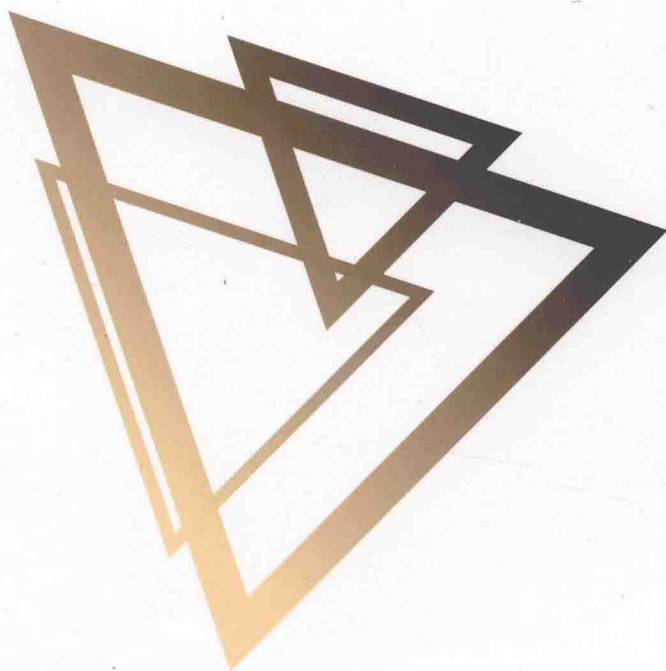


专业学位研究生教育系列教材

应用回归分析

Yingyong Huigui Fenxi

马立平 马乐 张蕊鑫 主编



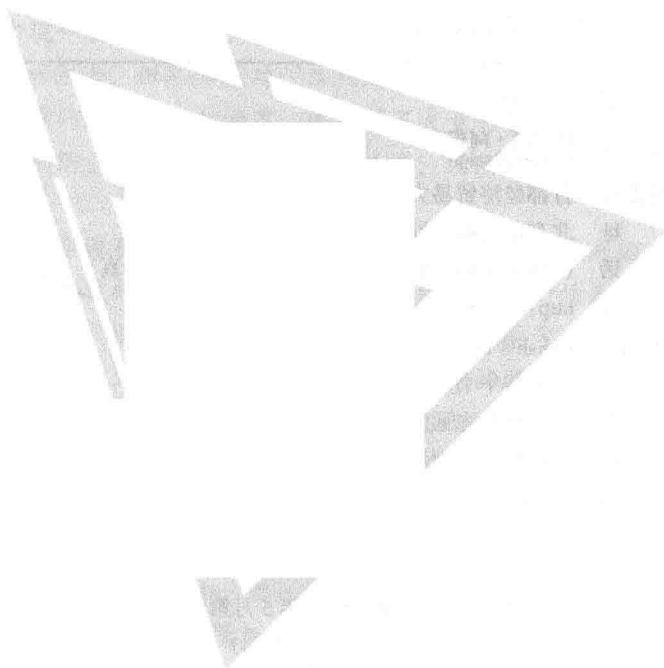
 首都经济贸易大学出版社
Capital University of Economics and Business Press

专业学位研究生教育系列教材

应用回归分析

Yingyong Huigui Fenxi

马立平 马乐 张蕊鑫 主编



 首都经济贸易大学出版社
Capital University of Economics and Business Press

• 北 京 •

图书在版编目 (CIP) 数据

应用回归分析/马立平, 马乐, 张蕊鑫主编. --北京: 首都经济贸易大学出版社, 2019. 5

ISBN 978-7-5638-2917-0

I. ①应… II. ①马… ②马… ③张… III. ①回归分析—研究生—教材
IV. ①O212. 1

中国版本图书馆 CIP 数据核字 (2019) 第 040691 号

应用回归分析

马立平 马乐 张蕊鑫 主编

责任编辑 雪 莲

封面设计 风得信·阿东
Fondesy Design

出版发行 首都经济贸易大学出版社

地 址 北京市朝阳区红庙 (邮编 100026)

电 话 (010) 65976483 65065761 65071505 (传真)

网 址 <http://www.sjmcb.com>

E-mail publish@cueb.edu.cn

经 销 全国新华书店

照 排 北京砚祥志远激光照排技术有限公司

印 刷 人民日报印刷厂

开 本 710 毫米×1000 毫米 1/16

字 数 334 千字

印 张 19

版 次 2019 年 5 月第 1 版 2019 年 5 月第 1 次印刷

书 号 ISBN 978-7-5638-2917-0/O · 47

定 价 39.00 元

图书印装若有质量问题, 本社负责调换

版权所有 侵权必究

前言 PREFACE

回归分析是统计学中的一个重要分支。它是研究变量之间数量依存关系的一类统计方法,是实际数据分析工作中最常用的统计工具之一,其应用领域不仅包括自然科学领域,也包括管理科学、社会科学和经济等各个领域。加强回归分析理论的推广、学习和方法的掌握,不仅成为统计教育界的共识,在实际部门相关问题的分析与研究中也得到了越来越多的应用和重视。

要在实际应用中发挥回归分析的重要作用,不仅要较好地理解回归分析的基本原理与方法,还需要掌握回归分析方法应用中的技巧,具有解决应用中所遇到的各种问题的能力。在教学与实际应用中,我们注意到,由于对分析对象认识的片面性或对回归分析原理与方法的理解和掌握不到位,研究者和学习者在方法的使用过程中常常会出现一些问题,例如,学习了回归分析的方法,但不知道如何应用;学习了回归分析方法,但不知道对什么问题在什么情况下应用;此外,在实际应用中存在着大量的方法误用的情况,像出现了大量的伪回归的问题、错误的模型等,形成统计陷阱;缺乏回归分析的应用技能,无法有效地将理论与方法应用于实践。

为了加强应用统计学课程的建设,笔者在总结多年统计学教学经验、实际工作经验的基础上编写了本教材。本教材的编写旨在介绍回归分析的基本理论、基本方法和方法的基本原理,以及此基础上如何应用,从而培养学生有效地使用回归分析方法分析、研究的能力,使他们能够透过数据发现问题。

全书共12章,包括基本的线性回归模型,如一元线性回归模型和多元线性回归模型;线性回归模型的推广、回归模型中自变量的选择和模型的设定与改进;经典回归模型假定的诊断与假定不满足时参数的估计方法;简单的时间数列回归模型、面板数据回归模型和分位数回归模型等。

本教材的编写着重突出以下几个方面的特点:

(1) 具有针对性,即针对经济管理类院校统计学专业、应用统计学专业、经济管理类学生的特点,在不失体系结构完整、内容严谨的前提下,借助一些简单的案例讨论强化最基本的回归分析方法的原理,突出应用。

(2) 突出实用性,即强调回归分析、统计思想的渗透,在系统阐述回归分析方法的基础上,以尽量避免或减少回归分析方法误用为目的,突出各种方法的

特点、应用条件与局限性的介绍，同时突出定性分析和定量分析的结合，强调回归分析的技术，提高学习者的应用能力。

(3) 注重系统性，即根据研究对象及内容，比较系统地介绍最基本的、最常用的回归分析的方法和基本理论，以体现教材结构的完整性。

(4) 强调实际操作性，在基本方法与原理的基础上，结合统计软件，系统地介绍回归分析的过程及技术实现，针对具体问题建立相应的回归模型，提高实际操作能力。

本教材的编写是在首都经济贸易大学研究生院的大力支持下完成的，同时还得到统计学院和首都经济贸易大学出版社的帮助，部分案例借鉴了相关的教材和著作，在此一并感谢。希望本教材的出版能够为学习者提供帮助，当然，由于作者水平有限，书中难免有不足和偏颇之处，敬请读者批评指正，我们衷心的感谢。

目 录 CONTENTS

1 回归分析概述 / 1

- 1.1 回归的释义与回归分析的作用 / 1
- 1.2 回归分析的基本过程 / 6
- 1.3 回归分析的基本概念与一般模型 / 13

2 一元线性回归模型 / 17

- 2.1 一元线性回归模型概述 / 17
- 2.2 模型的参数估计 / 25
- 2.3 回归模型的检验与评价 / 36
- 2.4 回归模型的修正 / 46
- 2.5 一元线性回归模型的应用 / 47
- 2.6 案例 / 51

3 多元线性回归模型 / 60

- 3.1 多元线性回归模型概述 / 60
- 3.2 多元线性回归的参数估计 / 64
- 3.3 回归模型的评价 / 69
- 3.4 回归模型的统计检验 / 72
- 3.5 标准化回归方程 / 77
- 3.6 案例 / 78

4 线性回归的推广 / 84

- 4.1 非线性回归 / 84
- 4.2 定性被解释变量的回归分析 / 93

5 关于自变量的选择 / 108

- 5.1 自变量选择的基本原则 / 108
- 5.2 增加一个自变量的“边际”贡献分析 / 118

5.3 自变量选择的逐步回归法 / 121

5.4 虚拟变量的选择 / 134

6 回归模型的设定与改进 / 142

6.1 回归模型的设定标准 / 142

6.2 模型设定错误的主要类型 / 146

6.3 回归模型设定问题的诊断与检验 / 149

7 共线性数据模型的建立与有偏估计 / 159

7.1 解释变量共线性定义、产生原因与后果 / 159

7.2 多重共线性的诊断 / 163

7.3 多重共线性的处理 / 169

8 方差齐性诊断与模型的加权最小二乘估计 / 179

8.1 异方差的定义、形成原因与后果 / 179

8.2 异方差的诊断 / 183

8.3 异方差的处理 / 186

9 误差独立性的诊断与模型的广义最小二乘法 / 195

9.1 自相关的定义、形成的原因与后果 / 195

9.2 自相关的诊断 / 197

9.3 自相关的处理 / 203

10 动态回归分析 / 211

10.1 动态回归模型的基本概念 / 211

10.2 自回归分布滞后模型的估计方法 / 213

10.3 格兰杰因果关系的检验 / 223

11 面板数据模型 / 230

11.1 面板数据模型概述 / 230

11.2 变截距模型 / 234

11.3 面板数据模型设定检验方法 / 242

11.4 案例 / 244

12 分位数回归模型 / 256

12.1 基本概念 / 257

12.2 分位数回归模型的估计 / 260

12.3 分位数回归模型的检验 / 266

12.4 分位数的计算与分位数回归的 Eviews 操作 / 271

12.5 分位数回归的案例分析 / 281

参考文献 / 295

1 回归分析概述

1.1 回归的释义与回归分析的作用

1.1.1 “回归”一词的历史渊源

“回归”一词是由英国著名统计学家弗朗西斯·高尔顿 (Francis Galton, 1822—1911) 首先提出的。高尔顿被誉为现代回归和相关技术的创始人, 主要致力于生物遗传问题的研究, 他在对父母身高和儿女身高之间关系的研究中发现, 虽然存在父母高, 儿女也高, 父母矮, 儿女也矮的总体趋势, 但从平均意义上说, 父母身高确定后, 其儿女的身高则趋向于或者说回归于总人口的平均身高。换句话解释, 尽管父母双亲是异常高或异常矮的, 但其儿女的身高并非普遍地异常高或异常矮, 而是具有回归于人口总的平均身高的趋势。更通俗地讲, 一群身高特别高的父辈, 其子辈的平均身高在同龄人中仅为高个子; 一群身高为高个子的父辈们, 其子辈的平均身高在同龄人中仅为略高个子; 相应地, 一群身高特矮的父辈们, 其子辈的平均身高在同龄人中呈现为矮个子; 一群矮个子的父辈们, 其子辈的平均身高在同龄人中为略矮个子。总体上, 子代的平均高度具有向人类平均身高回归的趋势, 这就是“回归”一词的由来。

正是因为子代的身高有回归到同龄人平均身高的这种趋势, 才使人类身高在一定时期内相对稳定, 未出现人类身高两极分化的现象。同样, 从平均水平看, 在第一次考试中成绩最差的那些学生在第二次考试中更倾向于有比上次好的成绩 (更加接近所有学生的平均成绩), 而在第一次考试中成绩最好的那些学生在第二次考试中则倾向于比上次差的成绩 (同样比较接近所有学生的平均成绩)。

如果把父辈和子辈的身高分别看作两个变量 X 和 Y , 若能分析出这两个变量之间的关系, 并建立适当的数学模型, 就可以根据父辈的身高预测子辈的身高, 这就是经典的回归分析要解决的问题, 它是处理变量之间关系的一种统计方法和技术。但是, 其所研究的变量之间的关系是一种统计关系, 即当给定变量 X 值

时, Y 的值不是唯一确定的, 它只能通过一定的概率分布来描述。于是, 我们称 X 条件下 Y 的数学期望 $f(x) = E(y|x)$ 为随机变量 Y 对 X 的回归函数, 该式从平均意义上刻画了变量 X 与 Y 之间的统计规律。

在实际问题分析时, 我们称 X 为自变量 (或解释变量), Y 为因变量 (或被解释变量), 若两变量之间表现为线性关系, 则可考虑用线性函数即直线方程 $E(y|x) = \alpha + \beta x$ 来描述。

高尔顿在研究身高的遗传问题时, 观察了 1 078 对夫妇, 以每对夫妇的平均身高作为 X , 取他们的一个成年儿子的身高作为 Y , 得到的回归直线方程为: $\hat{y} = -33.73 + 0.516x$ 。这一回归方程总体上表明父母平均身高 X 每增加一个单位时, 其成年儿子的身高 Y 将平均增加 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。

1.1.2 回归分析的发展与现代释义

尽管“回归”这个名称刚出现时具有特定的含义, 其最初的应用也只是在很窄的领域进行。科学技术的发展, 实践中生物、医学、农业、林业、经济、管理、金融、社会等领域的许多新问题的提出, 有力地推动了回归分析的发展。人们在研究大量实际问题时发现, 变量 X 与 Y 之间的关系并不总是具有上述传统“回归”的含义, 使用“回归”一词, 将研究变量 X 与 Y 之间统计关系的量化方法与分析称为“回归分析”更重要的是对高尔顿的纪念。

自高斯 (Gauss) 提出最小二乘法之后的 200 年来, 回归分析的应用非常广泛, 我们基本上很难找到不用它的领域。从国际上看, 如果关注 1969 年设立诺贝尔经济学奖的近 40 年以来的经济学得奖情况, 我们可以看到: 大部分获奖者都是统计学家、计量经济学家和数学家, 而其获奖成果都充分显示了回归分析方法在其研究中的作用。当然在我国, 回归分析方法的应用也日益普遍。在研究人员进行课题研究、实际部门工作人员进行经济活动分析时, 抑或是在撰写经济学硕士论文时, 如果没有回归模型等统计方法的应用与实证分析, 往往会给人以缺乏分量的感觉。

较早的、比较成熟的回归模型是经典回归模型, 包括线性回归模型和非线性回归模型。其中, 线性回归模型是最基本的, 也是最简单的形式。19 世纪初, 高斯首先提出了在线性关系下的回归方程的最小二乘法, 可以说是回归分析的起点, 其虽然简单, 但比较实用, 在实际应用中发挥了很大的作用, 但受计算手段的限制, 直到 20 世纪 60 年代回归分析都无显著的发展。同时, 在实际应用中严

格符合线性回归模型规律的问题并不多见,虽然大多数问题可以近似为线性回归模型,但在不少情形下,用非线性回归模型可能更加符合实际。从回归分析的发展看,非线性回归模型是线性回归模型的自然推广,目前已发展成为近代回归分析的一个重要研究分支。

在经典回归模型的研究中,通常首先认为变量之间呈线性关系,在对随机扰动项及变量提出若干假设条件下,进行模型的参数估计和推断。但是应用中,这些假设往往无法满足。随着满足基本假设的回归模型的理论研究与应用的逐渐成熟,人们对于违背基本假设的回归模型的参数估计问题进行了更多的、更广泛的研究。在对实际问题的研究中,人们发现最小二乘法估计参数的结果并不是在任何时候都令人满意的,尤其是对不满足基本假设的回归模型,其参数性质得不到有效的保证。于是统计学家从多方面进行了努力,逐步解决了经典回归模型与方法的不足。例如,提出了以岭估计、主成分估计、偏最小二乘估计等多种有偏估计的方法。

在对回归分析的应用与研究中,为了解决自变量个数较多的大型回归模型分析中自变量选择所存在的困难与各种问题,统计学家提出了许多关于回归自变量选择的准则和算法;为了克服最小二乘法对异常值的敏感性,统计学家研究并提出了各种稳健回归方法;为了研究模型假设条件的合理性及样本数据对统计推断的影响,统计学家研究并使用了相应的回归诊断方法;针对非线性回归模型,统计学家利用数学规划理论研究了非线性回归参数估计方法、微分几何方法等;为了分析和处理高维数据,尤其是高维非正态数据,统计学家提出了投影寻踪回归方法、切片回归方法等。

此外,为了解决实践中遇到的具体问题,回归分析新的研究方法不断涌现,如非参数统计、经验贝叶斯估计方法等,上述研究与应用在回归分析中占有重要地位,其对回归分析的发展起着重要的推动作用。

近 200 年来,回归分析的理论与方法研究不断发展,对统计学的发展与统计方法的进步具有重要的促进作用。例如,时间序列分析方法、多元统计分析方法中的判别分析方法、主成分分析法、因子分析法等都与回归分析具有密切的联系。这些方法的出现极大地丰富了统计学的研究方法。

总之,随着回归模型技术本身的不断完善和发展以及其应用领域的不断扩展,回归分析在统计学中越来越重要,并在很多领域的分析与研究中起到了重要且独到的作用。

当然,回归分析的发展离不开相关学科与技术的支持,例如,矩阵理论和计算机技术的发展为回归分析方法的实践与应用提供了极大的方便。如果没有计算机技术的发展,回归分析只能更多地停留在理论的探讨与研究上,其在各领域的

应用就会受到很大的限制。计算方法的改进和计算机技术的发展,使得过去不能完成,甚至不可想象的研究分析工作都得以推进,例如,对于复杂的宏观经济问题,可能需要建立涉及几十个,甚至几千个变量和方程的联立方程模型,其计算量巨大,若没有现代计算机,几乎无法运算。

从现代统计分析应用的角度看,回归分析是关于研究一个因变量对另一个或多个自变量的数量的依存关系,其目的与用意在于通过后者的已知值或给定值,去估计和(或)预测前者的(总体)均值。例如,在经济领域,经济学家想研究个人消费支出对个人可支配收入的依赖关系,这种分析会有助于估计边际消费倾向,即实际收入的变化所引起的消费支出的平均变化;具有决定市场价格或产出的垄断商,希望知道产品需求对价格变化的反应,通过回归分析,也许能够估计出产品需求的价格弹性,从而有助于确定最有利的价格策略。企业的销售经理肯定想知道广告宣传工作的市场反应,即企业产品销售量与广告费开支的关系,为此可以通过回归分析方法计算出相对于广告费用支出的需求弹性,即广告费用支出每变化1%时的销售百分比变化。

从方法论的角度看,回归分析主要包括回归模型的参数估计、假设检验、模型选择等理论和有关计算方法。

1.1.3 回归分析的主要作用

确定并估计回归方程是回归分析的最重要内容。回归方程较直接地、概括地反映了 Y 与一组变量 $X_1, X_2, \dots, X_p$ 之间的数量依存关系。利用所估计的回归方程,可评价单个预测变量的重要性;可分析预测变量的值改变之后, Y 值的变化或可能导致的模型系统的相关后果以及达到的效果;也可预测一组给定的预测变量值所对应的 Y 值。尽管回归方程是回归分析的最终产品,但同时也可获得许多其他的成果,如帮助我们理解某一特定环境中的变量间相互关系等。

回归分析方法被用于几乎所有的研究领域,包括社会科学、物理、生物、商业、科技和人文科学等。总体上,进行回归分析、建立回归模型,并对所得的回归模型进行统计检验与评价,可起到描述、预测及控制等作用。

1.1.3.1 对研究对象进行定量描述与分析

【例 1.1】 对两个地区的生活水平进行对比分析,如果说 A 地区生活水平没有 B 地区的生活水平高,这只是一种定性的描述。我们可以对两地区近 20 年的时间数列资料,以恩格尔系数为因变量,以时间 t 为自变量,采用回归分析方法得到下面回归模型:

A 地区:

$$\text{Engel} = 0.60 - 0.0077t$$

$$(69.9) \quad (-8.9)$$

$$R^2 = 0.83, \quad DW = 0.86, \quad F = 79.9$$

B 地区:

$$\text{Engel} = 0.29 - 0.0043t$$

$$(24.0) \quad (-12.1)$$

$$R^2 = 0.97, \quad DW = 1.2, \quad F = 372$$

通过以上回归模型,我们可以更直观、更深刻、更具体地看到:

(1) 初始阶段, A 地区的恩格尔系数高于 B 地区。

(2) 从恩格尔系数的下降速度看,近 20 年来, A 地区的恩格尔系数平均每年下降 0.0077, 快于 B 地区, A 地区的恩格尔系数的年下降速度是 B 地区的 1.79 倍。

(3) 结合收入水平等相关统计资料,验证了经济理论,即随着收入的增加,恩格尔系数的下降速度要逐步减缓。

可见,通过定量分析,对这一问题的了解要比只做定性分析清晰得多。

1.1.3.2 建立回归模型

通过建立回归模型寻找研究对象系统中各变量发展变化的规律及其之间的关系,通过模型及参数的可靠估计值(如边际系数、弹性系数、技术系数、比率、速率等),为预测与控制工作提供依据,这都是回归分析的重要作用。

【例 1.2】为充分展示改革开放前后 M_0 与 GDP 之间关系的变化,利用某地区 1952 年以后的现金需求量 (M_0) 和地区生产总值 (GDP) 数据,建立回归模型如下:

$$M_0 = 0.062GDP + 0.078GDPD_1$$

$$(2.4) \quad (3.0)$$

$$R^2 = 0.99, \quad DW = 0.67$$

其中:1978 年及以前, $D_1 = 0$; 1979 年以后, $D_1 = 1$ 。

上述模型也可以写成如下形式:

1978 年及以前: $M_0 = 0.062GDP$ 。

1979 年及以后: $M_0 = 0.140GDP$ 。

通过模型可以看到:

(1) 市场经济与计划经济期间的 M_0 与 GDP 之间的关系有明显不同;

(2) 改革开放后, GDP 对现金的边际需求比改革开放前增加了 1.26 倍。

1.1.3.3 预测

有时某些变量的数值在事前不可能获得,因为它们多数是未来值,所以现在

无法获得。但是为了决策,我们想要在获得它们真正数据前就能知道它们的值大约是多少。例如,作为企业的营销管理人员来说,未来的销售量在事前并不知道,但是经验使我们知道一般广告宣传费用与销售量之间存在一定的关系,如能有效地建立并估计两者之间的回归模型,就可以利用广告宣传费用的多少来预测产品的大致销售量,进而做好企业的生产计划。再如,若根据经验判断某产品未来的市场需求取决于居民收入水平与产品的价格水平,而假设居民收入水平在未来一段时间内保持相对稳定,那么产品的价格将会是影响企业产品市场竞争力和销售量及企业经济效益的重要因素。这时,如果有历史数据或相关的统计数据,我们就可以借助回归分析建立回归模型,通过模拟来预测在不同价格水平下的产品可能的销售量,并据此测算企业的预期利润,从而帮助企业确定产品的价格。可以说,预测是对企业、对单位、对决策者个人来说都是很重要的课题,因为它可以帮助决策者、帮助企业做规划。当然这也是利用回归模型所要解决的最重要内容,也是最困难的内容之一。

当然,之所以我们要通过回归模型进行预测,往往是由于真正的因变量 Y 值可能需要花费很高的代价才能获得,但是要得到影响 Y 的主要因素即自变量 X 值的花费却很少,因此用成本低的 X 值去预测 Y 的值,这是很划算的。当然,利用回归模型进行科学的预测,要求自变量选择得要正确,同时回归模型的形式要合理。

1.1.3.4 控制

控制与预测刚好相反,预测是给定 X 预测 Y ,而控制是想得到 Y 的某一结果,去确定 X 的取值应是多少。如某公司年底想做促销活动,利用媒体做广告,但到底要投入多少广告费才能使销售数量达到 10 000 件,根据过去 10 多年的经验,建立销售量 Y 与广告费 X (万元) 的回归模型为:

$$y = 2\,000 + 16x$$

则可以根据此模型得到:若想 $y = 10\,000$ 件,则 x 应为 500 万元,即需要投入 500 万元的广告费才能得到 10 000 件的销售量。

1.2 回归分析的基本过程

1.2.1 回归分析的主要内容

回归分析的研究对象是由多个有相关关系的现象组成的客观系统。回归分析是建立在对客观事物进行大量试验、观察和调查的基础上,通过建立回归模型,研究变量间相互关系的密切程度、结构状态、数量依存关系,以寻找不确定现象背后存在的统计规律的理论与方法。

其主要内容包括：

(1) 按照自变量和因变量之间的关系类型，即按回归模型的形式，回归分析可划分为线性回归与非线性回归；按照回归模型涉及的自变量的多少划分，回归分析可分为一元回归分析和多元回归分析。如果在回归分析中，只包括一个自变量和一个因变量，且二者之间的关系可用一条直线近似表示，这种回归分析称为一元线性回归分析。如果回归分析中包括两个或两个以上的自变量，且因变量和自变量之间是线性关系，则称为多元线性回归分析。这样，按变量的多少和回归模型的形式，回归分析可做如下划分（见表 1.1）。

表 1.1 按变量的多少和回归模型的形式分类

回归模型形式	变量的多少	回归类型
线性回归	一个因变量，一个自变量	一元线性回归
	一个因变量，多个自变量	多元线性回归
	多个因变量，多个自变量	多个因变量与多个自变量的回归
非线性回归	一个因变量，一个自变量	一元非线性回归，分段回归
	一个因变量，多个自变量	多元非线性回归

(2) 按变量的性质与类型，回归分析方法包括方差分析、协方差分析和 Logistic 回归等（见表 1.2）。

表 1.2 按变量的性质和类型分类

回归的类型	变量的类型与条件
方差分析	所有自变量均为定性变量
协方差分析	部分自变量为定性变量，另一部分为定量变量
Logistic	因变量为定性变量

(3) 按回归分析所涉及的内容，回归分析包括变量的选择、回归模型的设计和回归模型的诊断等，具体包括的内容如表 1.3 所示。

表 1.3 回归分析主要内容

主要内容	具体内容	方法
回归变量的选择	变量选择的方法	向前法
		向后法
		逐步回归
		主成分法
	变量的类型	自变量含有定性变量的回归模型
		因变量是定性变量的回归模型

续表

主要内容	具体内容	方法
参数估计	最小二乘法	普通最小二乘法
		加权最小二乘法
		广义最小二乘法
	基于最小二乘法的有偏估计	岭估计
		主成分回归
		偏最小二乘法
	似然估计及其他	最大似然估计
		其他
回归模型的设定	数据驱动建模	
	理论引导建模	
	回归函数形式的选择	
回归模型的诊断	回归模型基本假设的检验	
	回归模型的修正	
	回归模型的评价	

1.2.2 回归分析的基本流程

回归分析的基本流程大体可分为三个阶段，即理论模型的构造与数据收集阶段、模型参数估计与模拟阶段及模型的应用阶段。具体的建模过程如图 1.1 所示。

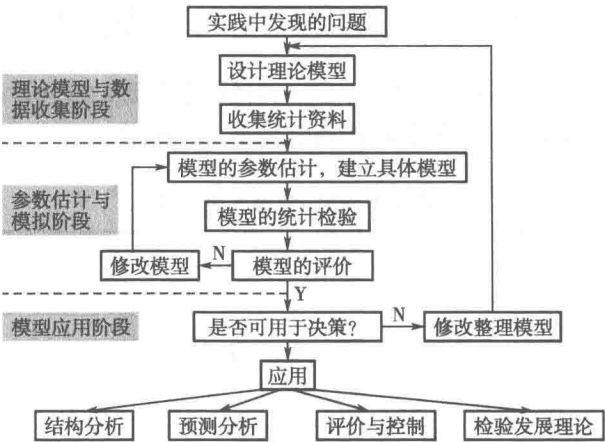


图 1.1 具体建模过程

1.2.2.1 理论模型构造与数据收集阶段

(1) 理论模型的构造。理论模型的构造包括变量的选择、模型形式的假设等基本内容。

①设置研究问题的指标、变量。回归分析的主要目的是揭示所研究对象系统中相关变量的数量依存关系。这就需要在确定研究对象及研究目的的基础上,首先,选用恰当的指标(变量)描述所研究的事物及其各主要影响因素。其次,根据各指标(变量)之间的关系确定其性质。例如,在用回归分析方法研究居民消费问题时,首先用居民消费支出作为衡量消费水平的变量,即因变量;根据消费理论我们知道影响居民消费支出的最主要的影响因素是居民收入,那么可以用居民可支配收入作为自变量。

总体上,选择解释变量需要理论支持,需要正确把握所研究现象背后的基本理论,例如,对某一经济现象建立回归模型,变量选择的依据之一就是该经济现象中暗含的经济学理论和经济行为规律,这是正确选择变量的基础。也就是说,建立模型和选择变量需要理论抽象。例如,建立产出回归模型要考虑到供需关系,如供给不足时,影响产出量的因素应该是投入要素方面的因素,这时的自变量一般是技术、资本、劳动等;而当需求不足时,影响产出量的因素应该在需求要素方面,主要是消费需求、投资需求和出口需求方面的因素,而不是投入要素,这时若研究消费品产出量应选择居民收入作自变量,而若研究生产资料产出量,则应选择固定资产投资作为自变量等。

确定指标(变量)需要研究者对所研究的问题及其背景有足够的了解,必要时,需要吸收一些专门领域的专家,与他们合作,共同选择确定变量。例如,研究金融模型时,最好吸收具有金融背景的专家学者合作;研究农业问题,最好吸收具有农业知识背景的专家合作等。只有这样,才有可能选取最恰当的指标变量。当然,由于人们对问题的认识存在一定的局限性,确定的指标变量并不一定是最佳的。

通常,被解释变量相对容易确定,而对被解释变量有影响的解释变量,即自变量的确定,难度相对要大一些。一方面,建立模型时,要求不能遗漏掉重要的解释变量,但研究者可能并不一定全面地了解影响被解释变量的重要因素,因而可能会遗漏掉重要的解释变量;另一方面,为了保证模型的简洁,要求模型的自变量不能过多,同时为了保证估计模型参数的有效性,参数估计方法要求设置的解释变量之间应该不具有高度的线性相关关系,但在实际问题分析时,我们很难保证多个自变量之间不相关。此外,我们经常遇到在理论上是重要的解释变量,但实际建模时无法找到该变量的统计数据,这时只能忽略该变量或用其他相近的变量代替,从而会影响到模型的效果。