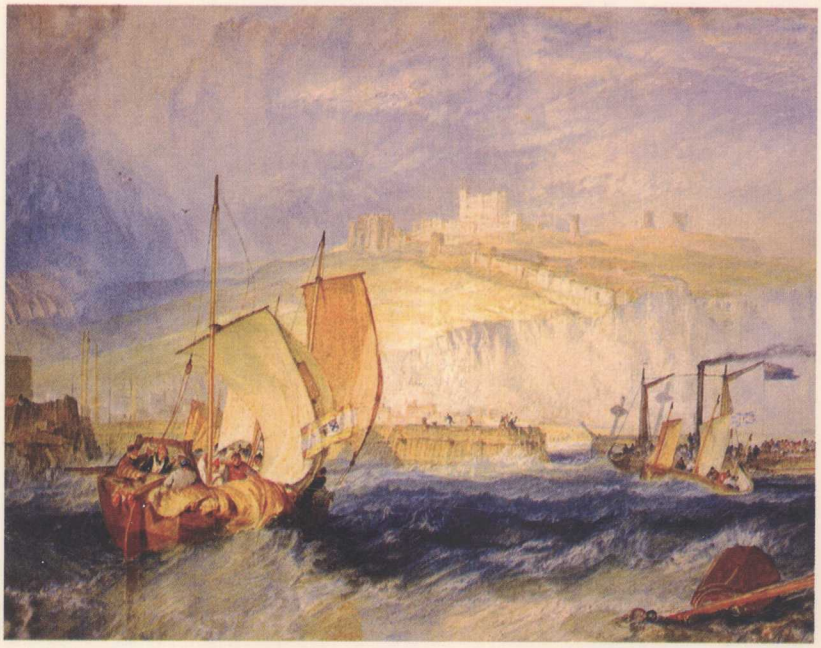
北京市高等教育精品教材立项项目

HZ BOOKS
华章教育

回归分析

Regression Analysis

马立平 编著



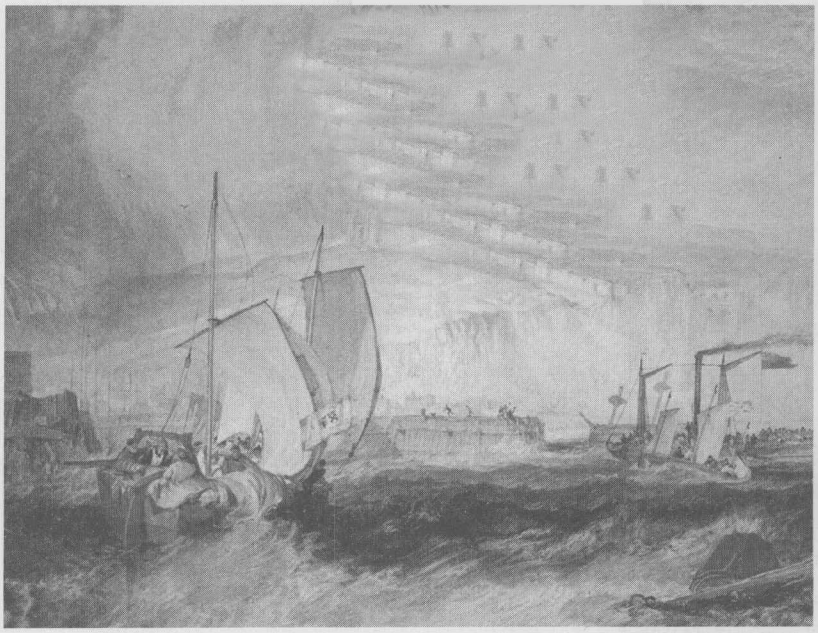
机械工业出版社
China Machine Press

北京市高等教育精品教材立项项目

回归分析

Regression Analysis

马立平 编著



机械工业出版社
China Machine Press

图书在版编目（CIP）数据

回归分析/马立平编著．—北京：机械工业出版社，2014.1
（北京市高等教育精品教材立项项目）

ISBN 978-7-111-45366-6

Ⅰ．回… Ⅱ．马… Ⅲ．回归分析－高等学校－教材 Ⅳ．O212.1

中国版本图书馆 CIP 数据核字（2014）第 001632 号

本书主要介绍回归分析的基本原理、基本方法及其在经济领域中的应用，具体内容包括回归分析概述、变量间的相关关系分析、一元线性回归分析、多元线性回归分析、方差齐性诊断与模型的加权最小二乘估计、误差独立性的诊断与模型的广义最小二乘法、共线性数据模型的建立与有偏估计、关于自变量的选择、动态回归分析、线性回归的推广、回归模型的设定与改进等。此外，书中还介绍了 EViews 和 SPSS 软件的基本使用方法，并在每章后提供了难度适宜的思考题和练习题，便于读者动手实践，巩固所学知识。

本书叙述通俗易懂，可以作为高等院校统计及相关专业回归分析课程的本科生教材和教学参考书，也可供相关研究人员阅读与参考。

机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码 100037）
责任编辑：王春华
北京瑞德印刷有限公司印刷
2014 年 3 月第 1 版第 1 次印刷
185mm×260mm·15.5 印张
标准书号：ISBN 978-7-111-45366-6
定 价：39.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：(010) 88378991 88361066 投稿热线：(010) 88379604
购书热线：(010) 68326294 88379649 68995259 读者信箱：hzjsj@hzbook.com

前　言

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，其应用领域十分广泛，包括自然科学、管理科学和社会、经济等各个领域。加强回归分析理论与方法的学习和应用不仅成为统计教育界的共识，也越来越受到实际应用部门相关分析人员的重视。

当然，回归分析受到普遍关注不仅仅在于对方法的研究，还包括对回归分析方法的使用、应用中的技术与技巧。但是由于使用者对分析对象认识的片面性或对回归分析方法掌握的不到位，研究者或学习者对方法的使用常常会出现一些问题，例如：1）学习了很多的回归分析方法，但不知如何应用；2）学习了方法，但不知对什么问题在什么情况下应用；3）实践中对所学习方法的误用，出现大量的伪回归，形成统计陷阱；4）缺乏回归分析技能，无法有效地将理论正确地应用于实践。

为了进一步加强统计学专业的课程建设，作者在总结多年教学经验、实际工作经验的基础上编写了本教材。本书旨在通过介绍回归分析的基本原理、基本方法及其在经济领域中的应用，培养学生使用回归分析方法解决实际问题的能力。全书包括回归分析基础、经典线性回归分析、违背经典假设的线性回归方程参数估计和实践中的回归分析四个部分，共11章。

本教材的编写着重突出了以下几方面的特点：

1）有针对性，即针对经济类院校统计专业、经济管理类专业学生的特点，在不失回归分析方法、体系结构完整、内容严谨的前提下，借助典型的相关经济案例论述回归分析的原理与方法，突出实际案例的应用。

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，突出各种方法的特点和局限性的介绍，以避免或减少对回归分析方法的误用，同时突出定性分析与定量分析的结合，强调回归分析的技术，以提高学习者分析的技能。

3）注重系统性，即根据研究对象及内容，全面介绍各种主要回归分析方法的理论，体现本教材结构的完整性。

4）结合统计软件全面、系统地介绍回归分析过程及技术实现，针对具体问题建立相应的回归模型，提高学生的实际操作能力与水平。

本书的编写得到了北京市属高等学校高层次人才引进与培养计划项目的资助和首都经济贸易大学、机械工业出版社华章公司的大力支持，部分案例借鉴了相关的教材和著作，在此一并表示感谢。作者期望本书的出版能够为学习者提供帮助，当然，由于水平有限，书中难免有不足和偏颇之处，敬请读者批评指正，作者表示衷心的感谢。

马立平

2013年11月

教学建议

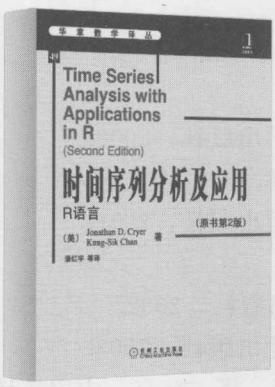
教学章节	教学要求	课时
第 1 章	了解回归分析的含义和作用 掌握回归分析的基本流程 了解回归模型的基本类型	2
第 2 章	了解变量间的关系 掌握变量相关分析的基本方法 掌握相关分析的应用	2
第 3 章	了解经典回归模型的基本问题 掌握一元线性回归模型的估计方法 掌握回归模型的统计检验与应用	3
第 4 章	了解多元回归模型的一般形式与基本假设 掌握参数估计的方法 了解参数估计的性质 掌握多元回归模型的检验与评价方法	4
第 5 章	掌握方差齐性的概念 掌握异方差的后果与诊断方法 掌握加权最小二乘估计的原理与应用	2
第 6 章	掌握误差独立性的概念 掌握误差项自相关的后果与诊断方法 掌握广义最小二乘估计的原理与应用	2
第 7 章	掌握共线性的概念 掌握多重共线性的后果与诊断方法 掌握多重共线性情况下的处理方法及有偏估计方法的原理与应用	4
第 8 章	掌握自变量选择的基本原则 了解自变量选择在回归分析中的重要作用 掌握回归分析中自变量选择的主要方法与应用	4
第 9 章	了解分布滞后模型、自回归模型的背景 掌握自回归分布滞后模型的估计方法 掌握自回归分布滞后模型的应用	3
第 10 章	了解几种典型的非线性回归模型 掌握定性因变量回归分析的基本模型 了解广义线性模型	3
第 11 章	了解回归模型的设定标准、原则与思路 了解模型设定误差的类型 掌握回归模型设定问题的诊断与检验	3
附录 A	掌握 EViews 软件的基本使用方法	1
附录 B	掌握 SPSS 软件的基本使用方法	1
总课时	第 1～11 章建议课时 综合实训建议课时	32 16

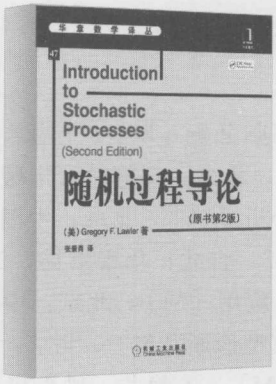
说明：

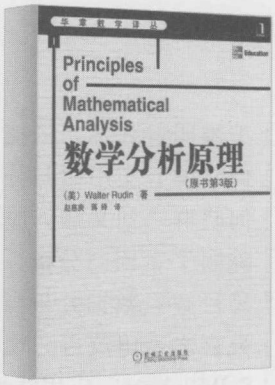
1）建议课堂教学“讲-练”结合。

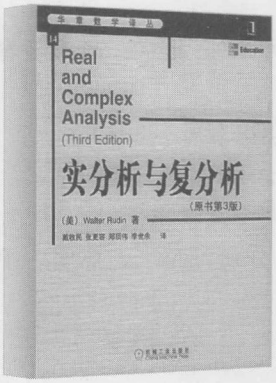
2）建议教学分为核心知识技能模块（前 11 章的内容）和综合实训模块。其中核心知识技能模块建议教学学时为 32，综合实训模块建议学时为 16，不同学校可以根据各自的教学要求和计划学时数对教学内容进行取舍。

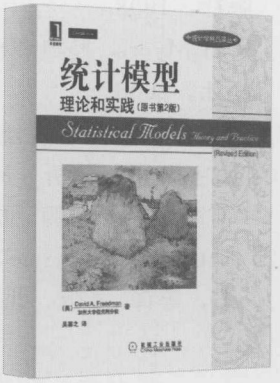
推荐阅读











■ **时间序列分析及应用：R语言（原书第2版）**
作者：Jonathan D. Cryer　Kung-Sik Chan
ISBN：978-7-111-32572-7
定价：48.00元

■ **随机过程导论（原书第2版）**
作者：Gregory F. Lawler
ISBN：978-7-111-31544-5
定价：36.00元

■ **数学分析原理（原书第3版）**
作者：Walter Rudin
ISBN：978-7-111-13417-6
定价：28.00元

■ **实分析与复分析（原书第3版）**
作者：Walter Rudin
ISBN：978-7-111-17103-9
定价：42.00元

■ **数理统计与数据分析（原书第3版）**
作者：John A. Rice
ISBN：978-7-111-33646-4
定价：85.00元

■ **统计模型：理论和实践（原书第2版）**
作者：David A. Freedman
ISBN：978-7-111-30989-5
定价：45.00元

推荐阅读

书名	书号	定价	出版年	作者
概率统计（英文版·第4版）	978-7-111-38775-6	139	2012	（美）Morris H. DeGroot等
数值分析（英文版·第2版）	978-7-111-38582-0	89	2013	（美）Timothy Sauer
数论概论（英文版·第4版）	978-7-111-38581-3	69	2013	（美）Joseph H. Silverman
数理统计学导论（英文版·第7版）	978-7-111-38580-6	99	2013	（美）Robert V. Hogg等
代数（英文版·第2版）	978-7-111-36701-7	79	2012	（美）Michael Artin
线性代数（英文版·第8版）	978-7-111-34199-4	69	2011	（美）Steven J. Leon
商务统计：决策与分析（英文版）	978-7-111-34200-7	119	2011	（美）Robert Stine等
多元数据分析（英文版·第7版）	978-7-111-34198-7	109	2011	（美）Joseph F. Hair, Jr等
统计模型：理论和实践（英文版·第2版）	978-7-111-31797-5	38	2010	（美）David A. Freedman
实分析（英文版·第4版）	978-7-111-31305-2	49	2010	（美）H. L. Royden
概率论教程（英文版·第3版）	978-7-111-30289-6	49	2010	（美）Kai Lai Chung
初等数论及其应用（英文版·第6版）	978-7-111-31798-2	89	2010	（美）Kenneth H. Rosen
数学建模（英文精编版·第4版）	978-7-111-28249-5	65	2009	（美）Frank R. Giordano
复变函数及应用（英文版·第8版）	978-7-111-25363-1	65	2009	（美）James Ward Brown
数学建模方法与分析（英文版·第3版）	978-7-111-25364-8	49	2008	（美）Mark M. Meerschaert
数学分析原理（英文版·第3版）	978-7-111-13306-3	35	2004	（美）Walter Rudin
实分析与复分析（英文版·第3版）	978-7-111-13305-6	39	2004	（美）Walter Rudin
泛函分析（英文版·第2版）	978-7-111-13415-2	42	2004	（美）Walter Rudin

目　录

第一部分
回归分析基础

第1章 Chapter

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，因此，高尔顿被誉为现代回归和相关技术的创始人。他主要致力于生物遗传问题的研究，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。换句话说，尽管父母是异常高或异常矮的，但其儿女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。有人更通俗地讲，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，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，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，其子辈的平均身高在同龄人中为略矮个子。总体上，子辈的平均身高具有向人类平均身高回归的趋势，这就是“回归”一词的由来。

正是因为子辈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，从总体上的平均水平看，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有比上次差的成绩（同样比较接近所有学生的平均成绩）。

把父辈和子辈的身高分别看做两个变量 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，这就是经典回归分析要解决的问题。经典回归分析是处理变量之间关系的一种统计方法和技术，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 $f(X)=E(Y \mid X)$ 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 $E(Y \mid X)=\alpha+\beta X$）来描述。

例如，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，而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的平均身高 x 每增加一个单位，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”这个名称的由来具有特定的含义，其最初的应用也只是在很窄的领域进行，但是随着科学技术的发展，生物、医学、农业、林业、经济、管理、金融、社会等领域许多实际新问题的提出，有力地推动了回归分析的发展。人们在研究大量的实际问题时发现，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（C. F. Gauss）提出最小二乘法之后，回归分析的应用越来越广泛，现在我们基本上很难找到不用它的领域。从国际上看，如果我们关注 1969 年设立诺贝尔经济学奖四十多年以来的经济学得奖情况，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出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，或者在撰写经济学论文时，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，它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，但比较有用，在实际应用中发挥了很大的作用，但 20 世纪 60 年代以前回归分析受计算手段的限制，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模型的自然推广，其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现线性关系，同时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，同时，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、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，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努力，逐步解决了经典回归模型与方法的不足。例如，提出了岭估计、主成分估计、偏最小二乘估计等多种有偏估计方法。

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，研究并提出了各种稳健回归；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，研究并使用了相应的回归诊断方法；针对非线性回归模型，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，提出了投影寻踪回归、切片回归等。

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重要地位，其对回归分析的发展起着很重要的推动作用。

回归分析理论与方法的不断发展，对统计学的发展与统计方法的进步起到了重要的促进作用。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的地位越来越重要，其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，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。由于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解决，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，没有现代计算机的使用几乎无法运算。

从现代统计分析应用的角度看，回归分析是研究一个因变量对另一个或多个自变量的数量依存关系，其目的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，在经济领域中，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政策；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，即广告费支出每变化百分之一时的销售百分比变化。

从方法论的角度看，回归分析主要是研究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的内容。回归方程较直接而概括地反映了 Y 与一组变量 X_1，X_2，…，X_p 之间的数量依存关系。利用所估计的回归方程，可以评价单

个预测变量的重要性，分析预测变量的值改变之后 Y 值的相应变化或可能导致的模型系统的相关后果以及达到的效果，也可以预测一组给定的预测变量值所对应的 Y 的值。尽管回归方程是回归分析的最终产品，但是从中我们也可获得许多其他的成果，如有助于理解某一特定环境中变量间的相互关系，等等。

回归分析方法几乎被用于所有的研究领域，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后，可起到描述、预测及控制等作用。具体地讲，回归分析的主要作用如下：

1）对研究对象进行定量描述与分析。例如，对两个地区的生活水平进行对比分析，如果说 A 地区的生活水平没有 B 地区的生活水平高，这只是一种定性的描述。若对 A、B 两地区 20 年恩格尔系数的时间序列资料，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如下回归模型：

A 地区：Engel＝0.60－0.007 7 t

B 地区：Engel＝0.29－0.004 3 t

据此进行的分析，将会使我们对此问题的描述更直观、更深刻、更具体。通过以上回归模型，我们可以看到：

- 初始阶段，A 地区的恩格尔系数高于 B 地区。
- 从恩格尔系数的下降速度看，20 年中，A 地区的恩格尔系数平均每年下降0.007 7，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。
- 若结合收入水平等相关统计资料，可以验证随着收入的增加，恩格尔系数的下降速度要逐步减缓的经济理论。

由此可见，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。

2）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。

例如，为充分展示改革开放前后 M_0 与 GDP 之间关系的变化，利用某地区现金需求量（M_0）和地区生产总值（GDP）时间序列数据，建立回归模型如下：

$$M_0 = 0.062\ \mathrm{GDP} + 0.078\ \mathrm{GDP}\ D_1$$

$$(2.4) \qquad (3.0)$$

$$R^2 = 0.99, \quad \mathrm{DW} = 0.67$$

其中：1978 年及以前，D_1＝0，1979 年及以后 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：

- 1978 年前后的 M_0 与 GDP 之间的关系有明显不同。
- 1979 年以后，GDP 对现金的边际需求比改革开放前增加了 1.26 倍。

3）预测。有时某些变量的数值在事前不可能获得，因为它们经常是未来值，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，未来的销售量在事前并不知道，但是通过经验我们知道一般广告宣传费与销售量之间存在一定的关系，如果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的高低与产品的价格水平，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的市场竞争力和销售量，进而影响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，我们就可以借助于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产品可能的销售量，并据此测算企业的预期利润，从而最终帮助确定产品的价格。对企业、单位和决策者个人来说，预测都是很重要的课题，因为它可以帮助决策者和企业做规划。当然，这也是利用回归模型所要解决的重要问题，也是最困难的内容之一。

之所以我们要通过回归模型进行预测，往往是因为真正的因变量 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，这时用成本低的 X 值来预测 Y 值是很划算的。当然，要利用回归模型进行科学的预测，需要选择正确的自变量，同时回归模型的形式要合理。

4）控制。控制与预测刚好相反，预测是给定 X 预测 Y，而控制是为得到 Y 的某一结果来确定 X 的取值。例如，某公司年底想做促销活动，利用媒体做广告，但到底需要投入多少广告费才能使销售达到 10 000 件？根据过去的经验，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　回归分析的基本类型与主要内容

回归分析的研究对象是由多个存在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。

从不同的角度，回归分析的分类如下：

1）按照自变量和因变量之间的关系类型，即回归模型的形式，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，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，且二者的关系可用一条直线近似表示，则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，回归分析的划分如表 1-1 所示。