产出的弹性. 但是计量经济学的观点认为,变量之间的关系并不像上面表达的那样精确,而是存在随机偏差. 若记随机偏差为 u,则上式变为

$$y=ak^{\alpha}l^{\beta}u.$$

对上式两边取对数就变成如下的线性回归模型:

$$\ln y=\ln a+\alpha\ln k+\beta\ln l+\ln u.$$

其次是应用散点图,将数据点描绘在同一个坐标系里,分析它

们之间的关系.

第四步,对模型参数的估计.

一般情况下,建立的回归模型都是有未知参数的.为了能够使用这一模型,必须估计出未知参数.后面的章节将介绍参数的最小二乘估计、极大似然估计、岭估计等估计方法.

第五步,模型的检验和修改.

当一个模型建好以后,我们要问一个问题:这个模型是否比较好地描述了问题中变量之间的关系?那么,我们就要检验这个模型.检验的方法一般有两种:一种是放在实践中去检验,一个好的模型必须能够很好地反映客观实际,如果该模型可以反映客观实际,那它就是一个好的模型,反之,它就不是一个好的模型,而是不可用的;另一种是统计检验,统计检验包含模型检验和回归系数的检验,这将在后面的章节里讲解.

如果经过检验,发现所建立的模型是一个比较差的模型,那么,就要对该模型进行修改,要回到第一步重新考虑问题,看哪一步出现了问题,以便对该模型进行修改.

第六步,回归模型的应用.

当一个好的模型建立起来以后,就可以用它来进行分析、控制和预测.由模型我们可以分析出变量之间的关系,特别是可以看出影响因变量的主要因素,如果它们是可以控制的,我们就可以对它们实行控制,从而达到我们的目的.一个好的模型还可以给出好的预测,一个好的预测可以为我们的决策提供有力依据.

小结

本章主要介绍了一般回归模型的定义及其特殊情况——线性回归模型,讨论了回归模型的主要任务和回归模型的建立过程.

第二章　一元线性回归分析

§2.1　一元线性回归模型

在研究实际问题时，经常需要研究某一现象与影响它的某一最主要因素的关系．比如，影响粮食产量的因素很多，但是在众多因素中，施肥量是一个最重要的因素．因此人们往往研究施肥量这一因素与粮食产量之间的关系．又如，保险公司在研究火灾损失的规律时，把火灾发生地与最近的消防站的距离作为最重要的因素，研究火灾损失与火灾发生地与最近的消防站的距离之间的关系．

以上 2 个例子都是研究 2 个变量之间的关系，且 2 个变量有着密切的关系，但是，它们之间的密切程度并不能由一个变量惟一确定另一个变量．下面再看一个具体的例子．

例 2.1　从常识上理解，一个家庭的消费支出主要受这个家庭收入的影响．一般而言，家庭收入高的，其家庭消费支出也高；家庭收入低的，其家庭消费支出也低．为了研究它们的关系，取家庭消费支出 y（元）为被解释变量，家庭收入 x（元）为解释变量．为此，调查得到如下数据（表 2.1）．

表 2.1　家庭收入与消费支出

家庭编号	1	2	3	4	5	6	7	8	9	10
家庭收入（元）	800	1 200	2 000	3 000	4 000	5 000	7 000	9 000	10 000	12 000
消费支出（元）	770	1 100	1 300	2 200	2 100	2 700	3 800	3 900	5 500	6 600