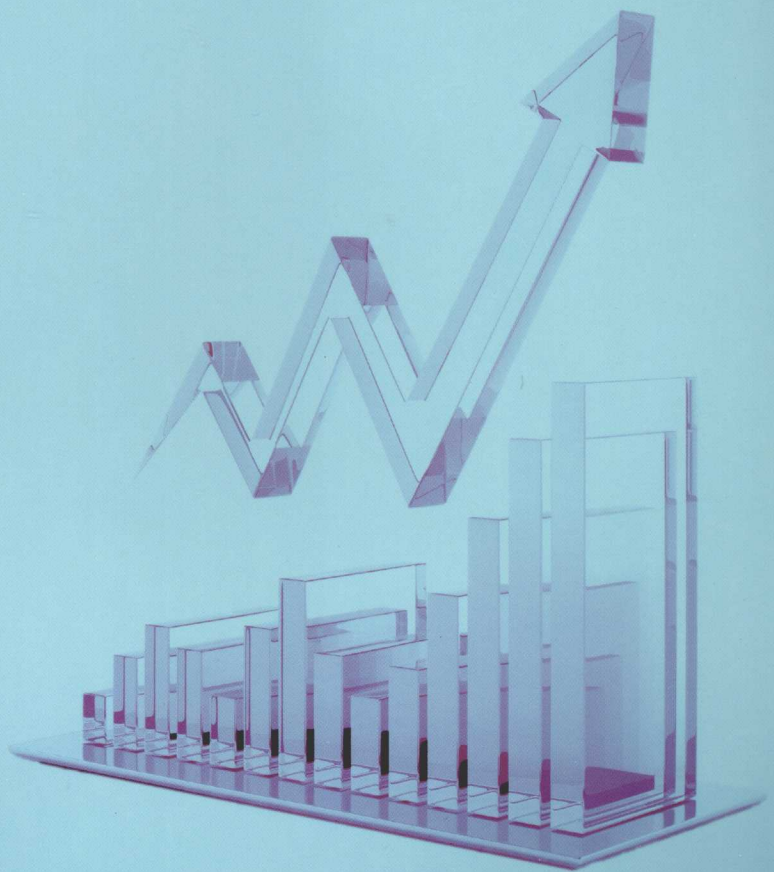
# 应用

唐年胜　李会琼　编著

# 回归分析



科学出版社

西部数学规划教材系列丛书

# 应用回归分析

唐年胜　李会琼　编著



科学出版社
北京

## 内 容 简 介

本书基于R软件系统介绍回归分析的理论和方法，包括一元线性回归模型与多元线性回归模型的参数估计理论和方法以及自变量选择，影响点和异常点的识别及处理，异方差性诊断和自相关性问题及处理，多重共线性问题及处理，多元线性回归模型的有偏估计，非线性回归模型和含定性变量的回归模型的参数估计理论、方法及算法，广义线性回归模型和缺失数据模型的统计推断等。此外，还收集了大量的实际例子，并配有相应的R程序来介绍这些回归分析方法在社会学、经济学、教育学和心理学等领域的具体应用.

本书可作为统计学专业本科生、应用统计专业硕士生的教学用书，也可作为社会学、教育学、心理学、经济学、金融学、人口学、生物医学和临床研究等领域的理论研究者和实际应用者的参考书.

**图书在版编目(CIP)数据**

应用回归分析/唐年胜，李会琼编著. —北京：科学出版社，2014.1
(西部数学规划教材系列丛书)
ISBN 978-7-03-039375-3

Ⅰ. ①应… Ⅱ. ①唐… ②李… Ⅲ. ①回归分析–高等学校–教材 Ⅳ. ①O212.1

中国版本图书馆 CIP 数据核字(2013)第 309597 号

责任编辑：郝玉龙 / 责任校对：韩 杨
责任印制：邝志强 / 封面设计：墨创文化

科 学 出 版 社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

成都创新包装印刷厂 印刷

科学出版社发行 各地新华书店经销
*
2014 年 1 月第 一 版 开本：787×1092 1/16
2014 年 1 月第一次印刷 印张：14.75
字数：350 000

**定价：29.00 元**

(如有印装质量问题，我社负责调换)

# 前　言

随着计算机技术的快速发展与统计软件的开发使用, 统计学在各行各业的应用越来越广泛. 在这些应用中, 如何用统计的理论和方法对给定的数据建立一个与之相符的回归模型呢? 这是数据分析人员极为关心的一个重要问题. 为了回答这个问题, 本书首先从数据和变量的概念入手, 深入浅出地介绍建立回归模型的一般步骤, 一元线性回归模型与多元线性回归模型的参数估计理论和方法以及自变量选择, 影响点和异常点的识别及处理, 异方差性诊断和自相关性问题及处理, 多重共线性问题及处理, 多元线性回归模型的有偏估计, 非线性回归模型和含定性变量的回归模型的参数估计理论、方法及算法, 以及广义线性回归模型和缺失数据模型的统计分析等. 这些内容为数据分析人员提供了一个完整的数据处理过程以及建立统计回归模型的技巧和方法.

尽管国内已有一些介绍回归分析的专著和教材, 但他们大都用常见的统计软件, 如SPSS、Excel、Matlab 等来介绍其回归分析的理论和方法. 由于 R 软件不仅免费使用, 而且它还拥有世界各地统计学家贡献的大量最新软件包且这些软件包的代码都是公开的, 因此, R 软件备受各国统计学家的广泛关注. 目前国内也有一些基于 R 软件来介绍数据分析的教材, 但没有系统地介绍回归分析的理论和方法. 而本书所有的分析都是通过 R 软件来实现的, 这就大大地增加了本书的实用性, 这也是本书的一大特色.

为使学生了解回归分析的最新发展和适应新时期下社会对统计学发展的新需要, 本书增加了一些国内其他回归分析教材中没有的, 但是新近发展的且学生不难理解并富有实用价值的内容, 如缺失数据模型的自变量选择、参数估计及其应用、广义线性回归模型及其参数估计和应用等. 这些内容在社会学、经济学、教育学、心理学和抽样调查等领域有着广泛的应用.

本书收集、编写大量的实际例子, 所用的数据例子都可以在《中国统计年鉴》网站上找到, 并且包括最新的数据, 如 2011 年的数据等, 每一数据例子都配有相应的 R 程序. 这些例子还反映了回归分析方法应用的很多方面的问题. 同时, 本书各章还附有习题. 这对培养学生的动手能力和应用所学知识解决实际问题的能力都是非常有益的.

本书力求理论结合实际例子讲授回归分析方法的直观意义、来龙去脉、什么问题用什么方法解决以及证明的思路. 有的证明放在本书习题中, 请学生参阅有关书目或自行完成.

本书除了作为统计学专业本科生的教学用书, 还可作为应用统计硕士的教学用书, 也可作为从事统计理论研究和实际应用的统计工作者、教师和学生的教学参考书. 此外, 本书还可作为从事社会学、教育学、心理学、经济学、金融学、人口学、生物医学以及临床研究等领域的理论研究者和实际应用者的参考书.

感谢科学出版社成都有限责任公司郝玉龙编辑.

由于编写时间紧且编者水平有限, 书中难免有不足之处, 敬请读者和同行批评指正.

唐年胜　李会琼

2013年9月17日 于昆明

# 目　　录

# 第 1 章　一些基本概念

## 1.1　数据和变量

在生活中, 我们随时随地都在与数据打交道, 可以说数据遍布于我们生活中的每一个角落. 例如, 从学生宿舍到上课地点的距离和所需时间, 某一宿舍的床位数, 某一学生的家庭人口数, 某一学生的年龄、身高、体重等. 统计数据是统计工作活动过程中取得的反映国民经济和社会现象的数字资料以及与之相联系的其他资料的总称. 统计研究客观事物的数量特征, 离不开统计数据, 统计数据是对客观现象进行计量的结果.

数据按其取值可分为以下四种类型.

(1) 计量数据, 如人的身高、体重, 物体的长度、质量, 粮食的产量、价格, 室内外的气温、湿度, 水库的容量、储水量, 家庭每月的用电量, 银行存款利率, 房价, 国民生产总值(GNP), 国内生产总值(GDP), 商品零售价格指数, 居民消费价格指数(CPI)等. 这些数据的取值可以是某一区间内的任一实数.

(2) 计数数据, 如某一学校的学生人数、教师人数、班级数, 某一农户的牲口数, 某一城市的汽车数, 某一时间段内接听电话次数, 书的页数等. 这些数据在整数范围内取值, 而且绝大多数还只能在非负整数范围内取值.

(3) 属性数据(又称为名义数据), 如人的性别(男、女), 民族(汉族、回族、彝族、傣族、苗族等), 季节(春季、夏季、秋季、冬季), 婚姻状况(未婚、有配偶、丧偶、离婚等), 国籍(中国、美国、英国、法国、德国等), 种族(黑色人种, 白色人种、黄色人种、红色人种、棕色人种等)等. 在属性数据分析中, 通常用数来表示属性的分类. 例如, 用数“1”和“2”表示男和女, 用“1”“2”“3”和“4”表示春季、夏季、秋季和冬季. 这些数只起一个名义的作用, 只是一个代码, 没有大小关系, 也不能进行运算.

(4) 有序数据, 如人的文化程度由低到高可分为文盲、小学、初中、高中、中专、大专、大学、硕士、博士等, 可用数 0, 1, 2, 3, 4, 5, 6, 7, 8 表示. 又如学生对老师教学效果的评价可分为“不好”“好”和“非常好”三类, 可用数“1”“2”和“3”表示; 顾客对银行营业员服务态度的评价可分为“不满意”“基本满意”“满意”和“非常满意”四类, 可用数“0”“1”“2”和“3”表示(也可用数“1”“2”“3”和“4”表示或用数“1”“3”“5”和“7”表示); 人的身体状况可分为“有病”“亚健康”“健康”. 不管用什么数表示, 这些数仅起一个顺序作用而没有通常数的意义, 类与类之间的差别是不能运算的. 例如, “小学”文化程度的人比“文盲”的人更有知识, 但他们的知识相差多少呢, 这是不能用数来计算的, 即用“1–0”来表示他们的文化程度的差异是没有意义的.

通常把**计量数据**和**计数数据**统称为**定量数据**, 而把**属性数据**和**有序数据**统称为**定性数据**.

数据按其形态可分为以下三种类型.

(1) 时间序列数据, 是指在不同时间点上对某一个体的某一指标或某些指标进行观测, 并将得到的数据按其时间先后顺序排列而成的一组数列, 也称为动态数据. 这类数据反映了某一个体的某一指标或某些指标随时间的变化状态或程度. 如某一商店(或超市)每月的销售额, 某一家庭每年末的存款额, 1980 年以来我国历年的国内生产总值(或国民生产总值或商品零售价格指数等), 某一年内每月消费者价格指数. 非随机性时间序列包括: 平稳性时间序列、趋势性时间序列和季节性时间序列三种. 由于经济变量或社会现象的前后期之间存在相关性, 所以时间序列数据容易产生模型中随机误差项的序列相关性.

(2) 横截面数据, 是指对若干个体在同一时间截面上进行观测得到的数据, 也称为静态数据. 这类数据反映了不同事物或现象在同一时间截面上的变化状况或程度. 如第六次全国人口普查数据, 年度工业普查数据、经济普查数据, 某一学期某一班全体同学某一门课程的成绩, 某一奶牛场某一天所有奶牛的产奶量, 2011 年 3 月全国 32 个大中城市的物价指数等都是横截面数据.

(3) 面板数据(panel data)或纵向数据(longitudinal data), 是同时在时间和截面上取得的二维数据, 它是横截面数据与时间序列数据综合起来的一种数据类型. 它是指对若干个体在不同时间点上进行重复观察得到的数据. 从横截面看, 面板数据是由若干个体在某一时间点构成的横截面观测数据, 而从纵剖面看, 每一个体都是一个时间序列. 面板数据根据个体观测次数的不同可以分为平衡面板数据和非平衡面板数据. 平衡面板数据是指面板数据中每一个体的重复观测次数是一样的, 而非平衡面板数据是指面板数据中每一个体的重复观测次数不完全一样. 例如, 1978~2011 年我国东北、华北、华东 15 个省级地区的居民家庭人均收入数据.

数据按其是否存在缺失可以分为: 观测数据和缺失数据. 缺失数据在我们的问卷调查中是普遍存在的. 例如, 在家庭问卷调查中被访者可能拒绝回答其收入情况, 在民意调查中被访者可能拒绝表达他们对某些敏感的或令人尴尬的问题的态度, 在药物药效研究中由于药物的副作用致使患者放弃该药物的治疗, 在心理学实验中由于机械故障而导致某些观测结果丢失, 等等. 为了得到更好的统计推断结果, 在进行统计分析时我们必须考虑缺失数据的影响.

在日常生活中, 我们经常发现有一些量是确定的, 而有些量是不确定的. 譬如, 一架飞机有多少个座位是一个确定的数, 通常称为**常数**(constant)或者常量. 然而, 某一天乘坐此架飞机从甲地到乙地的旅客数就不确定了, 这与季节、航班时间、甲乙两地的地理位置等有关. 我们把某一次乘坐此架飞机的旅客数称为**变量**(variable). 又如一个学校的教师人数是一固定值, 它是一个常量. 然而, 该校每天给本科学生上课的教师人数则是一个变量. 因此, 统计学中把一个可取两个或更多个可能值的特征、特质或属性称为**变量**(或**随机变量**). 例如, 人的性别是可取两个值的变量, 汽车每升油耗所能行驶的里程数是一个可取很多值的变量. 变量的值被称为数据. 变量和数据是统计学中很重要的两个概念. 它们既相互联系又相互区别. 一般情况下可以通用. 在数据分析中, 我们通常需要确定感兴趣的变量及其取值范围. 我们可以把变量的取值想象为散布在一条直线上的点, 而直线本身代表了这个变量.

变量按其取值不同可分为定量变量和定性变量两大类. 如果变量的取值是一些数量值, 则称该变量为**定量变量**或**数值变量**(quantitative variable); 因为是随机的, 也称为**随机变量**(random variable). 定量变量包括连续型变量和离散型变量两种, 如人的身高、体重是连续型变量, 而购买某商品的人数则是离散型变量. 离散型变量和连续型变量在数轴上的一点分别是, 离散型变量是数轴上的一个点, 而连续型变量是数轴上的一段距离. 如性别、民族、观点之类的取非数量值的变量就称为**定性变量**或**属性变量**(qualitative variable)或**分类变量**(categorical variable). 定性变量可分为名义变量和有序变量. 一些定性变量也可以由定量变量来描述, 如男女生的数目, 持有某观点的人数比例, 定性变量只有用数量来描述时, 才有可能建立数学模型, 并使用计算机来分析.

变量按其在回归分析中的作用和地位不同可分为**协变量**(或**解释变量**)和**因变量**(或**被解释变量**). 一般地, 在统计建模过程中, 我们把影响某一变量的量称为协变量, 而被影响的量称为因变量. 当协变量或因变量为定性变量时, 我们通常用**哑元**(dummy variable)来表示它们, 如性别通常用 0 和 1 分别表示女性和男性; 用 1, 2, 3 和 4 分别表示春、夏、秋和冬季, 等等.

## 1.2 变量之间的关系

### 1.2.1 定量变量间的关系

现实世界各现象之间是相互联系、相互制约、密不可分的. 一个现象的发生总是由与之相联系的其他现象的变化所引起的. 譬如, 房价的变动是由与之相联系的建筑材料、地价、劳动力成本、供求关系等因素的变化所引起的. 现实世界中任何一个现象不仅同与之相联系的现象构成一个关联体, 而且该现象各内部因素之间也存在着各种各样的联系, 在一定的条件下, 一些因素的变化引起另外一些与之相联系的因素发生变化. 因此, 正确认识影响某一现象的各因素之间的相关关系是我们正确把握或揭示这一现象的内在规律性的一个重要工作基础. 这就要求我们探讨变量之间的相互关系. 譬如, 一个地区想通过招商引资的方式来改善该地区的经济状况, 此时, 该地区的管理部门就想知道投资方式和经济发展之间的关系; 另一个地区想通过发展旅游来提高该地区的经济发展, 则需研究旅客人数和经济发展之间的内在联系等. 所有这些例子表明, 为了研究某一社会经济现象或自然现象就要讨论变量之间的关系, 否则就无从谈起任何有深度的应用. 下面我们来考察两个定量变量之间关系的例子.

**例1.1** 广告投入和销售额之间的关系. 表 1.1 显示了某企业的广告投入 $x$ 和销售额 $y$ 之间的关系(单位: 万元).

**表1.1 某企业的广告投入 $x$ 和销售额 $y$ 之间的关系**

| 广告投入$x$ | 1.0 | 3.2 | 3.2 | 5.5 | 5.9 | 7.1 | 7.3 | 9.2 | 10.8 | 12.1 |
|---|---|---|---|---|---|---|---|---|---|---|
| 销售额$y$ | 9.4 | 31.8 | 33.2 | 52.4 | 53.5 | 56.0 | 56.9 | 59.2 | 60.1 | 63.5 |

在此例中, 我们考虑下面的问题:

(1) 变量 $x$ 和 $y$ 是否相关?

(2) 如果相关, 它们之间的关系是否显著?

(3) 它们之间的关系能否用一个模型来描述?

(4) 这个关系是否带有普遍性?

(5) 这个关系是不是因果关系?

为了研究变量 $x$ 和 $y$ 之间的关系, 我们在图 1.1 中描绘出了它们的散点图. 从图 1.1 可以看出, 随着该企业广告费用投入的增加, 其销售额也在不断地增长. 由此表明：它们之间有很强的相关关系, 但是否存在因果关系呢? 这需要我们首先明确因果关系和相关关系的概念.

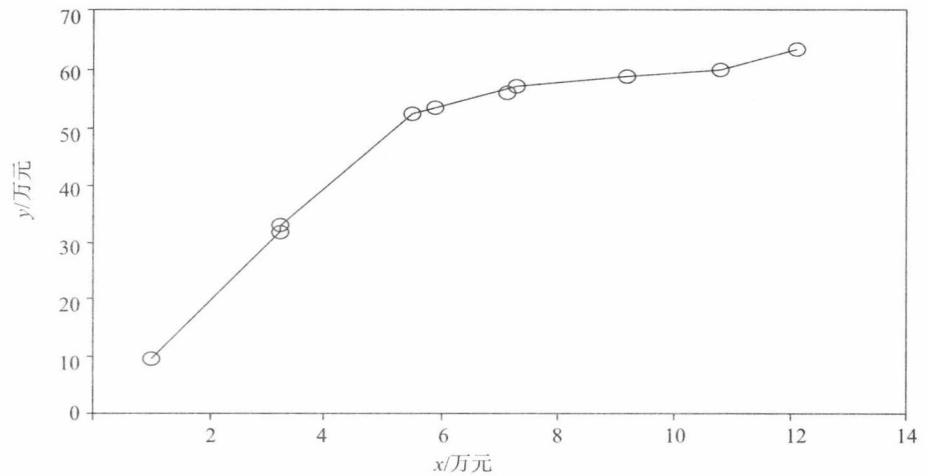



图 1.1 广告投入与销售额的散点图

所谓因果关系是指某一因素的存在一定会导致某一特定结果发生. 因果关系中最常见的是一因一果, 也存在一因多果、一果多因、多因多果等现象. 而相关关系是统计学上的一个概念, 它是指某个因素的变化会导致另外一个因素的变化, 但这个因素的变化是不是另外一个因素变化的原因, 是不能确定的. 一般来说, 变量之间有关系并不意味着它们之间一定存在明确的因果关系. 譬如, 夏天太阳镜的销售量和雪糕的销售量存在相关性, 但这不是说因为太阳镜卖多了, 雪糕就会卖得多. 它们呈相关关系, 仅仅是因为它们受同一因素——日光辐射强度的影响, 它们都是日光辐射强度的共同的结果. 又如, 天气冷和下雪的关系. 下雪的时候通常会伴随着气温的下降, 但是究竟是气温下降导致了下雪呢还是下雪导致了气温下降呢, 这是一个很难界定的问题, 只有根据具体情况而定了. 再如, 努力与学习成绩之间的关系. 我们经常遇见努力而学习成绩无法提高的现象, 而不努力但学习成绩未必下降(有时还会上升)的现象, 即是表明：努力与提高学习成绩之间只有相关关系但并无因果关系. 在可控试验中, 较容易找到变量之间的因果关系. 譬如, 治疗方式和疗效的关系等. 然而, 在许多实际问题研究中, 因果关系的确定是很复杂的, 要根据研究的问题本身来确定. 因果关系与相关关系是说明事物之间联系的两种形式, 也是常被人们混淆的两种关系, 因此, 正确理解它们的定义对我们研究变量之间的关系是非常有益的.

只要有关系, 即使不是因果关系也不妨碍人们利用这种关系来进行统计推断. 譬如,

利用公鸡打鸣来预报太阳升起；虽然公鸡打鸣绝对不是日出的原因(虽然打鸣发生在先). 简单的办法(诸如画图)可以得到一些信息, 但不一定能够给出满意的答案, 需要更多的工具和手段来进行数值分析得到更加严格和精确的解答. 因此, 需要继续我们的研究.

### 1.2.2 定性变量间的关系

定性变量可以用定量变量来进行表述, 下面我们看一个例子.

**例1.2** 表 1.2 中的数据来自于 123 名有关贯彻执行某项政策的调查研究. 这是一个基于收入(定量变量)、性别(定性变量)和观点(定性变量)的简单的三维表, 它显示了人们的收入和性别与其对该项政策的观点之间的关系.

**表1.2 不同收入和不同性别人群对某项政策的观点**

| 性别 | 观点:反对 | | | 观点:赞成 | | |
|---|---|---|---|---|---|---|
| | 低收入 | 中收入 | 高收入 | 低收入 | 中收入 | 高收入 |
| 男 | 5 | 8 | 10 | 20 | 10 | 5 |
| 女 | 2 | 7 | 9 | 25 | 15 | 7 |

基于表 1.2, 我们希望可以看出收入和性别对该项政策的观点是否有影响及如何影响. 如果要想得到更加精确的结论, 就要进行进一步的分析和计算. 具体的方法将在后面的列联表分析或多项分布对数线性模型中进行详细说明.

### 1.2.3 定性和定量变量间的混合关系

在研究过程中, 我们经常发现有些数据不仅只有定性变量或只有定量变量, 而且是同时包含两类变量. 在数据分析时, 通常想知道同时包括定性变量和定量变量的数据集中各变量之间的关系. 下面的数据就是这一问题的一个例子.

| | sex | polut | age | count |
|---|---|---|---|---|
| 1 | 1 | 1 | 12.2 | 7 |
| 2 | 1 | 1 | 18.6 | 3 |
| 3 | 1 | 1 | 25.3 | 8 |
| 4 | 1 | 1 | 33.6 | 11 |
| 5 | 1 | 1 | 38.6 | 7 |
| 6 | 1 | 1 | 43.8 | 5 |
| 7 | 1 | 1 | 52.8 | 9 |
| 8 | 1 | 1 | 60.1 | 6 |
| 9 | 1 | 1 | 64.7 | 6 |
| 10 | 1 | 1 | 72.7 | 6 |
| 11 | 1 | 2 | 8.9 | 12 |
| 12 | 1 | 2 | 19.6 | 5 |
| 13 | 1 | 2 | 25.9 | 13 |
| 14 | 1 | 2 | 31.5 | 10 |
| 15 | 1 | 2 | 39.2 | 12 |

该数据有 2 个定性变量(性别 sex、污染程度 polut)、2 个定量变量(年龄 age、发生哮喘的人数 count). 我们希望通过分析, 能够知道哮喘这一变量与其他 3 个变量之间的关系, 具体的方法将在 Poisson 对数线性模型中作详细讨论.

## 1.3 回归分析与相关分析

### 1.3.1 回归分析

“回归”一词的由来归功于英国著名遗传学家、统计学家高尔顿(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, 观测了 1078 对夫妇的平均身高(用 $x$ 表示)和他们的一个成年儿子的身高(用 $y$ 表示). 将这些结果在平面直角坐标系上绘成散点图后, 发现 $x$ 与 $y$ 几乎呈一条直线, 其计算出来的回归直线方程为: $\hat{y} = 33.73 + 0.516x$. 这一结果表明, 父母平均身高 $x$ 每增加一个单位时, 其成年儿子的身高也平均增加 0.516 个单位. 高尔顿在对试验数据进行深入分析后, 发现了一个很有趣的现象: 高个子父辈确有生高个子儿子的趋势, 但父辈身高每增加一个单位, 其儿子身高仅增加半个单位左右; 矮个子父辈确有生矮个子儿子的趋势, 但父辈身高每减少一个单位, 其儿子身高仅减少半个单位左右. 这表明, 当父辈身高较高时, 他的成年儿子的身高一般不会比父亲身高更高; 同样当父辈个子较矮时, 他的成年儿子的身高一般不会比父辈身高还矮, 而是向一般人的均值靠拢. 高尔顿和皮尔逊把这一现象称为**回归效应**, 即回归到一般高度的效应. 高尔顿依据试验数据推算出父辈身高与其成年儿子身高的关系式的过程就是著名的“回归分析”, 其关系式所代表的这条直线称为回归直线. 由此可以看出, 所谓“回归分析”就是指对具有相关关系的两个或多个变量之间的数量变化进行数量测定, 配合一定的数学方程(或模型), 以便由自变量的数值对因变量的可能值进行估计或预测的一种统计方法. 根据数学模型描绘出来的几何图形称为回归线.

回归分析按所涉及的自变量的个数不同, 可分为一元回归分析和多元回归分析. 当只有一个自变量时, 称为**一元回归分析**; 当自变量有两个或多个时, 则称为**多元回归分析**. 按自变量和因变量之间的关系类型, 可将回归分析分为**线性回归分析**和**非线性回归分析**. 如果回归分析所得到的回归方程关于未知参数是线性的, 则称为**线性回归分析**; 否则, 称为**非线性回归分析**.

### 1.3.2 相关分析

社会经济各现象之间在数量上的依存关系通常有两种类型: 一是函数关系, 二是相关关系.

**函数关系**是指变量之间存在确定性的数量对应关系. 我们可以把变量 $y$ 与 $p$ 个变量 $x_1, \cdots, x_p$ 之间存在着的某种函数关系写成下面的形式: $y = f(x_1, \cdots, x_p)$. 在此函数关系中, 当 $p$ 个变量的取值一定时, 与其相对应的变量 $y$ 的值也就随之而定了. 例如, 圆的面积与半径的关系可表示为: $s = \pi r^2$, 其中, $r$ 是圆的半径, $s$ 为圆的面积; 假设银行一年期定期存款利率为年息 3%, 则存入本金 $x$ 与到期的本息 $y$ 之间的关系可表示为: $y = (1 + 3\%)x$.

1. 相关关系

相关关系是指变量之间客观存在非确定性的数量对应关系(因果关系). 在相关关系

中, 当一个或几个变量取一定值时, 与其相对应的另一个变量的值不完全确定, 而是有多个值与其对应. 例如, 学习成绩与学习时间的关系, 收入与消费支出的关系都是相关关系. 在社会经济现象中, 这种相关关系是大量存在的. 如提高劳动生产率会使成本降低, 提高劳动生产率会使利润增加、粮食的亩产量与施肥量之间也存在着关联等.

相关关系是一种不完全确定的随机关系, 在相关关系的情况下, 因素标志的每个数值都可能有若干个结果标志的数值与之对应. 因此, 相关关系是一种不完全的依存关系. 现象之间之所以会存在这种不完全的依存关系, 是因为除了被分析的影响因素, 还有诸多其他的因素在发挥着作用. 如学习成绩的高低除了受到学习时间长短的影响外, 还要受到学习效率、学习基础、智力等因素的影响.

2. 相关关系与函数关系的区别与联系

相关关系与函数关系的不同之处表现在: ①函数关系指变量之间的关系是确定的, 而相关关系中两变量的关系则是不确定的, 可以在一定范围内变动; ②函数关系变量之间的依存可以用一定的方程 $y = f(x)$ 表现出来, 可以给定自变量来推算因变量, 而相关关系则不能用一定的方程表示. 函数关系是相关关系的特例, 即函数关系是完全的相关关系, 相关关系是不完全的相关关系.

函数关系与相关关系虽然有明显的区别, 但两者之间并不存在不可逾越的界限. 由于存在测算误差等原因, 函数关系在实际中往往通过相关关系表现出来. 而在研究相关关系时, 为了找到现象间数量关系的内在联系和表现形式, 又常常需要借助于函数关系的形式加以描述. 因此, 相关关系是相关分析的研究对象, 函数关系是相关分析的工具.

### 1.3.3 相关分析的内容

从狭义的角度来看, 相关分析以现象之间是否相关、相关的方向和密切程度等为主要研究内容, 它不区别自变量与因变量, 因为变量 $x$ 与变量 $y$ 是否相关和变量 $y$ 与变量 $x$ 是否相关是同一个问题; 另外, 狭义的相关分析对各变量的构成形式(关系的表现形态)也不关心.

从广义的角度来看, 相关分析就是研究两个或两个以上变量之间相关方向和相关密切程度大小以及用一定函数来表达现象相互关系的方法. 也就是说, 广义的相关分析除了包括对现象间数量关系的密切程度的测定, 还包括具体的相关形式的分析, 即**回归分析**.

### 1.3.4 相关关系的种类

1. 按相关的因素多少可分为单相关和复相关

这是按所考虑的变量数的多少对相关关系进行的分类. 所谓单相关(又称一元相关), 是指两个变量之间的相关关系, 即因素标志只有一个, 研究一个自变量与一个因变量之间的相关关系.

复相关(又称多元相关), 是指三个或三个以上变量之间的相关关系, 即因素标志不只一个, 有两个或两个以上, 研究一个因变量与多个自变量之间的相关关系.

2. 按相关的表现形式可分为线性相关和非线性相关

线性相关(又称直线相关)是指如果自变量数值发生变动, 因变量数值随之发生大致均等的变动, 从平面图上观察其各点的分布近似地表现为一直线, 这种相关关系就称为直线相关(也称线性相关).

非线性相关(又称曲线相关)是指如果自变量发生变动, 因变量数值也随之发生变动, 但这种变动不是沿着一个方向发生均等变动, 从图形上看, 其分布表现为各种不同的曲线形式, 这种相关关系称为曲线相关.

3. 按相关的方向可把直线相关分为正相关和负相关

正相关是指当自变量 $x$ 的数值增加(或减少)时, 因变量 $y$ 的数值也将随之相应地增加(或减少), 即因变量和自变量的变动方向是一致的, 这种相关关系称为正相关. 例如, 消费收入越多, 则消费支出也增加; 儿童数量增加, 玩具的销售量也会增加等.

负相关是指当自变量 $x$ 的数值增加(或减少)时, 因变量 $y$ 的数值随之减少(或增加), 即自变量与因变量的变动方向是相反的, 这种相关关系称为负相关. 例如, 劳动生产率提高, 产品成本降低; 商品价格降低, 销售量增加等.

4. 按相关的程度可分为完全相关、不完全相关和不相关

完全相关是指两个变量之间的相关关系, 当自变量改变一定量时, 因变量的改变量是一个确定的量, 则这两个变量间的关系称为完全相关, 此种关系实际上就是函数关系. 如前面提到的圆的面积与半径之间的关系、商品销售额与价格、销售量的关系等都是完全相关关系.

不相关是指研究的两个变量之间没有任何关系, 而是各自独立或互不影响, 则称为不相关(或零相关). 如一年中天气晴好所占的比率与同学们的学习成绩之间没有什么关系, 这两种现象就不相关.

不完全相关是指若变量之间的关系介于完全相关与不相关之间, 则称为不完全相关. 不完全相关是相关分析的主要对象, 也就是我们一般意义上所讲的相关关系.

回归分析和相关分析是互相补充、密切联系的. 回归和相关都是研究两个变量相互关系的分析方法. 它们的差别主要有以下两点: ①相关分析研究两个变量之间相关的方向和相关的密切程度. 但是相关分析不能指出两变量相互关系的具体形式, 也无法从一个变量的变化来推测另一个变量的变化关系. 回归方程则是通过一定的数学方程来反映变量之间相互关系的具体形式, 以便从一个已知量来推测另一个未知量. 为估算预测提供一个重要的方法. ②相关分析既可以研究因果关系的现象也可以研究共变的现象, 不必确定两个变量中哪个是自变量, 哪个是因变量. 而回归分析是研究两个变量具有因果关系的数学形式, 因此必须事先确定变量中自变量与因变量的地位. 计算相关系数的两变量是对等的, 可以都是随机变量, 各自接受随机因素的影响, 改变两变量的地位并不影响相关系数的数值.

相关分析需要回归分析来表明现象数量相关的具体形式, 而回归分析则应该建立在

相关分析的基础上. 依靠相关分析表明现象的数量变化具有密切相关, 进行回归分析求其相关的具体形式才有意义. 在相关程度很低的情况下, 回归函数的表达式代表性就很差.

## 1.4 建立回归模型的步骤

一般来说, 对一个实际问题建立回归模型, 需要考虑下面六个步骤.

**第一步: 根据研究目的, 设置指标变量**

回归模型主要是用来揭示事物间相关变量的数量关系. 首先要根据所研究的问题设置因变量 $y$, 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. 尤其是在研究具体实际问题时, 我们必须根据实际问题的研究目的, 确定实际问题中各因素之间的因果关系.

对于一个具体的问题, 当研究目的确定后, 被解释变量容易确定, 被解释变量一般直接表达、刻画研究目的. 另外, 不要认为一个回归模型所涉及的解释变量越多越好. 一个经济模型, 如果把一些主要变量漏掉肯定会影响模型的应用效果, 但如果引入的变量太多, 可能会选择一些与问题无关的变量, 还可能由于一些变量的相关性很强, 它们所反映的信息有严重的重叠, 这就有可能出现共线性问题. 当变量太多时, 计算工作量太大, 计算误差就大, 估计的模型参数精度自然不高.

总之, 回归变量的确定是一个非常重要的问题, 是建立回归模型最基本的工作. 这个工作一般一次并不能完全确定, 通常要反复比较, 最终选出最适合的一些变量.

**第二步: 收集、整理统计数据**

回归模型的建立是基于回归变量的样本统计数据. 当确定好回归模型的变量之后, 就要对这些变量进行收集、整理和统计数据. 数据的收集是建立回归模型的重要环节, 数据质量如何, 对回归模型有至关重要的影响.

常用的样本数据分为**时间序列数据**和**横截面数据**.

**时间序列数据**, 就是按时间顺序排列的统计数据. 如最近 10 年的 CPI、PPI 统计数据. 时间序列数据容易产生模型中随机误差项的序列相关, 这是因为许多经济变量的前后期之间总是有关系的. 如在建立需求模型时, 人们的消费习惯、商品短缺程度等具有一定的延续性, 它们对相当一段时间的需求量有影响, 这样就产生随机误差项的序列相关. 对于具有随机误差项序列相关的情况, 最常用的处理方法是差分法, 我们将在后面章节中详细介绍.

**横截面数据**, 即为在同一时间截面上的统计数据. 如同一年份全国 35 个大中城市的物价指数等都是横截面数据. 当用截面数据作样本时, 容易产生异方差性. 这是因为一个回归模型往往涉及许多解释变量, 如果其中某一因素或一些因素随着解释变量观测值的变化而对被解释变量产生不同影响, 就产生异方差性. 对于具有异方差性的建模问题, 数据整理就要注意消除异方差性, 这常与模型参数估计方法结合起来考虑.

不论是时间序列数据还是横截面数据的收集, 样本容量的多少一般要与设置的解释变量数目相配套. 通常为了使模型的参数估计更有效, 要求样本容量 $n$ 大于解释变量的个数 $p$. 样本容量的个数小于解释变量数目时, 普通的最小二乘法失效. $n$ 与 $p$ 到底应该