21世纪统计学系列教材

应用回归分析

何晓群
刘文卿 编著

民大学出版社

21世纪统计学系列教材

应用回归分析

何晓群　刘文卿　编著

中国人民大学出版社

图书在版编目（CIP）数据

应用回归分析/何晓群，刘文卿编著
北京：中国人民大学出版社，2001
21 世纪统计学系列教材

ISBN 7-300-03757-7/F·1128
Ⅰ. 应…
Ⅱ. ①何…②刘…
Ⅲ. 回归分析-教材
Ⅳ. O212.1

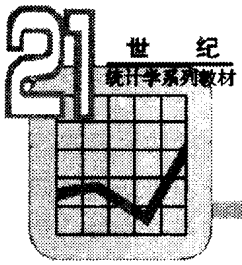
中国版本图书馆 CIP 数据核字（2001）第 15244 号

21 世纪统计学系列教材
应用回归分析
何晓群　刘文卿 编著

出版发行：中国人民大学出版社
（北京中关村大街 31 号　邮编 100080）
邮购部：62515351　门市部：62514148
总编室：62511242　出版部：62511239
E-mail：rendafx@public3.bta.net.cn
经　　销：新华书店
印　　刷：三河市新世纪印刷厂

开本：787×980 毫米 1/16　印张：16
2001 年 6 月第 1 版　2001 年 6 月第 1 次印刷
字数：291 000

定价：19.00 元
（图书出现印装问题，本社负责调换）



前 言

回归分析是统计学中的一个非常重要的分支，它在自然科学、管理科学和社会、经济等领域应用十分广泛。从20世纪90年代起，随着统计学在中国被确立为一级学科，统计专业的课程设置已有较大变化，加强推断统计内容的学习和应用已成为中国统计界的共识。本书正是为了适应新的统计学学科体系和财经类统计专业教学的需要而编写的。

本书写作的指导思想是：在不失严谨的前提下，明显不同于纯数理类教材，努力突出实际案例的应用和统计思想的渗透，结合统计软件较全面地系统介绍回归分析的实用方法。为了贯彻这一指导思想，本书在系统介绍回归分析基本理论和方法的同时，尽力结合中国社会、经济、自然科学等领域的研究实例，把回归分析方法与实际应用结合起来，注意定性分析与定量分析的紧密结合，努力把同行们以及我们在实践中应用回归分析的经验和体会融入其中。几乎每种方法都强调它的优缺点和实际运用中应注意的问题，并对每章的内容给予综述性评注。为使读者掌握本书内容，又考虑到这门课程的应用性和实践性，我们在每章后面给出一些简单的思考与练习，鼓励读者自己利用一些经济数据去实现这些方法。回归分析的应用离不开计算机，本书的案例主要运用在我国已很流行的SPSS软件实现，部分内容用Excel和SAS软件完成。本书一个显著的特点是在每种方法之

后结合实例概要介绍了 SPSS 或 Excel、SAS 软件的实际操作过程。本书的一些重要应用案例和结论都注明了参考文献，有兴趣的读者可作进一步的阅读和探讨。

本书共分九章。第一章为了给读者一个整体印象，对回归分析的研究内容和建模过程给出综述性介绍；第二章和第三章详细介绍了一元和多元线性回归的参数估计、显著性检验及其应用；第四章对违背回归模型基本假设的异方差、自相关和异常值等问题给出了诊断和处理方法；第五章介绍了回归变量选择与逐步回归方法；第六章对多重共线性从共线性问题的产生背景、诊断方法 、处理方法等方面结合实际经济问题给予讨论；第七章介绍解决共线性问题的一种岭回归估计方法；第八章介绍了可化为线性回归的曲线回归、多项式回归，以及不能线性化的本质非线性回归模型的计算；第九章结合案例分别介绍了自变量中含有定性变量和因变量是定性变量的回归问题，以及 Logistic 回归模型。

本书可作为统计学专业本科生的回归分析课程教材。对非统计专业的学生，书中打 * 号的章节可以不讲。由于本书的内容较多，教师在选用此书作教材时可以灵活选讲。本书还可作为非统计专业研究生量化分析教材。根据我们多年的教学实践，本书讲授 54 课时较为合适，能有计算机和投影设备的配合，教学将会更为方便和有效。

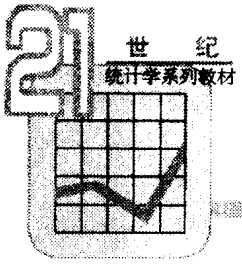
在本书的写作过程中，始终得到 21 世纪统计学系列教材编委会和中国人民大学出版社的支持。编写大纲经过教材编写委员会的认真讨论，教材初稿得到张尧庭、吴喜之两位教授的认真审阅，提出不少中肯意见。在此基础上我们对教材作了认真修改，若现在书中仍有不妥之处，责任当由笔者自负。本书的大部分案例是我们多年教学和科研工作的积累，有部分案例为体现其典型性引用他人著作。在此，我们谨向对本书出版给予帮助的师长和朋友表示衷心的感谢。

本书的完成是我们多年友好合作的结果。我们期望它的出版能为回归分析在中国的应用起到抛砖引玉的作用。由于我们水平所限，书中难免有不足之处，尤其是在一些应用研究的体会性讨论中，恐有偏颇之处，恳切希望读者批评指正。

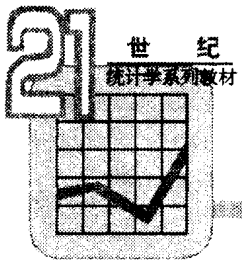
何晓群　刘文卿

2001 年 5 月

于中国人民大学科研楼



目 录



第 1 章

回归分析概述

为了在系统学习回归分析之前对该课程的思想方法、主要内容、发展现状等有个概括的了解，本章将由变量间的统计关系，引申出社会经济与自然科学等现象中的相关与回归问题，并扼要介绍“回归”名称的由来及近代回归分析的发展、回归分析研究的主要内容，以及建立回归模型的步骤与建模过程中应注意的问题。

1.1 变量间的统计关系

社会经济与自然科学等现象之间的相互联系和制约是一个普遍规律。例如社会经济的发展总是与一定的经济变量的数量变化紧密联系的。社会经济现象不仅同和它有关的现象构成一个普遍联系的整体，而且在它的内部也存在着许多彼此关联的因素，在一定的社会环境、地理条件、政府决策影响下，一些因素推动或制约另外一些与之联系的因素发生变化。这种状况表明，在经济现象的内部和外部联系中存在着一定的相关性，人们往往利用这种相关关系来制定有关的经济政策，以指导、控制社会经济活动的发展。要认识和掌握客观经济规律就必须探求经济现象间经济变量的变化规律，变量间的统计关系是经济变量变化规律的重要特征。

互有联系的经济现象及经济变量间关系的紧密程度各不一样。一种极端的情况是一个变量的变化能完全决定另一个变量的变化。例如，一个保险公司承保汽车5万辆，每辆保费收入为1 000元，则该保险公司汽车承保总收入为5 000万元。如果把承保总收入记为 y，承保汽车辆数记为 x，则 $y=1\,000x$。x 与 y 两个变量间完全表现为一种确定性关系，即函数关系。如图 1.1 所示。

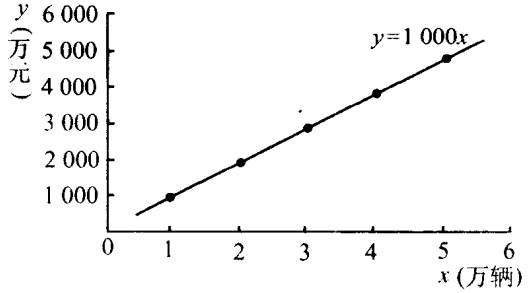


图 1.1　函数关系图

再如，银行的一年期存款利率为年息 2.55%，存入的本金用 x 表示，到期的本息用 y 表示，则 $y=x+2.55\%x$。这里 y 与 x 仍表现为一种线性函数关系。对于任意两个变量间的函数关系，我们可以表述为下面的数学形式

$$y=f(x)$$

再如，工业企业的原材料消耗总额用 y 表示，生产量用 x_1 表示，单位产量消耗用 x_2 表示，原材料价格用 x_3 表示，则

$$y=x_1x_2x_3$$

这里的 y 与 x_1，x_2，x_3 仍是一种确定性的函数关系，但它们显然不是线性函数关系了。我们可以将变量 y 与 p 个变量 x_1，x_2，…，x_p 之间存在着的某种函数关系用下面的形式表示

$$y=f(x_1,x_2,\cdots,x_p)$$

经济问题中还有很多函数关系的例子。物理学中的自由落体距离公式、初等数学中许多计算公式等都是变量间的函数关系。

然而，现实世界中还有不少情况是两事物之间有着密切的联系，但它们密切的程度并没有到由一个可以完全确定另一个的程度。下面举几个例子。

1. 我们都知道某种高档消费品的销售量与城镇居民的收入密切相关，居民收入高了这种消费品的销售量就大。但是由居民收入 x 并不能完全确定某种高档消费品的销售量 y，因为这种高档消费品的销售量还受着人们的消费习惯、心理因素、其他商品的吸引程度及价格的高低等诸多因素的影响。这样变量 y 与

变量x就是一种非确定的关系，见图1.2。

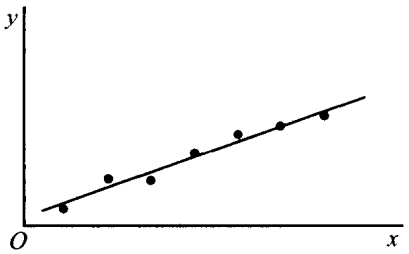


图1.2　y与x非确定性关系图

2. 粮食产量y与施肥量x之间有着密切的关系，在一定的范围内，施肥量越多，粮食产量就越高。但是，施肥量并不能完全确定粮食产量，因为粮食产量还与其他因素的影响有关，如降雨量、田间管理水平等。因此粮食产量y与施肥量x之间不存在确定的函数关系。

3. 储蓄额与居民的收入密切相关，但是由居民收入并不能完全确定储蓄额。因为影响储蓄额的因素很多，如通货膨胀、股票价格指数、利率、消费观念、投资意识等。因此尽管储蓄额与居民收入有密切的关系，但它们之间并不存在一种确定性关系。

再如：广告费支出与商品销售额、保险利润与保费收入、工业产值与用电量等。这方面的例子不胜枚举。

以上变量间关系的一个共同特征是它们之间有密切关系，但它们是一种非确定性关系。由于经济问题的复杂性，有许多因素因为我们的认识以及其他客观原因的局限，并没有包含在内。或者由于试验误差、测量误差以及其他种种偶然因素的影响，使得另外一个或一些变量的取值带有一定的随机性。因而当一个或一些变量取定值后，不能以确定值与之对应。

从图1.1看到确定性的函数关系，各对应点完全落在一条直线上。而由图1.2看到，各对应点并不完全落在一条直线上，即有的点在直线上，有的点在直线的两边。对于这种对应点不能分布在一条直线上的变量间的关系，也就是变量x与y之间有一定的关系，但是又没有密切到可以通过x惟一确定y的程度，这种关系正是统计学中研究的重要内容。在推断统计中，我们把上述变量间具有密切关联而又不能由某一个或某一些变量惟一确定另外一个变量的关系，称为变量间的统计关系或相关关系。这种统计关系规律性的研究是统计学中研究的主要对象，现代统计学中关于统计关系的研究已形成两个重要的分支，它们叫相关分析和回归分析。

回归分析和相关分析都是研究变量间关系的统计学课题。在应用中，两种分析方法经常相互结合和渗透，但它们研究的侧重点和应用面不同。它们的差别主要有以下几点：一是在回归分析中，变量 y 称为因变量，处在被解释的特殊地位。在相关分析中，变量 y 与变量 x 处于平等的地位，即研究变量 y 与变量 x 的密切程度与研究变量 x 与变量 y 的密切程度是一回事。二是相关分析中所涉及的变量 y 与 x 全是随机变量。而回归分析中，因变量 y 是随机变量，自变量 x 可以是随机变量，也可以是非随机的确定变量。通常的回归模型中，我们总是假定 x 是非随机的确定变量。三是相关分析的研究主要是为刻画两类变量间线性相关的密切程度。而回归分析不仅可以揭示变量 x 对变量 y 的影响大小，还可以由回归方程进行预测和控制。

由于回归分析与相关分析的研究侧重不同，使得它们的研究方法也大不相同。回归分析已成为现代统计学中应用最广泛、研究最活跃的一个独立分支。

1.2 回归方程与回归名称的由来

回归分析是处理变量 x 与 y 之间的关系的一种统计方法和技术。这里所研究的变量之间的关系就是上述的统计关系。即当给定 x 的值，y 的值不能确定，只能通过一定的概率分布来描述。于是，我们称给定 x 时 y 的条件数学期望

$$f(x)=E(y|x) \tag{1.1}$$

为随机变量 y 对 x 的回归函数，或称为随机变量 y 对 x 的均值回归函数。(1.1)式从平均意义上刻画了变量 x 与 y 之间的统计规律。

在实际问题中，我们把 x 称为自变量，y 称为因变量。如果要由 x 预测 y，就是要利用 x，y 的观察值，即样本观测值

$$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n) \tag{1.2}$$

来建立一个公式，当给定 x 值后，就代入此公式中算出一个 y 值，这个值就称为 y 的预测值。如何建立这个公式，这就要从样本观测值 (x_i,y_i) 出发，观察 (x_i,y_i) 在平面直角坐标系上的分布情况，图 1.2 就是居民收入与商品销售额的散点图。由这个图可看出样本点基本上分布在一条直线的周围，因而要确定商品销售额 y 与居民收入 x 的关系，可考虑用一个线性函数来描述。图 1.2 中的直线即为线性方程

$$y=\alpha+\beta x \tag{1.3}$$

方程(1.3)式中的参数 α，β 尚不知道，这就需要由样本数据(1.2)式去进行

估计。具体如何去估计参数 α，β，我们在第二章中将详细介绍。

当我们由样本数据(1.2)式估计出 α，β 的值后，以估计值 $\hat{\alpha}$，$\hat{\beta}$分别代替(1.3)式中的 α，β，得方程

$$\hat{y}=\hat{\alpha}+\hat{\beta}x \tag{1.4}$$

(1.4)式方程就称为回归方程。这里因为因变量 y 与自变量 x 的关系呈线性关系，故我们称(1.4)式为 y 对 x 的线性回归方程。又因(1.4)式的建立依赖于观察或试验积累的数据(1.2)式，所以我们又称(1.4)式为经验回归方程。相对这种叫法，我们把(1.3)式称为理论回归方程。理论回归方程是设想把所研究问题的总体中每一个体的(x，y)值都测量了，利用其全部结果而建立的回归方程(1.3)式，这在实际中是办不到的。理论回归方程中的 α 是方程(1.3)式所画出的直线在 y 轴上的截距，β 为直线的斜率，它们分别称为回归常数和回归系数。而方程(1.4)式中的参数 $\hat{\alpha}$，$\hat{\beta}$被称为经验回归常数和经验回归系数。

回归分析的基本思想和方法以及“回归”名称的由来归功于英国统计学家 F. 高尔顿(F.Galton:1822～1911)。高尔顿和他的学生、现代统计学的奠基者之一 K. 皮尔逊(K.Pearson:1856～1936)在研究父母身高与其子女身高的遗传问题时，观察了1 078对夫妇，以每对夫妇的平均身高作为 x，而取他们的一个成年儿子的身高作为 y，将结果在平面直角坐标系上绘成散点图，发现趋势近乎一条直线。计算出的回归直线方程为

$$\hat{y}=33.73+0.516x \tag{1.5}$$

这种趋势及回归方程总的表明父母平均身高 x 每增加一个单位时，其成年儿子的身高 y 也平均增加 0.516 个单位。这个结果表明，虽然高个子父辈确有生高个子儿子的趋势，但父辈身高增加一个单位，儿子身高仅增加半个单位左右。反之，矮个子父辈确有生矮个子儿子的趋势，但父辈身高减少一个单位，儿子身高仅减少半个单位左右。通俗地说，一群特高个子父辈(例如排球运动员)的儿子们在同龄人中平均仅为高个子，一群高个子父辈的儿子们在同龄人中平均仅为略高个子；一群特矮个子父辈的儿子们在同龄人中平均仅为矮个子，一群矮个子父辈的儿子们在同龄人中平均仅为略矮个子，即子代的平均高度向中心回归了。正是因为子代的身高有回到同龄人平均身高的这种趋势，才使人类的身高在一定时间内相对稳定，没有出现父辈个子高其子女更高，父辈个子矮其子女更矮的两极分化现象。这个例子生动地说明了生物学中“种”的概念的稳定性。正是为了描述这种有趣的现象，高尔顿引进了“回归”这个名词来描述父辈身高 x 与子代身高 y 的关系。尽管“回归”这个名称的由来具有其特定的含义，人们在研究大量

的问题中，其变量 x 与 y 之间的关系并不总是具有这种“回归”的含义，但借用这种名词把研究变量 x 与 y 间统计关系的量化方法称为“回归”分析也算是对高尔顿这个伟大的统计学家的纪念。

1.3 回归分析的主要内容及其一般模型

一、回归分析研究的主要内容

回归分析研究的主要对象是客观事物变量间的统计关系，它是建立在对客观事物进行大量试验和观察的基础上，用来寻找隐藏在那些看上去是不确定的现象中的统计规律性的统计方法。回归分析方法是通过建立统计模型研究变量间相互关系的密切程度、结构状态、模型预测的一种有效的工具。

回归分析方法在生产实践中的广泛应用是它发展和完善的根本动力。如果从19世纪初(1809年)高斯(Gauss)提出最小二乘法算起，回归分析的历史已有190多年。从经典的回归分析方法到近代的回归分析方法，它们所研究的内容已非常丰富。如果按研究的方法来划分，回归分析研究的范围大致如下：

- 回归分析
  - 线性回归
    - 一元线性回归
    - 多元线性回归
    - 多个因变量与多个自变量的回归
  - 回归诊断
    - 讨论如何从数据推断回归模型基本假设的合理性
    - 当基本假设不成立时如何对数据进行修正
    - 判定回归方程拟合的效果
    - 选择回归函数的形式
  - 回归变量的选择
    - 自变量选择的准则
    - 逐步回归分析方法
  - 参数估计方法的改进
    - 岭回归
    - 主成分回归
    - 偏最小二乘法
  - 非线性回归
    - 一元非线性回归
    - 分段回归
    - 多元非线性回归
  - 含有定性变量的回归
    - 自变量含定性变量的情况
    - 因变量是定性变量的情况

二、回归模型的一般形式

如果变量 x_1，x_2，…，x_p 与随机变量 y 之间存在着相关关系，通常就意味着每当 x_1，x_2，…，x_p 取定值后，y 便有相应的概率分布与之对应。随机变量 y 与相关变量 x_1，x_2，…，x_p 之间的概率模型为

$$y=f(x_1,x_2,\cdots,x_p)+\varepsilon \tag{1.6}$$

其中，随机变量 y 称为被解释变量(因变量)；x_1，x_2，…，x_p 称为解释变量(自变量)。在计量经济学中，也称因变量为内生变量，自变量为外生变量；$f(x_1,x_2,\cdots,x_p)$为一般变量 x_1，x_2，…，x_p 的确定性关系，ε 为随机误差。正是因为随机误差项 ε 的引入，才将变量之间的关系描述为一个随机方程，使得我们可以借助随机数学方法研究 y 与 x_1，x_2，…，x_p 的关系。由于客观经济现象是错综复杂的，一种经济现象很难用有限个因素来准确说明，随机误差项可以概括表示由于人们的认识以及其他客观原因的局限而没有考虑的种种偶然因素。随机误差项主要包括下列因素的影响：

1. 由于人们认识的局限或时间、费用、数据质量等制约未引入回归模型但又对回归被解释变量 y 有影响的因素；

2. 样本数据的采集过程中变量观测值的观测误差的影响；

3. 理论模型设定误差的影响；

4. 其他随机因素的影响。

模型(1.6)式清楚地表达了变量 x_1，x_2，…，x_p 与随机变量 y 的相关关系，它由两部分组成：一部分是确定性函数关系，由回归函数 $f(x_1,x_2,\cdots,x_p)$给出；另一部分是随机误差项 ε。由此可见模型(1.6)式准确地表达了相关关系那种既有联系又不确定的特点。

当概率模型(1.6)式中回归函数为线性函数时，即有

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\varepsilon \tag{1.7}$$

其中，β_0，β_1，β_2，…，β_p 为未知参数，常称它们为回归系数。线性回归模型的“线性”是针对未知参数 $\beta_i(i=0,1,2,\cdots,p)$而言的。对于回归解释变量的线性是非本质的，因为解释变量是非线性时，常可以通过变量的替换把它转化成线性的。

如果 $(x_{i1},x_{i2},\cdots,x_{ip};y_i)$，$i=1,2,\cdots,n$ 是(1.7)式中变量 $(x_1,x_2,\cdots,x_p;y)$的一组观测值，则线性回归模型可表示为

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_px_{ip}+\varepsilon_i,\quad i=1,2,\cdots,n \tag{1.8}$$

为了估计模型参数的需要，古典线性回归模型通常应满足以下几个基本假设：

1. 解释变量 x_1，x_2，…，x_p 是非随机变量，观测值 x_{i1}，x_{i2}，…，x_{ip}是常数。

2. 等方差及不相关的假定条件为

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

这个条件称为高斯-马尔柯夫(Gauss-Markov)条件，简称 G-M 条件。在此条件下，便可以得到关于回归系数的最小二乘估计及误差项方差 σ^2 估计的一些重要性质，如回归系数的最小二乘估计是回归系数的最小方差线性无偏估计等。

3. 正态分布的假定条件为

$$\begin{cases} \varepsilon_i \sim N(0,\sigma^2), \quad i=1,2,\cdots,n \\ \varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n \text{ 相互独立} \end{cases}$$

在此条件下便可得到关于回归系数的最小二乘估计及 σ^2 估计的进一步的结果，如它们分别是回归系数及 σ^2 的最小方差无偏估计等，并且可以作回归的显著性检验及区间估计。

4. 通常为了便于数学上的处理，还要求 $n>p$，即样本容量的个数要多于解释变量的个数。在整个回归分析中，线性回归的统计模型最为重要。一方面是因为线性回归的应用最广泛；另一方面是只有在回归模型为线性的假定下，才能得到比较深入和一般的结果；再就是有许多非线性的回归模型可以通过适当的转化变为线性回归问题进行处理。因此，线性回归模型的理论和应用是本书研究的重点。

对线性回归模型我们通常要研究的问题有：

1. 如何根据样本$(x_{i1},x_{i2},\cdots,x_{ip};y_i), i=1,2,\cdots,n$ 求出$\beta_0,\beta_1,\beta_2,\cdots,\beta_p$ 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，以及如何进行实际问题的结构分析。

1.4 建立实际问题回归模型的过程

在实际问题回归分析模型的建立和分析中有几个重要的阶段，为了给读者一个整体印象，我们以经济模型的建立为例，先用逻辑框图表示回归模型的建模过程。见图 1.3。

下面我们按逻辑框图顺序叙述每个阶段要做的工作以及应注意的问题。

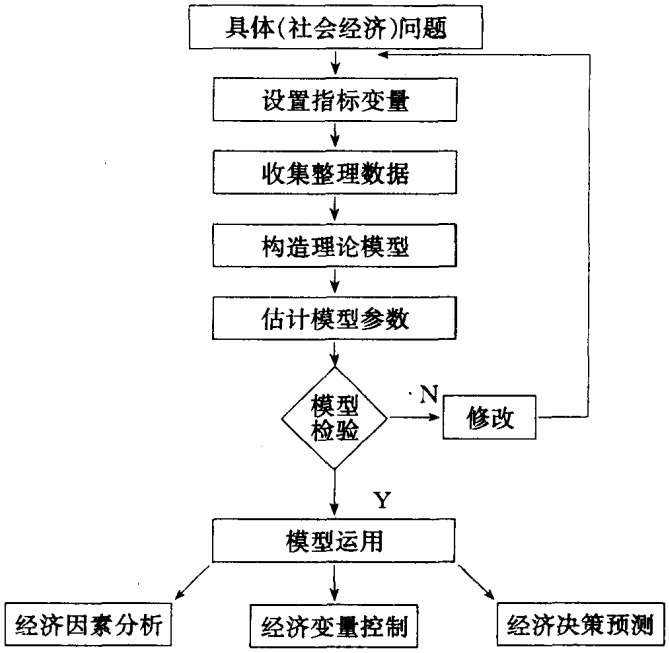


图 1.3　回归建模步骤流程图

一、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间相关变量的数量联系。首先要根据所研究问题的目的设置因变量 y，然后再选取与 y 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。尤其是在研究某种经济活动或经济现象时，我们必须根据具体的经济现象的研究目的，利用经济学理论，从定性角度来确定某种经济问题中各因素之间的因果关系。当我们把某一经济变量作为“果”之后，接着更重要的是要正确选择作为“因”的变量。在经济问题回归模型中，前者被称为“内生变量”或“被解释变量”，后者被称为“外生变量”或“解释变量”。变量的正确选择关键在于能否正确把握所研究的经济活动的经济学内涵。这就要求研究者对所研究的经济问题及其背景要有足够的了解。例如，要研究中国通货膨胀问题，必须懂得一些金融理论。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被解释变量，影响全国零售物价指数的有关因素就作为解释变量。参考文献[9]在研究中国通货膨胀问题时，曾把国民收入、居民存款、工农业总产值、全民所有制单位固定资产投资、货币流通量、职工平均工资、社会商品零售总额等 18 个